

Massimo Marchiori  
Francisco José Domínguez Mayo  
Joaquim Filipe (Eds.)

LNBIP 494

# Web Information Systems and Technologies


18th International Conference, WEBIST 2022  
Valletta, Malta, October 25–27, 2022  
Revised Selected Papers


 Springer

# Lecture Notes in Business Information Processing

494


## Series Editors

Wil van der Aalst , *RWTH Aachen University, Aachen, Germany*

Sudha Ram , *University of Arizona, Tucson, AZ, USA*

Michael Rosemann , *Queensland University of Technology, Brisbane, QLD, Australia*

Clemens Szyperski, *Microsoft Research, Redmond, WA, USA*

Giancarlo Guizzardi , *University of Twente, Enschede, The Netherlands*

LNBIP reports state-of-the-art results in areas related to business information systems and industrial application software development – timely, at a high level, and in both printed and electronic form.

The type of material published includes

- Proceedings (published in time for the respective event)
- Postproceedings (consisting of thoroughly revised and/or extended final papers)
- Other edited monographs (such as, for example, project reports or invited volumes)
- Tutorials (coherently integrated collections of lectures given at advanced courses, seminars, schools, etc.)
- Award-winning or exceptional theses

LNBIP is abstracted/indexed in DBLP, EI and Scopus. LNBIP volumes are also submitted for the inclusion in ISI Proceedings.

Massimo Marchiori ·  
Francisco José Domínguez Mayo ·  
Joaquim Filipe  
Editors

# Web Information Systems and Technologies

18th International Conference, WEBIST 2022  
Valletta, Malta, October 25–27, 2022  
Revised Selected Papers

*Editors*

Massimo Marchiori  
University of Padua (UNIPD)  
Padua, Italy

Francisco José Domínguez Mayo  
University of Seville  
Seville, Spain

Joaquim Filipe  
Polytechnic Institute of Setúbal/INSTICC  
Setubal, Portugal

ISSN 1865-1348

ISSN 1865-1356 (electronic)

Lecture Notes in Business Information Processing

ISBN 978-3-031-43087-9

ISBN 978-3-031-43088-6 (eBook)

<https://doi.org/10.1007/978-3-031-43088-6>

© The Editor(s) (if applicable) and The Author(s), under exclusive license  
to Springer Nature Switzerland AG 2023

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

# Preface

This volume includes extended and revised versions of a set of selected papers from the 18th International Conference on Web Information Systems and Technologies (WEBIST 2022), held in Valletta, Malta, on October 25–27, 2022.

WEBIST 2022 received 62 paper submissions from 24 countries, of which 21% were included in this book.

The papers were selected by the event chairs and their selection is based on a number of criteria that include the classifications and comments provided by the program committee members, the session chairs' assessment and also the program chairs' global view of all papers included in the technical program. The authors of selected papers were then invited to submit a revised and extended version of their papers having at least 30% innovative material.

The purpose of the International Conference on Web Information Systems and Technologies (WEBIST) is to bring together researchers, engineers and practitioners interested in the technological advances and business applications of web-based information systems. The conference has four main tracks, covering different aspects of Web Information Systems, namely Internet Technology, Web Intelligence and the Semantic Web, Social Network Analytics, HCI in Mobile Systems, and Web Interfaces.

The papers selected to be included in this book contribute to the understanding of relevant trends of current research on Web Information Systems and Technologies, including: Human Computer Interaction, UX and User-Centric Systems, Applications, Research Projects and Web Intelligence, Deep Learning, Human Factors, Natural Language Processing, Research Projects and Internet Technology, Data Web Mining, Interaction Design, and Web Information Filtering and Retrieval.

We would like to thank all the authors for their contributions and also the reviewers who helped ensuring the quality of this publication.

October 2022

Massimo Marchiori  
Francisco José Domínguez Mayo  
Joaquim Filipe

# Organization

## Conference Chair

Joaquim Filipe  
Polytechnic Institute of Setubal/INSTICC,  
Portugal

## Program Co-chairs

Stefan Decker  
RWTH Aachen, Germany  
Francisco José Domínguez Mayo  
University of Seville, Spain  
Massimo Marchiori  
University of Padua, Italy

## Program Committee

Marco Aiello  
University of Stuttgart, Germany  
Jesús Arias Fisteus  
Universidad Carlos III de Madrid, Spain  
Elarbi Badidi  
United Arab Emirates University, UAE  
Antonio Balderas  
Universidad de Cádiz, Spain  
Demis Ballis  
University of Udine, Italy  
Faiza Belala  
LIRE Laboratory, Constantine 2 University,  
Algeria  
Devis Bianchini  
University of Brescia, Italy  
Fernando Bobillo  
University of Zaragoza, Spain  
Pasquina Campanella  
University of Bari “Aldo Moro”, Italy  
Dickson Chiu  
University of Hong Kong, China  
Soon Chun  
City University of New York, USA  
Christophe Cruz  
University of Burgundy, France  
Daniel Cunliffe  
University of South Wales, UK  
Clodoveu Davis Júnior  
UFMG, Brazil  
Martine De Cock  
University of Washington, Tacoma, USA  
Toon De Pessemier  
Ghent University - Imec, Belgium  
Enrico Denti  
Alma Mater Studiorum - Università di Bologna,  
Italy  
Martin Drlik  
Constantine the Philosopher University in Nitra,  
Slovak Republic  
Vítor E. Silva Souza  
Universidade Federal do Espírito Santo, Brazil  
Larbi Esmahi  
Athabasca University, Canada

Przemyslaw Falkowski-Gilski	Gdansk University of Technology, Poland
Luís Ferreira Pires	University of Twente, Netherlands
Josep-Lluís Ferrer-Gomila	Balearic Islands University, Spain
Karla Fook	Instituto Tecnológico de Aeronáutica-ITA/IEC, São José dos Campos, Brazil
Xiang Fu	Hofstra University, USA
Matteo Gaeta	University of Salerno, Italy
Ombretta Gaggi	Università di Padova, Italy
Francisco García-Sánchez	University of Murcia, Spain
John Garofalakis	University of Patras, Greece
Ilche Georgievski	University of Stuttgart, Germany
Jose Gonzalez	University of Seville, Spain
Julián Grigera	Universidad Nacional de La Plata, Argentina
Francesco Guerra	University of Modena and Reggio Emilia, Italy
Shanmugasundaram Hariharan	Vardhaman College of Engineering, India
Ioannis Hatzilygeroudis	University of Patras, Greece
Jose Herrero Agustin	University of Extremadura, Spain
Hanno Hildmann	TNO, Netherlands
Andreas Hinderks	Universidad de Sevilla, Germany
Sergio Ilarri	University of Zaragoza, Spain
Edmond Jajaga	UBT- Higher Education Institution, Albania
Monique Janneck	Luebeck University of Applied Sciences, Germany
Karel Jezek	University of West Bohemia, Czech Republic
Georgia Kapitsaki	University of Cyprus, Cyprus
Sokratis Katsikas	Norwegian University of Science and Technology, Gjøvik, Norway
Ilan Kirsh	ObjectDB Software, UK
Andreas Klein	University of Seville, Spain
Matthias Klusch	German Research Center for Artificial Intelligence (DFKI) GmbH, Germany
Hiroshi Koide	Kyushu University, Japan
Fotios Kokkoras	University of Thessaly, Greece
Sylvain Kubler	University of Lorraine, France
Martin Llamas-Nistal	University of Vigo, Spain
Francisco J. Lopez-Pellicer	Universidad de Zaragoza, Spain
Andrea Mauri	Delft University of Technology, Netherlands
Mirosław Mazurek	Rzeszów University of Technology, Poland
Inmaculada Medina-Bulo	Universidad de Cádiz, Spain
Michele Melchiori	University of Brescia, Italy
Santiago Meliá	Universidad de Alicante, Spain
Ingo Melzer	Daimler Truck North America, USA



Marzal Miguel Ángel	Universidad Carlos III De Madrid, Spain
Akiyo Nadamoto	Konan University, Japan
Alex Norta	Tallinn University of Technology, Estonia
Kalpdrum Passi	Laurentian University, Canada
David Paul	University of New England, Australia
Bhanu Prasad	Florida A&M University, USA
Jim Prentzas	Democritus University of Thrace, Greece
Birgit Pröll	Johannes Kepler University Linz, Austria
Davide Rossi	University of Bologna, Italy
Gustavo Rossi	Lifia, Argentina
Lloyd Rutledge	Open University of the Netherlands, Netherlands
Comai Sara	Politecnico di Milano, Italy
Claudio Schifanella	Università degli Studi di Torino, Italy
Georg Schneider	Trier University of Applied Sciences, Germany
Wieland Schwinger	Johannes Kepler University, Austria
Weiming Shen	NRC Canada, Canada
Marianna Sigala	School of Management, University of South Australia Business School, Australia
Stian Soiland-Reyes	University of Manchester, UK
Eliza Stefanova	Sofia University, Bulgaria
Sergio Tessaris	Free University of Bozen-Bolzano, Italy
Dirk Thissen	RWTH Aachen University, Germany
Christos Troussas	University of West Attica, Greece
William Van Woensel	Dalhousie University, Canada
Costas Vassilakis	University of the Peloponnese, Greece
Guillermo Vega-Gorgojo	University of Valladolid, Spain
Jari Veijalainen	University of Jyväskylä, Finland
Victoria Vysotska	Lviv Polytechnic National University, Ukraine
Jason Whalley	Northumbria University, UK

## **Additional Reviewers**

Abdul Aziz	Universidad de Zaragoza, Spain
Dagoberto Herrera Murillo	Universidad de Zaragoza, Spain

## **Invited Speakers**

Mike Thelwall	University of Wolverhampton, UK
Wolfgang Nejdl	L3S and University of Hannover, Germany
Riccardo Rosati	Sapienza Università di Roma, Italy
Diana Maynard	University of Sheffield, UK

# Contents

Automated SLR with a Few Labeled Papers and a Fair Workload Metric . . . . .	1
<i>Allan Victor Almeida Faria, Maísa Kely de Melo, Flávio Augusto R. de Oliveira, Li Weigang, and Victor Rafael Rezende Celestino</i>	
Shift Toward Value-Based Learning: Applying Agile Approaches in Higher Education . . . . .	24
<i>Eva-Maria Schön, Ilona Buchem, Stefano Sostak, and Maria Rauschenberger</i>	
Improving the Representation Choices of Privacy Policies for End-Users . . . . .	42
<i>Michalis Kaili and Georgia M. Kapitsaki</i>	
Scaffolding Process-Aware Information Systems with the AKIP Platform . . . . .	60
<i>Ulisses Telemaco Neto, Toacy Oliveira, Raquel Pillat, Paulo Alencar, Don Cowan, and Glaucia Melo</i>	
Grammar-Based Question Classification Using Ensemble Learning Algorithms . . . . .	84
<i>Alaa Mohasseb and Andreas Kanavos</i>	
Leveraging Transfer Learning for Long Text Classification with Limited Data . . . . .	98
<i>Carlos Alberto Alvares Rocha, Li Weigang, Marcos Vinícius Pinheiro Dib, Allan Victor Almeida Faria, Daniel Oliveira Cajueiro, Maísa Kely de Melo, and Victor Rafael Rezende Celestino</i>	
Integrating Linguistic and Citation Information with Transformer for Predicting Top-Cited Papers . . . . .	121
<i>Masanao Ochi, Masanori Shiro, Jun'ichiro Mori, and Ichiro Sakata</i>	
Soft Web Intelligence with the J-CO Framework . . . . .	142
<i>Paolo Fosci and Giuseppe Psaila</i>	
An NLP Approach to Understand the Top Ranked Higher Education Institutions' Social Media Communication Strategy . . . . .	166
<i>Alvaro Figueira and Lirielly Nascimento</i>	

**Influence of Demographic Variables and Usage Behaviour on the Perceived User Experience** ..... 186  
*Jessica Kollmorgen, Martin Schrepp, and Jörg Thomaschewski*

**Categorizing UX Aspects for Voice User Interfaces Using the Kano Model** .... 209  
*Kristina Kölln, Andreas M. Klein, Jana Deutschländer, Dominique Winter, and Maria Rauschenberger*

**How We Evaluate the Accessibility of an Infographic: A Pilot Study Through SUS Questionnaire** ..... 229  
*Alessio Caccamo*

**Evaluating the Quality Characteristics of Space Geeks** ..... 248  
*Abdelbaset Assaf, Lana Issa, and Mohammed Eshtay*

**Author Index** ..... 261



# Automated SLR with a Few Labeled Papers and a Fair Workload Metric

Allan Victor Almeida Faria<sup>1,2</sup> , Maísa Kely de Melo<sup>1,3</sup> , Flávio Augusto R. de Oliveira<sup>1</sup> , Li Weigang<sup>1,2</sup> , and Victor Rafael Rezende Celestino<sup>1,2</sup>  

<sup>1</sup> LAMFO - Laboratory of ML in Finance and Organizations, University of Brasilia  
Campus Darcy Ribeiro, Brasilia, Brazil

maisa.melo@ifmg.edu.br, flaviooliveira@lamfo.unb.br,  
{weigang, vrcelestino}@unb.br

<sup>2</sup> University of Brasilia Campus Darcy Ribeiro, Brasília, Brazil

<sup>3</sup> Federal Institute of Minas Gerais Campus Formiga, Formiga, Brazil  
<http://www.lamfo.unb.br/>

**Abstract.** Citation screening is a crucial stage in conducting a Systematic Literature Review, where reviewers must read hundreds, if not thousands, of papers. Natural Language Processing-based models using Transformers have been successfully employed to automate this process and minimize the chances of missing relevant papers. In our research, we proposed three variations of these Transformer models, each with different pre-training techniques. With our models, reviewers only need to read 16 papers to train the model, thus saving as much as 80% of the workload. In addition, we revisited the AWSS@R metric, which normalized the WSS@R index and provided a fair way to estimate the workload saved using the different datasets.

**Keywords:** Automation of systematic literature review · Few-shot learning · Meta-learning · Transformers

## 1 Introduction

The Systematic Literature Review (SLR) is a key tool for comprehending the status quo of a research area. It is a type of study that summarizes all available data fitting pre-specified criteria to answer precise research questions providing evidence of directions taken in recent years and the next steps indicated by the scientific community [19]. SLR is a means of identifying, evaluating, and synthesizing available research relevant to a particular research question [8]. Citation screening is the stage where reviewers need to read and comprehend hundreds (or thousands) of documents and decide whether or not they should be included in the systematic review [19]. The collection, extraction, and synthesizing of the required data for systematic reviews are known to be highly manual, error-prone, and labor-intensive tasks [9, 19]. The workload involved in the SLR process is enormous and the process is slow, which motivates the effort to automate this process.

Once the SLR process is labor-intensive and subordinated to twelve steps [8], it is worth providing support to automate some of these steps, mainly those that consume the most time. Despite the benefits obtained by the automation of SLR, relatively little research has applied machine learning to this problem [9]. Natural language processing (NLP) has stood out to assist with this automation procedure due to its ability to understand texts and spoken words in much the same way human beings can [15]. To operate the NLP, the Pre-trained language models (PLMs) based on artificial neural networks (ANN) such as Embeddings from Language Model - ELMo [24], or in transformers such as Bidirectional Encoder Representations from Transformers - BERT, [3] have been successful in leading with text classification.

A challenge to automate the citation screening step in an SLR is to save time and miss as few relevant papers as possible in the classification [19]. Then, creating an artificial model that uses as few papers as possible and gets a good classification accuracy is necessary. Aiming this specific task, Melo et al. (2022) [22] developed a Model-Agnostic Meta-Learning (MAML) using Few-Shot Learning for SLR. They proposed an Adapted Work Saved over Sampling (AWSS@R) metric to make a fair comparison using the MAML crossing datasets from different research areas. The success obtained by combining a meta-learning model and a few-shot learning approach comes from the machine learning algorithms' applicability. The mechanism for learning to learn (or meta-learning) should be general to the task and the computation required to complete the task. The model or learner is trained during a meta-learning phase on a set of tasks, such that the trained model can quickly adapt to new tasks using only a small number of examples or trials [11]. Learning from just a few labeled examples while making the best use of a large amount of unlabeled data is a long-standing problem in machine learning [4]. Few-shot learning is well-studied in supervised tasks, where the goal is to learn a new function from only a few input/output pairs for that task, using primary data from similar tasks for meta-learning [10, 11, 32].

This chapter plays around with automating the "Selection of primary studies (citation screening)" of an SLR using a few labeled papers to train the model. The contributions are summarised as follows: We extend the study of Melo et al. (2022) [22] by providing a more robust demonstration for the AWSS metric and conducting a new numerical experiment. In the new numerical experiment, we evaluated the model proposed by Melo et al. (2022) using different PLM encoders subjected to an investigative study of model optimization to understand under which perspectives (referring to parameters model) the model achieves its best performance. We evaluated adapters as a few-shot learner, comparing different pre-trained models concerning a sparse perspective, and proposed a model that achieves a similar performance of AWSS compared to values obtained in [22] but, demanding less processing time. Our method yields significant workload savings. AWSS@R metric scores of up to 0.8 validate our model using 32 datasets; only a few resulted in scores below 0.1. This achievement is meaningful once these results considered only 16 papers to train the model.

Finally, our project is publicly available and open source.<sup>1</sup>

The rest parts of this paper are organized as follows. Section 2 presents the related work and background of the research. Section 3 shows the adopted research methodology, including dataset, model architecture, training framework, and AWSS derivation. Section 4 describes the meta-learning experiments selecting domains with over 50 entries for both included and excluded classifications addressing the issue of domain task imbalance. Section 5 verifies the effectiveness of the suggested model’s approach compared with some literature baseline. Section 6 explores hyperparameters’ impact on the model. Section 7 outlines the main topic achieved in this chapter and asks for ones that need to be studied in more detail. Finally, Sect. 8 presents the conclusion, limitations and future work.

## 2 Related Works

In an SLR, the citation screening step is admittedly the most time-consuming step [1, 8, 27, 30]. As we indicated above, automating the SLR process or part of it is mandatory when dealing with research at the frontier of knowledge. Great performances on this automation have been obtained using a deep learning algorithm combined with NLP. Recently, a deep learning algorithm was applied to automate the citation screening process [17]. van Dinter et al. (2021) [8] presented the first end-to-end solution to citation screening with a deep neural network. Both models claim to yield significant workload savings of at least 10% on most benchmark review datasets [19]. The BERT model and its variants have pushed state-of-the-art for many NLP tasks [7].

BERT is based on a multi-layer bidirectional transformer. Its training is done by conditioning both left and right contexts, simultaneously optimizing for tasks of a masked word and next sentence prediction. BERT-base has an encoder with twelve transformer blocks, twelve self-attention heads, a hidden size of 768, and a maximum input sequence of 512 tokens. A classification head is included with a simple softmax classifier to perform a classification task to return labels’ probabilities [29]. SciBERT is a BERT model specifically optimized to learn scientific text, making it an ideal choice for research tasks in the scientific domain [2]. oBERT is a variant of the BERT model that has undergone pruning to optimize inference while keeping some weights equal to zero [18]. This approach reduces the model’s size and improves inference time without compromising performance.

The Sentence Transformer (ST) is an efficient framework for pre-training a Transformer-based LLM model on pairs of texts in a Siamese manner to achieve a model that can perform semantic textual similarity, as described in (Reimers and Gurevych, 2019) [25]. The resulting model is then used to generate rich text embeddings, which are utilized for training a classification head. The ST utilized in this study is based on the MPNet model [28], and its code and weights are

---

<sup>1</sup> <https://github.com/BecomeAllan/ML-SLRC/tree/main/book>.

available at <https://huggingface.co/sentence-transformers/paraphrase-MPNet-base-v2>.

Finn et al. (2017) [11] proposed a MAML. It is compatible with any model trained with gradient descent and applicable to various learning problems, including classification, regression, and reinforcement learning. An approximation to this MAML can be obtained by ignoring second-order derivatives, using a generalized first-order MAML [23]. Wang et al. (2021) [31] reformulated traditional classification/regression tasks as a textual entailment task. In practice, they proposed an Entailment as a Few-Shot Learner approach that can turn pre-trained language models into better a few-shot learners. The key idea of this approach is to reformulate potential NLP tasks into an entailment one and then fine-tune the model with as few as eight examples for each class.

Melo et al. (2022) [22] created a first-order MAML model with a few-shot learning approach. The main idea behind the model is to train the initial parameters to achieve maximal performance on a new task. This is done by updating the parameters through one or more gradient steps using a small amount of data from the new task. Reviewers only need to label a few papers beforehand due to the model’s reliance on a few-shot learning. Furthermore, the model incorporates an innovative approach that leverages SciBERT to enhance training.

Houlsby et al. (2019) [13] built a system that performs well on all downstream tasks without training an entirely new model for every new task. They proposed a transfer learning strategy that yields compact and extensible downstream models. Their key innovation was designing and integrating a practical adapter module with the base model. To optimize these models for citation screening, the methodology employed in this research paper focused on training only the “Adapters” and the NormLayers while fixing the parameters of the encoder transformers according to [13].

To enhance the performance of the models, one can add a linear layer known as the “Feature map” to the pooling layer output of every encoder transformer. A pooling layer calculates the average of the outputs and produces a single vector as output. Chen et al. (2020) [4] applied the Feature map to unsupervised or self-supervised pretraining models in their research. This study proceeds similarly by using a Feature map technique to adapt an unsupervised model to a few-shot supervised scenario. This additional layer assists the model in improving the ability to extract essential features from the input data, leading to improved accuracy in citation screening tasks. These models, called “SLRC” or “SLR Classifier”, combine a PLM Encoder with a Feed Forward strategy; they were introduced in [22] to deal with classification on Systematic Literature Review. Overall, the methodology employed in this research paper provides an efficient and practical approach to training transformer models for specific NLP tasks about citation screening. By using three transformer models and training only the “Adapters” and the NormLayers while fixing the parameters of the encoder transformers and utilizing four transformer layers per model, researchers can achieve high accuracy in their tasks while minimizing the computational resources required for

training the models, making the approach particularly efficient for researchers with budget constraints or limited computational resources.

Ablation is a process inspired by neuroscience applied to seek a better understanding of neural networks, as its complexity increases and the model explainability becomes an open question [33]. Usually, an ablation study investigates the model performance sensitivity to small changes (*ceteris paribus*), for example, by removing some components in order to evaluate the contribution of such model components. In order to pass an ablation study, the model should show resilient performance to these changes. An ablation study resembles other procedures, such as hyperparameter tuning and pruning, but it seeks more explainability than enhanced training or performance. While hyperparameter tuning aims at better performance, pruning systematically removes model parameters to speed up training and inference, with similar performance [12]. Here, a customized ablation study methodology is proposed based on benchmarking of multiple experiments by different word vectorization preprocessing with SciBERT [2], oBERT [18], ST MPNet [25, 28], and selected training strategies, maintaining the same datasets, loss functions, and optimizers fixed [16].

This chapter presents an extension to the study conducted by Melo et al. (2022) [22], incorporating the use of adapters in the PLM Encoders proposed by [13] for efficient training, using the same model architecture with different backbones called as skulls and the same dataset of the experiments. The three different PLM Encoders, namely oBERT, SciBERT, and ST MPNet, were utilized in an ablation study to assess the effectiveness of the embedding paradigm of each model. For the ablation study, we considered different combinations of parameters to optimize the model concerning the number of Transformer layers (2/4/6/12), the number of inner epochs (4/6/16), and the number of examples to train in the domain phase (2/4/8/16).

## 3 Methodology

### 3.1 Dataset

To achieve a broad range of predictability, this chapter uses a domain-agnostic datasets proposed by Melo et al. (2022), consisting of data from 64 topics in SLRs as listed in Table 2 in [22]. According to van Dinter et al. (2021c) [8], most studies rely on domain-specific document metadata from the medical research field. However, our approach is unique in that we utilize data from research fields other than medical, such as the ASReview Project [26], Sciome Workbench for Interactive computer-Facilitated Textmining (SWIFT-Review) [14], and Cereals and Leafy Greens [3].

To create the dataset, we reduced the selected data to include only titles, abstracts, and labels that indicate whether a paper is included or excluded based on the revision criteria. For clarity, we use the terms “included” and “excluded” to refer to the papers classified as such in the SLR. By utilizing a diverse range of research fields and reducing the data to only the most relevant information, we believe that the dataset offers a robust and reliable basis for conducting research



in natural language processing tasks, such as citation screening, across a wide range of domains.

### 3.2 Model Architecture

The architecture methodology employed uses three transformer models: oBERT, SciBERT, and ST MPNet. Despite their differences, all three models used in this work have the same architecture of an encoder transformer, which is commonly used in natural language processing tasks. The “Adapters” architecture [13], consisting of a Feedforward down-project to 64 outputs, Nonlinearity layer, and Feedforward up-project, was introduced inside each transformer layer before all the NormLayer. This approach reduced the number of trainable weights while retaining essential language knowledge inside the attention heads, making it particularly efficient for researchers with budget constraints or limited computational resources.

To further reduce the computational resources required to train the models, we also reduced the number of transformer layers per model to four, similar to [22]. Despite the reduced number of transformer layers, using the “Adapters” architecture ensures that the models retain their ability to extract essential language features from the input data.

We also added a feature map to the output of the pooling layer for each encoder transformer. It outputs to a vector of 200 dimensions, followed by the hyperbolic tangent activation function. Finally, a linear layer with only one output with the sigmoid activation function is applied.

In general, our model architecture is inspired by the work [22] using the “Adapters” [13] technique. This approach offers an efficient and effective solution for researchers facing constraints on computational resources or budgets when performing natural language processing tasks such as citation screening. In total, the backbones have 53M of parameters, but only 156k are trainable and efficiently shareable for other researchers to have the same model skills.

### 3.3 Training Framework

Class imbalance is challenging for a model to train. Researchers’ most reoccurring challenge is class imbalance [9]. Regarding this work’s classification problem, class imbalance refers to unbalance between the amount of included and excluded papers. Generally, the amount of excluded articles is more considerable than included ones. Melo et al. (2022) [22] discussed the issue of severe data imbalance in classification tasks across various domains and presented a Meta-learning framework called ML-SLRC, which aims to address this challenge using a few-shot learning. The framework trains the SLRC model in the meta-learning phase using a small number of balanced domain data examples for each task, allowing the model to perform classification tasks in a given domain with minimal examples. The N-way-F-shots method is employed in the training process, where a support set consisting of batches of tasks is used to train the model, and multiple query batches are used to evaluate its performance [11].

The proposed framework follows the MAML procedure leveraged by Reptile Algorithm [23] to handle the computational challenges arising from the large number of parameters in the PLM. The training process adopts a 2-way-8-shots approach, with each batch of tasks used to train the model to learn a new specific task  $S_{k+1}$  with a few labeled data and infer on the  $k + 1$  on the unlabeled data task domain. Figure 2 presented in [22] illustrates the proposed meta-learner framework, which seeks to build a task training framework capable of handling imbalanced data, such as the SLR datasets, and producing a domain-independent Task Learner across various classification tasks.

### 3.4 WSS and AWSS Derivation

When dealing with SLR automation problems, especially those that involve optimizing the selection of primary studies, the most used metrics to evaluate model performance are precision, recall, and the F1-score [9]. However, these traditional metrics need to be revised for learner evaluation as they do not indicate how much effort was spared for the researcher by using the learner [19]. Furthermore, when creating a learner to optimize the literature review process, the cost of failing to detect relevant new literature is high, and such high recall is demanded [5]. These demands led to the creation of the WSS@R metric [5], defined as the fraction of work saved at a specific recall rate. The WSS@R score measures the percentage of papers that meet the original search criteria that the reviewers do not have to read (because the classifier has screened them out). Using automatic citation classification effectively reduces the workload of preparation of the systematic review.

The WSS@R score scale depends on the variation in sample imbalance that can be present in the SLR. In the context of the MAML (multi-task model), evaluating the model by testing it on different domain data with varying class imbalances is crucial to assess its impartiality. Normalizing the results is necessary to enable cross-comparison across different modalities. Melo et al. (2022) [22] proposed an alternative normalized metric, AWSS, which we are revisiting and providing a more rigorous mathematical definition. The requirement for a normalized metric was also brought out by Kusa et al. (2023) [20], who proposed to normalize WSS using min-max normalization.

The requirement for a normalized metric was also brought out by Kusa et al. (2023) [20], who proposed to normalize WSS using min-max normalization.

Let  $x_k$  represent a single paper randomly sampled from retrieved documents  $(x_1, \dots, x_n)$ . Each paper is labeled as included or excluded in the SLR process. Consider  $y_k$  indicating the labeling related to  $x_k$ : included ( $y = 1$ ) or excluded ( $y = 0$ ). So, it can be stated:

**Data:**  $D = \{(x_k, y_k)\}_{k=1, \dots, n}$ ,  
**Prediction model:**  $f_{\theta}(x_k) = \hat{y}_k$ ,  
**Labeling variables:**  $y, \hat{y} \in \{0, 1\}$ .

The probability of a prediction can be stated as

$$\mathbf{p}_{\hat{Y}|Y}(\hat{y}|y) := P(\hat{Y} = \hat{y}|Y = y, X = x),$$

where  $\mathbf{p}_{\hat{Y}|Y}(\hat{y}|y)$  expresses the conditional probability of an estimated value  $\hat{Y} = \hat{y} = f_{\theta}(x)$ , given the true label  $Y = y$ , for a randomly chosen example  $(x, y)$ . This randomly chosen example  $(x, y)$  also can be regarded as a random variable function of the random variable  $X$ , where  $f_{\theta}(X) = \hat{Y} \sim \text{Bernoulli}(p_{\theta})$ , being  $p_{\theta}$  an implicit parameter and  $\theta$  the model learned parameter.

Let  $p$  represent the included proportion in the sample. So,  $Y$  can be regarded as a random variable  $Y \sim \text{Bernoulli}(p)$ . Therefore,  $p$  is defined as

$$p = \frac{P}{N + P},$$

where  $P$  is the number of included, and  $N$  is the number of excluded in the sample. Therefore, the conditional probabilities can be derived from

$$\mathbf{p}_{\hat{Y}|Y}(0|0) = \mathbf{p}(0|0) = TN\% = \frac{TN}{N},$$

$$\mathbf{p}_{\hat{Y}|Y}(1|1) = \mathbf{p}(1|1) = TP\% = \frac{TP}{P},$$

where  $TP$  means the true positives and  $TN$  the true negatives.

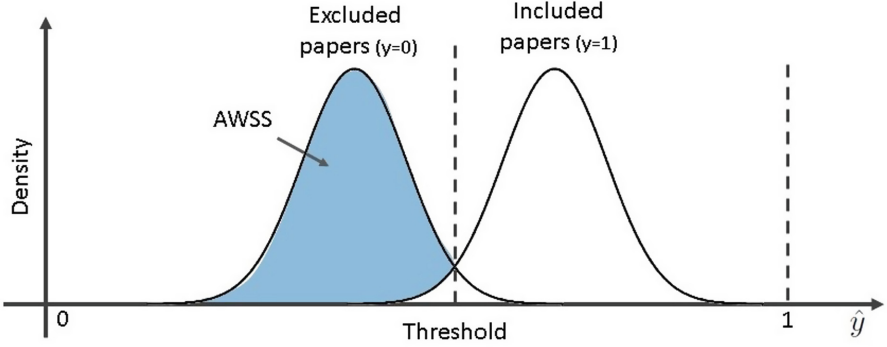
As proposed in [22], the Adjusted Work Saved Oversampling (AWSS) can be expressed as

$$AWSS = TN\% - (1 - TP\%) = \mathbf{p}(0|0) - (1 - \mathbf{p}(1|1)) = \mathbf{p}(0|0) - \mathbf{p}(0|1),$$

where  $\mathbf{p}(0|0)$  is the precision of excluded papers and  $\mathbf{p}(0|1)$  is the recall of false negatives. To force a 95% recall, we can set the threshold of the model and the AWSS results in

$$AWSS@95\% = \mathbf{p}(0|0) - 5\%.$$

The AWSS can be regarded as a statistical test significance over the sample's distribution of  $\hat{Y}$  given by  $Y$  about the included examples over the excluded examples. The  $\mathbf{p}(0|0)$  and  $\mathbf{p}(0|1)$  are the power test and error type 1, respectively. The  $AWSS@95\%$  metric evaluates how well the model rejects the excluded examples ( $N$ ), admitting 5% of the error of included examples over  $\hat{Y}$  given by  $Y$ . Figure 1 displays the hypothesis test considering the sample and the classification of each paper ( $\hat{Y}|Y$ ). The  $AWS@95\%$  represents the blue area of sample excluded papers ( $Y = 0$ ), that a certain threshold makes the proportion of predicted excluded samples ( $\hat{Y} = 0$ ) admit to missing 5% of the included sample ( $Y = 1$ ), as a measure of saving time about the proportion of selected papers not to be read by the reader. This figure is a representation only; the curves are not necessarily Gaussian distributions.



**Fig. 1.** Sample distribution over  $\hat{Y}|Y$  considering included and excluded papers. The blue area represents the power test, which is the AWSS metric in this case.

Once revisited AWSS definition, let us deduct its relationship to the original metrics, work saved oversampling (WSS) [6]. Derived as

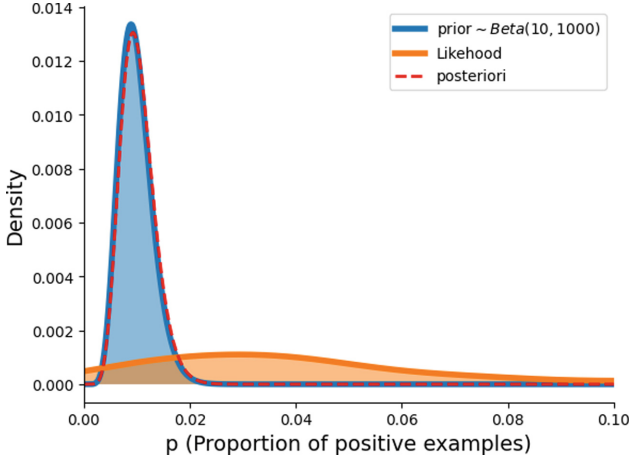
$$\begin{aligned}
 WSS &= \frac{TN}{N+P} + \frac{FN}{N+P} - (1 - TP\%) \\
 &= \mathbf{p}(0|0) \frac{N}{N+P} + \mathbf{p}(0|1) \frac{P}{P+N} - (1 - \mathbf{p}(1|1)) \\
 &= \mathbf{p}(0|0) \frac{N}{N+P} + \mathbf{p}(0|1) \frac{P}{P+N} - \mathbf{p}(0|1) \\
 &= \mathbf{p}(0|0) \frac{N}{N+P} + \mathbf{p}(0|1) \left( \frac{P}{P+N} - 1 \right) \\
 &= \mathbf{p}(0|0) \frac{N}{N+P} - \mathbf{p}(0|1) \frac{N}{P+N} \\
 &= \frac{N}{N+P} (\mathbf{p}(0|0) - \mathbf{p}(0|1)) \\
 &= (1 - p) (\mathbf{p}(0|0) - \mathbf{p}(0|1)) \\
 &= (1 - p) (AWSS)
 \end{aligned}$$

Based on above algebraic deduction, we obtain

$$AWSS = WSS \frac{1}{1 - p} = WSS \frac{N + P}{N},$$

which, as a preliminary finding, it can be stated that the relationship between WSS and AWSS varies with the proportion parameter  $p$  of the sample. If  $p \approx 0$ , then  $WSS \approx AWSS$ . If  $0 < p \leq 1$ , WSS is the relative assessment of model performance at the respective sample proportion (what were the proportion examples excluded given by the particular sample distribution?). In particular, Fig. 2 shows that  $p$  distribution among the 64 databases considered in this work as the orange line, with a maximum likelihood ( $\arg \max_{p \in (0,1)} \mathbf{p}(D|p)$ ) for the  $p$  estimated to 0.03. With a frequentist approach, it does not necessarily follows that  $p \approx 0$  for the relatively small sample of databases.

However, within the Bayesian framework,  $D$  is a group of SLR database datasets, and  $p$  is the proportion of included examples. The distribution of the parameter  $p$  is a random quantity, with a prior distribution  $\mathbf{p}(p)$ , from which it can be obtained the posterior distribution  $\mathbf{p}(p|D)$  via Bayes Theorem:



**Fig. 2.** Distributions of the proportion of the positive examples in a SLR considering the sample of 64 databases used in this study.

$$\mathbf{p}(p|D) = \frac{\mathbf{p}(D|p)\mathbf{p}(p)}{\mathbf{p}(D)}.$$

Now, suppose that  $D = \{D^{(1)}, \dots, D^{(n)}\}$  is a random sample, where  $\{y_1^{(k)}, \dots, y_{m_k}^{(k)}\} \in D^{(k)}$ , and the likelihood density ( $D|p$ ) follows some distribution estimated by gaussian kernel. Then, it is reasonable to assume that the prior distribution  $\mathbf{p}(p) \sim \text{Beta}(\alpha = 10, \beta = 1000)$ , where  $\alpha, \beta$  values are assumed by the authors. Intuitively,  $\alpha - 1$  also can be expressed as the number of positives examples, and  $\beta - 1$  is the number of negatives examples in an SLR. Figure 2 shows the prior Beta distribution of  $\mathbf{p}(p)$ , the likelihood of proportion of positive examples in the 64 datasets as an orange line, and the posterior resulted in distribution as a red dashed line. Since the posteriori and priori distributions are similar, shown in Fig. 2, a convergence to  $\beta \gg \alpha$  can be expected if the number of available publications increases.

**Table 1.** Statistical measures of the distributions.

Likelihood			Posteriori		
arg max	$Q_{95\%}$	$E(D p)$	arg max	$Q_{95\%}$	$E(p D)$
0.0292	0.00064	0.0001	0.0093	$\approx 0$	0.0001

Assuming the prior distribution over  $p$ , we can derive a posterior distribution  $p|D$ , and the expected value is given by Table 1, where  $E(p|D)$  can be used to estimate the proportion of included examples ( $p$ ). With the posterior distribution, 95% of the values of this distribution are close to zero ( $Q_{95\%} \approx 0$ ). So in this

database  $D$ , there is a 95% credibility interval for assuming that  $WSS \approx AWSS$  as obtained in [22]. Now we can apply the AWSS in different SLR datasets scenarios to evaluate the model’s performance, more than the specific distribution of the imbalanced data.

## 4 Split 50-50 Experiment

Meta-learning is a promising area that involves training models that can learn from a set of tasks and generalize that knowledge to new tasks, allowing for rapid adaptation and faster learning. One important consideration in meta-learning is domain task imbalance, which can occur when the proportion of included and excluded labels varies significantly between tasks. The class imbalance can impact the ability of the meta-learner to perform generically, and it is crucial to address this issue during the training and validation phase.

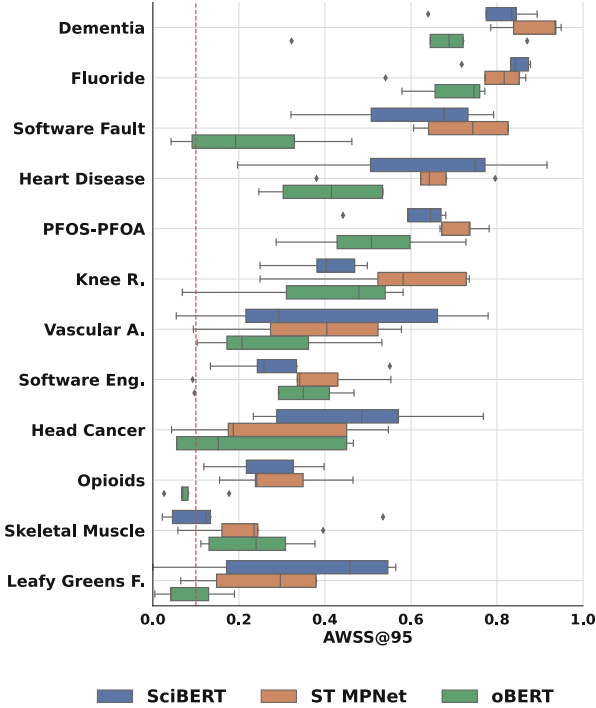
To address the issue of domain task imbalance, Melo et al. (2022) suggest selecting domains with over 50 entries for both included and excluded classifications from creating a training task set for the meta-learning phase [22]. During this phase, the learner is trained on six inner epochs and ten outer epochs, where a new batch of tasks containing 20 tasks with 16 examples (8 examples per class) for each task is randomly initialized for each outer epoch. The number 16 was empirically obtained. The inner phase is set to create batches of four examples to learn the task, while the outer phase is set to have five tasks to learn tasks. To compute the loss, the weight of included examples is set to 1.5 to retrieve more recall, and the learning rate of the inner and outer update step is set to  $5 \times 10^{-6}$  and  $5 \times 10^{-5}$  respectively. The learning rate is set to  $5 \times 10^{-3}$  during the domain learner phase. These strategies show promising results in improving the performance of meta-learner models and may help overcome task imbalance challenges.

Figure 3 presents the AWSS values for the oBERT, ST MPNet, and SciBERT models in the 50–50 split datasets. As can be seen, the AWSS values obtained by oBERT are smaller than those obtained by ST MPNet and SciBERT. This behavior is expected since oBERT works with BERT in a sparsing way, with multiple weights equal to zero. The most relevant contribution expected for oBERT concerns the inference time using the DeepSparse software of Neural Magic<sup>2</sup>.

SciBERT and ST MPNet have different pre-training principles. ST MPNet creates well-defined representative spaces where similar and dissimilar texts are allocated separately. On the other hand, SciBERT is trained to predict the word in the sentence or the following sentence in the phrase in scientific texts. Pre-training plays a crucial role in determining the model’s performance. As shown in Fig. 3, the ST MPNet generally achieved better AWSS values than SciBERT, indicating that its pre-training strategy is more suitable for this SLRC problem. Although SciBERT was trained on scientific texts from the Web of Science and Semantic Scholar databases, its pre-training mechanism is less impressive than

<sup>2</sup> <https://neuralmagic.com/deepsparse/>.

the sentence transformer pre-training strategy, even though on average, it has higher values in AWSS as in Table 2. The red dashed line in the Fig. 3 refers to the minimal expected value, 0.1, for the AWSS.



**Fig. 3.** Boxplot of five trials of the 50–50 split experiment on the validation dataset.

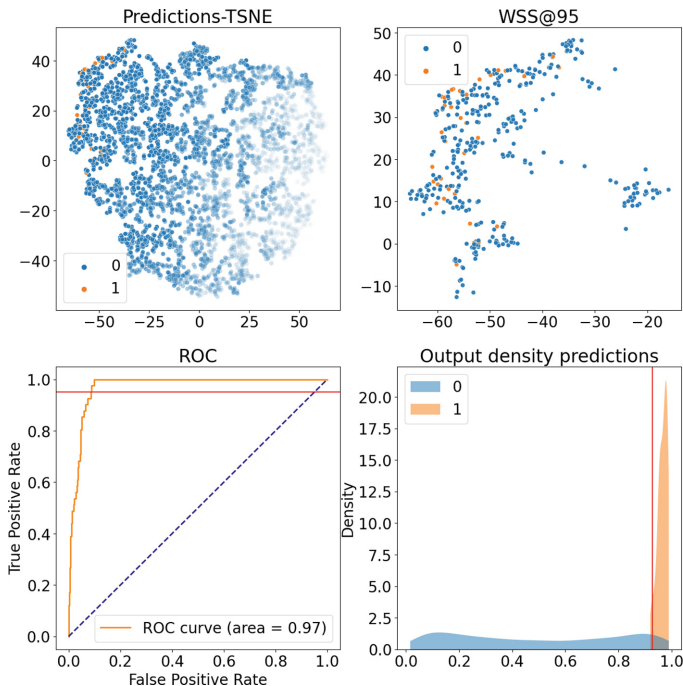
We desire to draw particular attention to the contribution and importance of the AWSS metric in constructing Fig. 3. This image shows the performance of the models considering several datasets, each presenting a different class imbalance arrangement. The traditional WSS metric condenses the information on the percentage of papers saved from reading into a single coordinate axis. Using the WSS metric here is not possible due to data imbalance, making the comparison unfair. However, with the new AWSS metric, which treats data in a normalized pattern, it is always possible to compare the saved work information, regardless of the database size and class imbalance.

To showcase the predictions of a learner who has been trained using 16 examples in a specific domain, Fig. 4 shows the ML-SLRC ST MPNet model outputs on the Fluoride dataset. The feature map layer output was transformed into two dimensions using the t-SNE technique [21] and presented as dispersion points. The confidence level of each point as opacity was determined using the sigmoid-activated classifier layer output. The blue color denoted examples excluded (0),

while orange represented those included (1), which were the examples’ true labels. This figure also shows the ROC curve and the Output density for the last layer of the model for this attempt on the test dataset on Fluoride. The red line is the recall at 95%, confirming the learner’s reasonable confidence in classifying the papers and reducing the workload of reading the example papers. In contrast, we can evaluate the WSS@95% plot, which shows the examples as points, separated by the model using t-SNE to be read, and the AWS@95% as the density curve in the (d) plot to explore the density margin of excluded papers before the red line by the model.

## 5 Benchmarks Experiment

To verify the effectiveness of our suggested models approach, we conducted a comparative analysis with an existing literature baseline [19]. For the ML-SLRC models, we used the exact baseline domains as test tasks with the learning rate as  $5 \times 10^{-3}$  during the domain learner phase, while for training the ML-SLRC models, we used the remaining tasks data of the 64 domains present in this work on the meta learner phase with the same training configuration of the “Split 50–50 validation”.



**Fig. 4.** ML-SLRC ST MPNet prediction in the test data of Fluoride training with eight examples for each class. Test data has 3846 examples (3805 excluded and 41 included).



Figure 5 is a reference point for analyzing the performance of different models used in the experiments. The results from 15 databases show that the metric AWSS@95% is generally above the threshold of 0.1 (indicated by the red reference line). However, only a handful of databases, namely Antipsychotics, Beta Blockers, Hypoglycemics, Proton Pump, and Calcium C., are underperforming. These databases suggest that some intrinsic characteristics, such as more complex structures or variables, can make predictions of similar data problematic when not using specialized models.

Table 3 presents the AWSS@95% and WSS@95% values following their respective standard deviation considering five validations of the models. The ST MPNet model exhibited the best performance, even though SciBERT obtained higher AWSS@95 values on average. The oBERT performed relatively worse than the others. Comparison of the overall average of AWSS@95% values from Table 3 with the results of the ML-SLRC SciBERT model without the adapter presented in [22] shows that the adapter-equipped ML-SLRC models perform similarly. That is an excellent achievement for the proposed model using adapters. The advantage of using the adapter lies specially in the number of trained parameters, demonstrating the efficiency and speed in training these models.

## 6 Ablation Experiments

Within NLP, the Transformer architecture has revolutionized various tasks such as machine translation, sentiment analysis, and language modeling. However, given the complexity of the architecture and pre-training framework, it can be challenging to ascertain which variables have the most significant impact on resulting outcomes. By evaluating outcomes sensitivity to different training schemes and manipulation of hyperparameters, ablation studies investigate which arrangement of variables favors the model most.

This chapter delves into three types of Transformers encoders: SciBERT, oBERT, and ST MPNet. These distinct architectures differ in their weights and how they were subsequently trained to be tuned for text classification. SciBERT is a post-training model of BERT tailored for scientific text. In contrast, oBERT is a post-training model of BERT aimed towards being a sparse model, and ST MPNet is a post-training model of MPNet designed to group similar text sentences.

Despite the differences in their post-training models, we validate them using Meta-Learning training (ML-SLRC) and without Meta-Learning (SLRC). The PFOS-PFOA, Fluoride, Software Eng., and Opioids datasets were used to conduct the validation, also exploring the following three hyperparameters:

- Number of Transformer layers: 2, 4, 6, 12
- Number of inner epochs (learning to classify): 4, 8, 16
- Number of training examples: 2, 4, 8, 16

In exploring the hyperparameters of the number of training examples and inner epochs, the models used for ML-SLRC resulted from the Split 50–50 validation section. For the number of transforming layers, the ML-SLRC models

use the same training setup as the 50–50 Split validation to train new models for this ablation section.

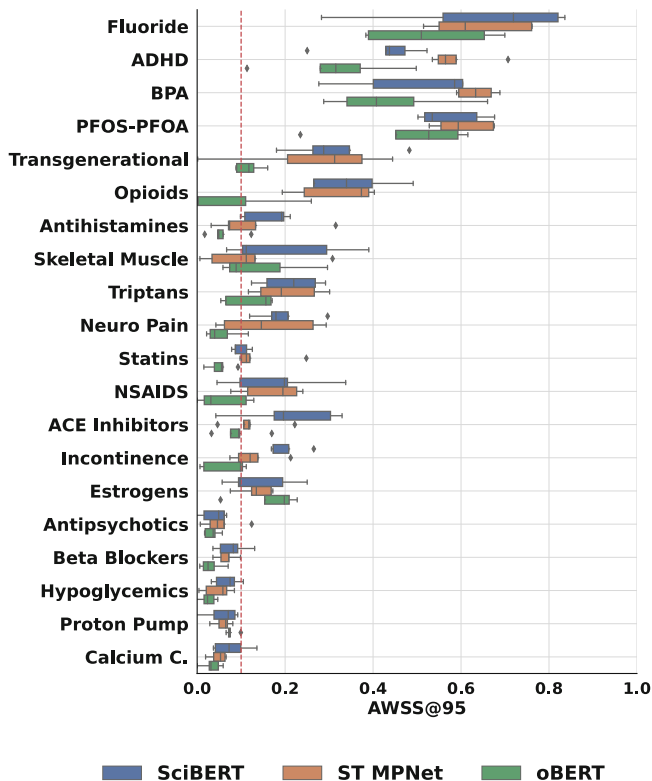


Fig. 5. Boxplot of five trials of the benchmark experiment on the validation dataset.

## 6.1 Transformer Layers

To explore the number of Transformers layers, Fig. 6 resumes five attempts for each model. There is a clear relationship between layer\_size and consistency in the AWSS@95% metric across validated datasets. Increasing the number of Transformers Layers results in greater values and low dispersions of AWSS@95% across the attempts. Additionally, when using the ML-SLRC model, there is less variance than the SLRC model. However, there is little difference between the results regarding the central tendency when using either the ML-SLRC or the SLRC model. These findings suggest that the choice between these models may depend on specific use cases and priorities of datasets topics.

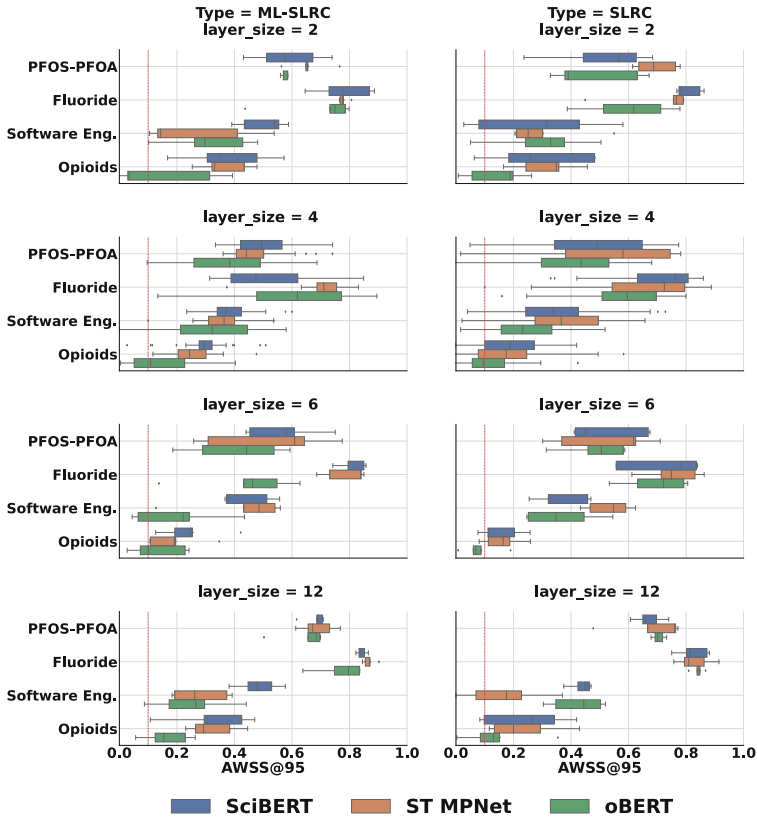


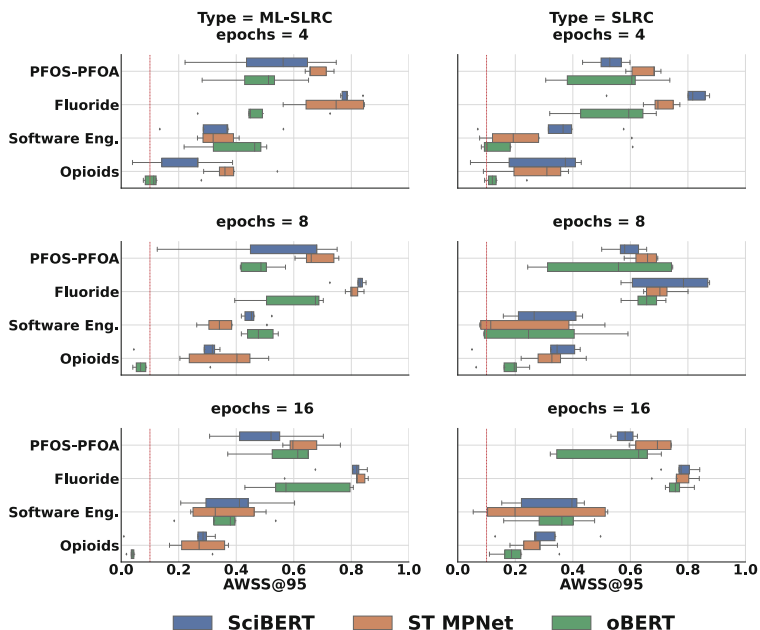
Fig. 6. Comparison of the AWSS concerning the number of Transformers layers on five attempts for each model.

### 6.2 Number of Inner Epochs

For the number of epochs required to train the few-shot model in the domain learning phase, there appears to be no clear and consistent pattern for the ML-SLRC model when analyzed using the AWSS@95% metric. However, for the SLRC model, there is a clear trend of improvement in this metric with an increase in the number of epochs. This is further supported by the results shown in Fig. 7.

### 6.3 Number of Training Examples

Figure 8 shows that the quantity of training examples is essential to consider in the domain learning phase. Our study shows that the ML-SLRC and SLRC models consistently improve the AWSS@95% metric with increased training examples. This finding underscores the significance of training data in achieving better performance in learning the task.



**Fig. 7.** Comparison of the AWSS concerning the number of epochs on five attempts for each model.

## 7 Discussion

In this study, we examined whether the proposed transformers models can effectively assist in reducing the number of articles that need to be read in an SLR. Our analysis revealed that the models SciBERT and ST MPNet demonstrated a significant workload reduction, achieving AWSS@95% values greater than 0.1 in most databases, as desired. In some cases, they achieved the incredible value of AWSS@95% = 0.89; in others, the desired minimum value of AWSS@95% was not reached.

For the example of the Benchmark experiment, the Oral Hypoglycemics (352 negatives versus 138 positives) and Neuropain (24,193 negatives versus 5,009 positive) datasets show that the models face difficulties in training some specific databases. The first one has no drastic class imbalances but is a relatively small dataset. The latter has a more aggressive class imbalance and is a relatively large dataset. One of the datasets that the models compromised best was Fluorine (4428 negatives versus 51 positives). It has a drastic class imbalance and is smaller than Neuropain. As these datasets address specific topics, they have precise and rare words, making it difficult to generalize the model in classification. So why didn't hypoglycemics and neuropain perform as well as the others? We keep this discussion open. We suggest an investigative study to answer these questions in future work.

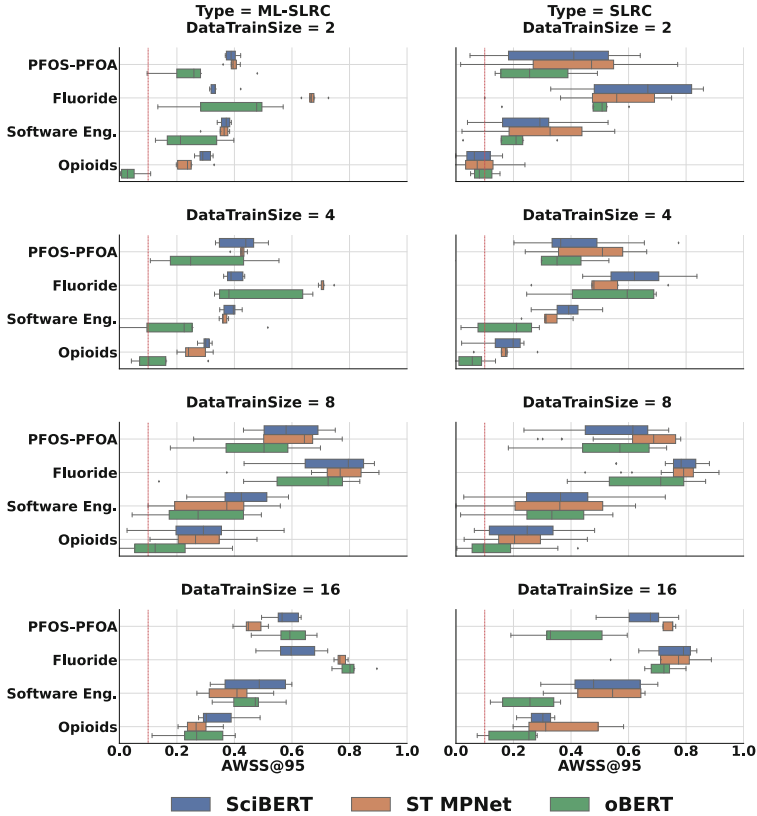


Fig. 8. Comparison of the AWSS concerning the number of training examples on five attempts for each model.

In the “50–50 split” experiment, the proposed datasets performed well on the AWSS@95% metric with the ML-SLRC meta-learning configuration. However, the oBERT backbone for ML-SLRC performed poorly on the metric, as expected when compared to SciBERT and ST MPNet. The experiments that oBERT performs worse than other models, such as SciBERT and ST MPNet in the AWSS@95% metric, can be justified as oBERT is indeed a sparse model [18]. On the other hand, it is important to highlight that the oBERT model can make predictions faster due to some of its parameters being zeros. Among the models, ST MPNet stands out in the metric due to its pre-training contribution [25] in naturally grouping similar sentences close to each other. We suggest an investigative study comparing the inference speed performance and energy efficiency among dense models on GPUs and sparse models using Neural Magic inference software on CPUs.

The ablation study found that ML-SLRC models using only trainable adapters [13] have similar results to SLRC models. It is observed that the models present greater consistency in the AWSS@95 metric with the increase in the number of epochs and training data, mainly for the SLRC models. However, for ML-SLRC models, increasing the number of epochs does not necessarily improve performance. The ablation study stated that models with more transformer layers have more consistency in the AWSS@95% metric, but there is also a slight increase in value. The models benefit more from the quality of the examples given as training data rather than the number of transformer layers.

The AWSS metric played a crucial role in analyzing the models' performance, as it relatively evaluates the model's abilities in its respective database, disregarding the data distribution. In contrast, the WSS metric provides the exact size in percentage of the excluded examples in the respective dataset. Without the consideration of AWSS, evaluating models in different databases would have been a challenge. This metric allowed a fair and accurate evaluation of the models, regardless of data distribution, making it a valuable tool in performance analysis. The AWSS is an evaluative metric of clustering performance among the classes. Here, it is the evaluation of the excluded cluster over the included. It can be extendable to multiple classes in the one-vs-all fashion. Implementing the AWSS metric in studies beyond the SLR spectrum but also demands measuring the work saved may provide a better understanding of model features and lead to better model development.

With the adapters [13], the researcher can download only the adapter's weights (just a few tens of MB) and plug in the base model to have a custom model for specific SLR tasks. It is the first step to have a custom search for the citation screening and leverage these results to a generative language model to bring insights about the citations or the main research question.

Overall, this study provides valuable insights into the performance and capability of different models in the given context. The findings suggest that these models help identify relevant SLR examples. In particular, the SLRC ST MPNet model has advantages due to its pre-training benefits in the Sentence Transformers framework. The authors explore efficient methods to train an LLM even with a scarcity of labeled data and imbalanced datasets. This study can benefit researchers aiming to automate the citation screening process with constrained budgets.

## 8 Final Considerations

This work investigates the efficient training of LLM models for Systematic Literature Reviews. It compares techniques such as first-order approximation of MAML for Meta-Learning and adapters to achieve a few-shot learning. Our findings have shown that LLMs, specifically models based on Sentence Transformers, can be a practical tool for automating citation screening. We demonstrate the usefulness of adapters in facilitating the disclosure of model parameters. Our

study is a starting point for developing an efficient citation tool to automate systematic literature reviews. Additionally, we confirm that the AWSS metric consistently evaluates the model’s ability across different databases without being limited by imbalanced data.

A limitation of the current work is that ML-SLRC models with adapters have limited flexibility to achieve better results in the domain learning phase. Furthermore, including a time metric to evaluate the quantitative gain between models is essential, especially since SLRC models using adapters efficiently learn a task with a few examples. Future research should explore the potential of sparse models like oBERT to be as efficient as sentence transformers in conducting SLR searches. Furthermore, it is crucial to evaluate the gain in speed and energy efficiency in inference between different SLRC models, including dense and sparse ones, using Neural Magic’s DeepSparse software or similar, as this software improves computation on CPUs.

**Acknowledgement.** We sincerely thank the Brazilian Ministry of Science, Technology, and Innovation, which partially supported this project.

## Appendix

**Table 2.** Summary of the mean and std. deviation of five validations, considering 16 examples (eight positive and eight negative) in the domain learner phase, after training the respective ML-SLRC in the meta learner phase (50–50 split).

Dataset	SciBERT		ST MPNet		oBERT	
	AWSS@95%	WSS@95%	AWSS@95%	WSS@95%	AWSS@95%	WSS@95%
Dementia	0.8(0.1)	0.8(0.1)	<b>0.89(0.07)</b>	0.89(0.07)	0.65(0.2)	0.65(0.2)
Fluoride	<b>0.83(0.06)</b>	0.82(0.06)	0.77(0.13)	0.76(0.13)	0.7(0.08)	0.7(0.08)
Head Cancer	<b>0.47(0.22)</b>	0.43(0.2)	0.28(0.21)	0.26(0.19)	0.24(0.21)	0.22(0.19)
Software Eng	0.3(0.16)	0.3(0.15)	<b>0.35(0.17)</b>	0.34(0.17)	0.32(0.14)	0.32(0.14)
Leafy Greens F	<b>0.35(0.25)</b>	0.11(0.08)	0.25(0.14)	0.08(0.04)	0.08(0.08)	0.03(0.02)
Opioids	0.28(0.11)	0.27(0.11)	<b>0.29(0.12)</b>	0.28(0.12)	0.08(0.06)	0.08(0.05)
PFOS-PFOA	0.61(0.1)	0.6(0.1)	<b>0.72(0.05)</b>	0.71(0.05)	0.51(0.17)	0.5(0.16)
Software Fault	0.61(0.19)	0.6(0.19)	<b>0.73(0.1)</b>	0.72(0.1)	0.22(0.17)	0.22(0.17)
Skeletal Muscle	0.17(0.21)	0.17(0.21)	0.22(0.12)	0.22(0.12)	<b>0.23(0.11)</b>	0.23(0.11)
Knee R	0.4(0.1)	0.38(0.09)	<b>0.56(0.2)</b>	0.53(0.19)	0.4(0.21)	0.37(0.2)
Vascular A	<b>0.4(0.31)</b>	0.39(0.3)	0.37(0.2)	0.37(0.19)	0.28(0.17)	0.27(0.17)
Heart Disease	<b>0.63(0.28)</b>	0.63(0.28)	0.62(0.15)	0.62(0.15)	0.41(0.13)	0.41(0.13)
Avg	0.49(0.26)	0.46(0.28)	0.50(0.26)	0.48(0.27)	0.34(0.26)	0.33(0.23)

**Table 3.** Summary of the mean and std. deviation of five validations, considering 16 examples (eight positive and eight negative) in the domain learner phase, after training the respective ML-SLRC in the meta learner phase (benchmarking).

Dataset	SciBERT		ST MPNet		oBERT	
	AWSS@95%	WSS@95%	AWSS@95%	WSS@95%	AWSS@95%	WSS@95%
ACE Inhibitors	<b>0.21(0.11)</b>	0.19(0.11)	0.12(0.06)	0.11(0.06)	0.09(0.05)	0.09(0.05)
ADHD	0.42(0.1)	0.38(0.09)	<b>0.59(0.07)</b>	0.54(0.06)	0.32(0.14)	0.29(0.13)
Antihistamines	<b>0.16(0.05)</b>	0.12(0.04)	0.12(0.11)	0.09(0.08)	0.06(0.04)	0.04(0.03)
Antipsychotics	0.04(0.03)	0.02(0.02)	<b>0.05(0.04)</b>	0.04(0.03)	0.03(0.02)	0.02(0.01)
BetaBlockers	<b>0.08(0.04)</b>	0.07(0.03)	0.07(0.02)	0.06(0.02)	0.03(0.03)	0.03(0.02)
Calcium C	<b>0.08(0.04)</b>	0.06(0.03)	0.05(0.02)	0.04(0.02)	0.03(0.02)	0.03(0.02)
Estrogens	0.14(0.08)	0.11(0.06)	0.13(0.04)	0.11(0.03)	<b>0.17(0.07)</b>	0.13(0.06)
NSAIDS	<b>0.18(0.11)</b>	0.14(0.09)	0.17(0.07)	0.13(0.06)	0.06(0.06)	0.05(0.05)
Opioids	0.3(0.19)	0.29(0.19)	<b>0.32(0.1)</b>	0.31(0.09)	0.07(0.12)	0.07(0.11)
Hypoglycemics	<b>0.07(0.03)</b>	0.05(0.02)	0.05(0.03)	0.03(0.02)	0.02(0.02)	0.02(0.02)
Proton Pump	0.06(0.04)	0.05(0.03)	0.06(0.02)	0.05(0.02)	<b>0.08(0.01)</b>	0.06(0.01)
Skeletal Muscle	<b>0.19(0.14)</b>	0.19(0.14)	0.12(0.12)	0.12(0.12)	0.14(0.1)	0.14(0.1)
Statins	0.1(0.02)	0.09(0.02)	<b>0.14(0.06)</b>	0.13(0.06)	0.05(0.03)	0.05(0.03)
Triptans	<b>0.21(0.07)</b>	0.14(0.05)	0.2(0.08)	0.14(0.05)	0.12(0.06)	0.08(0.04)
Incontinence	<b>0.2(0.04)</b>	0.15(0.03)	0.13(0.05)	0.1(0.04)	0.07(0.05)	0.05(0.04)
BPA	0.49(0.15)	0.49(0.15)	<b>0.63(0.04)</b>	0.63(0.04)	0.44(0.15)	0.43(0.14)
Fluoride	<b>0.64(0.23)</b>	0.64(0.23)	0.64(0.12)	0.63(0.12)	0.53(0.15)	0.52(0.14)
Neuro Pain	<b>0.19(0.07)</b>	0.16(0.05)	0.16(0.11)	0.13(0.09)	0.06(0.04)	0.05(0.03)
PFOS-PFOA	0.57(0.08)	0.57(0.08)	<b>0.61(0.07)</b>	0.6(0.07)	0.48(0.15)	0.48(0.15)
Transgenerational	<b>0.31(0.11)</b>	0.31(0.11)	0.27(0.19)	0.26(0.19)	0.12(0.03)	0.12(0.03)
Avg	0.23(0.20)	0.21(0.20)	0.23(0.22)	0.21(0.22)	0.15(0.17)	0.14(0.17)

## References

1. Bannach-Brown, A., et al.: Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. *System. Rev.* **8**(1), 1–12 (2019). <https://doi.org/10.1186/s13643-019-0942-7>
2. Beltagy, I., Cohan, A., Lo, K.: Scibert: pretrained contextualized embeddings for scientific text. *CoRR abs/1903.10676* (2019). <http://arxiv.org/abs/1903.10676>
3. van den Bulk, L.M., Bouzembrak, Y., Gavai, A., Liu, N., van den Heuvel, L.J., Marvin, H.J.: Automatic classification of literature in systematic reviews on food safety using machine learning. *Curr. Res. Food Sci.* **5**, 84–95 (2022). <https://doi.org/10.1016/j.crfs.2021.12.010>
4. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. *Adv. Neural Inf. Process. Syst.* **33**, 22243–22255 (2020). <https://doi.org/10.48550/arXiv.2006.10029>
5. Cohen, A.M., Hersh, W.R., Peterson, K., Yen, P.Y.: Reducing workload in systematic review preparation using automated citation classification. *J. Am. Med. Inf. Assoc.* **13**(2), 206–219 (2006). <https://doi.org/10.1197/jamia.M1929>
6. Collins, C., Dennehy, D., Conboy, K., Mikalef, P.: Artificial intelligence in information systems research: a systematic literature review and research agenda. *Int. J. Inf. Manag.* **60**(June), 102383 (2021). <https://doi.org/10.1016/j.ijinfomgt.2021.102383>







7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. ArXiv abs/1810.04805 (2019). <https://doi.org/10.18653/v1/N19-1423>
8. van Dinter, R., Catal, C., Tekinerdogan, B.: A multi-channel convolutional neural network approach to automate the citation screening process. *Appl. Soft Comput.* **112**, 107765 (2021). <https://doi.org/10.1016/j.asoc.2021.107765>
9. van Dinter, R., Tekinerdogan, B., Catal, C.: Automation of systematic literature reviews: a systematic literature review. *Inf. Softw. Technol.* **136**, 106589 (2021). <https://doi.org/10.1016/j.infsof.2021.106589>
10. Fei-Fei, L., Fergus, R., Perona, P.: A bayesian approach to unsupervised one-shot learning of object categories. In: *Proceedings Ninth IEEE International Conference on Computer Vision*, pp. 1134–1141. IEEE (2003). <https://doi.org/10.1109/ICCV.2003.1238476>
11. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. CoRR abs/1703.03400 (2017). <http://arxiv.org/abs/1703.03400>
12. Han, S., Pool, J., Tran, J., Dally, W.: Learning both weights and connections for efficient neural network. *Adv. Neural Inf. Process. Syst.* **28** (2015). <https://doi.org/10.48550/arXiv.1506.02626>
13. Houlsby, N., et al.: Parameter-efficient transfer learning for nlp. In: *International Conference on Machine Learning*, pp. 2790–2799. PMLR (2019). <https://doi.org/10.48550/arXiv.1902.00751>
14. Howard, B.E., et al.: Swift-review: a text-mining workbench for systematic review. *Syst. Rev.* **5**(1), 1–16 (2016). <https://doi.org/10.1186/s13643-016-0263-z>
15. IBM Cloud Education: Natural language processing (NLP) (2021). <https://www.ibm.com/cloud/learn/natural-language-processing>. Accessed 08 Mar 2022
16. Jackson, R.G., et al.: Ablations over transformer models for biomedical relationship extraction. *F1000Research* **9**, 710 (2020). <https://doi.org/10.12688/f1000research.24552.1>
17. Kontonatsios, G., Spencer, S., Matthew, P., Korkontzelos, I.: Using a neural network-based feature extraction method to facilitate citation screening for systematic reviews. *Expert Syst. Appl. X* **6**, 100030 (2020). <https://doi.org/10.1016/j.eswax.2020.100030>
18. Kurtic, E., et al.: The optimal bert surgeon: Scalable and accurate second-order pruning for large language models (2022). arXiv preprint [arXiv:2203.07259](https://arxiv.org/abs/2203.07259)
19. Kusa, W., Hanbury, A., Knoth, P.: Automation of citation screening for systematic literature reviews using neural networks: a replicability study (2022). arXiv preprint [arXiv:2201.07534](https://arxiv.org/abs/2201.07534)
20. Kusa, W., Lipani, A., Knoth, P., Hanbury, A.: An analysis of work saved over sampling in the evaluation of automated citation screening in systematic literature reviews. *Intell. Syst. Appl.* **18**, 200193 (2023). <https://doi.org/10.1016/j.iswa.2023.200193>
21. van der Maaten, L., Hinton, G.E.: Visualizing data using t-sne. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008). <https://www.jmlr.org/papers/v9/vandemaaten08a.html>
22. Melo, M., et al.: Few-shot approach for systematic literature review classifications. In: *18th International Conference on Web Information Systems and Technologies* (2022). <https://doi.org/10.5220/0011526400003318>
23. Nichol, A., Achiam, J., Schulman, J.: On first-order meta-learning algorithms. CoRR abs/1803.02999 (2018). <http://arxiv.org/abs/1803.02999>
24. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. CoRR abs/1802.05365 (2018). <http://arxiv.org/abs/1802.05365>

25. Reimers, N., Gurevych, I.: Sentence-bert: sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2019). <http://arxiv.org/abs/1908.10084>
26. van de Schoot, R., et al.: An open source machine learning framework for efficient and transparent systematic reviews. *Nat. Mach. Intell.* **3**(2), 125–133 (2021). <https://doi.org/10.1038/s42256-020-00287-7>
27. Sellak, H., Ouhbi, B., Frikh, B.: Using rule-based classifiers in systematic reviews: a semantic class association rules approach. In: Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services, pp. 1–5 (2015). <https://doi.org/10.1145/2837185.2837279>
28. Song, K., Tan, X., Qin, T., Lu, J., Liu, T.: Mpnnet: masked and permuted pre-training for language understanding. *CoRR abs/2004.09297* (2020). <https://arxiv.org/abs/2004.09297>
29. Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune BERT for text classification? In: Sun, M., Huang, X., Ji, H., Liu, Z., Liu, Y. (eds.) *CCL 2019. LNCS (LNAI)*, vol. 11856, pp. 194–206. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-32381-3\\_16](https://doi.org/10.1007/978-3-030-32381-3_16)
30. Tsafnat, G., Glasziou, P., Karystianis, G., Coiera, E.: Automated screening of research studies for systematic reviews using study characteristics. *Syst. Rev.* **7**(1), 1–9 (2018). <https://doi.org/10.1186/s13643-018-0724-7>
31. Wang, S., Fang, H., Khabsa, M., Mao, H., Ma, H.: Entailment as few-shot learner. *CoRR abs/2104.14690* (2021). <https://arxiv.org/abs/2104.14690>
32. Weigang, L., da Silva, N.C.: A study of parallel neural networks. In: *IJCNN 1999. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339)*, vol. 2, pp. 1113–1116. IEEE (1999). <https://doi.org/10.1109/IJCNN.1999.831112>
33. Wu, L., Won, Y.S., Jap, D., Perin, G., Bhasin, S., Picek, S.: Explain some noise: ablation analysis for deep learning-based physical side-channel analysis. *Cryptology ePrint Archive* (2021). <https://eprint.iacr.org/2021/717>



# Shift Toward Value-Based Learning: Applying Agile Approaches in Higher Education

Eva-Maria Schön<sup>1</sup> , Ilona Buchem<sup>2</sup> , Stefano Sostak<sup>2</sup> ,  
and Maria Rauschenberger<sup>3</sup> 

<sup>1</sup> Faculty Business Studies, University of Applied Sciences Emden/Leer, Emden, Germany  
eva-maria.schoen@hs-emden-leer.de

<sup>2</sup> Faculty I Economics and Social Sciences, Berlin University of Applied Sciences, Berlin,  
Germany

buchem@bht-berlin.de

<sup>3</sup> Faculty of Technology, University of Applied Sciences Emden/Leer, Emden, Germany  
maria.rauschenberger@hs-emden-leer.de

**Abstract.** Due to circumstances such as digital teaching during the coronavirus pandemic and the emergence of powerful artificial intelligence tools (*e.g.*, ChatGPT), digitization in higher education has increased rapidly in recent years. For this reason, innovative didactic concepts are being applied, and new teaching methods are being tested. One of these is value-based learning, an approach that aims to develop students' values alongside specialist knowledge. The objective of this research is to investigate how value-based learning can be implemented in higher education through agile practices and agile values. Thus, we have chosen a multiple case study research method that includes three case studies at different German universities of applied sciences. The results show that the application of agile practices and values varies by context and is individualized. Therefore, we developed a conceptual model that shows how value-based learning can be applied to higher education through agile practices and agile values. This conceptual model shows how courses and modules, as well as students and lecturers, evolve through continuous feedback over the course of a semester. Moreover, it allows students to be taught competencies that enable them to adapt to continuous change.

**Keywords:** Agile · Higher education · Value-based learning · Student-centered learning · Teaching

## 1 Introduction

In recent years, the education system has witnessed numerous transformations that have quickly flipped the script and accelerated digitization. For example, during the coronavirus pandemic, almost all courses were held online, and lecturers and students had to get used to the new conditions in a very short time [30]. In recent months, the increase in digitalization has been amplified by the emergence of powerful artificial intelligence (AI) tools such as ChatGPT, as referenced in [24, 29]. As a result of these changes, innovative ideas and concepts for teaching have been explored, including the integration of gamification frameworks into educational settings [28], the incorporation of new

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

M. Marchiori et al. (Eds.): WEBIST 2022, LNBIP 494, pp. 24–41, 2023.

[https://doi.org/10.1007/978-3-031-43088-6\\_2](https://doi.org/10.1007/978-3-031-43088-6_2)

technologies such as a robot acting as a teaching assistant and Scrum Master [9], and the adoption of agile methodologies [18,23,25,30]. These examples demonstrate the range of agile practices and values and how they can be used and combined in higher education. This observation is consistent with the findings shown in the annual *State of Agile* study [14], which shows that the combination of agile methods and agile practices depends on the context of their application. Moreover, these trends show how the context of higher education has evolved due to the application of agility. On the one hand, we see changes toward technical agility (*doing agile*), which means agile methods and practices are applied [34]. On the other hand, there are also changes in cultural agility (*being agile*), which means that an agile mindset has been internalized and practiced. This shift implies a transition toward a student-centered learning approach [38] where agile values such as self-organization, autonomy, mastery, and purpose play an important role. This is also known as a value-based learning approach.

We understand value-based learning as an educational approach that emphasizes developing students' values alongside specialist knowledge. It is essential for students to learn how to acquire new knowledge independently due to the constant evolution of digital media and tools and the knowledge explosion brought about by globalization and digitalization. As a result, students need to possess learning competencies that enable them to independently structure and reflect upon their learning process as well as set goals.

This paper examines the following research question: *How can value-based learning be implemented in higher education through agile practices and agile values?* As outlined above, the context of higher education has been accelerating in recent years, requiring lecturers to explore new ways to support students in acquiring knowledge and skills. To answer the research question, we analyze concepts for integrating agile practices and values in higher education using three case studies at German universities of applied sciences in Berlin, Hamburg, and Emden. Analyzing these three case studies enables us to derive a conceptual model that visualizes the application of value-based learning through agile practices and values.

This paper is structured as follows: Sect. 2 outlines the background and related work. In Sect. 3, we describe our research methodology and the contexts of the three case studies. Then, Sect. 4 presents the results of our three case studies and shows how agile practices and values are applied in each case. In Sect. 5, we answer our research question and present our conceptual model that describes how value-based learning is applied in higher education through agile practices and values. In addition, we discuss the limitations of this research. Finally, Sect. 6 concludes this work and presents suggestions for future work.

## 2 Background and Related Work

In the following, we provide a brief overview of the background of agile methods and agile practices, as well as didactic concepts and related work.

## 2.1 Agile Methods and Agile Practices

*“Agile is the ability to create and respond to change. It is a way of dealing with, and ultimately succeeding in, an uncertain and turbulent environment” [1].*

Agile methods and their adoption in the workplace (e.g., companies, organizations) have become a highly discussed and popular topic over the years. Agile practices are concrete procedures for implementing agile values and principles. Agile values refer to a value set that is used as a basis for the application of agile methods. The agile community distinguishes between the concepts of *doing* and *being* agile [17, 34]. Doing agile (technical agility) refers to the application of agile methods and agile practices (e.g., using Scrum or a Kanban board). Being agile (cultural agility) points to the agile values and principles that are central to an agile mindset.

Initially, the agile values were captured in the *Agile Manifesto* with four values (cf. *individuals and interactions, working software, customer collaboration, and responding to change*) [7]. Since then, agile values have evolved. For instance, Modern Agile includes the values *make people awesome, make safety a prerequisite, experiment and learn rapidly, and deliver value continuously* [22]. In addition, agile methods such as Scrum [33] have their own values. Agile methods such as Scrum [33] or Kanban [2] have their origins in software development. In computer science and IT businesses, these models have been used for decades to solve complex problems. The usage of agile methods and agile practices is intended to increase transparency and accelerate change, as well as minimize risks and errors in the development process. It is done to avoid a big upfront design and reduce the design phase to a minimum so that executable software is generated as quickly as possible. Compared to plan-oriented approaches, such as the waterfall model, agile methods focus on the iterative development and testing of incremental solutions and collecting feedback. However, this approach requires a change of mindset because solutions are not planned in detail in advance but are continuously developed and optimized based on feedback from relevant stakeholders and users [32].

Agile methods and agile practices are also being used more and more frequently in other areas outside IT, as presented in the annual *State of Agile* [14] study. Moreover, the study shows that organizational culture has an influence on the successful use of agility. Furthermore, it becomes clear that resistance to change and a lack of understanding of the agile mindset are often problematic for the introduction of agility within an organization. The agile mindset involves fundamental assumptions, such as believing in the competence and responsibility of individuals, encouraging collaboration, continuous learning and improvement, encouraging creativity, promoting innovation, and taking moderate risks cf. [37]. In recent years, it has been shown that the agile mindset and the adaptation of agile practices and values are also interesting for higher education didactics, as autonomous, project-based, and iterative learning in short cycles with continuous feedback can support the development of competencies in higher education.

## 2.2 Didactic Concepts and Related Work

Didactic concepts for agility in higher education are still a relatively young field of practice and research. For instance, there are concepts for agile didactics in the sense

of agile interactions of teachers and learners in the classroom, as well as didactic concepts for integrating agile practices from the field of software development into other subject areas of higher education. The book *Agile University Didactics* is often cited [4] as an example of applying agility in higher education. Arn [4] compares agile didactics to planned didactics and defines it as a mixture of planned and unplanned teaching, a didactic that emerges from communication and interaction, especially when learners and teachers not only meet at eye level but encounter each other openly. Lecturers play the dual role of teachers and coaches at the same time. They teach according to the principle of structured improvisation and react to the learners' feedback in a way analogous to the interaction with customers and users in agile software development [4].

Agile practices and values are used in different contexts and disciplines, such as in economics to improve lifelong learning and employability of students [12], in computer science to teach concepts of agility by means of low code platforms [18], in doctoral studies to support collaborative learning between doctoral students [31,35], and at the university to improve studying and teaching [19]. Moreover, agile practices can be used to integrate new technology in higher education, such as a robot acting as a teaching assistant and Scrum Master [9]. Other didactic concepts in higher education rely on agile software development methods and propose concepts and principles for the redevelopment of universities. For example, Baecker [5] emphasizes the conversion from primarily vertical to primarily horizontal organizational structures, acting in networks at universities in the sense of scientific communities, and having a stronger interlocking with professional practice.

Based on agile methods, didactic methods such as eduScrum® are being developed and used. eduScrum® is described as a framework for coaching learners in which the responsibility for the learning process is transferred to the learners [36]. Like Scrum, eduScrum® is based on the collaboration of teams with the associated descriptions of roles, ceremonies, artifacts, and rules. Neumann et al. [23] show an example of the application of eduScrum® in real-world settings in higher education. Another example of transferring agility in didactic concepts is the *Agile Manifesto for Teaching and Learning* by Krehbiel et al. [16]. The manifesto defines agile principles, concepts, and practices for higher education in a way that is comparable to the Agile Manifesto from software development. The objective is to increase student engagement, encourage students to take responsibility for learning, improve the level and quality of collaboration, and produce high-quality results in teaching. With a similar objective, the concept of agile learning with *Just in Time Teaching* (JiT) has been proposed. It builds on constructivism and self-determination theory principles and emphasizes adaptive teaching with coupled teaching-learning cycles and continuous feedback loops [20].

### 3 Research Methodology

This paper investigates the shift toward value-based learning and its implementation in higher education through agility. Therefore, our research is guided by the following research question: *How can value-based learning be implemented in higher education through agile practices and agile values?* To answer this question, we conducted three case studies at three universities of applied sciences in Germany (in Berlin, Hamburg,

and Emden). In doing so, we examine concepts for integrating agile practices and values in higher education. Case studies are a useful way to examine complex phenomena in their respective contexts [6, 39]. A case study allows us to collect data in practice to understand the application of agile practices and values in the context of higher education and to visualize a conceptual structure. In the following, we first describe the context of each case study, followed by an explanation of the data collection and analysis procedures.

### 3.1 Case Study 1 - Berlin University of Applied Sciences

*Berlin University of Applied Sciences* is a public, technical university of applied sciences with around 13,000 students and over 70 accredited bachelor's and master's degree programs in applied engineering, natural sciences, and economics. Essential qualifications such as the ability to work in a team and social skills play a central role in the studies programs. The use of digital technologies in teaching is part of the university's digitization strategy. The Berlin University of Applied Sciences is also a member of the *Virtuelle Fachhochschule (VFH)*, which provides different online degree programs such as business administration, media informatics, business informatics, and industrial engineering. The case study involves the module *Agile Project Management* (6 CP with 4 SWS), which is mandatory in the third semester of the degree program Business Administration Digital Economy (B. Sc.) and is in the Department of Economics and Social Sciences. The students of the course are interested in technology but generally have little prior knowledge of agile principles and methods.

### 3.2 Case Study 2 - HAW Hamburg

*HAW Hamburg* is a public university of applied sciences in northern Germany with over 70 accredited bachelor's and master's degree programs. In the winter semester of 2020/2021, there were a total of 17,125 enrolled students. HAW Hamburg aims to develop sustainable solutions for the social challenges of the present and the future. The case study involves the optional course *Agile Project Management* (6 CP with 4 SWS), which is offered by the Faculty of Technology and Information Technology. The course is primarily for students in the fifth or sixth semester of the bachelor's degree program Business Informatics (B.Sc.). Students of other study programs can also participate in the module, as far as its capacity allows. Thus, the target group of the module is rather technically inclined and already has some prior knowledge regarding agile methods.

### 3.3 Case Study 3 - University of Applied Sciences Emden/Leer

The public University of Applied Sciences Emden/Leer has around 4,100 students and over 40 bachelor's and master's degree programs in maritime sciences, social work, health, technology, and business. It is a founding member of the *Virtuelle Fachhochschule (VFH)*, which provides different online degree programs such as business administration, media informatics, business informatics, and industrial engineering. The online degree programs use interactive, multimedia learning materials and state-of-the-art collaboration and communication media to implement contemporary learning scenarios on the Internet. Digital tools and individual learning are also part of the teaching strategy.

The case study involves the mandatory module *Project Group module* (10 CP with 3 SWS) in the Media Technology degree program (B. Eng.) in the Department of Technology. Students in this program are fond of technologies (e.g., audio, video, animation), and part of their coursework includes programming, but they are not familiar with agile principles and agile methods.

### 3.4 Data Collection and Analysis

Two case studies were conducted during the summer semester of 2021, and one case study was in the winter semester of 20/21. Digital teaching was conducted during these semesters due to the coronavirus pandemic. Therefore, the teaching and learning concepts were tailored to the digital format. An analysis of the course material of the three case studies was carried out to collect relevant data. The course material was analyzed in terms of didactic goals, teaching concepts and methods, and learning controls. In addition, the extent to which agile practices and values were applied in the course was examined. The results of this analysis are presented as a narrative comparison in the following section. A table presents an overview of the implemented agile practices and values to compare the three case studies better.

## 4 Case Study Results

In the following, we describe the results of our three case studies and the analysis regarding agile practices and values.

### 4.1 Case Study 1 - Berlin University of Applied Sciences

The *Agile Project Management* module at *Berlin University of Applied Sciences* is a mandatory module in the third semester of the Business Digital Economy degree program and is offered entirely in English in order to strengthen the internationality in the degree program and prepare students to work on international projects. The *Agile Project Management* module has been offered every winter semester since 2015. In the following, descriptions of the didactic goals, the teaching concept and methods, and the learning assessments and digital awards are given. In addition, there will be an explanation of how agile practices and values have been implemented.

**Didactic Goals.** The learning objectives of the module were developed as learning outcomes in the sense of competence orientation to the revised learning objectives taxonomy of [3] and formulated in the module handbook as follows: (1) Students know the theoretical and methodological basics of agile project management and can classify agile project management as a methodological approach and compare it with other approaches. (2) Students have a general overview of the central frameworks, methods, instruments, and application areas of agile project management in business management practice. (3) Students can apply methods, instruments, and decision-making tools of agile project management in practice, taking into account agile values and principles. (4) Students are able to plan and implement projects according to the agile approach, as well as evaluate and present the results.



**Teaching Concepts and Didactic Methods.** The module *Agile Project Management* is based on teamwork in small groups. Students work on projects from the project marketing seminar and apply methods of agile project management in the course. The module consists of a *seminar class* (SC) and a *tutorial* (T) with integrated project work. The module is taught by two lecturers, a professor from the *Berlin University of Applied Sciences* (SC) and a lecturer from the business world (T). The grade for the module is composed of three subgrades. The first subgrade is assigned to the SC and accounts for 40% of the final grade. It is determined based on the results of the eight online quizzes in terms of continuous learning assessments. The second subgrade is assigned to the tutorial and also accounts for 40% of the final grade. It is determined based on team coaching sessions (eight sessions per team). In addition, students can earn five bonus points in team coaching. The third subgrade is considered a common subgrade in the SC and the T and accounts for 20% of the overall grade. It is based on the evaluation of the final video reflection (one video per team).

*Seminar Class* (SC): The content on agile project management is taught in the SC. Here, basic agile principles are learned, including project management in transition, characteristics and types of projects in the digital economy, project leadership in the digital age, agile values, mindset, and principles, and agile frameworks such as *Scrum*, *Kanban*, and *DSDM*. The instructional design from SC is based on the ARCS model, a motivational, instructional design approach from (Keller et al., 1987) with four basic principles: attention, relevance, confidence, and satisfaction. Various didactic methods are used in teaching, including flipped classroom (*i.e.*, preparation for SC with learning videos, application in SC, follow-up with learning scripts, weekly quizzes), game-based learning (*e.g.*, games for applying agile frameworks), and collaborative learning in project teams. Various digital learning materials are used to best support students with different learning styles and preferences, including interactive presentation slides in Google Drive, scripts in PDF format in the LMS *Moodle*, learning videos on *LinkedIn Learning*, and interactive learning materials created with H5P for Moodle.

*Tutorial* (T): In addition to the SC, there is a weekly T for the students. The aim of the 90-minute T is to deepen the knowledge gained in the SC and supplement it with practical experience. Agile working is to be made experienceable. This is done by presenting and applying methods from work with agile project teams in software companies, as well as creating a framework for agile collaboration of the students on the projects in the marketing seminar. The T is divided into five parts: warm-up, knowledge reinforcement, team time, *Lean Coffee*, and query of Return Of Time Invested (ROTI). The warm-up, which takes place at the beginning of each T, serves to activate the students and includes an activity to promote group interaction [27]. This common warm-up creates a positive working atmosphere in the group. It also increases the receptivity of the participants [21]. Typically, the warm-up lasts five to 15 min and includes a previously unfamiliar activity designed to cognitively stimulate the students. The knowledge-deepening subsection is about deepening the content learned in the SC, which is complemented by practical case studies. During team time, students work in their teams on their specific projects for the marketing project seminar. This gives students the opportunity to apply what they have learned directly to their project work. *Lean Coffee* is an agile practice that facilitates discussions with minimal planning. It uses innovative voting techniques

such as dot voting to support collaboration and the decision-making process [13]. In *Lean Coffee*, students have the opportunity to raise issues relevant to them and discuss them with the lecturer and other students in the course. At the end of each T, a survey of ROTI was conducted. This asked students to indicate their personal return on time invested in the T on a scale of one to five. A rating of five indicates a very high return on time invested. If they gave a rating below five, students were also asked to indicate what they thought could be better. This allows instructors to iteratively adjust and improve the structure of the T. Other methods include clarification of individual expectations and team retrospectives.

**Learning Assessments and Digital Awards.** In the SC, there are weekly quizzes in the LMS *Moodle* issued as continuous learning assessments to test the students' knowledge of the central topics in *Agile Project Management*. Different question formats are used, including multiple-choice, assignment, and drag & drop. In the exercise, starting from the answer to the ROTI survey, the students' participation in the exercise is checked. Students receive five points for each participation. In addition, students can earn five bonus points by facilitating a team retrospective. The end-of-semester video reflection is graded on a criterion-referenced basis. Each team creates a 10-minute video in which each team member reflects on the agile work in the team, including the use of agile methods and tools, according to the following criteria: (1) agile team, (2) agile principles, (3) agile methods and tools, and (4) takeaways.

In the *Agile Project Management* module, two additional digital awards based on Open Badges in the LMS *Moodle* are given to students who have met specific requirements. Students who have achieved the maximum score for the first subgrade (knowledge-based learning assessment) receive an agile expert digital badge. Students who have achieved the maximum score for the second subgrade (team coaching) receive an agile team digital badge. In addition, students are guided on how to use the digital badges for profiling on social media (e.g., on *LinkedIn*).

**Agile Practices and Agile Values.** Table 1 shows an overview of the agile practices and values implemented in Case Study 1 (*Berlin University of Applied Sciences*). The agile values and agile practices are ordered in the same order as in the Agile Manifesto.

## 4.2 Case Study 2 - HAW Hamburg

This section presents the results of the analysis of the optional course *Agile Project Management* at the HAW Hamburg. The module has been offered once at the Faculty of Technology and Information Technology primarily for bachelor students in the fifth or sixth semester of the degree program Business Informatics (B.Sc.). In the following, descriptions of the didactic goals, the teaching concept and methods, and the learning assessments are provided. In addition, there is an explanation of how agile practices and values were implemented.

**Table 1.** Overview of agile values and agile practices in Case Study 1, based on [30].

Didactic element	Agile values	Agile practices
Seminar class (SC)	individuals and interactions over processes and tools, customer collaboration over contract negotiation, responding to change over following a plan	Scrum Team, Scrum Events, product backlog, team board, timebox, user story, estimation, querying expectations, gathering feedback, iteratively responding to student needs and feedback
Tutorial (T)	individuals and interactions over processes and tools, customer collaboration over contract negotiation, responding to change over following a plan, working software over comprehensive documentation	team building, team phases, Scrum Events, lean coffee, retrospective, asking for expectations, collecting feedback, iteratively responding to student needs and feedback, timebox, project slicing, story mapping, team time to work on the marketing project
Exam	individuals and interactions over processes and tools	collaborative reflection

**Didactic Goals.** The learning objectives of the course have been formulated as learning outcomes within the framework of competency-based teaching. For the presentation, a user story [11] has been created and presented by means of a sketchnote (see Fig. 1). For the formulation of the acceptance criteria, the taxonomy levels according to [8] have been used. The goal is for students to be able to apply as many agile practices and values as possible during the course.

## LEARNING OBJECTIVES FOR AGILE PROJECT MANAGEMENT



**As a student, I want to understand the values and practices of agile project management so that I can apply it in practice.**

- Be able to name agile values and principles
- Be able to understand the agile mindset
- Be able to describe agile practices and process models
- Be able to apply a selection of agile practices

**Fig. 1.** Learning objectives in the user story format [30].

**Teaching Concepts and Didactic Methods.** The optional module *Agile Project Management* is divided into two SWS *seminar classes* (SC) and two SWS *tutorials* (T). The module is taught by a professor from HAW Hamburg. The professor brings both the expertise and the application knowledge from corporate practice. In addition, there are guest lectures by well-known personalities of the agile community from industry and science within the framework of the SC.

*Seminar class* (SC): In the SC, the theoretical basics of agile project management are provided. The following topics are covered: agile mindset, state of agile in practice, product discovery and product execution, agile estimation and planning, agile methods, scaling agile, and agile leadership. Theoretical concepts are introduced, and content is supplemented with videos and interactive discussions to implement activating teaching. In the summer semester of 2021, there were two guest contributions from people in industry and academia who reported on agility in practice and current research on the agile way of working during the coronavirus pandemic.

*Tutorial* (T): The tutorial consists of three exercise units, which are assessed as a prerequisite for the exam. The tutorial was implemented by means of a Sprint logic. The duration of a Sprint is three weeks. During the SC, the new tasks are presented (*planning*). The students then work on the tasks in self-organized teams (*doing*). A shared exercise date is then used for the teams to present the results to each other and receive feedback (*review*). At the end of the exercise, a *retrospective* takes place in which the participants reflect together on what went well, what could be optimized, and what was learned. The lecturer takes on the role of the Product Owner in the T, presents the tasks to be worked on, and accepts the solutions at the end. During the first T, the students conduct a product discovery and apply the agile practices personas, story maps, and user stories. During the second T, agile estimation and release planning occur. Here, the agile practices magic estimation, release planning using a story map, and minimum viable product (MVP) are applied. In the third T, a release retrospective is conducted with the entire course to reflect on learning outcomes, thereby consolidating the content in long-term memory.

**Learning Assessments.** Different methods are used to assess learning progress. Interactive quizzes are regularly included in the SC; these can be group discussions as well as smaller surveys or quizzes. In addition, the students apply the contents of the SC in the T. Another aspect is the exam of the semester, which is in the form of a presentation. The students independently choose a topic from the SC and create a scientific poster, which is then presented in a prerecorded audio presentation. The course *Agile Project Management* was conducted completely digitally due to the pandemic regulations that were in effect during the summer semester of 2021. The following tools were used to conduct the digital teaching: Miro, Trello, Retromat, MS Teams, Zoom, Whiteboard, and Mentimeter.

**Agile Practices and Agile Values.** Table 2 shows the results of the agile practices and values implemented for Case Study 2 (*HAW Hamburg*).

**Table 2.** Overview of agile values and agile practices in Case Study 2 based on [30].

Didactic element	Agile values	Agile practices
Seminar class (SC)	individuals and interactions over processes and tools, responding to change over following a plan	product backlog, Kanban board, timebox, user story, informal documentation, sketchnotes, storytelling
Tutorial (T)	individuals and interactions over processes and tools, working software over comprehensive documentation, openness, respect, courage	Product Owner, timebox, Sprint logic, planning meeting, review meeting, retrospective, user story, product discovery, personas, story maps, agile estimation and planning, magic estimation, story points, release planning, minimum viable product, release retrospective
Exam	individuals and interactions over processes and tools, autonomy, mastery, and purpose	timebox

### 4.3 Case Study 3 - University of Applied Sciences Emden/Leer

The *Project Group* module at the University of Applied Sciences Emden/Leer is a mandatory module in the fifth semester of the Media Technology degree program. It has been offered on a regular basis every winter semester since 2017.

**Didactic Goals.** The module is defined by the module handbook as a group activity that involves the following: (1) students learn theoretical and methodological basics within a group; (2) students are solution-orientated and self-organized toward an increment; (3) students are able to report to stakeholders according to the development stage they are at. Since this module is openly designed, the project groups for the winter semester 2020/2021 were announced as Scrum Teams. This means that the project team's aim is to plan and create a medium to help people learn how to use Scrum effectively, both in theory and in practice. A medium can be game development, radio play, or video creation. The students are to organize themselves as a Scrum Team with a Kanban board and work together to acquire the corresponding skills. This is to be achieved with easily understandable and motivating media-prepared learning content in the form of a serious game and/or as a video (à la "Scrum in a Nutshell") for their own project organization. Scrum is a process model of project and software development and is now also used in many other disciplines. The Kanban board helps teams to visualize their parallel tasks and to identify bottlenecks. Students had no previous knowledge of Scrum. They first learned within their project group, then presented their knowledge, and finally, discussed the roles, agile values, and agile principles. Since students had to organize themselves as a Scrum team, deeper discussions about roles and responsibilities took place.

**Teaching Concepts and Didactic Methods.** As the module is called *Project Group*, the course is organized into groups of five to seven students depending on the media (e.g., video, radio play, or game) they are interested in using and the responsibility they might take on. The module is divided into two SWS *seminar classes* (SC) and one SWS *tutorial* (T).

*Seminar Class* (SC): Theoretical basics of agile project management are provided in the SC. The following topics are discussed: agile mindset, Scrum, Kanban, agile planning, agile process, agile teams, (dis)advantages of agile teams, impediments, and agile leadership. Group discussions and supplementary materials (*Blended Learning*) accompany each theory part.

*Tutorial* (T): The discussions of roles, responsibilities, and leadership were lively as groups thought of ways to organize themselves as an agile team depending on the project context, project member needs, and skills. Students had to plan their projects and present their plans in each class to gather feedback from their peers regarding the activities (*planning, doing, and reviewing*). Students applied Kanban, release planning, regular daily standup meetings (i.e., dailies customized to the group's needs), and storyboards. The lecturer took on the role of mediator, stakeholder, or input-giver. During the semester, students experienced difficulties such as the gap between planning and actually developing technical solutions in Sprints, team member conflict, uncertainty about how to lead and how to solve technical issues and personal conflicts, release planning, and communicating ideas, plans, results, and rearrangements.

**Learning Assessments.** The module is designed as *constructive alignment*, which means everything is relevant for the oral exam. Students apply the content from the SC and the content they compiled (Blended Learning) in the T. In the oral exam at the end of the semester, students present their final product, discuss how they organized themselves as an agile team, and explain which agile values and agile practices they applied as well as the lessons learned. The oral presentation is accompanied by a project description to reflect on the agile group structure and the final product. The following tools are used to conduct the digital teaching: Conceptboard and BigBlueButton.

**Agile Practices and Agile Values.** Table 3 shows the results for the agile practices and values applied in Case Study 3 (*University of Applied Sciences Emden/Leer*).

## 5 Discussion

In this section, we answer our research question of how value-based learning can be implemented in higher education through agile practices and values. Therefore, we present our conceptual model that describes how value-based learning is applied in higher education. In addition, we discuss the implications of our findings and the limitations of this research.

**Table 3.** Overview of agile values and agile practices in Case Study 3.

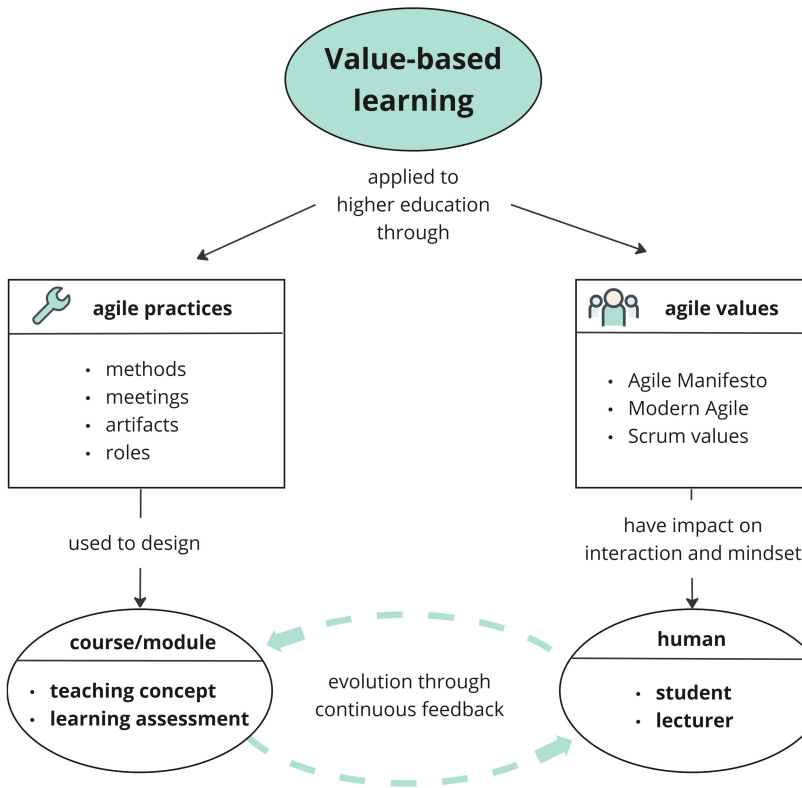
Didactic element	Agile values	Agile practices
Seminar class (SC)	individuals and interactions over processes and tools, responding to change over following a plan	product backlog, Kanban board, timebox, user story, informal documentation, storytelling, Scrum Team, Scrum Events, estimation, querying expectations, gathering feedback, iteratively responding to students' needs and feedback
Tutorial (T)	individuals and interactions over processes and tools, working software over comprehensive documentation, responding to change over following a plan, openness, respect	team building, team phases, Scrum Events, collecting feedback, Product Owner, timebox, Sprint logic, planning meeting, review meeting, retrospective, user story, agile estimation and planning, release planning, minimum viable product, release retrospective, iteratively responding to students' needs
Exam	individuals and interactions over processes and tools, autonomy, purpose	timebox

### 5.1 How Can Value-Based Learning Be Implemented in Higher Education Through Agile Practices and Agile Values?

The analysis of the three case studies (cf. Sect. 4) has shown how agile practices and values can be used in higher education. We can conclude that the combination of different agile practices and values is diverse and is very context-dependent in each case study. These experiences are in line with the findings of the application of agility in commercial enterprises [14].

However, some commonalities can be derived from the three case studies examined. We can use these findings to create a conceptual model that allows us to answer our research question. Figure 2 presents our conceptual model for value-based learning in higher education using agile practices and values. Value-based learning is applied to higher education through the usage of agile practices and values. The former develops the students' specialist knowledge and relates to technical agility (*doing agile*). The latter develops students' values and refers to cultural agility (*being agile*). Agile practices such as methods, meetings, artifacts, and roles are used in each case study to design seminar classes, tutorials, and exams. The components of seminar classes, tutorials, and exams are part of each case study and vary in their design. The adoption of agile values helps the students to focus on interacting with peers and working as a self-organized team. Moreover, agile values impact the interaction between lecturers and students and change the mindsets of the involved humans. Another point we can draw from the examined case studies is the evolution of the course/modules and the humans through continuous feedback loops implemented as part of the didactic concept. Instead

of a single evaluation at the end of the course, many possibilities for adapting the didactic concept were built in.



**Fig. 2.** Conceptual model of the application of value-based learning through agile practices and agile values.

The introduction of agility to higher education has changed the roles of lecturers and students and the interaction between these groups. Lecturers are seen as coaches who provide students with a roadmap (*e.g.*, didactic goals and course material) for acquiring knowledge and skills. They accompany the student's learning process and are available as advisors, experts, and mentors. In addition, lecturers motivate the students and support them in self-organized learning. In terms of agile practices, this role is known as the team coach [15]. The role of the learner also changes as changes in values and mindset take place. Teaching evolves into a student-centered approach in which the students, with their prior knowledge and attitudes regarding learning, are the focus. In the three case studies (see Sect. 4), continuous feedback was obtained from the students to adapt the subsequent learning units to the needs of the learners over the course of the semester. In this regard, care has been taken to ensure that the changes in the teaching



concept meet the requirements of the modules in terms of content. The continuous collection of feedback from the students serves as quality control of the iteration process and the evolution of the course or module (see Fig. 2).

Furthermore, the lecturers encourage the intrinsic motivation of the students and use didactic concepts such as growth mindset [10] to bring agile values such as autonomy, mastery, and purpose [26] into focus. This shift toward value-based learning supports competency-based teaching, as the focus is less on teaching subject knowledge and more on teaching competencies. In addition, the linking of the *Agile Project Management* module with other modules in Case Study 1 makes it possible to apply and deepen the teaching content across modules. Thus, students benefit from what they have learned in several ways. For example, they experience the value-creating character of an agile way of working through the theory and personal successes (e.g., positive feedback from customers and users in the marketing project seminar or a better grade). The Sprint logic in Case Study 2 allows students to improve their way of working iteratively over the semester and supports the development of reflective skills and critical thinking. Moreover, Case Study 3 shows that agile can be used for the organization of groups and includes the agile development of a product from the students.

## 5.2 Critical Review and Limitations

While using the taxonomies of cognitive learning objectives, we discovered that the affective level (e.g., attitudes and motivation) is currently missing in the description of the learning objectives. However, this level of the learning objectives is crucial if we aim to develop attitudes in relation to values more strongly and integrate them into the curricula. In future courses, we want to expand the descriptions of the learning objectives by using other taxonomies that specifically deal with attitudes and motivation. Thus, the applied student-centered approach can be further improved. The current results are based on an analysis of the authors' course materials and an evaluation of the learning assessments. However, the agile practices and values used (see Table 1, 2 and 3) might have been perceived differently by the students. Therefore, we have increased the objectivity of the analysis by discussing the results in the authors' group. Another limitation is that this research has so far been limited to higher education, as we have conducted case studies only in higher education institutions. The authors have already gained experience in how agile practices and values can be incorporated into teaching in other modules such as programming, information systems, and group work in health-care studies. However, this is not part of the scope of this work because comparable data have not yet been evaluated.

## 6 Conclusion and Future Work

This paper presents the results of three case studies in which the application of agility in higher education is investigated. The case studies show that the application of agile practices and values varies in different contexts. In addition, the analysis allows us to better understand similarities in the application of agile practices and values. The main contribution of this work is our conceptual model that shows how value-based learning

can be integrated into higher education using agile practices and values. We developed our conceptual model from the three case studies that examined the application of agile practices and values at different universities of applied sciences in Germany.

In the context of higher education, value-based learning is applied through the usage of agile practices and values. The application of agile practices develops the students' specialist knowledge and relates to technical agility (*doing agile*). The transformation toward agile values develops students' values and relates to cultural agility (*being agile*). The continuous feedback loop in our model shows that there is an evolution of the course or module and of the lecturers and students. The changes happening in the education system highlight the importance of teaching students competencies that enable them to adapt to continuous changes. The three case studies illustrate how agile practices and values help students to adapt to changes. However, we still need to better understand how to choose agile practices and values for the context of higher education.

In future work, we want to understand how other teachers and students assess our model and what elements should be added. Future studies could apply the conceptual model presented in this paper to analyze and compare the applications of agile principles in other university courses. Furthermore, future research could explore how value-based learning affects learning and academic achievement. It could also explore the suitability of value-based learning for other teaching and learning contexts. For instance, value-based learning is also suitable for other teaching and learning contexts, such as adult education and different types of schools. This should be evaluated in depth in future studies.

## References

1. Alliance, X.A.: What is agile? (2020). <https://www.agilealliance.org/agile101/>. Accessed 24 Mar 2023
2. Anderson, D.J.: Kanban - Successful Evolutionary Change for your Technology Business. Blue Hole Press (2010)
3. Anderson, L., Krathwohl, D.: A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, Complete Longman, London (2001)
4. Arn, C.: Agile Hochschuldidaktik, 3. auflage edn. Juventa Verlag GmbH (2020)
5. Baecker, D.: Agilität in der Hochschule (Agility in the university). Die Hochschule: J. für Wissenschaft und Bildung (2017). <https://www.fachportal-paedagogik.de/literatur/vollanzeige.html?Fid=1129050>
6. Baxter, P., Jack, S.: Qualitative case study methodology: study design and implementation for novice researchers (2008)
7. Beck, K., et al.: Manifesto for agile software development (2001). <https://www.agilemanifesto.org/>
8. Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H., Krathwohl, D.R.: Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain, David McKay Company, Philadelphia (1956)
9. Buchem, I., Baecker, N.: Nao robot as scrum master: results from a scenario-based study on building rapport with a humanoid robot in hybrid higher education settings. In: Training, Education, and Learning Sciences, vol. 59, pp. 65–73. AHFE International (2022). <https://doi.org/10.54941/ahfe1002385>

10. Claro, S., Paunesku, D., Dweck, C.S.: Growth mindset tempers the effects of poverty on academic achievement. *Proc. Natl. Acad. Sci. U.S.A* **113**, 8664–8668 (2016). <https://doi.org/10.1073/pnas.1608207113>
11. Cohn, M.: *User Stories Applied: For Agile Software Development*. Addison-Wesley, Boston (2004)
12. Cubric, M.: An agile method for teaching agile in business schools. *Int. J. Manag. Educ.* **11**, 119–131 (2013). <https://doi.org/10.1016/j.ijme.2013.10.001>
13. Dalton, J.: *Lean coffee*. In: *Great Big Agile*, pp. 191–192. Apress (2019)
14. Digital.ai: 16th state of agile report. Technical report, digital.ai (2022). <https://digital.ai/resource-center/analyst-reports/state-of-agile-report/>
15. Hawkins, P.: *Leadership Team Coaching: Developing Collective Transformational Leadership*. Kogan Page Publishers, London (2021)
16. Krehbiel, T., et al.: Agile manifesto for teaching and learning. *J. Effect. Teach.* **17**, 90–111 (2017)
17. Kuchel, T., Neumann, M., Diebold, P., Schön, E.M.: Which challenges do exist with agile culture in practice? In: *Proceedings of the 2023 In The 38th ACM/SIGAPP Symposium on Applied Computing (SAC 2023)*, Tallinn, Estonia, 27–31 March 2023 (2023). <https://doi.org/10.1145/3555776.3578726>
18. Lebens, M., Finnegan, R.: Using a low code development environment to teach the agile methodology. In: Gregory, P., Lassenius, C., Wang, X., Kruchten, P. (eds.) *XP 2021. LNBP*, vol. 419, pp. 191–199. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-78098-2\\_12](https://doi.org/10.1007/978-3-030-78098-2_12)
19. Mayrberger, K., Slobodeaniuk, M.: Gruppe. Interaktion. Organisation. Zeitschrift für Angewandte Organisationspsychologie (GIO) **48**(3), 211–216 (2017). <https://doi.org/10.1007/s11612-017-0376-4>
20. Meissner, B., Stenger, H.J.: Agiles lernen mit just-in-time-teaching. adaptive lehre vor dem hintergrund von konstruktivismus und intrinsischer motivation (agile learning with just-in-time teaching. adaptive teaching against the backdrop of constructivism and intrinsic motivation.). In: Zawacki-Richter, O., Kergel, D., Kleinefeld, N., Muckel, P., Stöter, J., Brinkmann, K. (eds.) *Teaching Trends 2014. Offen für neue Wege: Digitale Medien in der Hochschule*, pp. 121–136. Waxmann (2014)
21. Mesquida, A.-L., Karać, J., Jovanović, M., Mas, A.: A game toolbox for process improvement in agile teams. In: Stolf, J., Stolf, S., O’Connor, R.V., Messnarz, R. (eds.) *EuroSPI 2017. CCIS*, vol. 748, pp. 302–309. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-64218-5\\_25](https://doi.org/10.1007/978-3-319-64218-5_25)
22. *Modern Agile Community: Modern agile* (2023). <https://modernagile.org/>. Accessed 24 Mar 2023
23. Neumann, M., Baumann, L.: Agile methods in higher education: adapting and using eduscrum with real world projects. In: *Proceedings of the 2021 IEEE Frontiers in Education Conference (FIE)*, pp. 1–8 (2021). <https://doi.org/10.1109/FIE49875.2021.9637344>
24. Neumann, M., Rauschenberger, M., Schön, E.M.: “We Need To Talk About ChatGPT”: the future of AI and higher education. In: *Proceedings of the 2023 IEEE/ACM 5th International Workshop on Software Engineering Education for the Next Generation (SEENG)*, Melbourne, Australia, 14–20 May 2023 (2023). <https://doi.org/10.1109/SEENG59157.2023.00010>
25. Ozkan, N., Özcan-Top, Ö., Bal, S., Gök, M.Ş.: Teaching agile in an agile way: a case from the first iteration in a university. In: *2022 3rd International Informatics and Software Engineering Conference (IISEC)*, pp. 1–6 (2022). <https://doi.org/10.1109/IISEC56263.2022.9998281>
26. Pink, D.H.: *Drive: The Surprising Truth About What Motivates Us*. Riverhead Books, New York City (2009)

27. Przybyłek, A., Kotecka, D.: Making agile retrospectives more awesome. In: Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, FedCSIS 2017, vol. 11, pp. 1211–1216 (2017). <https://doi.org/10.15439/2017F423>
28. Rauschenberger, M., Willems, A., Ternieden, M., Thomaschewski, J.: Towards the use of gamification frameworks in learning environments. *J. Interact. Learn. Res.* **2019**(30(2)), 147–165 (2019)
29. Rudolph, J., Tan, S., Tan, S.: ChatGPT: bullshit spewer or the end of traditional assessments in higher education? *J. Appl. Learn. Teach.* **6** (2023). <https://doi.org/10.37074/jalt.2023.6.1.9>. <https://journals.sfu.ca/jalt/index.php/jalt/article/view/689>
30. Schön, E.M., Buchem, I., Sostak, S.: Agile in higher education: How can value-based learning be implemented in higher education? In: Proceedings of the 18th International Conference on Web Information Systems and Technologies (WEBIST 2022) (10 2022). <https://doi.org/10.5220/0011537100003318>
31. Schön, E.M.: How do agile practices support organizing a ph.d.? *IT Prof.* **20**, 82–86 (2018). <https://doi.org/10.1109/MITP.2018.2876927>. <https://ieeexplore.ieee.org/document/8617761/>
32. Schön, E.M., Winter, D., Escalona, M.J., Thomaschewski, J.: Key challenges in agile requirements engineering. In: Baumeister, H., Lichter, H., Riebisch, M. (eds.) *Agile Processes in Software Engineering and Extreme Programming*, pp. 37–51 (2017). [https://doi.org/10.1007/978-3-319-57633-6\\_3](https://doi.org/10.1007/978-3-319-57633-6_3)
33. Schwaber, K., Sutherland, J.: *The scrum guide* (2020)
34. Sidky, A., Arthur, J., Bohner, S.: A disciplined approach to adopting agile practices: the agile adoption framework. *Innov. Syst. Softw. Eng.* **3**(3), 203–216 (2007). <https://doi.org/10.1007/s11334-007-0026-z>
35. Stewart, J.C., DeCusatis, C.S., Kidder, K., Massi, J.R., Anne, K.M.: Evaluating agile principles in active and cooperative learning. Student-Faculty Research Day, CSIS, Pace University, pp. B3.1–B3.8 (2009). <https://csis.pace.edu/~ctappert/srd2009/b3.pdf>
36. Stolze, A., Fritsch, K.: *The eduscrum® guide* (2020)
37. Tolfo, C., Wazlawick, R.S., Ferreira, M.G.G., Forcellini, F.A.: Agile methods and organizational culture: reflections about cultural levels. *J. Softw. Maint. Evol. Res. Pract.* **23**, 423–441 (2011). <https://doi.org/10.1002/smr.483>
38. Wright, G.B.: Student-centered learning in higher education. *Int. Soc. Explor. Teach. Learn.* **23**(3), 92–97 (2011)
39. Yin, R.K.: *Case Study Research: Design and Methods*, vol. 5. SAGE Publications, Thousands Oaks (2003)



# Improving the Representation Choices of Privacy Policies for End-Users

Michalis Kaili<sup>1,2</sup> and Georgia M. Kapitsaki<sup>1</sup>(✉) 

<sup>1</sup> Department of Computer Science, University of Cyprus, Nicosia, Cyprus  
gkapi@ucy.ac.cy

<sup>2</sup> Chalmers University of Technology, Gothenburg, Sweden

**Abstract.** Privacy policies provide users the possibility to get informed about how their data are being used by specific services and vendors. Unfortunately their texts are usually long and users are not devoting the required time to read them and understand their content. Tools that bring the privacy policies closer to the users can assist towards enhancing users' privacy awareness. In this work, we are presenting the updated version of *Privacy Policy Beautifier*, our approach and accompanying tool that offers various representations of the privacy policy text, as a way to assist the users in better understanding the policy, devoting less time to explore its main content. Text highlighting, text summarization, word cloud, GDPR terms presence/absence are the techniques employed for the representations. The updated version of Privacy Policy Beautifier has been evaluated for its enhanced features via the participation of 32 users with promising results.

**Keywords:** Privacy policy · Text presentation · Text summarization · GDPR

## 1 Introduction

Different kinds of applications are used daily by users assisting in their daily activities, ranging from web applications to mobile and smart devices. Privacy policies are the central location where users can be informed about the terms of an organization concerning the collection, usage, sharing and maintenance of their personal data, as well as the handling of their rights under the EU General Data Protection Regulation (GDPR) and other relevant laws and regulations [21].

For users, it is important to be informed on the above concerning their personal data, as this can assist them in subsequent decisions concerning the use of specific services and applications, whereas they may even have the possibility to decide exactly which kind of personal information they are willing to disclose to such applications [19]. Privacy policies have the main problem of being long to read, whereas vague terms are used in many cases, making the process of understanding their content a difficult task [3]. Previous works have focused on using Artificial Intelligence (AI) techniques to provide summaries of texts or pinpoint to specific locations [8] or have focused on the representations of the

policy from a theoretical perspective [17] but they have not investigated different representations in an integrated technical platform. In the current work, we are focusing on this last aspect. Therefore, in comparison to previous works, we are experimenting with different kinds of visualizations, offering at the same time a technological approach to offer those visualizations to users in an automated way, accompanying the approach with a tool implementation.

In a previous publication, we have introduced the initial version of *Privacy Policy Beautifier* [9]. The aim of and purpose of that tool was to encourage users to read, at least partially, privacy policies from websites or applications that they use or intend to use in the future. The main idea is to alter the appearance of a privacy policy in a way that makes it easier to read, navigate, and extract information that interests the user. In order to accomplish this, the tool provides several tabs with each tab containing different visualisations either of the privacy policy text itself or information related/extracted from it. In the current work, we are presenting the extended version of *Privacy Policy Beautifier* that includes an updated list of GDPR terms presence or absence, a text summarization representation, as well as a user evaluation that focuses on the new functionality and the usability and acceptance of the system with the evaluation taking place mainly in a different cultural setting than the evaluation of the original version of the approach.

The remaining of the text has the following structure. Section 2 presents the background and the related work. Section 3 is dedicated to the process in which *Privacy Policy Beautifier* uses in order to improve the policy text presentation, as well as to how the new summarization feature was implemented. In Sect. 4, the prototype is presented along with a use demonstration of the system, whereas Sect. 5 analyzes the user evaluation process and its main outcomes. Section 6 looks into the threats to validity and, finally, Sect. 7 concludes the paper mentioning possible future directions of the work.

## 2 Background and Related Work

Relevant previous works can be found in the area of web policies and usable privacy. Earlier studies in the literature have suggested new ways of conveying privacy policies to users [10]. Giving people the option to make thoughtful decisions about the sharing of their personal information is an important topic, as noted by Angulo et al. [1]. The authors provide their research findings from the PrimeLife project, where the PrimeLife Policy Language (PPL) was developed and tested via a “Send Data?” prototype browser extension. The aim of the browser extension is to explain the essential components of a service provider’s privacy policy to the user.

A corpus of 6,278 distinct English-language privacy policies from inside and outside the European Union was produced in order to study the effects of GDPR on privacy policies, and their pre-GDPR versions were then compared [15]. It was noted that privacy policies are now considerably longer in relation to older privacy policies, most likely to address and satisfy the new standards and regulations. The updated privacy policies have improved visual representation in

addition to being more comprehensive, making them more appealing to consumers. However, it should be emphasized that prior rules altered the landscape of privacy policies by encouraging more websites to adopt or modify their privacy policies, some of which are now more comprehensive and informative. Yet doing so was always at the expense of the privacy policy’s readability and clarity. Although there has been an improvement in visual representation, policy texts are still too long for consumers to read quickly.

In another work relevant to GDPR, a tool called CompLicy [18] has been released. It examines the extent to which privacy rules indicate GDPR provisions. CompLicy’s mission is to demonstrate how well a privacy policy wording incorporates GDPR guidelines. The required text is extracted using a parser from the privacy policy web page. Following that, the content is examined and processed along with a list of terms and phrases that are related to GDPR. As a result a score is assigned to the privacy policy to indicate how well it addresses every need of the GDPR, and a further in-depth analysis identifies which requirements have been met and which are missing from the privacy policy. The authors provide the view of the current state of privacy policies on GDPR compliance for the web applications domain.

The OPP-115 dataset [20], which is also used in the context of the current work, has provided the most useful support in comprehending privacy policies, even if more work is still needed in the area. This work describes the production process, such as choosing a privacy policy, the dataset’s structure, its content, and preliminary experimentation. Deep learning and graphs for describing the necessary information are presented in Polisis, which makes use of the aforementioned OPP-115 dataset, as more automated ways to assess, process, and alter the privacy policy to make it simpler for the typical user to grasp [8]. In the same endeavour, Pribot was developed to aid in addressing both structured and free-form inquiries about the policy.

There have been attempts to use manual or automatic techniques to improve the visualization of privacy policies when they are presented and visualised. Instead of the conventional textual representation, a proposed study suggests different visualisation strategies for privacy regulations, but with a focus on how each style affects users and their privacy awareness [17]. Unchanged policy text, WordBridge tag clouds, and Document Cards were the three visualizations that were employed. The privacy policy of Instagram was utilised to be changed to the above representations, although no automatic technique was employed for this process. When compared to the conventional textual representation, the evaluation conducted with students found that employing the two visualization techniques increased the policy awareness of users. This study reinforces the drive for the ongoing work to employ various visuals, but our focus still remains on automating the methodology [9].

PrivacyCheck is a strategy that uses data mining to produce summaries of privacy regulations that give the user the overall notion in a smaller, more digestible chunk and colour coded symbols [22]. An overview of the privacy policy’s contents is provided to the user via the browser plugin PrivacyCheck. This approach makes use of classification techniques to examine the content of

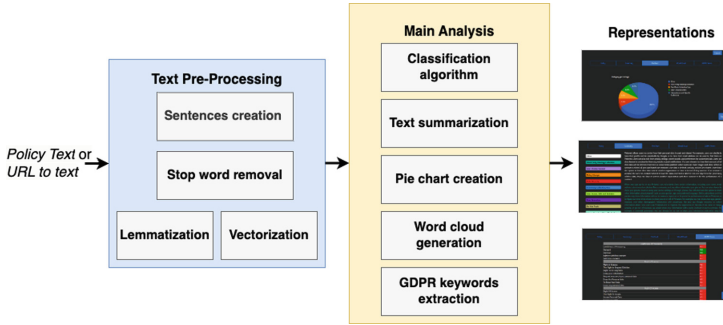


Fig. 1. Main phases and processing steps.

GDPR and user control in the policy text. Inspired by this and building on our previous work [9] our tool *Privacy Policy Beautifier* was expanded with a summarization feature. This decision was also inspired by the recent trend in AI usage (i.e. ChatGPT) [16].

### 3 Process and Representation Approaches

The goal of *Privacy Policy Beautifier* is to give people a more streamlined approach to view privacy policies that are available online. For this reason, it offers various visual representations of the text’s various sections, including: 1) highlighting passages that refer to particular topics of interest, such as categories like “data category” and “data retention”, 2) displaying a pie chart with the categories encountered, 3) displaying a word cloud of the original policy text, and 4) highlighting any GDPR-related terms that are mentioned in the text in a table format. The primary function of the proposed system is text highlighting, which is carried out using supervised machine learning techniques by applying categorization to the content of the policies. After being divided up and given labels, the material is presented to the user in a highlighted format so they can more easily navigate to particular sections of the policy text. This allows the user to read the information they need without having to spend additional time reading through the entire text to find a certain section.

The updated version of *Privacy Policy Beautifier* as presented in the current work expands on the tool by adding a more complete list of GDPR terms (updated with more terms stemming mainly from GDPR user rights) and by introducing the ability to summarize the previously highlighted sections into more bite-sized chunks. The underlying processing remains the same for the parts available in the initial implementation but with a few further steps added. These new additions are examined more closely in the following subsections. Figure 1 illustrates the steps of *Privacy Policy Beautifier*.



### 3.1 Text Pre-processing

This stage involves cleaning up the text by removing information that is not necessary for the classification process. For this purpose, the policy wording (i.e. whole text) is initially divided into sentences. Thereafter, common stop words (in English) are eliminated. The list of stop words utilised includes all 127 of the standard stop words from Python’s NLTK (Natural Language Toolkit) package. The terms are then lemmatized as a next step. It should be noted that while stemming was also investigated as shown in Fig. 1, lemmatization produced superior results in the experiments with various classifiers as stated below and this is why it was chosen as part of the text pre-processing. The text is finally transformed into a vector representation. The term frequency-inverse document frequency (TF-IDF), bag-of-words, and other strategies were used in this step and they were compared to attain the maximum accuracy, with TF-IDF being chosen as the final technique applied.

### 3.2 Text Representations

**Text Highlighting View Using Classification.** The primary function of *Privacy Policy Beautifier*, which uses supervised machine learning to build a model to categorize the various elements of the privacy policy text, is text highlighting. To help the user identify the information they are most interested in, these categorized segments are then presented to the user in various colours.

The OPP-115 dataset, which has been extensively used in prior publications [20], was used for the classifier’s training and testing. The following 10 categories from the OPP-115 dataset are used to emphasize the text. Given a sentence from a privacy policy text, the classifier places it in one of the ten categories listed below:

- *First Party Collection/Use*: indicates by which ways and for which reasons data are being collected by the service provider.
- *Third Party Sharing/Collection*: reflects how data are collected or shared with third parties, if any such collection/sharing is performed.
- *User Choice/Control*: indicates the choices and controls available for users.
- *User Access, Edit and Deletion*: mentions if and how users can perform the above actions.
- *Data Retention*: is related to the duration of storing data, as indicated by the service provider.
- *Data Security*: includes any security techniques adopted and applied on the data.
- *Policy Change*: informs about if and how users are informed about changes performed on the privacy policy.
- *Do not Track*: is related to how the “do not track” option is applied.
- *International and Specific Audiences*: may contain practices applicable to specific user groups.
- *Other*: includes any text not covered in other categories (covers also any introductory text the policy may have).

For the main classification process, the following widely used text categorization algorithms were used and compared in terms of accuracy [5]:

1. Multilayer perceptron (MLP) [7]: MLP is a feedforward neural network which is composed of multiple layers of perceptrons.
2. Naive Bayes classifier [12]: it is based on the Bayes Theorem and includes an assumption of independence among the predictors.
3. Random forest classifier [14]: it is a supervised learning algorithm that trains a set of decision tree classifiers.

The results of the training procedure are presented in Table 1, where the classifier accuracy, which counts the proportion of accurate predictions to all other predictions listed, is shown. The Random forest classifier was used in the development of the web platform for *Privacy Policy Beautifier* since it had the highest accuracy in comparison to the other algorithms used. Lemmatization as part of the preparation procedures gave the random forest classifier that had the greatest accuracy an accuracy of 74% compared to 70% with the use of stemming.

**Table 1.** Accuracy of classification algorithms.

Algorithm	Accuracy
MPL	50%
Naive Bayes	60%
Random forest	74%

**Text Summarization View.** The new version of *Privacy Policy Beautifier* includes a new visualization, the text summarization. As mentioned before this new visualization was partly inspired by past research, namely PrivacyCheck [22] and by the sudden influx in popularity of AI chatbots, such as ChatGPT. This new visualization feature aims to summarize the content of the previously divided classifications made in the text highlighting view.

In order to prepare the summarization, these previously divided categories are being sent to an existing API and with the use of its pre-trained models and the customizations performed by our implementation, we are able to retrieve a summary of each of the 10 categories mentioned in OPP-115, in case they exist in the respective privacy policy text. More specifically, the text from each category is sent to an API provided by OpenAI<sup>1</sup> along with a custom made prompt in order to create a summary of said text as accurately and as concisely as possible. We would not want to miss out on any important information from the initial text, since it would not be desirable to provide the user with wrong or misleading information but at the same time we do not want to end up with a text that is so long that ends up discouraging the user from reading it, resulting in the

<sup>1</sup> <https://openai.com/>.

same problem we intended to solve originally. This is a very delicate balance to achieve and one that we hope to improve upon in possible future iterations of the tool. For convenience and to avoid confusion while maintaining consistency in the tool the summaries presented to the user maintain the color coding that was assigned to them in the text highlighting view.

This new visualization aims to help the user gain a general understanding of what a certain part of the privacy policy text is containing. Even though the text is altered both in form and in structure, we believe that it can provide valuable information for the user as a first step to understand the privacy policy and maybe even give the user an incentive to go one step further and read the initial privacy policy text itself in order to visit the parts that caught his/her attention. It should be noted that any AI generated text cannot be used without assuming that it may include flaws, so users' discretion is advised when advising the privacy text summarization.

**Word Cloud View.** Word clouds are a text visualisation method that has been widely used in earlier publications and online [11]. This illustration was added so that the user could have a broad understanding of the privacy policy's contents and how much of each term they should anticipate to see without having to read the entire text. In the original policy text, stop words were not disregarded for this representation type.

**Presence/Absence of GDPR Terms View.** For this representation, we relied on a keyword list from [18] in order to discover which terms appear in the policy text. In the updated version of *Privacy Policy Beautifier*, we used an enlarged version of the keywords list. The following GDPR categories are covered by the terms:

- *Lawfulness of Processing* (example terms: Lawfulness of Processing, Consent, Contract, Right to Withdraw Consent, Right to Withdraw consent)
- *Right to Restriction of Processing* (example terms: Right to Requests the Restriction of Their Use, Right of data subjects to be informed about the restriction, Restriction of Processing, Restrict Your data)
- *Right to Data Portability* (example terms: Right to receive a copy of your personal information, Right to Transmit Data, Request the transfer of your personal data, Right to Receive the Personal Data)
- *Right to Rectification* (example terms: Right to demand processing restrictions, Right to request correction, Right to complete incomplete personal data, Right to have incomplete personal data)
- *Right to Object* (example terms: Right to Object at any time to Processing of Personal data, Right to Object to Processing, Right to Object at any time to Processing, Processing Objection, Object to processing)
- *Right to Erasure* (example terms: Right of Erasure, Right to Request Deletion, Right To be Forgotten, Right to Erase your Information, Right to Request erasure of your personal data)

- *Right of Access by the Data Subject* (example terms: The Right to Lodge a Complaint, Right Of Access, Right to request and receive information, Right to Request a Copy of your data, Right to File a Complaint)

This section of the work is thus, unique to GDPR rules and rights that are applicable to web platforms. The privacy policy text was searched for the precise terms that covered the indication of the same GDPR right but with different language used, as the creator of the privacy policy may slightly change the wording of the right. The user is told that the term's inclusion indicates that the policy text, and by extension the degree to which the applicable application complies with GDPR. At the same time, the users is informed on the absence of other terms from the policy text.

## 4 Privacy Policy Beautifier Demonstration

In this section, we are providing some examples of what the user can see when interacting with the *Privacy Policy Beautifier*. As a first step, the user can add a new privacy policy text to be analyzed either by copying and pasting the relevant content as shown in Fig. 2 or by pointing to a URL where the policy text is available.

Various tabs on the User Interface are available to the user for the five supported representations:

- Text highlighting tab.
- Summary tab, created using the results of the text highlighting process by summarizing the concatenated sentences.
- Pie chart tab, created using the results of the text highlighting process by indicating the percentage of appearance of each of the 10 categories.
- Word cloud tab.
- GDPR terms presence/absence tab.

Concerning the text highlighting outcomes, an example of which is illustrated in Fig. 3, the colour coded and dynamically inserted categorised segments provided by the classifier are added in a way so as not to alter the privacy policy's original structure. The user is taken to the specific sections of the text that include text from the selected category by selecting it from the filters on the left side of the page. This enables the user to concentrate on that particular element of the policy, as the font size of the relevant text is enlarged. When interacting with the *Privacy Policy Beautifier*, the user can activate and deactivate the category filtering as they wish as long as the category itself exists in the original privacy policy text. The pie chart is also used as a summary indicating the presence of the various components as displayed in Fig. 4 in order to convey to the user on which categories the privacy policy lays more focus and whether or not a given category is present. Furthermore, the user may see the word clouds of the policy's terms and the existence of GDPR terms in each of their respective tabs (Fig. 5 and 6 respectively). Each GDPR phrase is highlighted in green when it is



Fig. 2. Privacy policy text addition.

present and in red when it is absent. It should be noted that the choice to employ colours, images, and 2D tables was driven by research from earlier studies that showed how appealing and simple they are to read for users [17].

As aforementioned, in relation to the first version of *Privacy Policy Beautifier* [9], a new summary tab was added as well as an updated GDPR terms list. This new summary tab as seen in Fig. 7 provides the user with a summary of the contents of each of the categories recognized in the previous steps in the process.

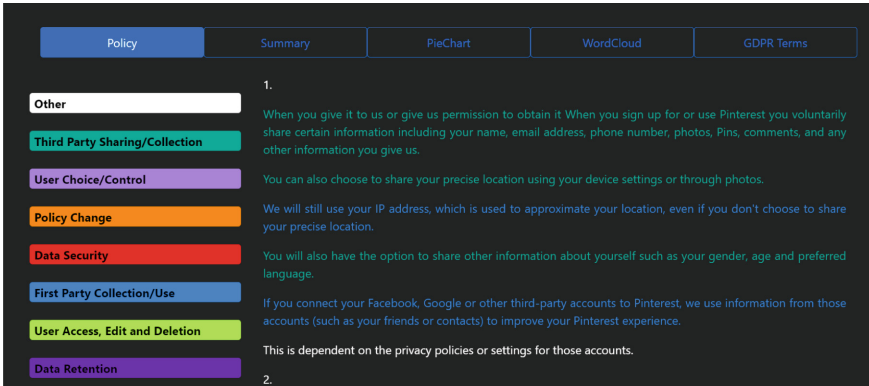


Fig. 3. Text highlighting view.

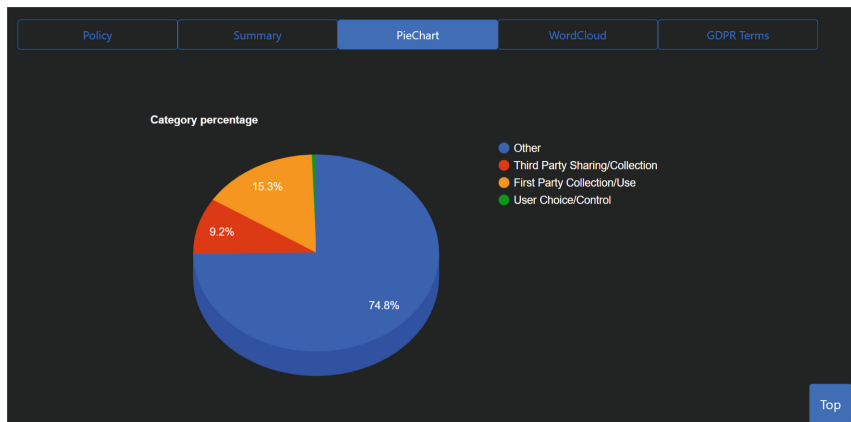


Fig. 4. Pie chart with policy categories view.

## 5 User Evaluation

Initial feedback collected by users with the participation of 90 individuals was presented in the existing publication [9]. The main conclusions that were drawn after analyzing all participants answers were that a tool like *Privacy Policy Beautifier* would be welcomed by users and that if provided they would gladly use it, possibly at a regular basis. These answers came from a questionnaire that was sent out to potential participants since we had chosen to use the survey methodology to assess the system’s usability and privacy awareness, among other factors. More details about the questionnaire, the responses and the conclusions from it can be found in publication of the initial version of the *Privacy Policy Beautifier* [9] although some comparisons are made also later in the text.

In the current work, we present an additional evaluation with users from a different cultural setting focusing on the new features of the platform and its usability characteristics, whereas we compare with the previous evaluation. As before, we created a questionnaire that was sent out to participants in order to assess the tool in its entirety, as well as the new features that were added. The questionnaire available online<sup>2</sup> begins with all participants submitting their consent after being told about the questionnaire’s purpose and the applicable use of the obtained data. Demographic data on the participants were acquired in the first section (i.e. age, gender, level of education, educational background). The questionnaire asked participants about their prior experiences reading privacy regulations. Finally, the participants were asked for feedback on their experience using *Privacy Policy Beautifier*, including some usability-related questions like how simple it was to use and whether they would use it again.

For this questionnaire some questions on usability were added following the System Usability Scale (SUS) [4, 13] that includes the following 10 questions:

<sup>2</sup> <https://forms.gle/bjARtmjkuZjSYkveA>.

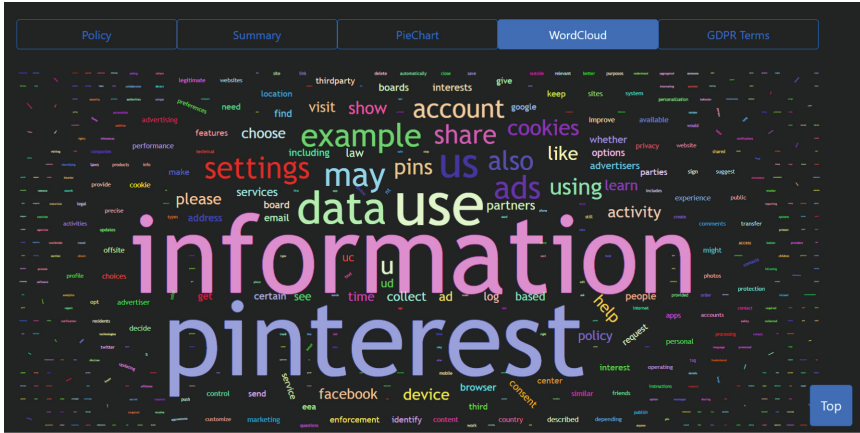


Fig. 5. Word cloud view.

GDPR Term	Status
Lawfulness of Processing	
Lawfulness of Processing	NO
Consent	YES
Contract	YES
right to withdraw consent	NO
Withdraw consent	NO
Right of Erasure	
Right of Erasure	NO
The Right to Request Deletion	NO
Right To be Forgotten	NO
Erase your Information	NO
Request erasure of your personal data	NO
Erase the Personal data	NO
To Erase Your Data	NO
Erase any personal data	NO
Right Of Access	
Right Of Access	NO
The Right To Access	NO
Access Personal Data	NO
Access your data	NO

Fig. 6. GDPR terms presence/absence view.

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

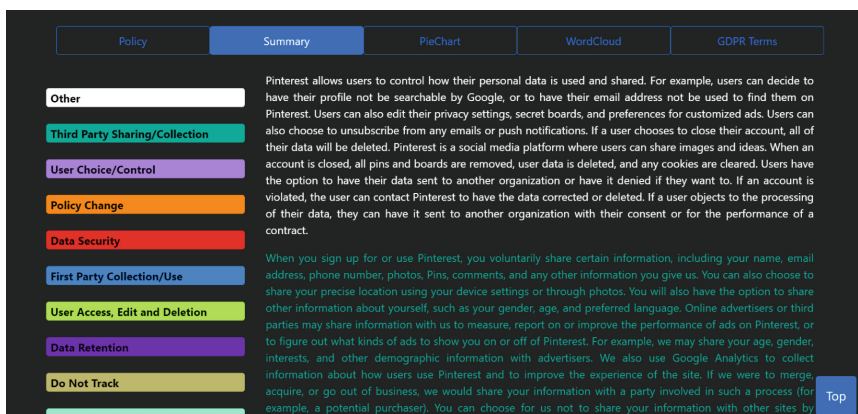


Fig. 7. Privacy policy summarising view.

This was done in order to put emphasis on the evaluation of the user friendliness of the tool, as it is very important to ensure that users will come back after using the tool once. Another notable difference for this questionnaire were the questions centered around the new summarization tab. We wanted to make sure that the new feature operated as expected and we gathered users' view on how they see it.

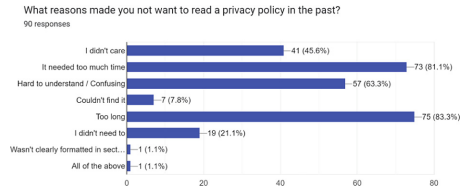
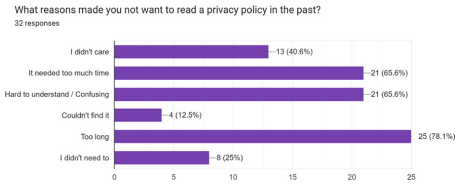
After filling in their demographic data and before continuing with the remaining of the questionnaire the users were asked to navigate to the *Privacy Policy Beautifier* tool and interact with it to any extend they wished before returning to answer the next questions regarding their experience with the tool.

In total 32 participants participated in the questionnaire after distribution of the survey by email communication and personal contact. About a third of them were the same users from the first iteration of the survey that used the initial version of the tool [9]. Despite the difference in a number of the participants, the ratios between the two questionnaires concerning the age and gender of our participants remains very similar: 62.5% males and 37.5% females participated in the current survey, versus 55.6% and 40% respectively in the first, whereas the dominant age group of the participants was above 60 years old for 43.8% of the participants in the current questionnaire versus 41.1% in the first (followed by the age group of 25–29 with 21.9% of participants in the current and 18.9% in the previous survey). The same is observed about the level of education (43.8% of users possess a master's degree and 43.8% a bachelor's degree versus 38.9% and 47.8% respectively in the initial survey), as well as other demographic data despite most of the participants coming from different countries than the first study (most participants of the new study are residing in Sweden, whereas most participants from the first study are residing in Cyprus).

But most importantly the trends of how many have read privacy policies before and how their experience was remains the same: most participants have read at least a privacy policy partially (62.5% have partially read a privacy



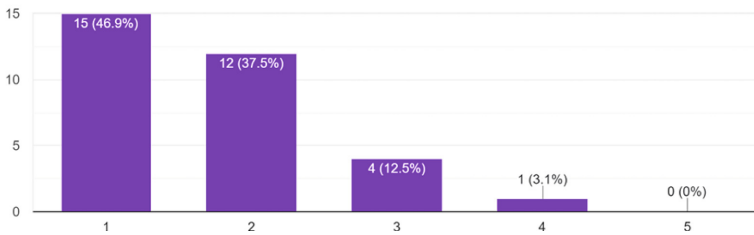
policy versus 54.4% in the previous survey) but the majority of them has had an unpleasant experience with it with even more participants indicating this now (44.8% of users mentioned that they did not enjoy the experience versus 37.3% in the first survey). We can also see that the reasons that the participants indicated for choosing not to read a privacy policy are also very similar as seen in Fig. 8 and 9 which show both the results of the second and of the first survey.



**Fig. 8.** Reasons not to read a policy in current study.

**Fig. 9.** Reasons not to read a policy in 1st study [9].

I think that I would need the support of a technical person to be able to use this system.  
32 responses

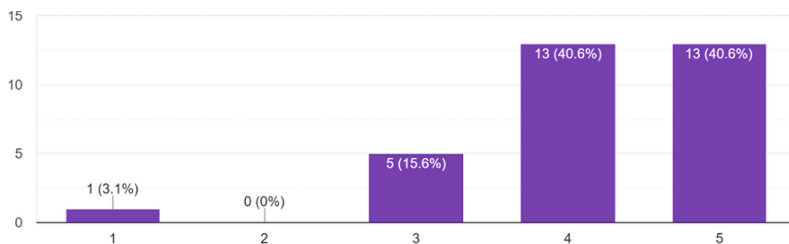


**Fig. 10.** Would you need technical assistance to operate the tool?

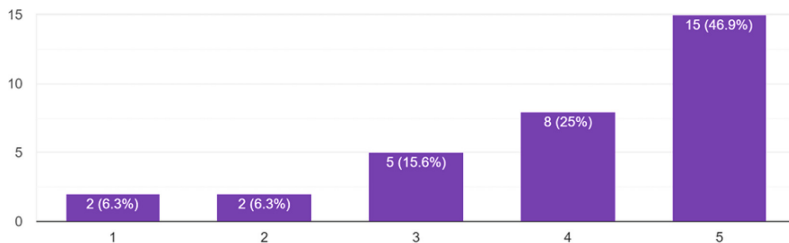
### 5.1 User Views on Usability

It is encouraging to see that the tool is easy to understand and easy to use according to the questionnaire participants, achieving the desired goal of a user friendly interface that anyone can use despite their technological skills and knowledge as seen in Fig. 10, 11 and 12. It is also important to note at this point that the new functionality of the tool, the newly added summary tab, also received positive feedback, as seen in Fig. 13 and 14.

Participants were also requested to submit feedback on the system, and some of their suggestions were incorporated into the finished product. A few suggestions have already been added, such as the ability to disable a filter by clicking it once more rather than having to click the “clear filter” button, the ability to



**Fig. 11.** Was the tool easy to use?



**Fig. 12.** Will people be able to learn how to use the tool quickly?

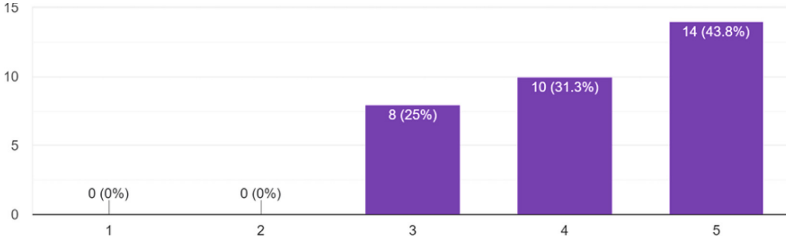
scroll until the first instance of the category the user clicked on is visible, and the ability to make the selected filter more obvious and distinct. Overall, the tool seems to be well received by participants, and with only a few minor complaints, issues or suggestions appearing in the open-ended questions. Some more examples of these suggestions in addition to the ones integrated in the system include:

1. Provision of the tool as Chrome browser plugin (indicated by 2 participants).
2. The text in policy should have a larger font.
3. Suggestion for use of different colours in the user interface.

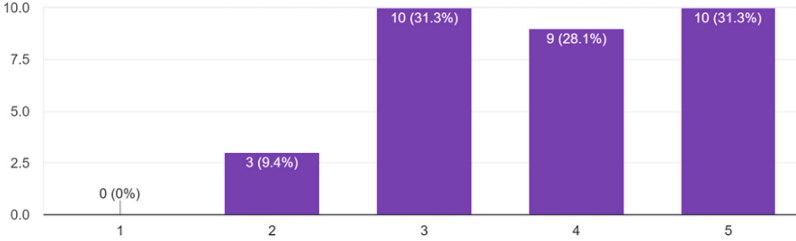
Regarding System Usability Scale, we calculated the overall score to 71.41 (out of 100) that is a good result since values over 70 are considered good for SUS [2]. The score for each of the 10 questions can be found in Fig. 15 with some questions scoring better than others: questions 4 and 10 have a score of 80 and above showing that the system is easy to use, whereas question 1 has the lowest score than all that may be logical considering that users may not need to use of *Privacy Policy Beautifier* on a very frequent basis.

## 5.2 User Views on Policy Representations

Another important question that was asked to the participants of the survey was: *Which presentation of the content of the privacy policy did you prefer?*. This question was aimed to see what visual representation the users preferred the most. The results of this question as seen in Fig. 16 reveal that the new



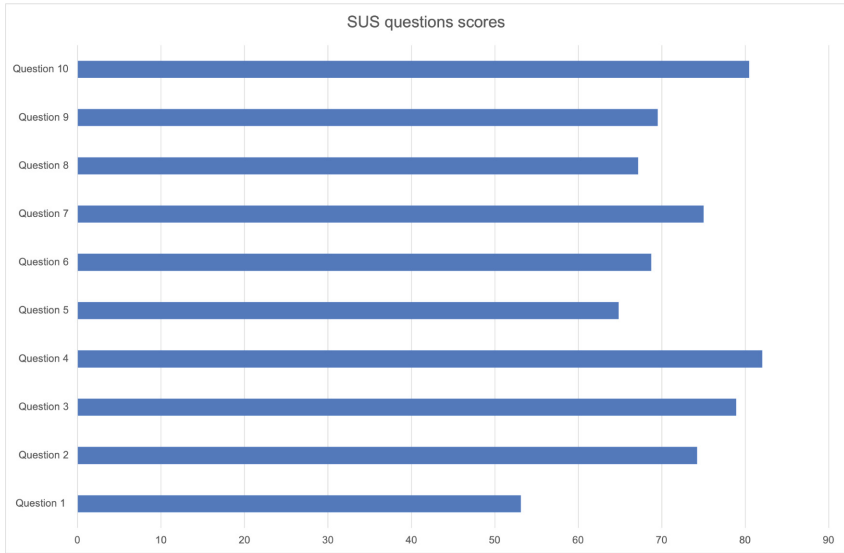
**Fig. 13.** Did you find the summary tab useful?



**Fig. 14.** Would you use the summary tab in the future?

summary visualization is the most popular by quite a margin, verifying that it was a good option to provide this representation to the users.

We also performed some limited statistical analysis in order to examine if there are differences in how participants perceive the summary tab. Although our sample size of users is small, this analysis can show us some tendencies in the results. We investigated whether the gender or the technical expertise of the participants affects how they view the summary tab in terms of: 1) how easy it is to use it, 2) how useful it is, and 3) whether they trust the results. We observed some statistically significant results only for the usefulness of the summary ( $p - value = 0.025$ ) with non-technical users finding the summary tab more useful: the mean value on the 5-likert scale ranging from Strongly disagree to Strongly agree is 4.86 for non-experts, 4.2 for technical experts and 3.87 for those that have some technical knowledge. The exact percentages for each group are shown in Table 2. Non-technical users are thus, finding the summary tab more useful, showing that technical users may be more familiar with the terminology of the privacy policies and more willing to read its full content.



**Fig. 15.** System Usability Scale questions scores.

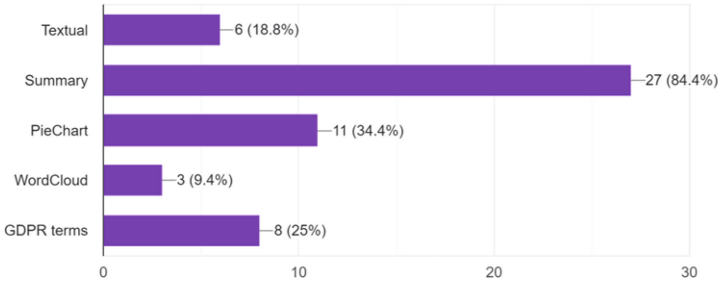
**Table 2.** Usefulness of summary tab for participants (*Did you find the summary tab useful?*).

Technical expert	# users	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Yes	10	0%	0%	20%	40.0%	40.0%
No	7	0%	0%	0%	14.3%	85.7%
I know some stuff	15	0%	0%	40.0%	33.3%	26.7%

## 6 Threats to Validity

There are a few threats to validity, in the form established by Feldt et al. [6], when it comes to the current research. Starting with *construct validity*, our tool is reliant on the accuracy of our classification model. The most significant restriction on the text highlighting approach’s accuracy is the modest size of the training data and the fact that the data lacks a wide variety of diversity. The zero frequency problem in the case of Naive Bayes, where the algorithm assigns a zero probability to a categorical variable whose category in the test dataset was not available in the training dataset, may have led to a lower accuracy in the results. Other specific characteristics of the classifiers may also have had an impact on the accuracy of each one.

The user study is impacted by *external validity*, which refers to how far we may apply our findings to more general populations. Even though the sample group taking part in our survey was different than our previous work [9] it was also significantly smaller. More users had the opportunity to interact with



**Fig. 16.** Results on question *Which visual representation did you prefer?*

the tool in total, especially with the pre-existing features, but the newly added summary feature was seen by a smaller number of users. Although the users in our user sample came from a variety of backgrounds (including ages, levels of competence, nationality), a bigger sample or repeating the study in users of different cultural backgrounds might yield slightly different findings.

## 7 Conclusions

Through the *Privacy Policy Beautifier* tool, which allows users to view policies in a variety of ways, including a textual format with text highlighting, as a pie chart, as a word cloud, as a table with indicators of the presence or absence of GDPR terms and the newly added summary tab, we have presented our work on more user-friendly representations of the text of privacy policies. The suggested classifier's classification accuracy (74%) demonstrates encouraging findings that can be further improved, while the user study revealed that users value the various representations, with many users interacting well with various representations while showing a preference for the summary tab.

The use of unsupervised techniques or a combination of supervised and unsupervised techniques may be able to increase classification accuracy, according to future research. As part of our ongoing work, we also plan to improve the *Privacy Policy Beautifier* by adding support for different languages.

## References

1. Angulo, J., Fischer-Hübner, S., Wästlund, E., Pulls, T.: Towards usable privacy policy display and management. *Inf. Manag. Comput. Secur.* **20**(1), 4–17 (2012)
2. Bangor, A., Kortum, P., Miller, J.: Determining what individual SUS scores mean: adding an adjective rating scale. *J. Usability Stud.* **4**(3), 114–123 (2009)
3. Bhatia, J., Breaux, T.D., Reidenberg, J.R., Norton, T.B.: A theory of vagueness and privacy risk perception. In: 2016 IEEE 24th International Requirements Engineering Conference (RE), pp. 26–35. IEEE (2016)
4. Brooke, J., et al.: SUS-a quick and dirty usability scale. *Usability Eval. Ind.* **189**(194), 4–7 (1996)

5. Dhar, A., Mukherjee, H., Dash, N.S., Roy, K.: Text categorization: past and present. *Artif. Intell. Rev.* **54**(4), 3007–3054 (2021)
6. Feldt, R., Magazinius, A.: Validity threats in empirical software engineering research—an initial survey. In: SEKE, pp. 374–379 (2010)
7. Gardner, M.W., Dorling, S.: Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmos. Environ.* **32**(14–15), 2627–2636 (1998)
8. Harkous, H., Fawaz, K., Lebret, R., Schaub, F., Shin, K.G., Aberer, K.: Polisris: automated analysis and presentation of privacy policies using deep learning. In: 27th {USENIX} Security Symposium ({USENIX} Security 2018), pp. 531–548 (2018)
9. Kaili, M., Kapitsaki, G.M.: Privacy policy beautifier: Bringing privacy policies closer to users. In: Decker, S., Mayo, F.J.D., Marchiori, M., Filipe, J. (eds.) Proceedings of the 18th International Conference on Web Information Systems and Technologies, WEBIST 2022, Valletta, Malta, 25–27 October 2022, pp. 54–63. SCITEPRESS (2022). <https://doi.org/10.5220/0011541600003318>,
10. Kelley, P.G., Bresee, J., Cranor, L.F., Reeder, R.W.: A “nutrition label” for privacy. In: Proceedings of the 5th Symposium on Usable Privacy and Security, pp. 1–12 (2009)
11. Kim, K., Ko, S., Elmqvist, N., Ebert, D.S.: Wordbridge: using composite tag clouds in node-link diagrams for visualizing content and relations in text corpora. In: 2011 44th Hawaii International Conference on System Sciences, pp. 1–8. IEEE (2011)
12. Leung, K.M.: Naive Bayesian classifier. Polytechnic University Department of Computer Science/Finance and Risk Engineering 2007, pp. 123–156 (2007)
13. Lewis, J.R.: The system usability scale: past, present, and future. *Int. J. Hum.-Comput. Interact.* **34**(7), 577–590 (2018)
14. Liaw, A., Wiener, M., et al.: Classification and regression by randomforest. *R News* **2**(3), 18–22 (2002)
15. Linden, T., Khandelwal, R., Harkous, H., Fawaz, K.: The privacy policy landscape after the GDPR. arXiv preprint [arXiv:1809.08396](https://arxiv.org/abs/1809.08396) (2018)
16. Lund, B.D., Wang, T.: Chatting about ChatGPT: how may AI and GPT impact academia and libraries? *Library Hi Tech News* (2023)
17. Soumelidou, A., Tsohou, A.: Effects of privacy policy visualization on users’ information privacy awareness level: the case of Instagram. *Inf. Technol. People* **33**(2), 502–534 (2019)
18. Vanezi, E., Zampa, G., Mettouris, C., Yeratziotis, A., Papadopoulos, G.A.: CompLicy: evaluating the GDPR alignment of privacy policies - a study on web platforms. In: Cherfi, S., Perini, A., Nurcan, S. (eds.) RCIS 2021. LNBP, vol. 415, pp. 152–168. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-75018-3\\_10](https://doi.org/10.1007/978-3-030-75018-3_10)
19. Wagner, C., Trenz, M., Veit, D.: How do habit and privacy awareness shape privacy decisions? (2020)
20. Wilson, S., et al.: The creation and analysis of a website privacy policy corpus. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1330–1340 (2016)
21. Wu, K.W., Huang, S.Y., Yen, D.C., Popova, I.: The effect of online privacy policy on consumer privacy concern and trust. *Comput. Hum. Behav.* **28**(3), 889–897 (2012)
22. Zaeem, R.N., German, R.L., Barber, K.S.: Privacycheck: automatic summarization of privacy policies using data mining. *ACM Trans. Internet Technol. (TOIT)* **18**(4), 1–18 (2018)



# Scaffolding Process-Aware Information Systems with the AKIP Platform

Ulisses Telemaco Neto<sup>1</sup>(✉) , Toacy Oliveira<sup>2</sup> , Raquel Pillat<sup>2</sup> , Paulo Alencar<sup>1</sup>,  
Don Cowan<sup>1</sup> , and Glauca Melo<sup>1</sup>

<sup>1</sup> David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Canada  
{utelemaco, palencar, dcowan, gmelo}@uwaterloo.ca

<sup>2</sup> System Engineering and Computing Program, Federal University of Rio de Janeiro,  
Rio de Janeiro, Brazil  
{toacy, rmpillat}@cos.ufrj.br

**Abstract.** Business Process Automation (BPA) refers to the automation of business processes through technology, with the goal of improving efficiency, reducing errors, and increasing productivity by eliminating manual and repetitive tasks. Over the past few years, the number of BPA platforms has increased, including both open-source and proprietary solutions. However, some challenges and limitations still exist related to the adoption of these solutions, such as vendor lock-in, limited UI/UX, limited integration, outdated technology stack, and lack of support for non-process features. To address these issues, this paper presents the AKIP platform, an open-source framework for developing process-aware information systems (PAISs) using code generation techniques. AKIP generates functional process-aware web applications from BPMN business process models, making it the only known software tool capable of generating fully functional process-aware web applications. To evaluate the effectiveness of the AKIP platform, a case study was conducted in the industry and six business processes automated into a process-oriented web application were analyzed in this paper. The study showed that the AKIP platform was able to generate a functional web application while supporting the automation of business processes with different model sizes and complexities. Furthermore, this study showed that even a team of professionals with little experience was able to produce such results by using the AKIP platform.

**Keywords:** Business process automation · Code generation · Process-aware information system · BPMN

## 1 Introduction

There are several reasons why investing in business process automation (BPA) can benefit a company:

1. **Increased Efficiency:** Automation can greatly increase the speed and efficiency of repetitive tasks, freeing up employees to focus on higher-level tasks.

The authors thank the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Centre for Community Mapping (COMAP).

2. **Improved Accuracy:** Automated processes are less prone to human error, leading to increased accuracy and consistency in the data and results.
3. **Cost Savings:** Automation can help to reduce labor costs and other expenses associated with manual processes.
4. **Increased Scalability:** Automated processes can be easily replicated and scaled up, making it easier to handle increased workloads and growth.
5. **Better compliance:** Automation can help companies meet regulatory requirements and maintain compliance with industry standards.
6. **Improved customer experience:** Automation can lead to faster and more efficient service, resulting in improved customer satisfaction.

Overall, investing in business process automation can lead to significant improvements in productivity, profitability, and customer satisfaction [3]. There are many solutions for BPA, including open-source and proprietary platforms. These solutions will be briefly described in Sect. 3. But for now, it is important to mention the main problems and limitations of these solutions including:

1. *Vendor Lock-In:* One of the main risks in adopting a proprietary platform is vendor lock-in where a company becomes dependent on a particular vendor for products or services and switching to a different vendor may be difficult or too expensive. Vendor lock-in may lead to issues such as limited flexibility to respond to changing market conditions, higher costs, reduced competition, less innovation, lack of control, and risk of vendor failure.
2. *Limited UI/UX:* Despite claiming that the systems provide features for building dynamic forms, the forms produced by the BPA platforms are still quite limited. In this way, the construction of complex components (such as auto-complete input fields or elements whose validation rule is complex) becomes challenging and often not even feasible on some low-code platforms.
3. *Limited Integration:* Many solutions for BPA provide extension mechanisms designed to integrate the platform with existing systems or third-party solutions. However, implementing such integrations is still costly, not efficient, and sometimes not feasible.
4. *Outdated Technology Stack:* Many of the existing BPA solutions are based on outdated techniques such as stateful [5] and server-side rendering [7] applications.
5. *Lack of Support for Non-Process Features:* Another limitation of current BPA platforms is that they focus only on process automation and often provide limited support for other features that are usually necessary for process execution. For example, customized authentication mechanisms, CRUD<sup>1</sup> for domain entities, and import/export data are typically features hard to develop in BPA solutions. There is a need for solutions that can support the automation of both process and web-related features.

One approach to mitigate these problems and limitations is to develop a traditional web application and integrate it with a process engine. The resulting software does not

---

<sup>1</sup> Acronym for *Create, Read, Update, and Delete*.



have the limitations of a platform focused on BPA and has the benefit of being controlled by a process engine. The main disadvantage of this approach is that integrating a web application with a process engine is not straightforward and can be very challenging and costly [16, 18, 19].

In this context, we present a solution that fills this gap by providing an open-source platform to facilitate the development of *process-aware* information systems based on code generation techniques and a reference architecture. The AKIP platform can generate a functional *process-aware* web application from a business process model defined in BPMN. To the best of our knowledge, there is no other software tool that generates fully functional *process-aware* web applications.

The framework was originally derived by the AgileKIP Research Group<sup>2</sup> that focuses on Knowledge-Intensive Agile Processes and Model-Driven Engineering techniques for Information Systems [1, 10, 11, 15, 21]. In 2018, we released the first version of the platform, and in 2019 we conducted the first POC (Proof of Concept) using industrial applications. In 2020, we launched the second version of the platform that has been used to automate dozens of processes. The third version has just been released and among the new features are support for dashboards and modularization. The AKIP platform has been used in academia and industry for the development of several non-profit and commercial process-aware applications.

In [22], we presented an overview of the platform and demonstrated its use through an illustrative example. Previously, we also presented some general data on a study that examined the automation of three business processes (specified in BPMN) using the proposed platform. This paper extends our previous work by addressing the following issues:

1. Illustrating an extended version of the previous example demonstrating new resources provided by the platform, such as decision gateways and complex user tasks as well as a domain information model that involves several related entities.
2. Introduction and application of four metrics that provide indicators about the size and complexity of BPMN business process models.
3. A case study from industry that illustrates the automation of three new business processes in a process-oriented web application.
4. Research questions addressed through the case study to investigate the size and complexity of the automated business process models.
5. Identification of concerns about the validity of the case study.

The remainder of this paper is organized as follows: Section 2 presents the background of this research. The related work is briefly described in Sect. 3. The platform overview is presented in Sect. 4. Sections 5 and 6 discuss, respectively, how to use the platform and how to use a generated process-aware application. The industry case study is presented in Sect. 7. Finally, Sect. 8 concludes the paper and presents future work.

<sup>2</sup> <https://agilekip.github.io/site/>.

## 2 Background

This section briefly introduces the main concepts that support our solution for business process automation (Sect. 2.1). Further this section describes the metrics for BPMN process models that have been used in this paper to indicate the size and complexity of process models automated in the case study presented in this paper (Sect. 2.2).

### 2.1 Business Process Automation

**Business Process Model and Notation (BPMN)** is an ISO [8] and OMG [14] standard and the leading technology for modeling business processes. Currently, BPMN is the business process notation most used in practice [6] and is supported by a large number of tools. BPMN models can be interpreted and manipulated by both technical and non-technical personnel, reducing the likelihood of erroneous knowledge transfer [14]. Moreover, BPMN can also express executable models that can be interpreted by process engines. In fact, systems such as Camunda, Flowable, and BonitaSoft deliver an integrated environment where users can design and execute BPMN models.

**Business Process Automation (BPA)** is defined as the automation of business processes and functions beyond conventional data manipulation and record-keeping activities, usually using advanced technologies [9]. BPA is based on three pillars: orchestration, integration, and dynamic automated process execution [13]. Based on these pillars, BPA can be enabled by developing a systematic solution supporting a given business process.

In this context, a **Process-Aware Information System (PAIS)** is a particular type of application that uses information technology to manage and execute operational processes involving people, applications, and information sources [3]. PAIS requirements are described using process models such as BPMN, where activities, resources, decisions, events, and their relationships can be used to represent the flow of work.

**Modern PAISs** have been built as process-aware web applications based on features that allow authorized users to manage (e.g., execute, view, query, delegate) their tasks and interact with the business processes seamlessly. It is almost imperative that these applications provide rich user interfaces (light, fast, and user-friendly) on top of a reliable and scalable software architecture based on cutting-edge technologies, focused on cloud-native distribution, and easy integration with third-party applications.

### 2.2 Metrics for BPMN Process Models

Some metrics for business process models have shown their practical importance in evaluating and identifying process models that are less error-prone, and easy to understand, maintain, and manage. In this paper, we are especially interested in metrics related to complexity to evaluate this aspect of the business process models automated in our case study. In this sense, we selected some complexity metrics frequently cited in the literature and empirically validated to provide data about the models [20].

In addition, we also intend to quantify the `size` of the process models used in the case study. For this, we selected two popular business process size metrics: the **Number of Activities (NOA)** [2] and the **Number of Nodes (NON)** [12]. In a BPMN process

model, we compute NOA as the total number of *Tasks*<sup>3</sup> and *Call Activities*<sup>4</sup> whereas NON is computed as the total number of Activities (NOA), *Gateways*, and *Events* in the process model.

The complexity of a process model quantifies how simple and easy it is to understand [12]. Complexity can be measured simply in terms of size, e.g. using the NOA or NON metric. Usually, larger models are more complex and hence difficult to understand (as well as to automate) [4]. However, two models can have the same size and yet one of them might be more difficult to understand because it contains too many gateways. Therefore, it is also common to measure the complexity of a BPMN process model using the metric called **Coefficient of Network Connectivity (CNC)** [12], which is the number of flows in a process model divided by the number of nodes. The higher the CNC is, the higher is the number of gateways relative to nodes, which makes the model more difficult to understand. However, this metric does consider that process gateways have different semantics. Thus, another complexity metric, named **Control-Flow Complexity (CFC)** [2], is often used to measure the complexity introduced by different types of split gateways. CFC is an additive metric, where each split gateway in the model adds a certain complexity value based on the number of states that follow it. In BPMN process models, we sum the corresponding complexity values of each following split gateway:

- *Exclusive or Event-based (XOR) Split*: its complexity value is  $n$ , which corresponds to the number of outgoing flows from the gateway;
- *Inclusive (OR) Split*: its complexity value is  $2^n - 1$ ;
- *Parallel (AND) Split*: its complexity is simply 1.

According to the CFC metric, *Inclusive Splits* add the highest complexity to a process model, while *Parallel Splits* have the lowest complexity. The higher the CFC value, the more complex is the process design since the developer must handle all the states between splits and their associated outgoing flows and activities [2].

To guide our analysis of the selected metrics for business process models, in this paper we will adopt the indicators of complexity (understandability) provided in [20] (see Table 1). We selected indicators from this research because it is the most recent that we have identified and they were obtained from empirical studies.

### 3 Related Work

Considerable research has been conducted in the domain of Business Process Management (BPM) and many tools have been proposed for different aspects of Business Process Automation (BPA). Because of the wide range of solutions, it is a challenge to compare our platform with existing tools. However, we briefly describe a set of platforms for BPA that can be either a source of evaluation and/or inspiration for our framework. We divided those solutions into two groups: proprietary and open-source platforms.

<sup>3</sup> Including *Subprocess*' internal tasks.

<sup>4</sup> Their internal elements are not computed because *Call Activities* denote reusable (sub)processes that are specified in independent models.

**Table 1.** Complexity indicators for business process model metrics (based in [20]).

<i>Business Process Model Metrics</i>	<i>Analysis Model</i>
Number of Nodes (NON)	If NON > 58, then the model is difficult to understand If 44 = < NON <= 58, then the model is moderately understandable If NON < 44, then the model is easy to understand
Number of Activities (NOA)	If NOA > 31, then the model is difficult to understand If 22 = < NOA <= 31, then the model is moderately understandable If NOA < 22, then the model is easy to understand
Control Flow Complexity (CFC)	If CFC > 21, then the model is difficult to understand If 10 < CFC <= 21, then the model is moderately understandable If CFC <= 10, then the model is easy to understand
Coefficient of Network Connectivity (CNC)	If CNC > 1.43, then the model is difficult to understand If 0.90 < CNC <= 1.43, then the model is moderately understandable If CNC <= 0.90, then the model is easy to understand

Proprietary low-code solutions for BPA include *Bizagi*<sup>5</sup>, *Aris*<sup>6</sup> [17], *Signavio*<sup>7</sup>, *Pipefy*<sup>8</sup>, *Heflo*<sup>9</sup>, *Nintex*<sup>10</sup>, *Process Street*<sup>11</sup>, and *SoftExpert BPM*<sup>12</sup>. These solutions support BPMN or provide modeling tools based on BPMN that can be integrated with a low-code platform where users can create processes and business rules, add functional roles, create interfaces, customize forms, and manage related content in an integrated approach.

Open-source platforms include *Camunda*<sup>13</sup>, *Flowable*<sup>14</sup>, *Bonita*<sup>15</sup>, and *JBPM*<sup>16</sup>. The *Camunda* Platform is an open-source workflow and decision automation platform. The solution is composed of tools that include a BPMN 2.0 process engine, a modeler, a cockpit, and a task-list manager. These tools can be used for creating workflow and decision models, operating deployed models in production, and allowing users to execute workflow tasks assigned to them. *Flowable* is a lightweight business process engine written in Java. The platform supports the deployment of BPMN 2.0 and can be used for creating process instances of those process definitions, running queries, accessing active or historical process instances and related data. *Bonita* is an open-source business process management and low-code development platform for BPA. The platform is composed of five main components: *Bonita Studio*, *Bonita BPM Engine*, *Bonita Portal*, *Bonita UI Designer*, and *Bonita Continuous Delivery*. *JBPM* (Java Business Process Model) is an open-source workflow engine written in Java that can execute business processes specified in BPMN 2.0 (or in jPDL, its own process definition language, in earlier

<sup>5</sup> <https://www.bizagi.com/>.

<sup>6</sup> <https://www.softwareag.com/>.

<sup>7</sup> <https://www.signavio.com/>.

<sup>8</sup> <https://www.pipify.com/>.

<sup>9</sup> <https://www.heflo.com/>.

<sup>10</sup> <https://www.nintex.com/>.

<sup>11</sup> <https://www.process.st/>.

<sup>12</sup> <https://www.softexpert.com/>.

<sup>13</sup> <https://www.camunda.org/>.

<sup>14</sup> <https://www.flowable.com/>.

<sup>15</sup> <https://www.bonitasoft.com/>.

<sup>16</sup> <https://www.jbpm.org/>.

versions). jBPM is a toolkit for building business applications to assist in automating business processes and decisions. It is maintained by Red Hat Inc., part of the JBoss community and closely related to the systems Drools and OptaPlanner projects in the KIE group. It is released under the ASL (or LGPL in earlier versions) by the JBoss company.

Limitations of these existing solutions were described in Sect. 1, and, as a reminder, are listed here as well: (1) *Vendor Lock-in*; (2) *Limited UI/UX*; (3) *Limited integration*; (4) *Outdated technology stack*; and (5) *Lack of support for non-process features*.

## 4 Platform Overview

The AKIP Process Automation Platform is an open-source project based on code generation techniques that is designed to facilitate process automation initiatives. The platform was built by developers and researchers aiming to disseminate the use and development of process-aware information systems, more specifically modern web applications that we call *KIPApps*. Figure 1 presents an overview of the platform. It consists of two components supporting KIPApp development: an *Application Generator* and a *Reference Architecture*.



Fig. 1. Platform overview.

### 4.1 KIPApp

In our solution, we call KIPApp (which stands for *Knowledge Intensive Process Applications*) a modern process-aware web application (such as described in Sect. 2.1) that executes on the top of an open-source process engine. Briefly, a KIPApp provides the following main features:

- **Web Application Management:** It includes the management of users, configuration properties, health checks, logs, and application metrics.
- **Process Management:** It includes the management of domain entities, deployed processes, process instances, and tenants.
- **Task List:** It provides an updated user tasks list showing tasks completed, assigned, and waiting to execute.

## 4.2 Application Generator

The application generator is a tool to generate, develop, and deploy KIPApps quickly. It generates code for the base reference architecture of the AKIP platform. The generator is a key tool in our solution because it accelerates the development process by scaffolding a huge amount of code that, without the generator, would have to be coded manually.

## 4.3 Reference Architecture

The reference architecture represents the backbone of a KIPApp. It is composed of front and back-end frameworks, a process engine, native features, extensions, and connectors. Details on technologies used for front and back-end frameworks and the process engine are presented in the following subsection. The process engine is the tool used in our solution to orchestrate a process workflow from an executable BPMN model. Next, we depict the remaining three elements of the reference architecture:

- **Native Features:** are common features already provided by the reference architecture, meaning there is no need to repeat the code of these features in each KIPApp. Examples of native features include *Advanced Tasks List*, *Start-Process*, *Management of Deployed Processes*, *Process-Instances Dashboard*, and *User Management*.
- **Extensions:** are components that allow seamless integration of the KIPApp with the process engine.
- **Connectors:** are components that allow the application to integrate quickly with the external world. Examples of connectors include an *Email Connector* used to send email automatically during the process execution, *RestAPI Connector* to integrate with other systems through Rest APIs, *JMS Connector* to carry out communication through messaging, and *AWS Connectors* that allow integration with main AWS services. The main idea behind these elements is to reuse the most common integration components through a sophisticated and highly configurable set of connectors.

## 4.4 Technical Notes

This section presents information about the technologies that support the platform's implementation. A KIPApp is a single web page application generated by the AKIP platform, which runs on a process engine. The AKIP Application Generator is an extension of the *JHipster*<sup>17</sup> generator (technically, it is a JHipster blueprint) that overrides many sub-generators and provides its own templates and functions. Our Reference Architecture is based on one of the most modern tool stacks in the software industry. Currently, the platform uses *VueJS*<sup>18</sup> as a front-end framework, *Spring Boot*<sup>19</sup> as a back-end framework, and the *Camunda Platform* as a process engine.

<sup>17</sup> <https://www.jhipster.tech/>.

<sup>18</sup> <https://www.vuejs.org/>.

<sup>19</sup> <https://spring.io/projects/spring-boot>.

## 5 Platform Walk-Through

The three main steps shown in Fig. 2 to use the platform are:

1. Generation of a KIPApp based on the Reference Architecture.
2. Technical implementation of one or more business processes.
3. Deployment of the business process(es) in the KIPApp.

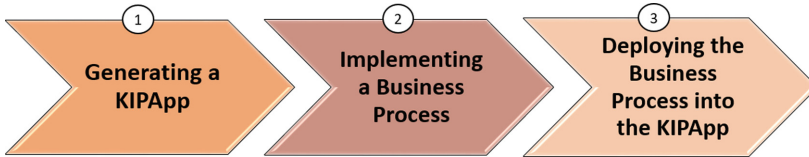


Fig. 2. Main steps in using the platform [22].

We briefly explain next how these general steps are performed. Details on how to install and use the platform can be found in the Github project public repository<sup>20</sup>.

### 5.1 Generating a KIPApp

The application generator was designed so that the user only needs to execute a single command and answer some questions. These questions are provided by a wizard related to the application configuration to get your web application running with several features already installed. For more details, the reader can access the platform's public documentation.

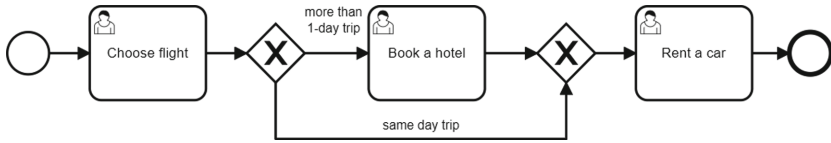
### 5.2 Implementing a Business Process

In this section, we focus on the development of a *Travel Plan Process*, a simple process which is based on an example in [22]. In a nutshell, this process has steps that users execute when planning to travel. Therefore, our goal is to develop a process-aware web application that aids a user in planning a trip. For illustrative purposes, we will use a very simple version of this process, composed of three user tasks, as shown in Fig. 3. The process starts with the user choosing a flight. The hotel booking task is skipped for same-day travel and executed for trips that last longer than one day. This decision is determined by evaluating the possible flows from the first gateway of the process (a split Exclusive Gateway). Finally, the user must rent a car (last process task).

To implement this or any other business process, we should perform the following steps:

1. Define the business process model.

<sup>20</sup> <https://agilekip.github.io/pap-documentation>.



**Fig. 3.** *Travel Plan Process: A running example.*

2. Define domain entities.
3. Generate domain entities.
4. Define process entities.
5. Generate process entities.
6. Customize the generated code.
7. Refine the business process model for automation.

**Defining the Business Process Model.** The business process model should be specified in BPMN using a modeling tool supported by the platform<sup>21</sup>. The model should also contain only typical tasks (usually User, Service, or Message Tasks) as illustrated by the process model in Fig. 3. Moreover, the process should have an associated identifier (id) and be defined as *executable*. Optionally, we can also specify a description for the process using markdown notation. The application will show this description when a new process instance is initiated.

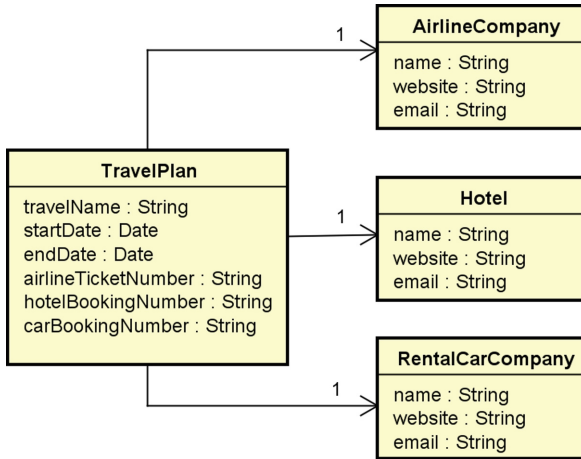
**Defining Domain Entities.** In Fig. 4, we present a domain model as a UML Class Diagram illustrating domain entities related to our running example. To keep our example simple, we will use only four domain entities, named *TravelPlan*, *AirlineCompany*, *Hotel*, and *RentalCarCompany*, that have a few properties as shown in the figure. These entities represent the data that are handled by the process. However, such a model is not an input artifact for the AKIP platform.

Our platform uses JSON (JavaScript Object Notation) to create domain entities. JSON is a standard data interchange format for the web. JSON files are lightweight, text-based, human-readable, and can be edited using a text editor. Figures 5 and 6 present the metadata of the domain entities *AirlineCompany* and *TravelPlan*, respectively, in JSON format. The `fields` element describes primitive elements of the entity and the `relationships` element should describe relationships (when they exist) with other domain entities (as is the case of the *TravelPlan* entity). JSON specifications for the domain entities *Hotel* and *RentalCarCompany* will not be presented here because their structures are the same as the *AirlineCompany*.

**Generating Domain Entities.** Once we have specified the JSON files for the domain entities, we must use the platform to generate the files that will support the manipulation of these entities in the web application (KIPApp). The “entity” sub-generator of the AKIP platform will create the necessary application files, including back-end files

<sup>21</sup> For now, only Camunda Modeler is supported.





**Fig. 4.** Simplified domain model.

```

{
  "fields": [
    { "fieldName": "name",      "fieldType": "String" },
    { "fieldName": "website",  "fieldType": "String" },
    { "fieldName": "email",    "fieldType": "String" }
  ],
  "relationships": [],
  "entityType": "domain",
  "service": "serviceClass",
  "dto": "mapstruct",
  "jpaMetamodelFiltering": false,
  "readOnly": false,
  "pagination": "no",
  "name": "AirlineCompany",
  "skipFakeData": true
}

```

**Fig. 5.** *AirlineCompany* domain entity metadata.

(database tables, the controller of basic CRUD operations, and services) as well as files from the front-end layer.

**Defining Process Entities.** In this step, we need to define KIPApp’s entities related to the process support. Each process entity type plays a different role in the application:

- **Process-Binding Entity.** This entity represents the binding between a business process (BPMN model) and its corresponding domain entity. In other words, the aim of this entity is to indicate the domain entity associated with the process.
- **User-Task Entity.** This entity is used to generate the User Interface (UI) form for a specific user task of the process.

```

{ "fields": [
  { "fieldName": "travelName", "fieldType": "String" },
  { "fieldName": "startDate", "fieldType": "LocalDate" },
  { "fieldName": "endDate", "fieldType": "LocalDate" },
  { "fieldName": "airlineTicketNumber", "fieldType": "String" },
  { "fieldName": "hotelBookingNumber", "fieldType": "String" },
  { "fieldName": "carBookingNumber", "fieldType": "String" }
],
"relationships": [
  { "relationshipName": "airlineCompany",
    "otherEntityName": "airlineCompany",
    "relationshipType": "many-to-one",
    "otherEntityField": "name"
  },
  { "relationshipName": "hotel",
    "otherEntityName": "hotel",
    "relationshipType": "many-to-one",
    "otherEntityField": "name"
  },
  { "relationshipName": "rentalCarCompany",
    "otherEntityName": "rentalCarCompany",
    "relationshipType": "many-to-one",
    "otherEntityField": "name"
  }
],
"entityType": "domain",
"service": "serviceClass",
"dto": "mapstruct",
"jpaMetamodelFiltering": false,
"readOnly": true,
"pagination": "no",
"name": "TravelPlan",
"skipFakeData": true
}

```

**Fig. 6.** *TravelPlan* domain entity metadata.

### Airline Companies

Refresh list
+ Create a new Airline Company

ID	Name	Website	Email	
1001	Air Canada	aircanada.ca	contact.us@aircanada.ca	<span style="background-color: #007bff; color: white; padding: 2px 5px; border-radius: 4px;">View</span> <span style="background-color: #007bff; color: white; padding: 2px 5px; border-radius: 4px; margin-left: 5px;">Edit</span> <span style="background-color: #dc3545; color: white; padding: 2px 5px; border-radius: 4px; margin-left: 5px;">X Delete</span>
1701	Frontier	flyfrontier.com	contact@flyfrontier.com	<span style="background-color: #007bff; color: white; padding: 2px 5px; border-radius: 4px;">View</span> <span style="background-color: #007bff; color: white; padding: 2px 5px; border-radius: 4px; margin-left: 5px;">Edit</span> <span style="background-color: #dc3545; color: white; padding: 2px 5px; border-radius: 4px; margin-left: 5px;">X Delete</span>

**Fig. 7.** *AirlineCompany* CRUD screen generated by the AKIP platform.

- **Start-Form Entity.** This entity is used to generate the UI form used to start the process.

```

"entityType": "process-binding",
"processBpmnId": "TravelPlanProcess",
"domainEntityName": "TravelPlan",
"name": "TravelPlanProcess",

```

**Fig. 8.** Fragment of the *TravelPlanProcess* process-binding entity metadata [22].

As in the case of domain entities, process entities also need to be specified as JSON files. In our example, the **process-binding entity** is named *TravelPlanProcess*. Its JSON specification is very similar to the *TravelPlan* domain entity presented in Fig. 6. For this reason, we will not show its complete specification, but only some of its details. *TravelPlanProcess* differs from the *TravelPlan* entity only in relation to the properties presented in Fig. 8. Specifically, we highlight the `processBpmnId` and `domainEntityName` properties. The first one should match the process Id defined in the BPMN process model (as mentioned in Sect. 5.2). The second one (`domainEntityName`) should match the main domain entity previously created (i.e., *TravelPlan*).

A **user-task entity** should be created for each user task present in the business process model. In the case of our example, we need to create three user-task entities (e.g., named *TaskFlight*, *TaskHotel*, and *TaskCar*), since the Travel Plan process contains three user tasks (see Fig. 3). Owing to space limitations, we will not present the JSON specification for all these entities, but only for the first one.

The specification of the *TaskFlight* user-task entity is presented in Fig. 9. The `fields` element of the JSON specification should contain properties from the *TravelPlan* domain entity that should appear in the KIPApp’s UI form used to execute this process’s user task. On the other hand, the `relationships` element describes relationships from the domain entity that should be included in this task form. In the case of the *TaskFlight* entity, it refers to the *AirlineCompany* domain entity. The `entityType` property should be `user-task-form`; `processBpmnId` matches the process id defined in the BPMN model; `processEntityName` matches the process-binding entity previously created; and `domainEntityName` matches the main domain entity previously created. In addition, it should contain the `taskBpmnId` property matching the corresponding task id defined in the BPMN model file.

Concerning the **start-form entity**, it is very similar to the task-user entity specification, as both entity types (user-task and start-form) are used to generate UI forms of the resulting KIPApp. However, its `entityType` property is `start-form` and it does not have the `taskBpmnId` property, since the start-form of a process is unique. Owing to space limitations, we will not present the JSON specification for the start-form entity and its corresponding UI form, but they can be found in the platform’s online tutorial and in [22].

**Generating Process Entities.** Now, we can generate the KIPApp’s code corresponding to the process entities defined previously. The “entity” sub-generator of the AKIP platform will create all the necessary application files to support such entities. Figure 10 shows the KIPApp’s UI form generated from the *TaskFlight* user-task entity (Fig. 9).

```

{
  "fields": [
    { "fieldName": "travelName", "fieldType": "String",
      "fieldReadOnly": true },
    { "fieldName": "startDate", "fieldType": "LocalDate",
      "fieldReadOnly": true },
    { "fieldName": "endDate", "fieldType": "LocalDate",
      "fieldReadOnly": true },
    { "fieldName": "airlineTicketNumber", "fieldType": "String" }
  ],
  "relationships": [
    { "relationshipName": "airlineCompany",
      "otherEntityName": "airlineCompany",
      "relationshipType": "many-to-one",
      "otherEntityField": "name"
    }
  ],
  "entityType": "user-task-form",
  "processBpmnId": "TravelPlanProcess",
  "processEntityName": "TravelPlanProcess",
  "taskBpmnId": "TaskFlight",
  "domainEntityName": "TravelPlan",
  "service": "serviceClass",
  "dto": "mapstruct",
  "jpaMetamodelFiltering": false,
  "readOnly": false,
  "pagination": "no",
  "name": "TaskFlight",
  "skipFakeData": true
}

```

**Fig. 9.** *TaskFlight* user-task entity metadata.

This form will be used when the user executes the first task of the process (*TaskFlight*) to input the airline ticket information into the system. The short description that appears on the top of the form (below the title) came from the *documentation* element of the BPMN process model. The form of this task shows the following fields: *Travel Name*, *Start Date*, *End Date* (read-only), *Airline Ticket Number*, and *Airline Company*. Note that the *Airline Company* field from the task form is a select input showing all the airline companies previously created from the *AirlineCompany* CRUD form (Fig. 7).

**Customizing the Generated Code.** In a real scenario, once the application code has been generated, it usually still needs to be customized by developers. The most common customizations that have been identified in practice are the addition of support for business rules and adjustments in UI form fields.

**Refining the Business Process Model for Automation.** Sometimes it is necessary to refine the BPMN model developed in the *Defining the Business Process Model* step to include information about entities that were defined after the business process specifi-

**Buying the Flight Tickets**

Airfare can easily be the largest expense of your trip. Expensive plane tickets mean you need to choose a more affordable destination or spend less money at your vacation stop to stay within your spending limit.

Travel Name  
Travel to Brazil

Start Date  
2021-12-01

End Date  
2021-12-08

Airline Ticket Number  
12345

Airline Company  
Air Canada  
Frontier

Assigned to  
admin

Process Definition  
Travel Plan Process

Process Instance  
TravelPlan#1504

Task Def Key  
TaskFlight

Task Id  
3146

Execution Id  
3133

← Back   Complete

**Fig. 10.** KIPApp’s UI form used to execute the *Choose flight* process task.

Community Edition  
**AgileKIP**  
Process Automation

**TravelPlan** vUNKNOWN

My Tasks

Process Definitions

Refresh List   Deploy a Process

ID	Name	Bpmn Process Definition Id	
1001	Travel Plan Process	TravelPlanProcess	Init   Deployments   Instances   View

**Fig. 11.** KIPApp screen showing a deployed business process (adapted from [22]).

cation. As an example, decision gateways are usually configured using expressions that rely on the domain model defined in the *Defining Domain Entities* step. Considering our *Travel Plan* illustrative process (Fig. 3), to specify when the process flow “*more than 1-day trip*” should be executed, we could associate it with the following formal expression:

$$\${processInstance.travelPlan.endDate} > \${processInstance.travelPlan.startDate}$$

The `processInstance` variable is automatically set by the Reference Architecture of the AKIP platform and represents an instance of the process-binding entity (in our example, *TravelPlanProcess*) created previously. As we have already mentioned, the process-binding entity has a reference to the domain entity (*TravelPlan*). Finally, the *TravelPlan* entity has two date fields (`startDate` and `endDate`) that are used in the expression associated with the process flow. The expression evaluates to true whenever the end date is after the start date. A similar configuration expression should also be associated with the process flow “*same day trip*”.

### 5.3 Deploying a Business Process into a KIPApp

Finally, the KIPApp is ready and all entities supporting the execution of the *Travel Plan* business process are generated and customized. Next, we need to deploy this process into the KIPApp.

To deploy the process, it is necessary to log into the application using an account with *admin* privileges, access the *Process Definitions Management* feature (Fig. 11), click on the *Deploy a Process* button, and choose the business process model file (BPMN) corresponding to the *Travel Plan* Process (detailed in Sect. 5.2). The bottom of Fig. 11 shows the *Travel Plan* process already deployed.

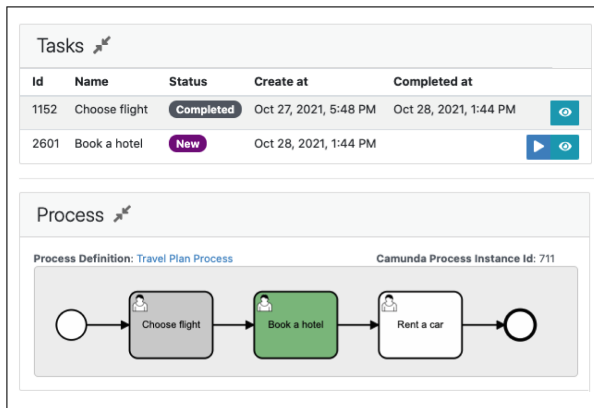


Fig. 12. User tasks from a process instance (adapted from [22]).

## 6 Using a KIPApp

This section briefly presents how to use a KIPApp to execute a business process. Owing to space limitations, we will not present other functions provided by the application. Enacting a business process involves at least (1) starting a new process instance, and (2) executing user tasks from this instance.

Once the *Travel Plan* process has been deployed, we can start the process by clicking on the *Init* button from the application screen shown in Fig. 11. The start form of this process (shown in [22] and in the platform's online tutorial) contains the fields defined in the corresponding start-form process entity. To see the instances of the process already created, we can click on the *Instances* button shown in Fig. 11.

A KIPApp user can see an updated process task list from the *My Tasks* application menu. In the case of the *Travel Plan* process, its tasks are available for any authenticated user, since they were not designed for specific process roles (that would be represented by lanes in the BPMN process model). Figure 12 shows the user tasks screen where one can note the *Choose flight* task has already been completed (owing to its status) and the *Book a hotel* task is enabled for execution (the BPMN process model in this screen shows in green the current task waiting to be executed). To execute it, the user only needs to click on the *Play* button on the right side of the task. A process instance will be concluded after all its tasks have been completed.

## 7 Case Study

The AKIP platform has been used in practice within a software development company as a support tool to build process-aware web applications (KIPApps). This section presents a case study conducted in the context of this enterprise. The following subsections present the goal of our study (Sect. 7.1), data about the company and automated business processes (Sect. 7.2), the procedure that was followed (Sect. 7.3), the results of the case study (Sect. 7.4), some considerations about the case study (Sect. 7.5), and, finally, threats to validity of this study (Sect. 7.6).

### 7.1 Goal

Our goal with this study was to assess the feasibility of using the AKIP platform. In other words, we intended to verify its effectiveness in supporting the automation of business processes into a process-aware web application in the context of a software development company. Specifically, the study aimed to answer the following research question (RQ):

**RQ 1.** Is the AKIP platform capable of supporting the automation of real-world business processes into process-aware web applications (KIPApps)?

To understand better the characteristics of business processes automated with the AKIP platform in this case study, we also considered the following research subquestions:

**RQ 1.1.** What is the variety of BPMN process elements covered by business processes automated in this case study?

**RQ 1.2.** What is the size and complexity of business process models automated in this case study?

### 7.2 Subject Data

This case study was conducted with a Brazilian software development company (with approximately 30 employees), located in the city of Rio de Janeiro. Its focus is on the development of IT solutions for the logistics and port sectors.

The case study was conducted using the *Process Automation* project, which is automating and integrating business processes from a client company by providing a customized process-aware web application. The project follows a Scrum-based agile development process that delivers an automated business process at each iteration (sprint). The project has a team composed of a project manager, a product owner, a process analyst, a software analyst, and two developers. When the process automation project started three years ago, all the analysts and developers assigned to the project had less than six months of experience in their roles. None of them knew or had ever used the AKIP platform. We evaluated six automated business processes in this project, which are from the logistics and seaport domains and from diverse business units within the client organization. However, as these processes are proprietary and confidential, we will only present general information about the models.

### 7.3 Procedure

First, the software architect generated the KIPApp for the client company using the AKIP platform. At each project's development process iteration (sprint), the team followed the sequence of steps presented in Sect. 5.2 for implementing a business process:

1. The BPMN business process model is built by the process analyst, interacting with the product owner and users of the process.
2. The domain model is created by the software analyst with the participation of the process analyst.
3. The domain and process entities metadata are specified by the developers in JSON and used by the AKIP platform to generate the code base for the given process.
4. The generated app's code is customized by developers (including support for business rules and integrations) and UI forms (screens) are customized by the software analyst.
5. The BPMN model is refined with expressions based on the domain model.
6. Finally, the business process is deployed in the KIPApp and the application is made available for client validation.

### 7.4 Results

Next, we answer the research questions (RQs) of this case study, starting with the more specific ones (RQs 1.1 and 1.2) and ending with the main research question (RQ 1).

#### **RQ 1.1: What Is the Diversity of BPMN Process Elements Covered by Business Processes Automated in This Case Study?**

Table 2 shows data on the six process models automated with the AKIP platform in this case study (identified by P1, P2, P3, P4, P5, and P6) in terms of used BPMN elements. As shown in Table 2, business processes considered in this case study are composed of lanes (representing process participants), call activities (independent subprocesses), gateways, events, and different types of BPMN tasks (User, Message, and Service<sup>22</sup>

<sup>22</sup> BPMN's Script tasks behave exactly like Service tasks in an automated BPMN model with the AKIP platform. The differences between these task types are the visual representation and the semantics for the model.



**Table 2.** BPMN elements used in the case study business process models. \*Call activities are subprocesses defined in independent models and their internal elements are not computed as part of the processes in which they are present.

Process Model Data	P1	P2	P3	P4	P5	P6
<b>Lanes</b>	<b>6</b>	<b>4</b>	<b>2</b>	<b>4</b>	<b>3</b>	<b>3</b>
<b>Call Activities*</b>	<b>0</b>	<b>1</b>	<b>0</b>	<b>1</b>	<b>0</b>	<b>0</b>
<b>Tasks</b>	<b>34</b>	<b>29</b>	<b>15</b>	<b>20</b>	<b>8</b>	<b>53</b>
User tasks	6	6	2	4	3	4
Message tasks	15	15	7	9	5	18
Service tasks	13	8	6	7	0	31
<b>Gateways (Splits and Joins)</b>	<b>11</b>	<b>10</b>	<b>5</b>	<b>10</b>	<b>1</b>	<b>17</b>
Parallel (AND) gateways	0	0	0	0	0	0
Exclusive (XOR) gateways	11	10	5	10	0	17
Inclusive (OR) gateways	0	0	0	0	1	0
<b>Events</b>	<b>7</b>	<b>7</b>	<b>4</b>	<b>6</b>	<b>3</b>	<b>11</b>
Start events	1	1	1	1	1	1
End events	3	5	2	5	2	3
Conditional events	3	1	1	0	0	4
Time events	0	0	0	0	0	3
Link events	0	0	0	0	0	7

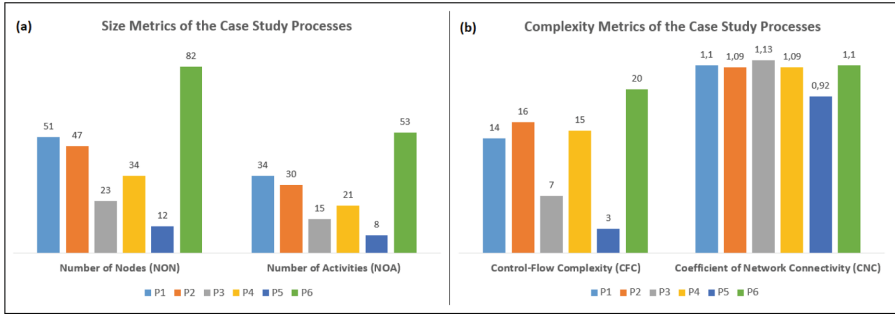
tasks) that allow testing of code-generation resources related to the implementation of diverse BPMN process elements. In this paper, we did not present the implementation of some of these resources, such as the support for Message and Service tasks, owing to space limitations, but they are covered in the platform’s online tutorial.

However, as one can observe in Table 2, process models considered in the case study include only two types of BPMN gateways (Exclusive and Inclusive) and three types of non-essential events (Conditional, Time, and Link events). Thus, the effectiveness of the AKIP platform in automating processes with other types of BPMN gateways (Parallel, Event-based, and Complex gateways) and events (Error, Escalation, Cancel, Compensation, and Signal) was not evaluated in this study. Moreover, no automated process includes embedded subprocesses. All subprocesses are Call Activities, i.e., they are defined in independent process models. Other BPMN process elements have not yet been tested using the platform, such as transactions, compensation activities, and business rule tasks. In future work, we intend to address this limitation.

**RQ 1.2: What Is the Size and Complexity of Business Process Models Automated in This Case Study?**

To guide our analysis concerning the size and complexity of the automated models, we adopt the metrics presented in Sect. 2.2 and their corresponding indicators (summarized in Table 1). As previously mentioned, size metrics can also be used to measure complexity in a simplified way.

Figure 13 shows the values computed for size metrics, Number of Nodes (NON) and Number of Activities (NOA), and exclusive metrics of complexity, Control-Flow Complexity (CFC), and Coefficient of Network Connectivity (CNC), from the business process models of the case study. As can be observed in Graph (a) from the figure, the size of automated process models ranges from 12 to 82 in terms of the Number of



**Fig. 13.** Size (a) and complexity (b) metrics of the case study business process models.

Nodes (NON) and from 8 to 53 in terms of the Number of Activities (NOA). When these size metrics are considered for evaluating the process model's complexity, their results are very similar considering the metric's analysis model presented in Table 1. Such metrics present divergence only regarding the process P1, which according to the NON analysis would be moderately understandable and according to the NOA analysis would be difficult to understand. The analysis of the other processes is as follows: P2 would be moderately understandable; P3, P4, and P5 would be easy to understand; and P6 would be difficult to understand.

Concerning the analysis of the CFC and CNC complexity values, presented in Fig. 13 (b), their results diverge with respect to processes P3 and P5, considering the metric's analysis model presented in Table 1. While the CFC analysis indicates that these processes are easy to understand, the CNC analysis indicates that they are moderately understandable. For all other processes (P1, P2, P4, and P6), the analysis of both metrics indicates that they are moderately understandable. Finally, we conclude this section by answering below the main research question that motivated this case study.

**RQ 1: Is the AKIP Platform Capable of Supporting the Automation of Real-World Business Processes into Process-Aware Web Applications (KIPApps)?**

All business processes considered in this study could be successfully automated into a generated KIPApp. Code customizations were necessary, but even so, the platform contributes to the automatic generation of a large volume of application code, which would have to be developed manually without the platform. The project team that used the platform reported that the solution is very useful in streamlining the work of application development.

The evaluation that was conducted allowed us to conclude that the AKIP platform was effective in generating a web application supporting the business processes implemented in the context of the evaluated company's project. Based on the cases we evaluated in this practical study, we can claim that the AKIP platform satisfactorily generates useful and modern application code supporting business process execution. However, more case studies should be analyzed to extend the platform's evaluation and consolidate its validity.

## 7.5 Discussion

In this case study, we have applied different metrics for quantifying the size and identifying the complexity of the automated process models because there is no single or best-accepted method for measuring these attributes in process models. We used the most cited metrics from the literature that present some empirical validation, but each metric adopts its own method to evaluate the same process attribute. For this reason, as we saw previously, their results differ with respect to the same processes. However, if we consider the general spectrum of the results of all four metrics used, we can say that the evaluated processes range from small to large and from simple (easy to understand) to complex (difficult to understand). We also identified that the process models automated in this case study cover a diversity of BPMN elements, although some specific types of gateways and non-essential events were not used and, consequently, could not be evaluated.

Despite the limited experience of the development team observed in this case study (developers and analysts had less than six months of experience when they joined the project), they were able to automate business processes with different levels of complexity and varied range of BPMN process elements by using the support provided by the AKIP platform, which generates a process-aware web application based on a modern architecture and with several semi-ready application files (database files, CRUD screens, among others). However, the case study also revealed a significant effort to customize the generated code (especially the UI forms) to meet user experience/interface (UX/UI) requirements and third-party integrations.

In addition, we can state that the case study (as well as our experience in automating business processes with the platform) shows the AKIP Process Automation Platform can be used successfully in conjunction with an agile software development process such as Scrum to aim the development of *process-aware* information systems. The project team that participated in the case study reported that the platform helped them to achieve more agility in implementing business processes.

In summary, our results are promising, as demonstrated by several independent uses, and we are still working to improve the platform to support more BPMN elements and include features geared towards the low-code movement.

## 7.6 Threats to Validity

The case study presented in this paper is subject to the following main threats:

- *Limited Number of evaluation Scenarios.* It is difficult to obtain data to drive research in the domain of business process modeling because companies which adopt process modeling usually consider process artifacts extremely sensitive and confidential. We obtained access to people and artifacts from the Process Automation project of a software development company that has been using the AKIP platform in practice to automate business processes into process-aware web applications for three years, but to date, we could only obtain and evaluate six of the automated business process models.

- *Data Coming from a Single Project and Company.* The business process models and AKIP platform’s usage scenarios presented in this case study may not be representative of those occurring in other realistic settings. Different business process models, software development processes, domains and organizations may lead to slightly different results. Thus, our platform should be tested further on processes and usage scenarios from other organizations and domains.
- *Evaluated Business Process Models Do Not Include some BPMN Process Elements.* The business process models evaluated in this case study do not include important BPMN process elements such as Parallel and Event-based gateways and Timer, Error, and Signal events. Thus, code-generation resources of the AKIP platform related to these process elements could not be evaluated.

## 8 Conclusions and Future Work

This paper presented the AKIP Process Automation Platform, an open-source project designed to simplify business process automation through the use of code generation techniques and a reference architecture. The platform generates a functional process-aware web application called KIPApp automating business processes from their specifications in executable BPMN models. This paper presents a more comprehensive example to illustrate features of the platform compared to our previous work. The new resources showcased in this paper include decision gateways, complex user tasks, and a domain information model that incorporates multiple related entities.

Notably, to the best of our knowledge, no other software tool can generate fully functional process-aware web applications. Our solution stands out because the generated application runs on top of a process engine responsible for orchestrating the execution of business processes, and its architecture is based on modern technologies commonly used in the software industry.

The AKIP platform has been used in both realistic and academic scenarios to support the development of process-aware web applications. In this paper, we presented a case study from industry to demonstrate the usefulness of the platform. The study involved using the platform to automate six real-world business processes specified as executable BPMN models. Although some types of gateways and non-essential events were not used and could not be evaluated, we observed that these process models cover a diverse range of BPMN elements. Moreover, we identified that such process models have varied sizes (in terms of their numbers of nodes and activities) as well as varying levels of complexity (from easy to difficult to understand) when different complexity metrics (each with its own quantification method) are applied. Therefore, this case study demonstrated that the AKIP platform effectively generated a web application automating business processes specified in BPMN models of varying size and complexity. Moreover, the study showed that even a team of professionals with limited experience in system development and analysis could achieve these results using the AKIP platform.

Future work includes conducting more case studies in different business domains to broaden the platform’s validation, as well as developing a new component for the AKIP platform called SCRUB4PA (SCRUB for Process Automation), which is a software

development process based on agile practices for building KIPApps. This process will encompass phases, tasks, roles, artifacts, templates, and tools used throughout the development of a KIPApp. Overall, the AKIP Process Automation Platform is a unique and useful tool for facilitating business process automation, and the addition of SCRUB4PA will further enhance its capabilities.



## References

1. Basso, F.P., Pillat, R.M., Oliveira, T.C., Roos-Frantz, F., Frantz, R.Z.: Automated design of multi-layered web information systems. *J. Syst. Softw.* **117**, 612–637 (2016). <https://doi.org/10.1016/j.jss.2016.04.060>
2. Cardoso, J., Mendling, J., Neumann, G., Reijers, H.A.: A discourse on complexity of process models. In: Eder, J., Dustdar, S. (eds.) *BPM 2006*. LNCS, vol. 4103, pp. 117–128. Springer, Heidelberg (2006). [https://doi.org/10.1007/11837862\\_13](https://doi.org/10.1007/11837862_13)
3. Dumas, M., Van der Aalst, W.M., Ter Hofstede, A.H.: *Process-Aware Information Systems: Bridging People and Software Through Process Technology*. Wiley (2005)
4. Dumas, M., La Rosa, M., Mendling, J., Reijers, H.A.: *Fundamentals of Business Process Management*, 2nd edn. Springer, Heidelberg (2018). <https://doi.org/10.1007/978-3-662-56509-4>
5. Dwyer, G.: Stateful vs stateless architecture: why stateless won (2021). <https://www.virtasant.com/blog/stateful-vs-stateless-architecture-why-stateless-won>
6. Harmon, P.: The state of business process management 2016. Technical report (2016). <https://www.bptrends.com/bpt/wp-content/uploads/2015-BPT-Survey-Report.pdf>
7. Iskandar, T.F., Lubis, M., Kusumasari, T.F., Lubis, A.R.: Comparison between client-side and server-side rendering in the web development. *IOP Conf. Ser. Mater. Sci. Eng.* **801**, 012136 (2020)
8. ISO: ISO/IEC 19510:2013: Information technology - object management group business process model and notation. Technical report, Organization for Standardization (2013). <https://www.iso.org/standard/62652.html>
9. Kirchmer, M., Scheer, A.W.: Business process automation-combining best and next practices. In: Scheer, A.W., Abolhassan, F., Jost, W., Kirchmer, M. (eds.) *Business Process Automation*, pp. 1–15. Springer, Heidelberg (2004). [https://doi.org/10.1007/978-3-540-24702-9\\_1](https://doi.org/10.1007/978-3-540-24702-9_1)
10. Lins, L.F., Melo, G., Oliveira, T., Alencar, P., Cowan, D.: PACAs: process-aware conversational agents. In: Marrella, A., Weber, B. (eds.) *BPM 2021*. LNBI, vol. 436, pp. 312–318. Springer, Cham (2022). [https://doi.org/10.1007/978-3-030-94343-1\\_24](https://doi.org/10.1007/978-3-030-94343-1_24)
11. Lucas, E.M., Oliveira, T.C., Schneider, D., Alencar, P.S.C.: Knowledge-oriented models based on developer-artifact and developer-developer interactions. *IEEE Access* **8**, 218702–218719 (2020). <https://doi.org/10.1109/ACCESS.2020.3042429>
12. Mendling, J.: *Metrics for Process Models*. LNBI, vol. 6. Springer, Heidelberg (2008). <https://doi.org/10.1007/978-3-540-89224-3>
13. Mohapatra, S.: *Business Process Automation*. PHI Learning Pvt. Ltd. (2009)
14. OMG: *Business Process Model and Notation (BPMN)*, Version 2.0.2. Technical report, Object Management Group (2013). <https://www.omg.org/spec/BPMN/2.0.2>
15. Pillat, R.M., Oliveira, T.C., Alencar, P.S., Cowan, D.D.: BPMNT: a BPMN extension for specifying software process tailoring. *Inf. Softw. Technol.* **57**, 95–115 (2015). <https://doi.org/10.1016/j.infsof.2014.09.004>
16. Samland, F., Tuting, W.: Monolith to microservice, waterfall to agile, success with camunda (2019). <https://camunda.com/customer/deutsche-telekom/>

17. Scholz, T., Wagner, K.: Aris process platform<sup>TM</sup> and sap NetWeaver<sup>TM</sup>: next generation business process management. In: Scheer, A.W., Abolhassan, F., Jost, W., Kirchmer, M. (eds.) *Business Process Automation*, pp. 29–37. Springer, Heidelberg (2004). [https://doi.org/10.1007/978-3-540-24702-9\\_3](https://doi.org/10.1007/978-3-540-24702-9_3)
18. Shan, D.: Finance back-office billing engine using Camunda, May 2022. <https://camunda.com/customer/atlassian/>
19. Straube, C., Horn, D.: Scaling process automation with a modular open source platform (2021). <https://camunda.com/customer/city-of-munich/>
20. Sánchez-González, L., García, F., Ruiz, F., Piattini, M.: A case study about the improvement of business process models driven by indicators. *Softw. Syst. Model.* **16**(3), 759–788 (2015). <https://doi.org/10.1007/s10270-015-0482-0>
21. Telemaco, U., Oliveira, T., Alencar, P., Cowan, D.: A catalogue of agile smells for agility assessment. *IEEE Access* **8**, 79239–79259 (2020). <https://doi.org/10.1109/ACCESS.2020.2989106>
22. Telemaco., U., Oliveira., T., Pillat., R., Alencar., P., Cowan., D., Melo., G.: AKIP process automation platform: a framework for the development of process-aware web applications. In: *Proceedings of the 18th International Conference on Web Information Systems and Technologies - WEBIST*, pp. 64–74. SciTePress (2022). <https://doi.org/10.5220/0011550000003318>



# Grammar-Based Question Classification Using Ensemble Learning Algorithms

Alaa Mohasseb<sup>1</sup>  and Andreas Kanavos<sup>2</sup> 

<sup>1</sup> School of Computing, University of Portsmouth, Portsmouth, UK  
alaa.mohasseb@port.ac.uk

<sup>2</sup> Department of Informatics, Ionian University, Corfu, Greece  
akanavos@ionio.gr

**Abstract.** Question Classification is one of the important applications of information retrieval, as it plays a crucial role in improving the performance of question-answering systems. Differentiating between factoid and non-factoid questions is a particularly difficult task. Different methods have been suggested to improve the identification and classification of factoid questions. Most of these methods rely on semantic features and bag-of-words. This research paper explores the utilization of a Grammar-based framework for Questions Categorization and Classification (GQCC) to identify question types. This framework incorporates features such as grammatical features, domain-specific features, and patterns. These features leverage the question structure. By employing Ensemble Learning models, experimental findings demonstrate that the integration of question grammatical features with Ensemble Learning models contributes to achieving a good level of accuracy.

**Keywords:** Question classification · Grammatical features · Factoid questions · Information retrieval · Machine learning · Ensemble learning

## 1 Introduction

Question Classification, an essential application in the field of information retrieval, greatly impacts the performance of question-answering systems [34]. Properly distinguishing between different question types is crucial for generating accurate answers to user queries.

The classification and distinguishing between factoid and non-factoid questions presents a challenge. As indicated by [20], categorizing “wh-” questions into semantic categories is particularly difficult compared to other question types in question-answering systems. Additionally, selecting appropriate classifier features is vital for obtaining precise question classification [16]. Various studies have employed features such as bag-of-words [21, 27, 44, 45] semantic and syntactic features [13, 19, 40, 44], uni-gram and word shape features [16], as well as grammatical and domain-specific grammatical features [29, 31, 32] to classify questions.

Ensemble learning algorithms and techniques have gained significant attention in recent decades, both within the scientific and industrial communities [4, 35, 36]. These methods combine a diverse set of prediction models to create a composite global model, which yields accurate and reliable predictions or estimates. Theoretical and empirical

evidence has shown that ensemble models outperform individual models in terms of prediction performance [8]. Consequently, various ensemble learning methodologies and techniques have been proposed and applied to classification and regression problems in real-world scenarios [23,24].

In previous research [33], an ensemble learning grammar-based approach was utilized for question identification in which several experiments have been conducted to investigate the impact of combining grammatical features and domain-specific grammatical features with ensemble learning models.

In this paper, further experiments have been conducted using different ensemble learning algorithms to enhance the question types prediction and the model accuracy and to assess the performance of grammatical features and ensemble learning classifiers. The aim of the research presented in this paper is to:

- Evaluate the influence of using grammatical features and domain-specific grammatical features on the classification performance.
- Investigate the impact of using ensemble learning algorithms on the classification performance and compare the classification performance of different machine learning algorithms.

The rest of the paper is organized as follows: Sect. 2 provides an overview of previous work on question classification using various machine learning algorithms. Section 3 describes the approach and the grammatical features utilized, while Sect. 4 discuss the obtained results. Finally, Sect. 5 concludes the paper and outlines potential directions for future research.

## 2 Related Work

This section provides a comprehensive overview of previous research on question identification methods. Recent studies have proposed various approaches to question classification using different machine learning algorithms.

In one study [10], a method combining feature selection, ensemble classification, and the Gravitational Search Algorithm was proposed. Similarly, in [42] a feature selection algorithm was introduced to determine appropriate features for different question types. The authors designed a new type of feature based on question patterns and employed a feature selection algorithm to determine the most suitable feature set for each question type. The proposed approach was tested on the TREC benchmark dataset using SVM for the classification algorithm.

In [16], the authors introduced head word features, which are single words that specify the object of the question. They employed two approaches to enhance the semantic features of these headwords using WordNet. Furthermore, they augmented other standard features such as the wh-word, unigram feature, and word shape feature. Authors in [44] proposed a framework that integrates a question classifier with a document/passage retriever and context-ranking models. This framework offers flexible features, including word forms, syntactic features, and semantic word features. The context-ranking model, based on sequential labeling tasks, combines rich features such as full parsers, predefined syntactic patterns, and additional training materials to predict the relevance



of input passages to question types. While authors in [12] proposed a hybrid approach called ATICM, which utilizes dependency tree analysis for automated answer type identification and classification, incorporating both syntactic and semantic analysis. They employed a compact WordNet-based hypernym expansion strategy to classify question target words into question target categories. The ATICM approach achieved high accuracy on the UIUC and TREC10 datasets.

Furthermore, authors in [17] utilized methods such as word segmentation, Part-Of-Speech (POS) Tagging, and Named Entity Recognition (NER) for feature extraction. Support Vector Machine (SVM) and Random Forest algorithms were employed for question classification. The results indicated that SVM and Random Forest methods achieved favorable outcomes compared to ensemble learning and hierarchical classification approaches. In [21], authors utilized machine learning approaches, including different classifiers and multiple classifier combination methods. They incorporated a dependency structure from Minipar and linguistic knowledge from WordNet into question representation. Features such as Dependency Structure, Wordnet Synsets, Bag-of-Words, and Bi-gram were employed, and various kernel functions were explored. The precision of question classification was analyzed by evaluating different ways of combining classifiers, such as Voting, adaboost, Artificial Neural Networks (ANN), and Transition-Based Learning (TBL).

Additionally, a statistical classifier based on SVMs was proposed in [26], authors proposed a statistical classifier based on Support Vector Machines (SVM) that incorporates prior knowledge about correlations between question words and types. They developed question word-specific classifiers using SVM. This statistical framework can be used with any dataset, question ontology, or set of features.

SVMs were also utilized in [25,27,43]. For instance, in [25], SVM was used in conjunction with classifiers like MaxEnt, Naive Bayes, and Decision Tree for primary and secondary classification. Another study [27] employed SVM along with k-Nearest Neighbor and Naive Bayes, incorporating features such as bag-of-words, n-grams, as well as lexical, syntactic, and semantic features. Similarly, [43] introduced an SVM-based approach that incorporated dependency relations and high-frequency words for question classification. Lastly, Bidirectional Long-Short Term Memory (Bi-LSTM) was employed in [1] for question classification. The classification results demonstrated that Bi-LSTM achieved higher accuracy compared to basic LSTM and Recurrent Neural Network (RNN) models. While in [46] five machine learning algorithms were employed (KNN, NB, Decision Tree, Sparse Network of Winnows, and SVM) using similar features. In addition, in [5] SVM was employed for the classification of open-ended questions. They trained SVM to recognize specific keywords or phrases associated with question classes, allowing accurate identification based on keyword recurrence. [11] also used SVM for question classification. They represented questions as frequency-weighted vectors of salient terms, replacing regular expression-based classifiers.

Authors in [38,41] employed Neural Networks as the machine learning algorithm. In [38] a neural network-based question-answering system was proposed with multiple layers and networks. While in [41] two machine learning tasks were formulated; entity detection in the question and question classification into relation types in the knowledge base. Two recurrent neural networks were trained, outperforming state-of-the-art

approaches. Finally in [22], authors proposed an ensemble learning-based classification method for community question-answering systems. They employed supervised and semi-supervised learning with different feature extraction methods and classifiers, achieving enhanced classification accuracy.

### 3 Grammar-Based Ensemble Learning Approach

Our proposed approach for Question Categorization and Classification (QCC) utilizes a grammar-based framework, as introduced in a previous work [31]. The framework consists of three phases:

*Phase I: Question Analysis:* Initially, the question undergoes analysis to identify keywords and phrases, which serve as the basis for generating grammatical rules. Subsequently, a question domain-specific is created by combining a simplified version of English grammar with domain-specific grammatical categories.

*Phase II: Parsing and Mapping:* Each question is parsed and tagged using grammatical features in conjunction with domain-related information to transform it into its grammatical structure.

*Phase III: Question Classification:* In this phase, an automatic classification model is built and evaluated.

For this research, we employed the dataset and grammatical features generated in [31].

#### 3.1 Dataset

The dataset consists of 1,160 questions that were randomly selected from the following three different sources:

1. Yahoo Non-Factoid Question Dataset<sup>1</sup>
2. TREC 2007 Question Answering Data<sup>2</sup>
3. Wikipedia dataset<sup>3</sup> [39]

Each question in this dataset is classified into six different categories, which are: causal, choice, confirmation (Yes-No questions), factoid (“Wh-” questions), hypothetical and list. These categories are based on the question types in English and the classification is based on the types of questions asked by users and the answers given.

For the objective of investigating the impact of the ensemble learning model to distinguish between Factoid and Non-Factoid questions, a new label was created [33], entitled non-factoid which consists of the five question types, namely causal, choice, confirmation, hypothetical and list. Their distribution is given in Table 1.

<sup>1</sup> <https://ciir.cs.umass.edu/downloads/nfL6/>.

<sup>2</sup> [http://trec.nist.gov/data/qa/t2007\\_qadata.html](http://trec.nist.gov/data/qa/t2007_qadata.html).

<sup>3</sup> <https://www.cs.cmu.edu/~ark/QA-data>.

**Table 1.** Data Distribution [33].

Question Type	Number of Questions
Non-Factoid	473
Causal	31
Choice	12
Confirmation	32
Hypothetical	7
List	101
Factoid	687

### 3.2 Question Grammatical Structure

The primary objective of utilizing the grammatical features of questions is to leverage their structural aspects, encompassing both general and domain-specific grammatical categories [31]. One limitation of the aforementioned methodologies introduced thus far is their reliance on feature selection approaches to reduce the number of input variables. Consequently, these approaches overlook the grammatical structure of the questions. Given that question characteristics can vary significantly (e.g., some questions may be concise while others may possess multiple meanings, leading to ambiguity), using a limited set of features is inadequate. Additionally, two questions could share the same set of terms but exhibit different intents. Therefore, incorporating the grammatical structure of questions, along with domain-specific grammatical categories, can enhance the accuracy of the classification process [33].

The transformation of questions using grammatical features involves representing them as a sequence of grammatical terms. The grammatical features consist of *Verb*, *Noun*, *Determiner*, *Adjective*, *Adverb*, *Preposition*, and *Conjunction* in addition to question words such as *How*, *Who*, *When*, *Where*, *What*, *Why*, *Whose* and *Which*. Furthermore, the grammatical features consist of word classes like Nouns and Verbs. Furthermore, nouns can have sub-classes, such as *Common Nouns*, *Proper Nouns*, *Pronouns*, and *Numeral Nouns*; the same stands for verbs, which can have sub-classes, such as *Action Verbs*, *Linking Verbs* and *Auxiliary Verbs*. In addition, the grammatical features consist of other features as well, such as *Singular terms* (e.g. Common Noun - Other - Singular) and *Plural terms* (e.g. Common Noun - Other - Plural). Table 2 provides the list of the grammatical terms and their abbreviation.

Furthermore, domain-specific grammatical features related to question-answering were taken into consideration, which correspond to topics such as *Events*, *Entertainment*, *History and News*, *Health Terms*, *Geographical Areas*, *Places and Buildings* as shown in Table 3 [31]. These grammatical features and structures will be used in the question type identification since each factoid and non-factoid question type has a certain structure. The different feature representations help in distinguishing between different question types as shown in Table 4.

**Question Grammatical Structure Example.** The following example “*what are the symptoms of COVID*” will illustrate how these features are used:

**Table 2.** Grammatical Features [33].

Grammatical Feature	Abbreviation
Verbs	<i>V</i>
Action Verbs	<i>AV</i>
Auxiliary Verb	<i>AuxV</i>
Linking Verbs	<i>LV</i>
Adjective	<i>Adj</i>
Adverb	<i>Adv</i>
Determiner	<i>D</i>
Conjunction	<i>Conj</i>
Preposition	<i>P</i>
Noun	<i>N</i>
Pronoun	<i>Pron</i>
Numeral Numbers	<i>NN</i>
Ordinal Numbers	<i>NN<sub>O</sub></i>
Cardinal Numbers	<i>NN<sub>C</sub></i>
Proper Nouns	<i>PN</i>
Common Noun	<i>CN</i>
Common Noun - Other - Singular	<i>CN<sub>OS</sub></i>
Common Noun - Other - Plural	<i>CN<sub>OP</sub></i>
Question Words	<i>QW</i>
How	<i>QW<sub>How</sub></i>
What	<i>QW<sub>What</sub></i>
When	<i>QW<sub>When</sub></i>
Where	<i>QW<sub>Where</sub></i>
Who	<i>QW<sub>Who</sub></i>
Which	<i>QW<sub>Which</sub></i>

All terms in the questions will be extracted by parsing the following question:

*Question: What are the symptoms of COVID?*

The terms extracted are “What”, “are”, “the”, “symptoms”, “of”, “COVID”.

After the parsing process, each term in the question will be tagged to one of the grammatical features and domain-specific grammatical features, such as:

- *What* = *QW<sub>What</sub>*
- *are* = *LV*
- *the* = *D*
- *symptoms* = *CN<sub>OP</sub>*
- *of* = *P*
- *COVID* = *CN<sub>HLT</sub>*

After tagging each term in the question, the pattern is formulated as illustrated below:

*Pattern: QW<sub>What</sub> + LV + D + CN<sub>OP</sub> + P + CN<sub>HLT</sub>*

The question grammatical feature in each question will be used to identify the question type. As a result, this will produce the final classification of each question. In the given example, the question will be classified as *Non-Factoid*.

**Table 3.** Domain Specific Grammatical Features [33].

Domain specific Features	Abbreviation
Celebrities Name	$PN_C$
Entertainment	$PN_{Ent}$
Newspapers, Magazines, Documents, Books	$PN_{BDN}$
Events	$PN_E$
Companies Name	$PN_{CO}$
Geographical Areas	$PN_G$
Places and Buildings	$PN_{PB}$
Institutions, Associations, Clubs, Foundations and Organizations	$PN_{IOG}$
Brand Names	$PN_{BN}$
Software and Applications	$PN_{SA}$
Products	$PN_P$
History and News	$PN_{HN}$
Religious Terms	$PN_R$
Holidays, Days, Months	$PN_{HMD}$
Health Terms	$PN_{HLT}$
Science Terms	$PN_S$
Database and Servers	$CN_{DBS}$
Advice	$CN_A$
Entertainment	$CN_{Ent}$
History and News	$CN_{HN}$
Site, Website, URL	$CN_{SWU}$
Health Terms	$CN_{HLT}$

### 3.3 Question Types Identification

The following algorithms were investigated and utilized to address three specific aspects: identifying factoid and non-factoid questions, evaluating the use of domain-specific grammatical features with ensemble learning models, and assessing the classification accuracy comparing the classification performance obtained previously in [33].

- **Random Forest (RF)** is a type of ensemble learning technique that involves constructing a multitude of decision trees at the training time in which each tree depends on the values of a random vector sampled which are independently and identically distributed across all trees. For classification tasks, the output of the random forest is the class selected by most trees [3, 15].
- **Naive Bayes (NB)** is a probabilistic classifier that applies Bayes' theorem, assuming the independence of features within a class [37]. This classifier has been widely utilized in text classification because it is fast and easy to implement [28].
- **Support Vector Machine (SVM)** uses a hyperplane to separate the data. this algorithm aims to identify a hyperplane that maximizes the margin between support vectors within the dataset [7]. SVM has proven to be highly effective in text categorization and prediction tasks, as it eliminates the need for feature selection, simplifying the process of text categorization significantly. Furthermore, it does not require any parameter tuning as it can automatically determine suitable parameter settings [18].

**Table 4.** Grammatical Features that Identify Question Types [33].

Questions	Grammatical Features
Factoid	Question Words such as What, Where, When, Which, Why, Who, How
Non-Factoid	Conjunction (OR), Linking Verbs, Auxiliary Verbs, Plural Common Nouns, Question Words such as What, Which, Who, Why, How

- **Decision Tree (DT)** is a non-parametric method, which learns simple decision rules inferred from data features to create a model that predicts the value of a target variable.
- **K-Nearest Neighbour (KNN)** also constitutes a non-parametric and instance-based learning algorithm. This algorithm is based on a similarity measure, namely the distance function. KNN forms a majority vote between the  $k$  points and then, similarity is defined according to a distance metric between two data points. In the experiments, the value of  $k$  was equal to 3.
- **Voting** The approach combines various machine learning classifiers and employs a majority vote or the average predicted probabilities to determine the class labels. Specifically, in this paper, the majority vote (MV) method was employed. The aim of this approach is to enhance the model’s performance by leveraging multiple models. In a majority vote, the predicted class label corresponds to the class label that receives the highest number of predictions from each individual classifier.
  - **Voting Model 1 (VM1)** which consists of the following algorithms; Decision Tree, Support Vector Machine and K-Nearest Neighbour (DT, SVM, KNN).
  - **Voting Model 2 (VM2)** which consists of the following algorithms; Naive Bayes, Decision Tree and K-Nearest Neighbour (NB, DT, KNN).
  - **Voting Model 3 (VM3)** which consists of the following algorithms; Naive Bayes, Support Vector Machine and K-Nearest Neighbour (NB, SVM, KNN).
  - **Voting Model 4 (VM4)** which consists of the following algorithms; Naive Bayes, Decision Tree and Support Vector Machine (NB, DT, SVM).
  - **Voting Model 5 (VM5)** which consists of the following algorithms; Naive Bayes, RandomForest and Support Vector Machine (NB, RF, SVM).
  - **Voting Model 6 (VM6)** which consists of the following algorithms; Naive Bayes, RandomForest and K-Nearest Neighbour (NB, RF, KNN)
  - **Voting Model 7 (VM7)** which consists of the following algorithms; RandomForest, Support Vector Machine and K-Nearest Neighbour (RF, SVM, KNN).
- **Boosting** is an ensemble learning method that combines a set of weak learners into a strong learner to minimize training errors. The boosting process involves selecting a random subset of data, fitting a model to it, and then sequentially training subsequent models to compensate for the limitations of their predecessors. The following boosting models were built and examined in the experiments:
  - **AdaBoost** is a meta-estimator, which fits a sequence of weak learners on repeatedly modified versions of the data. In the following, it combines the predictions through a weighted majority vote (or sum) to produce the final prediction [9].

- **Gradient Boosting Classifier (GB)** is a machine learning technique utilized in various tasks, including regression and classification. It constructs a prediction model in the form of an ensemble, typically consisting of decision trees. The construction of a gradient-boosted trees model follows a stage-wise approach, similar to other boosting methods, however, this method enables the optimization of an arbitrary differentiable loss function, therefore offering greater flexibility [14].
  - **XGB Classifier** is an ensemble learning method that combines the predictions of multiple weak models to produce a strong prediction. The XGB Classifier is widely used in various domains due to its ability to handle complex datasets and its efficiency in producing accurate results in numerous machine learning tasks including classification and regression [6].
- **Bagging** is an ensemble learning approach employed to mitigate variance in noisy datasets. It involves randomly selecting data samples from a training set with replacement, allowing individual data points to be chosen multiple times. Subsequently, these selected samples are used to train independent weak models. The specific task, whether regression or classification, determines the subsequent steps in the bagging process. Bagged DT is a bagging model that was built and examined in the experiments. It is a meta-estimator that fits each base classifier on random subsets of the original dataset. This method generates multiple versions of a predictor and uses these in order to produce an aggregated predictor [2]. This method can be as well used to reduce the variance of a decision tree.

## 4 Experimental Study and Results

In the experimental study, Grammar-Based Ensemble Learning Approach was utilized. the objective of the experimental study is to investigate the ability of the ensemble learning models to distinguish between different question types based on grammatical features. To assess the performance of grammatical features and ensemble learning classifiers, several experiments have been conducted. The experiments were set up using the typical 10-fold cross-validation.

Table 5 presents the accuracy of the classification performance of the ensemble learning models. Additionally, Table 6 outlines the classification performance details, which are Precision, Recall and F-Score, of the classifiers that have been examined. The findings demonstrate that employing grammatical features in combined with ensemble learning algorithms achieves a high level of accuracy.

According to the results achieved previously in [33], Bagged DT achieved the highest accuracy, with a value equal to 89% in distinguishing between factoid and non-factoid questions. However, based on the current experiments XGB classifier achieved similar results to Bagged DT with a value equal to 89% while VM2 remain the lowest accuracy, e.g. 79%. Moreover, algorithms such as GB classifier, XGB classifier, and VM7 achieved high accuracy in identifying and classifying Factoid questions while VM5 and VM 6 achieved high accuracy in identifying and classifying non-factoid questions. In addition, based on previous experiments RF, VM1, Bagged DT, and VM7 are also more effective in the identification and classification of factoid questions, whereas

**Table 5.** Accuracy of the Ensemble Learning Models.

Ensemble Learning Model	<i>Non-Factoid</i>	<i>Factoid</i>	<i>Avg/Total</i>
Random Forest	78%	93%	88%
VM 1 (DT, SVM, KNN)	71%	93%	85%
VM 2 (NB, DT, KNN)	87%	74%	79%
VM 3 (NB, SVM, KNN)	73%	84%	80%
VM 4 (NB, DT, SVM)	83%	81%	82%
VM 5 (NB, RF, SVM)	85%	83%	83%
VM 6 (NB, RF, KNN)	86%	78%	81%
VM 7 (RF, SVM, KNN)	71%	91%	83%
AdaBoost	80%	90%	86%
Bagged DT	80%	93%	89%
GB Classifier	77%	95%	88%
XGB Classifier	79%	95%	89%

VM2 and VM4 classifiers are more accurate in the identification and classification of non-factoid questions as shown in Table 5.

In addition, regarding XGB Classifier, this classifier achieved a good performance in classifying factoid questions with a value equal to 95%; however, it achieved lower recall performance for non-factoid questions with a value equal to 79%. Similarly, GB Classifier and VM7 achieved good results in classifying factoid questions but achieve lower recall values, e.g. 77% and 71% respectively, for non-factoid questions. On the contrary, VM5 and VM6 achieved better results in classifying non-factoid questions, whereas lower recall values, e.g. 83% and 78% respectively, for factoid questions, were obtained.

In comparison to previous works, [30,31], algorithms such as KNN, SVM and NB were combined with grammatical features and domain-specific grammatical features. Specifically, in [30], KNN achieved an accuracy value equal to 83.7%, while in [31], SVM and NB achieved an accuracy of 88.6% and 83.5% respectively. These findings demonstrate that by combining domain-specific grammatical features with ensemble learning algorithms, the classification accuracy was improved, enabling the machine learning algorithms to better differentiate between factoid questions and non-factoid questions. Moreover, nearly all the algorithms achieved high performance and accurate classification accuracy. The following points summarise the above observations:

- After conducting further experiments using different ensemble learning algorithms the identification and classification of non-factoid questions have improved however, non-factoid questions remain the most difficult question type to predict.
- The imbalanced distribution of dataset categories, as shown in Table 1, had an impact on both the classification accuracy and the ability to predict non-factoid questions.
- The identification accuracy of factoid and non-factoid questions was influenced by shared grammatical features, particularly question words such as what, which, who, why, and how. These question words serve as primary grammatical indicators for identifying factoid questions, as demonstrated in Table 4.



**Table 6.** Classification Performance Details.

Question Type	<i>Precision</i>	<i>Recall</i>	<i>F1-Score</i>
Random Forest			
Non-Factoid	87%	78%	82%
Factoid	88%	93%	90%
VM 1 (DT, SVM, KNN)			
Non-Factoid	85%	71%	78%
Factoid	85%	93%	89%
VM 2 (NB, DT, KNN)			
Non-Factoid	66%	87%	75%
Factoid	91%	74%	82%
VM 3 (NB, SVM, KNN)			
Non-Factoid	73%	73%	73%
Factoid	84%	84%	84%
VM 4 (NB, DT, SVM)			
Non-Factoid	72%	83%	77%
Factoid	89%	81%	85%
VM 5 (NB, RF, SVM)			
Non-Factoid	74%	85%	79%
Factoid	90%	83%	86%
VM 6 (NB, RF, KNN)			
Non-Factoid	69%	86%	77%
Factoid	90%	78%	84%
VM 7 (RF, SVM, KNN)			
Non-Factoid	82%	71%	76%
Factoid	84%	91%	87%
AdaBoost			
Non-Factoid	82%	80%	81%
Factoid	89%	90%	89%
GB Classifier			
Non-Factoid	89%	77%	82%
Factoid	87%	95%	91%
XGB Classifier			
Non-Factoid	90%	79%	84%
Factoid	88%	95%	92%
Bagged DT			
Non-Factoid	88%	80%	84%
Factoid	89%	93%	91%

- By conducting additional experiments, high levels of accuracy were achieved using various ensemble learning algorithms, including GB, XGB, VM5, VM6, and VM7.
- XGB classifier Bagged DT achieved the highest accuracy, with a value equal to 89% in distinguishing between factoid and non-factoid questions.
- Our results showed that algorithms such as GB, XGB, RF, VM1, VM7 and Bagged DT are more suitable for the identification of factoid questions, whereas VM2, VM4, VM5 and VM6 classifiers are more suitable for the identification of non-factoid questions.

- The utilized grammar-based question classification approach using Ensemble Learning Algorithms improved the classification accuracy and domain-specific grammatical features helped in differentiating between factoid questions and non-factoid questions.

## 5 Conclusions and Future Work

In this paper, grammar-based Question Classification using Ensemble Learning Algorithms was utilized and examined for the prediction of factoid and non-factoid questions. This approach includes components such as grammatical features, domain-specific features, and patterns. These components aid in leveraging the question structure. Furthermore, we examined the effectiveness of various ensemble learning algorithms. The results demonstrate that our approach achieved good performance compared to existing methods. It is also proven that ensemble models provide considerably better prediction performance than single models.

As part of our future research, we intend to explore the influence of incorporating different levels of grammatical features on the prediction accuracy and conduct a comparative analysis of the results. In addition, we aim at testing our approach using different datasets and extend the analysis of the different types of questions and explore how applying imbalance methods to non-factoid questions would affect the prediction results.

## References

1. Anhar, R., Adji, T.B., Setiawan, N.A.: Question classification on question-answer system using bidirectional-LSTM. In: 5th International Conference on Science and Technology (ICST), pp. 1–5 (2019)
2. Breiman, L.: Bagging predictors. *Mach. Learn.* **24**(2), 123–140 (1996)
3. Breiman, L.: Random forests. *Mach. Learn.* **45**(1), 5–32 (2001)
4. Brown, G.: Ensemble learning. In: *Encyclopedia of Machine Learning*, pp. 312–320 (2010)
5. Bullington, J., Endres, I., Rahman, M.A.: Open-ended question classification using support vector machines. In: MAICS (2007)
6. Chen, T., et al.: XGBoost: extreme gradient boosting. R package version 0.4-2 **1**(4), 1–4 (2015)
7. Cortes, C., Vapnik, V.: Support-vector networks. *Mach. Learn.* **20**(3), 273–297 (1995)
8. Dietterich, T.G.: Ensemble learning. *Handb. Brain Theory Neural Netw.* **2**(1), 110–125 (2002)
9. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**(1), 119–139 (1997)
10. Golzari, S., Sanei, F., Saybani, M.R., Harifi, A., Basir, M.: Question classification in question answering system using combination of ensemble classification and feature selection. *J. Artif. Intell. Data Min. (JAIDM)* **10**(1), 15–24 (2022)
11. Hacioglu, K., Ward, W.H.: Question classification with support vector machines and error correcting codes. In: *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*. The Association for Computational Linguistics (2003)

12. Hao, T., Xie, W., Wu, Q., Weng, H., Qu, Y.: Leveraging question target word features through semantic relation expansion for answer type classification. *Knowl. Based Syst.* **133**, 43–52 (2017)
13. Hardy, H., Cheah, Y.N.: Question classification using extreme learning machine on semantic features. *J. ICT Res. Appl.* **7**(1), 36–58 (2013)
14. Hastie, T., Tibshirani, R., Friedman, J.: Boosting and additive trees. In: *The Elements of Statistical Learning*, pp. 337–387 (2009)
15. Ho, T.K.: Random decision forests. In: *3rd International Conference on Document Analysis and Recognition (ICDAR)*, pp. 278–282 (1995)
16. Huang, Z., Thint, M., Qin, Z.: Question classification using head words and their hypernyms. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 927–936 (2008)
17. Jiang, Y., Zhang, X., Jia, W., Xu, L.: Answer classification via machine learning in community question answering. *J. Artif. Intell.* **3**(4), 163–169 (2021)
18. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) *ECML 1998. LNCS*, vol. 1398, pp. 137–142. Springer, Heidelberg (1998). <https://doi.org/10.1007/BFb0026683>
19. Kanavos, A., Makris, C., Plegas, Y., Theodoridis, E.: Ranking web search results exploiting Wikipedia. *Int. J. Artif. Intell. Tools* **25**(3), 1650018:1–1650018:26 (2016)
20. Li, F., Zhang, X., Yuan, J., Zhu, X.: Classifying what-type questions by head noun tagging. In: *22nd International Conference on Computational Linguistics (COLING)*, pp. 481–488 (2008)
21. Li, X., Huang, X., Wu, L.: Question classification using multiple classifiers. In: *5th Workshop on Asian Language Resources and First Symposium on Asian Language Resources Network (ALR/ALRN@IJCNLP)* (2005)
22. Li, Y., Su, L., Chen, J., Yuan, L.: Semi-supervised learning for question classification in CQA. *Nat. Comput.* **16**(4), 567–577 (2017)
23. Livieris, I.E., Kanavos, A., Tampakas, V., Pintelas, P.E.: A weighted voting ensemble self-labeled algorithm for the detection of lung abnormalities from x-rays. *Algorithms* **12**(3), 64 (2019)
24. Livieris, I.E., Kiriakidou, N., Kanavos, A., Tampakas, V., Pintelas, P.E.: On ensemble SSL algorithms for credit scoring problem. *Informatics* **5**(4), 40 (2018)
25. May, R., Steinberg, A.: Building a question classifier for a TREC-style question answering system. *The Stanford Natural Language Processing Group, Final Projects* (2004)
26. Metzler, D., Croft, W.B.: Analysis of statistical question classification for fact-based questions. *Inf. Retrieval* **8**(3), 481–504 (2005)
27. Mishra, M., Mishra, V.K., Sharma, H.R.: Question classification using semantic, syntactic and lexical features. *Int. J. Web Semant. Technol. (IJWesT)* **4**(3), 39 (2013)
28. Mitchell, T.M.: *Machine Learning, International Edition*. McGraw-Hill Series in Computer Science, McGraw-Hill, New York (1997)
29. Mohasseb, A., Bader-El-Den, M., Cocea, M.: Classification of factoid questions intent using grammatical features. *ICT Express* **4**(4), 239–242 (2018)
30. Mohasseb, A., Bader-El-Den, M., Cocea, M.: Detecting question intention using a k-nearest neighbor based approach. In: *14th International Conference on Artificial Intelligence Applications and Innovations (AIAI)*, vol. 520, pp. 101–111 (2018)
31. Mohasseb, A., Bader-El-Den, M., Cocea, M.: Question categorization and classification using grammar based approach. *Inf. Process. Manag.* **54**(6), 1228–1243 (2018)
32. Mohasseb, A., Bader-El-Den, M., Cocea, M.: Domain specific grammar based classification for factoid questions. In: *15th International Conference on Web Information Systems and Technologies (WEBIST)*, pp. 177–184 (2019)

33. Mohasseb, A., Kanavos, A.: Factoid vs. non-factoid question identification: an ensemble learning approach. In: 18th International Conference on Web Information Systems and Technologies (WEBIST), pp. 265–271 (2022)
34. Moldovan, D.I., Pasca, M., Harabagiu, S.M., Surdeanu, M.: Performance issues and error analysis in an open-domain question answering system. *ACM Trans. Inf. Syst. (TOIS)* **21**(2), 133–154 (2003)
35. Pintelas, P.E., Livieris, I.E.: Special issue on ensemble learning and applications. *Algorithms* **13**(6), 140 (2020)
36. Polikar, R.: Ensemble learning. In: *Ensemble Machine Learning*, pp. 1–34 (2012)
37. Rennie, J.D.M., Shih, L., Teevan, J., Karger, D.R.: Tackling the poor assumptions of Naive Bayes text classifiers. In: 20th International Conference on Machine Learning (ICML), pp. 616–623 (2003)
38. Sagara, T., Hagiwara, M.: Natural language neural network and its application to question-answering system. *Neurocomputing* **142**, 201–208 (2014)
39. Smith, N.A., Heilman, M., Hwa, R.: Question generation as a competitive undergraduate course project. In: *NSF Workshop on the Question Generation Shared Task and Evaluation Challenge* (2008)
40. Song, W., Wenyin, L., Gu, N., Quan, X., Hao, T.: Automatic categorization of questions for user-interactive question answering. *Inf. Process. Manag.* **47**(2), 147–156 (2011)
41. Türe, F., Jojic, O.: Simple and effective question answering with recurrent neural networks. *CoRR abs/1606.05029* (2016)
42. Van-Tu, N., Anh-Cuong, L.: Improving question classification by feature extraction and selection. *Indian J. Sci. Technol.* **9**(17) (2016)
43. Xu, S., Cheng, G., Kong, F.: Research on question classification for automatic question answering. In: 2016 International Conference on Asian Language Processing (IALP), pp. 218–221 (2016)
44. Yen, S., Wu, Y., Yang, J., Lee, Y., Lee, C., Liu, J.: A support vector machine-based context-ranking model for question answering. *Inf. Sci.* **224**, 77–87 (2013)
45. Zhan, W., Shen, Z.: Syntactic structure feature analysis and classification method research. In: *International Conference on Audio, Language and Image Processing (ICALIP)*, pp. 1135–1140 (2012)
46. Zhang, D., Lee, W.S.: Question classification using support vector machines. In: 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 26–32 (2003)



# Leveraging Transfer Learning for Long Text Classification with Limited Data

Carlos Alberto Alvares Rocha<sup>1</sup>(✉) , Li Weigang<sup>1,2</sup>(✉) ,  
Marcos Vinícius Pinheiro Dib<sup>1,2</sup> , Allan Victor Almeida Faria<sup>1</sup> ,  
Daniel Oliveira Cajueiro<sup>1</sup> , Maísa Kely de Melo<sup>1,3</sup> ,  
and Victor Rafael Rezende Celestino<sup>1</sup>

<sup>1</sup> LAMFO, University of Brasilia, Brasilia, Brazil  
carlosrochacaar@gmail.com, weigang@unb.br

<sup>2</sup> TransLab, University of Brasilia, Brasilia, Brazil

<sup>3</sup> Federal Institute of Minas Gerais, Campus Formiga, Formiga, Brazil

**Abstract.** Natural language processing (NLP) has emerged as a significant area of research within the field of artificial intelligence, receiving increased attention in recent years, which has prompted the Brazilian Ministry of Science, Technology, and Innovation to launch a project aimed at finding international funding opportunities for Brazilian researchers through its Research Financing Products Portfolio. However, the challenge of classification in this context is exacerbated by the scarcity of high-quality labeled data, which is a requirement for state-of-the-art NLP implementations. In this study, we employ machine learning strategies to classify long, unstructured, and irregular texts obtained by scraping funding institutions' websites. Given the limited availability of labeled training data, we adopt an incremental approach to identify a suitable method with optimal performance. In order to alleviate the challenge of data scarcity, we use pre-training technology to learn word context from other data sets with significant similarities and larger scales. Then, we combine transfer learning with deep learning models to enhance sentence comprehension. We also conduct pre-processing experiments to address text irregularities. Comparative analysis with the baseline model reveals that our proposed approach yields promising results, with most trained models achieving over 90% accuracy. Our Longformer + CNN model has achieved 94% accuracy with 100% precision, while the Word2Vec + CNN model has achieved 93.55% accuracy. These findings highlight the successful application of artificial intelligence in public administration.

**Keywords:** Deep learning · Long text classification · Transfer learning · Limited size datasets · CNN · LSTM · DNN · Word embeddings · Longformer · Word2Vec

---

Supported by Ministry of Science, Technology and Innovation (MCTI).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023  
M. Marchiori et al. (Eds.): WEBIST 2022, LNBIP 494, pp. 98–120, 2023.  
[https://doi.org/10.1007/978-3-031-43088-6\\_6](https://doi.org/10.1007/978-3-031-43088-6_6)

# 1 Introduction

In Natural Language Processing (NLP), text classification is a traditional problem that involves predicting or assigning predefined categories to input texts. One of the challenges in text classification, particularly for long texts, is the limited availability of labeled training data. Collecting large amounts of labeled data can be expensive and time-consuming, especially in domains where data is scarce, such as research financing. Additionally, traditional approaches may not be effective in handling long texts due to the attention cost limitations of Transformer-based models, which typically restrict inputs to a fixed number of tokens [4, 7, 23].

More recent studies have demonstrated the effectiveness of pre-trained language models, such as Generative Pre-trained Transformer (GPT) [4, 23] and Bidirectional Encoder Representations from Transformers (BERT) [7], which utilize a large amount of unlabeled data to learn common language representations. However, the quantity and quality of the data used for training can influence the results, and in real-world applications, these models may yield unexpected and unsatisfactory results, affecting the robustness of the solution [11]. Despite the potential limitations of using pre-trained language models, they offer promising opportunities for transfer learning and leveraging knowledge from other domains to improve classification performance in the context of research financing opportunities, which are crucial for advancing scientific and technological innovation.

Motivated by the need to automate the Research Financing Products Portfolio (FPP), we propose leveraging transfer learning for long-text classification. The FPP is supported by the Brazilian Ministry of Science, Technology, and Innovation (MCTI) and consists of a collection of financing opportunities offered by various institutions worldwide outside of the Union budget. However, manually curating and managing such portfolios can be time-consuming and labor-intensive. Therefore, there is a need for automated approaches to streamline this process and ensure the efficient allocation of resources. The problem description and conceptual model of FPP/MCTI are shown in Fig. 1. The main idea presented is to collect text information on opportunities available within a selected list of institutions worldwide and then utilize our classification model to identify and filter opportunities eligible for Brazilian projects to construct the FPP.

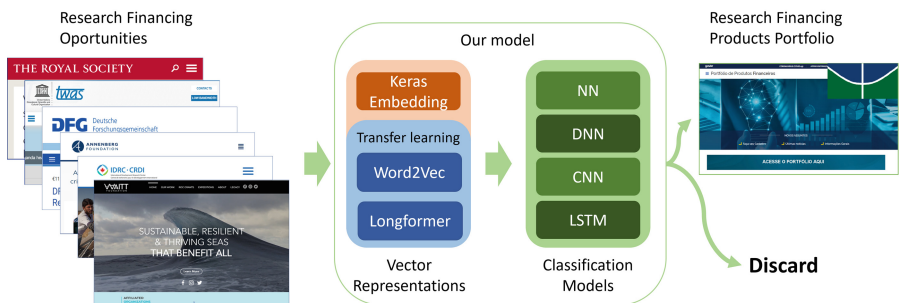


Fig. 1. FPP/MCTI classification model [25].

In this paper, we build upon the work conducted in [25] by exploring different NLP strategies to address the challenge of long-text classification with limited training data, leveraging transfer learning. Our goal is to identify the most effective method with optimal performance, with the ultimate aim of automating the FPP. We introduce a new pre-processing strategy to boost the performance of the classification model and present a new dataset with 928 rows of data that allowed us to achieve up to 94% classification accuracy. The remainder of this paper is organized as follows: in the next section, we review related work in the field of NLP and text classification, including recent approaches for handling long texts and utilizing transfer learning. We also describe the datasets used in our experiments, including a new dataset that we introduce in this study. The following section presents our proposed methodology, including the pre-processing strategy and the transfer learning approach. Next, we present our experimental results and discuss the findings. Finally, we conclude with a summary of our contributions and directions for future research.

## 2 Related Work

In NLP, classification involves predicting or assigning a predefined category to a text entry. Representation of the text is a fundamental step in this process. In the literature, several models of neural networks have been proposed for learning from textual representations, including convolutional models, recurrent models, and attention mechanisms. Another alternative is using pre-trained models on a large corpus of data, which have shown excellent performance for classification and other NLP tasks, potentially eliminating the need to train a new model from scratch. Examples of such pre-trained models include Word Embeddings like Word2Vec [16] and GloVe [21], as well as contextualized Word Embeddings like CoVe [15] and ELMo [22]. These Word Embeddings, whether contextualized or not, are often used as additional features to aid the main task.

Although several high-performance models are available for text classification, the literature does not provide a definitive answer on which model best suits the needs of a given project. In this section, we will present some models that served as a reference for the architecture of the model developed in this project.

### 2.1 Deep Learning Models

Recent advances in computational power and data availability have allowed Deep Learning a resurgence [30], more specifically, the progress of Artificial Neural Networks (ANN) such as Convolutional Neural Networks (CNN) and Long Short-term Memory (LSTM). These deep learning models have been successful in classifying texts and documents [18].

Zhou [36] uses, in LSTM and CNN models, Term Frequency-inverse Document Frequency (TF-IDF) to remove features with lower weights, extract key features in the text, extract the corresponding word vector through Word2Vec, and then insert it into the CNN-LSTM model. [14], to overcome Word2Vec's failure to capture the semantic information of the text entirely, proposes a hierarchical Transformer CNN model and

applies it to multi-label classification. A Transformer-CNN model is built to capture semantic information from different levels of the text (word and sentence level), using a multi-headed self-attention mechanism and a convolutional neural network to extract key semantic features.

[31, 34] used LSTM networks associated with Word Embeddings to achieve excellent results in text classification. [27] applied DNN with LSTM, performing tests with different approaches for textual representation, and highlighted the results of Word2Vec compared to Bag of Words (BOW) techniques. [12] performed a series of experiments using CNN trained from pre-trained vector representations for classification tasks and achieved excellent results.

## 2.2 Long Texts

Although approaches using Transformers have reached the state of the art in NLP tasks, they are more suitable for relatively short texts. These models usually limit input to  $n = 512$  tokens/words due to the  $O(n^2)$  cost related to the attention mechanism, making it challenging to use them to classify long texts [1].

Extended Transformer Construction (ETC) [1] and Longformer [2] were proposed to handle these kinds of problems and obtained state-of-the-art results through a balance between performance and memory use. The Longformer approach [2] uses a pattern of attention that combines local and global information while also scaling linearly with the sequence length. It can perform a wide range of document-level NLP tasks without segmenting/shortening long input and without complex architecture to combine information between these fragments, achieving state-of-the-art results in the character-level language modeling task.

Similarly, ETC [1] also uses an attention mechanism but differs from Longformer by combining global-local attention with relative position encodings and flexible masking, allowing it to encode inputs structured similarly to graph neural networks. While these implementations performed well when using Transformers, they used a large amount of training data to achieve their results. [29] presented an approach that uses fragmentation and text fractions to allow using BERT in long texts.

## 2.3 Few-Shot Learning

In 1998, the “Once learning” mechanism was proposed as a method for clustering images based on a single example, aiming to simulate human learning behavior [32] using Self-Organization Map (SOM) algorithms. This paradigm was also applied to identifying Radar images [33].

The researchers [17] defined a process that leverages shared densities in transformations to learn from a single training example for each class, using “prior knowledge” to develop a classifier.

Further advancements in this field were made by Li FeiFei and colleagues [10], who developed “One-shot learning” to utilize knowledge about object categories and classify new objects in a manner similar to human cognition. Subsequently, this concept was generalized as Few-shot Learning (“Learning in a few shots”), gaining acceptance in the research community and finding successful applications in the field of NLP [4].



## 2.4 Transfer Learning

Another way to train models with little data is Transfer Learning. The use of transfer learning in NLP tasks is not new [9, 19] and has achieved excellent results over the years.

The main idea is to transfer knowledge from different source domains to a target domain. A common approach is using word vector representations [24, 26], such as Word Embeddings. It is best used when sufficient training data is only available in another domain of interest. In this case, knowledge transfer could significantly improve learning performance, avoiding costly data labeling efforts [20].

BERT [7] is still presented as the state of the art for most NLP tasks [5]. To process long documents, BERT truncates text to the maximum input size (1024 for large BERT). However, this neural network generally only works when extensive information is available in the dataset [8, 13].

## 2.5 Research Financing Opportunities

In line with the government's digital transformation strategies, the Secretariat of Financial Structures and Projects (SEFIP) organized a working group to develop a project for automating the process of searching, classifying, and recommending funding opportunities and scientific research using Data Science and Artificial Intelligence (AI) techniques [3]. In this project, the classification task was performed using Naive Bayes, Support Vector Machines, Random Forest, and BOW and TF-IDF as representation techniques. A cross-analysis was conducted to evaluate the classification performance of the three methodologies and two vector representation techniques. The best accuracy results from each combination were selected and presented [28]. Their results can be seen in Table 1 and represent the baseline for this study.

**Table 1.** Overall baseline results [28].

	Accuracy	F1-Score	Precision	Recall
NB + BOW	0.82	0.81	0.81	0.82
NB + TF-IDF	0.86	0.85	0.85	0.85
SVM + BOW	0.78	0.76	0.76	0.76
SVM + TF-IDF	0.82	0.80	0.81	0.79
RF + BOW	0.84	0.82	0.85	0.80
RF + TF-IDF	0.82	0.80	0.81	0.79

The results were positive and promising in evaluating and classifying funding opportunities, particularly considering that the project had only 200 data points for training. The Naive Bayes methodology yielded the best result, achieving an accuracy of 86%, which is notable for small sample data. The work also suggests future proposals, such as incorporating deep learning into the classification process, which is the focus of this ongoing project.

## 2.6 Section Overview

After analyzing related works, it was found that several deep learning models have achieved excellent results in text classification. Efficient techniques to overcome the challenges of limited data and long texts have also been observed.

The effectiveness of DNN, CNN, and LSTM models in NLP for classification tasks can be highlighted. Additionally, transfer learning has proven beneficial in addressing the issue of limited training data. Transformer-based models have also shown to be promising options for these tasks, although some of them have limitations in terms of sentence size and require high computational power to execute.

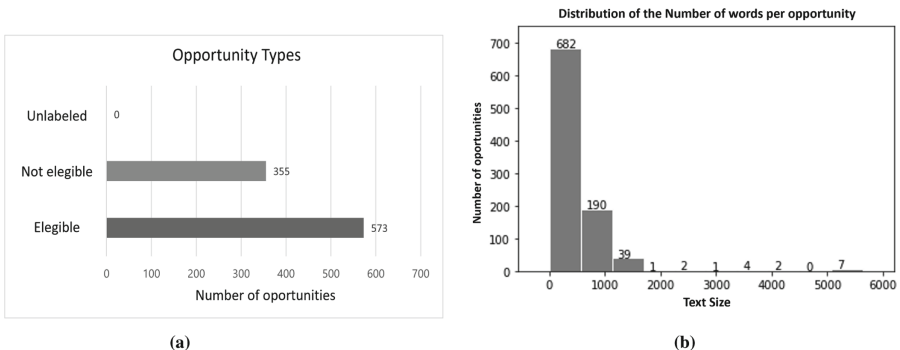
Considering the limited resources of this study, described in Sect. 4, it was decided to conduct tests using Word Embeddings as textual representation and DNN, CNN, and LSTM for the classification models. Transfer learning with Word2Vec and Longformer was also evaluated in the tests.

## 3 Datasets

In [25], the MCTI dataset (*FPP-base*) was introduced, consisting of 357 opportunities, of which 260 were categorized as eligible or ineligible, and the other 97 were unlabeled. The average text length in this dataset was 800 tokens but with significant variation, ranging up to 5000 tokens.

In addition, an unlabeled dataset (*FPP-extra*) was used, which was obtained from the scraping team and provided a preview of the *FPP-extended* dataset with opportunities previously scraped that still needed to be labeled. These opportunities can be used for unsupervised pre-training and comprised 519 opportunities with similar text length, all of which were unlabeled.

For this study, a larger dataset (*FPP-extended*) was introduced, which was delivered by the scraping team and consisted of 928 opportunities scraped from various institutions such as The Royal Society and The Waterloo Foundation, among others.



**Fig. 2.** *FPP-extended* dataset details - (a) Number of eligible, non-eligible, and non-labeled opportunities; (b) Text size in number of words (without pre-processing).

**Table 2.** Examples of opportunity texts.

opo_texto	opo_texto_ele
The Call offers two 12- month grants for Visiting Professors/Researchers who fled Ukraine. Grants will be allocated until funds are available (rolling application). Total budget: €40,000 Contacts: international.cooperation @ateneo.univr.it	The Call offers two 12- month grants for Visiting Professors/Researchers who fled Ukraine. Grants will be allocated until funds are available (rolling application). Total budget: €40,000 Contacts: international.cooperation @ateneo.univr.it
The Tinker Foundation’s Institutional Grants program provides project funding to organizations working to improve the lives of Latin Americans, with an emphasis on support for organizations in the region	Organization Status The Tinker Foundation provides grants only to organizations that are charitable in nature, i.e., with a United States 501(c)(3) tax status or its equivalent if the organization is located outside the U.S. Organizations from Latin America do not need to have United States 501(c)(3) status. Geographic Focus The project must be focused on the Spanish and Portuguese-speaking countries of Latin America, including: Argentina Bolivia Brazil Chile Colombia Costa Rica Cuba Dominican Republic Ecuador El Salvador Guatemala Honduras Mexico Nicaragua Panama Paraguay Peru Uruguay Venezuela

Compared to the *FPP-extra* dataset, the *FPP-extended* dataset contains 682 new opportunities acquired during the final scraping stage, using the most up-to-date scraping scripts. This new data is essential in the context of the project as it provides additional unique training and validation data for the network. These opportunities were manually labeled by the project team in collaboration with the MCTI team, and they formed the basis for training the classification model. The dataset, as shown in Fig. 2a, consists of 573 eligible and 355 non-eligible opportunities, with an average text length of 450 words. The distribution of the number of words per opportunity can be seen in Fig. 2b.

Table 2 provides examples of opportunities (non-eligible and eligible, respectively) to facilitate a closer examination of the dataset. For classification purposes, the columns of interest are `opo_texto` and `opo_texto_ele`, which correspond to the full text of the opportunity and the text segment containing eligibility information, respectively. The column `opo_texto_ele` was created based on a simple keyword search and often did not contain additional information relevant to understanding the opportunity.

From the analysis of Table 2, it is evident the presence of encoding issues that occur during the HTML scraping process, resulting in strange characters within some texts. Additionally, it is apparent that in certain opportunities, it is impossible to differentiate between the columns `opo_texto` and `opo_texto_ele`.

## 4 Methodology and Implementation

After reviewing the relevant literature, it became apparent that several approaches and models have the potential to achieve satisfactory classification results, each with its strengths and weaknesses.

Employing the most advanced techniques described in the literature could be an overestimation of the problem, as it may require unnecessary computational resources in addition to implementation efforts.

For this project, Google Colab Pro was used as the computational resource for model training. The environment utilized the Tesla P100 graphics card with 16 GB of video memory and 25 GB of RAM available in the “high RAM environment.” However, since this high RAM environment has a monthly hour limit, it had to be used judiciously.

An incremental approach was adopted, conducting tests with simpler architectures and strategies. The results obtained at each stage were analyzed, and improvements were suggested based on the diagnosis.

This section provides context for the related project and describes the modeling and implementation of the solutions used at each stage, highlighting the observations and challenges encountered.

### 4.1 Related Project

The FPP is a tool created and maintained by SEFIP and MCTI to promote scientific research and attract non-budgetary resources and financing. However, the current system for searching and updating information in the FPP is manual, time-consuming, and prone to errors. The project aims to modernize the system using an agile approach to implement automated searches in known electronic addresses of institutions that provide scientific opportunities. It will enable analysis and categorization of offers to provide recommendations, and, with user interactions, it will continuously improve the tool’s usability using reinforcement learning. The automation process is expected to improve customer service and timely service delivery, allowing for greater focus on strategic tasks. The project involves data scraping, classification, summarization, and recommendation. Additionally, the technology being developed has potential applications in other areas of public administration.

### 4.2 Vector Representations

In order to apply Deep Learning techniques in NLP, it is essential to convert the text into numerical information, known as vector representation, which captures the contextual use of a word (a set of characters) and its relationship with other words. This information is then introduced into the modeled neural network. For this purpose, Word Embeddings and Document-Long Embeddings were utilized.

In the proposed incremental approach, the initial step involved using embeddings trained along with the overall neural network, which was achieved by incorporating the Keras Embedding layer. Subsequently, pre-trained Word Embeddings were utilized using the Word2Vec technique as the next incremental step. Lastly, a Transformer was

employed to construct representations of the entire text instead of just words or sentences, utilizing the attention mechanism. Considering the project objectives, the Longformer model was selected. Table 3 provides a comparative summary of the vector representation techniques employed.

**Table 3.** Comparison of the vector representation techniques used.

	<i>Keras Embedding</i>	<i>Word2Vec</i>	<i>Longformer</i>
<b>Dimensionality</b>	8 dimensions/token	300 dimensions/token	768 dimensions/token
<b>Training Strategy</b>	Training with the model	Pre-training with unlabeled data	Pre-trained Base ( <i>longformer-base-4096</i> )
<b>Computational Cost</b>	Low	Medium	High

**Keras Embedding + Deep Learning.** The first approach is the simplest, using an Embedding layer with a representation of 8 dimensions for the embeddings to keep the computational cost low in the initial stage.

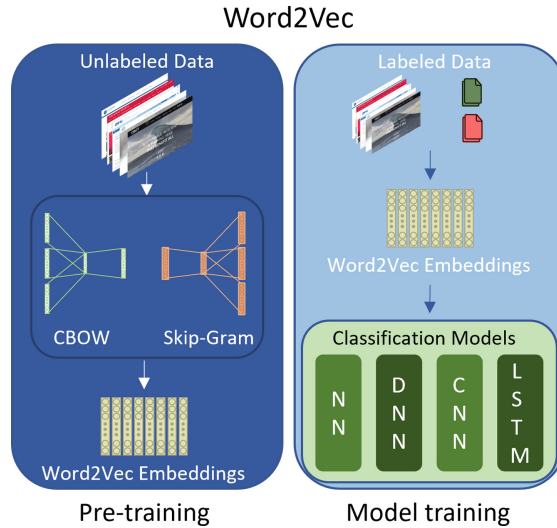
Since the Embedding layer is integrated with the neural network, it learns alongside the supervised training of the entire network. As a result, it requires labeled data to define and adjust its weights. It is the only technique discussed that does not involve transfer learning. Based on the literature review conducted, it was observed that most text classification architectures rely on a large amount of data for supervised training. Therefore, considering the limited amount of cataloged data available for training in this project, it is presumed that some form of transfer learning may be necessary to achieve the best possible results.

Despite being the most straightforward implementation, this approach may yield superior results compared to the BOW technique, as the Word Embeddings are updated with each training epoch, meaning that during each cycle in which all the data passes through the model, the weights are adjusted to better fit the proposed solution.

**Word2Vec + Deep Learning.** The second approach utilizes Word2Vec Embeddings for word representation. Unlike the first approach, the training of this model is unsupervised, allowing for the use of unlabeled data to train Word Embeddings. It can enable better training of the embeddings, leading to improved classification performance compared to the Keras Embeddings used in the previous step. However, to achieve better representations through pre-training, as shown in Fig. 3, it is crucial to use a robust dataset that has a high level of similarity with the type of text the model will process. In order to determine dataset similarity, Equation 1 was used as a comparative criterion among different datasets considered.

$$P_C = \frac{\#\{\{Base\} \cap \{New\}\}}{\#\{Base\}} * 100 \quad (1)$$

where  $\{Base\}$  represents the set of unique words from the MCTI dataset,  $\{New\}$  represents the set of unique words from the target dataset, and  $\#$  denotes the number of



**Fig. 3.** Word2Vec model training [25].

elements in a set. This formula calculates the percentage of words in the original dataset present in the new dataset, with a maximum value of 100%, indicating that all tokens are present in the new dataset.

In this work, the default number of dimensions, 300, was used for Word2Vec embeddings to take advantage of the model as designed.

The pre-trained Word2Vec model serves as a dictionary, translating words into their corresponding vector representations that will be used as input for the classification deep neural network. Alternatively, another option would be to couple the Word2Vec model to the network and fine-tune the weights, but this process requires higher computational resources.

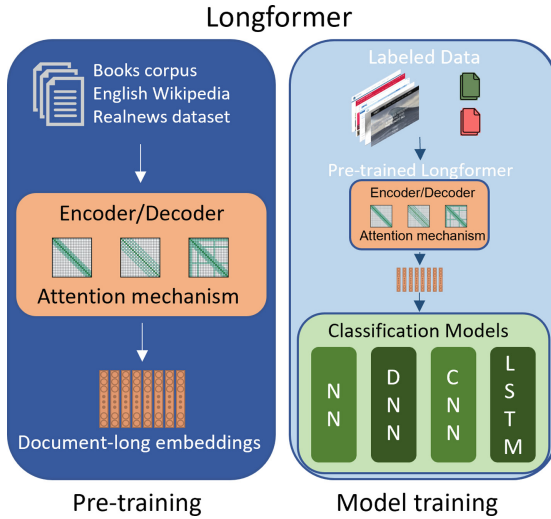
**Longformer + Deep Learning.** The third approach used for vector representations was Longformer [2], which was chosen due to the limitations of first-generation Transformers and BERT-based architectures that are restricted to a maximum of 512 tokens per input sequence.

Longformer overcomes this limitation by allowing sequences of up to 4096 characters to be processed without facing the memory bottleneck of BERT-type architectures. It has achieved state-of-the-art performance in several benchmarks.

Although Longformer has optimized memory usage compared to BERT-based models, it still requires substantial computational power to be used. In the experiments conducted by [2], RTX8000 GPU cards with 48 GB of GDDR6 RAM were used, which have higher computational power than the Tesla P100 GPU available in Colab Pro, which has only 16 GB of RAM.

Due to the limitation of computational power, the Longformer model was not trained from scratch in this work. Instead, a pre-trained base (available at the link<sup>1</sup>)

<sup>1</sup> <https://huggingface.co/allenai/longformer-base-4096>.



**Fig. 4.** Longformer model training [25].

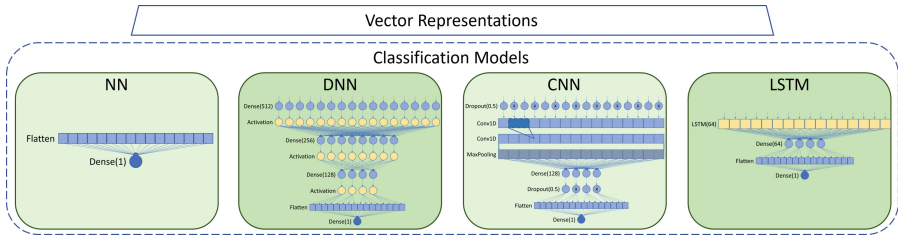
that was previously trained with a combination of vast datasets, including Book Corpus [37] plus Wikipedia in English and Realnews Dataset [35], was used as input to the model, as illustrated in Fig. 4.

Unlike previous techniques, Longformer performs document-long embeddings, which means it works with the full text to define its vector representation. In this case, the texts were converted to their contextualized vector representations and then used as input for the classification deep neural networks. Although training the Longformer model from scratch was not possible, it is expected that the model’s results will be superior to other approaches because Longformer is state-of-the-art in classifying long texts precisely because it incorporates the entire text by using the mechanism of attention in its vector representation.

### 4.3 Classification Models

After obtaining the vectorial representation of the texts, which are loaded with information, these vectors can be used as input to a classification neural network to analyze the text and classify it based on its eligibility for Brazilian research funding. The main focus of the research was to identify the classification methods that achieve the highest accuracy among the measured metrics, which include accuracy, precision, recall, and F1-score.

There are several types of neural network architectures with different performance characteristics, training times, and computational resource requirements. In this project, after analyzing the main architectures available in the literature and considering the project’s scope, the following neural networks were studied: Shallow Neural Network (SNN), DNN, LSTM, and CNN. The goal was to empirically analyze each architecture and determine which best fits the project objectives.



**Fig. 5.** Classification models [25].

**Table 4.** Comparative classification models analyzed.

	SNN	DNN	CNN	LSTM
<b>Layers</b>	Flatten and Dense	Dense, Activation and Flatten	Dropout, Conv1D, MaxPooling, Dense and Flatten	LSTM, Dense and Flatten
<b>Activation Functions</b>	Sigmoid	Sigmoid and ReLU	Sigmoid and ReLU	Sigmoid and ReLU
<b>Computational Cost</b>	Low	Medium	Medium	High

Figure 5 illustrates the structure and classification models analyzed, and Table 4 presents a comparative summary between them.

**SNN.** The analysis begins with the most straightforward neural network, the SNN, which consists of a Flatten layer and a dense classification layer. The main objective of this network is to validate the pre-processing, formatting, and data validation steps in a sequential neural network using Keras/TensorFlow. Since the architecture of the SNN is extremely simple, it allows for isolating the prior systems from the classification network without necessarily coupling them into a more complex network with more potential points of failure in the implementation. Furthermore, it can be trained quickly and with low computational costs due to its simplicity.

**DNN.** Continuing the analysis with a more elaborate neural network, the DNN, which is a Multi-Layer Perceptron (MLP) with multiple dense and activation layers, providing a more significant number of parameters for learning. The modeled architecture alternates between dense and ReLU activation layers, with varying dimensionality from 512 to 256, tapering down to 128. Finally, a Flatten layer was added to feed the final classification dense layer.

**CNN.** The architecture of the CNN network used in this study consists of a 50% dropout layer, followed by two 1D convolution layers associated with a MaxPooling layer. After MaxPooling, a dense layer with the size 128 is added, connected to a 50%



dropout layer, which ultimately connects to a Flatten layer and the final classification dense layer. Dropout layers are incorporated to prevent overfitting by masking part of the data, allowing the network to learn redundancies in the analysis of inputs.

**LSTM.** The LSTM model used in this study consists of a single LSTM layer with a dimension of 64, connected to a dense layer with the same dimension. After that, a Flatten layer and the final classification dense layer were added to the architecture.

#### 4.4 Pre-processing

In [25] the only pre-processing performed was removing special characters and tokenization. However, based on the results presented in [25], it was identified that more data and improved data treatment were needed. It was recognized that implementing better pre-processing techniques could enhance the training speed of the models and their overall performance.

Additional pre-processing techniques were incorporated to achieve this. Numerous techniques are available for reducing the number of unique words in a text without necessarily impacting its comprehension and interpretation by the computer. The following techniques were chosen for implementation in the project: Expansion of contractions; Converting text to lowercase; Punctuation removal; Stemization; Lemmatization; and Stopword removal.

A total of 8 experiments (Table 5) were conducted to identify the best approach for pre-processing the data, considering the reduction in the number of unique tokens, sentence size, and performance and accuracy improvement. These techniques were combined to compose the experiments, which were performed on the *FPP-extended* dataset and coupled with the simplest classification neural network (Keras Embedding + SNN) to validate the use of pre-processed data for training a classification network.

The first phase involved evaluating the treatment of punctuation and capitalization, resulting in the construction and evaluation of four experiments (Xp1, Xp2, Xp3, Xp4). Subsequently, the simplification of content was assessed based on the database resulting from the Xp4 experiment, considering stemming (Xp5), stemming (Xp6), stemming + stopwords removal (Xp7), and stemming + stopwords removal (XP8).

**Table 5.** Pre-processing experiments carried out.

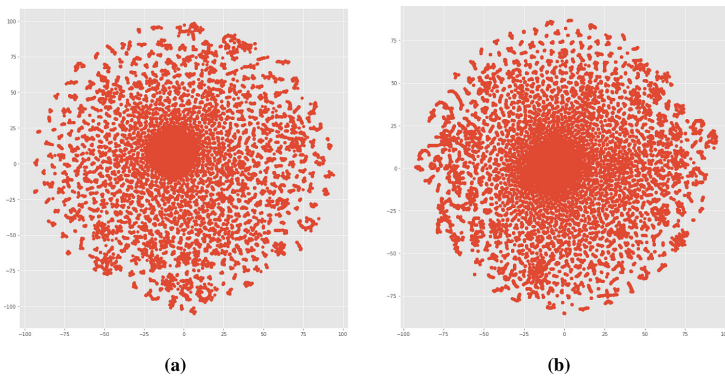
Base	Original texts
Xp1	Expansion of contractions
Xp2	Expansion of contractions + Converting text to lowercase
Xp3	Expansion of contractions + Punctuation removal
Xp4	Expansion of contractions + Punctuation removal + Converting text to lowercase
Xp5	Xp4 + Stemization
Xp6	Xp4 + Lemmatization
Xp7	Xp4 + Stemization + Stopword removal
Xp8	Xp4 + Lemmatization + Stopword removal

Based on the experiments detailed in Table 5, the most effective techniques for the type of data used were identified, considering the improvement in system performance and the effectiveness in reducing the size of input texts while ensuring minimal loss of information.

#### 4.5 Classifying *FPP-Extended*

In an attempt to classify the *FPP-base* dataset, deep learning models capable of classifying opportunity texts were identified. In the pre-processing stage (Subsect. 4.4), the best data treatment techniques were identified to optimize system performance. With the introduction of the *FPP-extended* dataset, the objective is to consolidate the best techniques on the more extensive dataset to obtain improved results. Since the new dataset contains 928 cataloged data, better results than those found in [25] are expected as the network will be able to learn in a supervised manner from a larger number of examples. Similarly to the initial classification stage, a cross-analysis was conducted using vector representation techniques and deep learning models to compare and identify the most suitable model for the final solution.

The application of embeddings using the Keras layer was straightforward and trouble-free. Notably, the SNN model was tested multiple times during the pre-processing stage for validation, which increased the likelihood of obtaining better initializations and improved results. For Word2Vec embeddings, pre-training was performed with the 928 data from the *FPP-extended* dataset, which approached the video memory limit of the available card in Google Colab Pro and required optimizations to free up memory for training. As the weights were trained with the same data as Keras embeddings, the difference in results between the two representation types is expected to be smaller than in the preliminary analysis, where Word2Vec had more training data. Another set of Word2Vec weights was trained with a dataset that combined opportunities from the *FPP-base*, *FPP-extra*, and *FPP-extended* datasets to enhance the Word Embeddings of the model. Figure 6, generated using the t-SNE technique, displays a



**Fig. 6.** Pre-trained Word2Vec weights: a) trained with the *FPP-extended* dataset; b) trained with the combined dataset (*FPP-base* + *FPP-extra* + *FPP-extended*).

representation of the trained weights, showcasing an enrichment of the weights with the addition of new words within the connected area, as evidenced by the increased density of points near the center, resulting from pre-training with this combined dataset of over 1500 opportunity texts.

The two sets of pre-trained weights were utilized for training the networks, and their results are presented and compared in Subsect. 5.2, enabling a comparison of the outcomes and an evaluation of the gains achieved through pre-training with a larger dataset.

The application of Longformer posed a significant challenge due to the model's size. The Keras Embedding representations have their size included in the final network and it takes less than 1 Mb to load in memory. The Word2Vec representations use a bit more memory, 56.1Mb for the model trained with the *FPP-extended* dataset and 85.2 Mb for the model trained with the combined dataset (*FPP-base + FPP-extra + FPP-extended*). Compared to the other models the Longformer takes up considerably more memory. Loading the Longformer representations into memory requires a whopping 11 GB of space just for the model, not considering the weight of the deep learning network. As a result, training using Longformer necessitates a GPU with more than 20 GB of memory, surpassing the project's available resources that utilizes a Tesla P100 board with 16 GB of memory.

As a result of this limitation, training the networks required utilizing a CPU environment. Specifically, the high RAM environment of Google Colab Pro, which provides 25 GB of memory for CPU usage, was employed. However, this does mean that training time was increased, as GPUs perform matrix operations more quickly.

The models were trained 20 times with 100 training epochs, except for the models utilizing Longformer, which were trained only five times due to their lengthy runtime.

## 5 Experiments and Results

The results obtained can be replicated using the notebooks in the project repository<sup>2</sup>. In this section, the results obtained from each of these experiments will be presented, followed by a brief discussion.

### 5.1 Pre-processing

The results of this stage, considering the metrics: accuracy, precision, recall, F1-score, training times, number of unique tokens (N\_tokens), and the maximum sentence length after pre-processing, are given in Table 6.

Table 6 reveals that the last two experiments (Xp7 and Xp8) yielded the best results regarding accuracy, f1-score, and precision. Additionally, these experiments had shorter training times and smaller sentence lengths. As such, both techniques demonstrate outstanding performance and can be selected as pre-processing steps.

The results obtained in the pre-processing step were analyzed in detail. Experiment Xp7 showed shorter training times and fewer unique tokens (N\_tokens), while Xp8

<sup>2</sup> Project repository: <https://github.com/chap0lin/PPF-MCTI>.

**Table 6.** Pre-processing results.

	Accuracy	F1-score	Recall	Precision	Avg. time	N_tokens	Máx. length
Base	0.8978	0.8420	0.7909	0.9095	417.77 s	23788	5636
Xp1	0.8871	0.8159	0.7154	0.9733	414.72 s	23768	5636
Xp2	0.9032	0.8564	0.7719	0.9744	368.38 s	20322	5629
Xp3	0.9194	0.8773	0.7966	0.9872	386.65 s	22121	4950
Xp4	0.9086	0.8661	0.8085	0.9425	326.83 s	18616	4950
Xp5	0.9194	0.8768	0.7847	1.0000	257.96 s	14319	4950
Xp6	0.8978	0.8506	0.7966	0.9187	282.65 s	16194	4950
Xp7	0.9247	0.8846	0.7966	1.0000	210.32 s	14212	2817
Xp8	0.9247	0.8846	0.7966	1.0000	225.58 s	16081	2726

had smaller maximum lengths. Although Xp7 had shorter training times, the difference compared to Xp8 in this aspect is so small that it is insignificant for decision-making.

Considering the methods and models to be applied throughout the work, the following points are relevant for analyzing which experiment is best suited:

**Keras Embedding.** A higher number of unique tokens in Keras Embedding only represents a more extensive vocabulary in one-hot encoding, but it does not necessarily increase the size of the network. On the other hand, the size of the largest sentence modifies the size of the input layer required for training, and this is also reflected in the number of weights in the next layer.

**Word2Vec.** In Word2Vec, a higher number of unique tokens means more time for pre-training the network. However, since this pre-training is done only once, the interference in the system is not significant. Additionally, the pre-trained Word2Vec weight file will be larger and must be loaded into memory to translate the input to the vector representation. However, it can also be unloaded from memory after translation and should not influence memory usage during training. The size of the most extensive sentence has little impact on Word2Vec training. However, it increases the input of the final network and the amount of data loaded into memory after vector representation.

**Longformer.** The Longformer used by [2] has been pre-trained with a much larger number of tokens, so the number of unique tokens is irrelevant in the analysis. The size of the largest sentence also has minimal impact, as the two values obtained in Xp7 and Xp8 (2817 and 2726, respectively) are below the maximum allowed size of 4096 for the network. However, it should be considered that when the model is applied for use in the MCTI, having a model capable of reducing the input size thoughtfully and effectively results in less loss of information being truncated or disregarded when it exceeds the maximum limit of the network.

Based on the issues raised, it was determined that the most suitable pre-processing model for this project's stage is experiment 8 (Xp8: XP4 + lemmatization + stopword removal), as it has a smaller input size for training and tends to reduce the input size further for future use, thus preventing information loss when it exceeds the maximum limit of the network.

## 5.2 Classifying *FPP-Extended*

Table 7 presents the results obtained by utilizing pre-trained Word2Vec embeddings with data from the *FPP-extended* dataset in combination with the deep learning models discussed in Subject. 4.3.

**Table 7.** Results for Word2Vec trained with the *FPP-extended* dataset, combined with deep learning models for classification of the *FPP-extended* dataset.

Model	Accuracy	F1-score	Precision	Recall
SNN	0.8925	0.8382	0.9710	0.7415
DNN	0.9032	0.8652	0.8870	0.8518
CNN	0.9247	0.8842	0.9872	0.8085
LSTM	0.8978	0.8436	0.9581	0.7536

The experiment was also conducted using the pre-trained Word2Vec model with the combined dataset (*FPP-base + FPP-extra + FPP-extended*) and the deep learning models. The results of this model are presented in Table 8.

**Table 8.** Results for Word2Vec trained with the combined dataset (*FPP-base + FPP-extra + FPP-extended*), combined with deep learning models for classification of the *FPP-extended* dataset.

Model	Accuracy	F1-score	Precision	Recall
SNN	0.9301	0.8964	1.0000	0.8125
DNN	0.9247	0.8830	0.9688	0.8125
CNN	0.9355	0.9040	0.9878	0.8333
LSTM	0.9247	0.8844	0.9563	0.8229

There was a substantial improvement in the accuracy of the models when additional data was included for training the Word2Vec model. Furthermore, the Word2Vec + CNN model achieved an accuracy of 93.5% and an f1-score exceeding 90%, a remarkable performance level. The final results of the classification are presented in Table 9.

The results of classifying the *FPP-extended* dataset, including the pre-processing step, are presented in Table 9. This table includes performance metrics, and the time taken for training each epoch and validating the results, specifically for ranking the top 20% of the validation set from the *FPP-extended* dataset.

The results reveal that these findings surpassed the outcomes achieved in the first stage of classification [25]. The Longformer + CNN model achieved an improved accuracy of 94%, while the Word2Vec + CNN model achieved 93.5%, and the Keras + CNN and Keras + LSTM models achieved 93%. This improvement can be attributed to the more significant amount of labeled data used for supervised training and the superior quality of data obtained through up-to-date scraping scripts.

It is worth mentioning that the unsupervised training of Word Embeddings using Word2Vec was not necessarily superior to embeddings trained with Keras, as evident from Table 7, which suggests that without additional data, Word2Vec embeddings are not necessarily better than simple embeddings trained along with the network. This is

**Table 9.** Final classification results.

Model	Accuracy	F1-score	Recall	Precision	Epoch time	Validation time
Keras + SNN	0.9247	0.8846	0.7966	1.000	0.2 s	0.7 s
Keras + DNN	0.8978	0.8441	0.7781	0.9257	1 s	1.4 s
Keras + CNN	0.9301	0.8991	0.8518	0.9569	0.4 s	1.1 s
Keras + LSTM	0.9301	0.8894	0.8332	0.9554	1.4 s	2 s
Word2Vec + SNN	0.9301	0.8964	0.8125	1.0000	1.4 s	1.2 s
Word2Vec + DNN	0.9247	0.8830	0.8125	0.9688	2 s	6.8 s
<b>Word2Vec + CNN</b>	<b>0.9355</b>	<b>0.9040</b>	<b>0.8333</b>	<b>0.9878</b>	1.9 s	3.4 s
Word2Vec + LSTM	0.9247	0.8844	0.8229	0.9563	2.6 s	14.3 s
Longformer + SNN	0.8924	0.8554	<b>0.8895</b>	0.8377	28 s	2.5 s
Longformer + DNN	0.9193	0.8762	0.8037	0.9762	81 s	8.4 s
<b>Longformer + CNN</b>	<b>0.9409</b>	<b>0.9069</b>	0.8341	<b>1.0000</b>	57 s	4.5 s

because that the new dataset has more labeled data, enhancing supervised joint learning. However, with the combination of datasets for pre-training the Word2Vec weights, the models demonstrate excellent results, as shown in Table 8.

Another important aspect is the difference in training times among the models. The models utilizing Keras embeddings took, on average less than 1 s per epoch, resulting in an average training time of 25 min when accounting for memory deallocation and metric calculations. The models utilizing Word2Vec showed slightly higher training times, still below 2 s per epoch, resulting in an average training time of 35 min. On the other hand, the models utilizing Longformer, as mentioned in Subsect. 4.5, needed to be trained using the CPU and presented an average time of 55 s per epoch, leading to an average training time of 5 h.

Despite using a high RAM environment to enable the training of models associated with Longformer, optimizations such as reducing the batch size and training in a small number of epochs at a time were necessary. Despite these efforts, the LSTM network model was not able to learn. The average training time of 5 h made adjustments and validation challenging, especially considering the monthly limit on resource usage. However, it is worth mentioning that although the training time is significantly higher for the Longformer model, the system execution time (validation time) does not differ significantly from the other models, indicating that the application with this model may not necessarily be slower compared to the others.

## 6 Discussions

The main objective of this research was to apply NLP and deep learning models for text classification of funding opportunities for the FPP. These texts were long, unstructured, non-uniform, and scarce, posing a clear challenge requiring state-of-the-art transfer learning and network training strategies. Initial classification attempts validated that using contextualized vector representations in conjunction with deep learning networks outperformed the baseline results obtained by [28]. This led to identifying improvements for the system and adjusting the models to maximize the results.

One effective strategy that significantly improved the system was implementing a structured pre-processing process. Tests that combined several well-known NLP and text normalization techniques resulted in a minimum increase of 2% in classification accuracy (Table 6). The strategy applied in experiment Xp8, selected based on the criteria described in Subsect. 5.1, involved the combination of contraction expansion, character standardization to lowercase, punctuation removal, lemmatization, and stopword removal. In addition to enhancing the network performance evaluation metrics, this strategy also reduced the maximum sentence size by more than half, optimizing the network size required for the application and significantly reducing the number of unique tokens that needed to be learned by the network.

Furthermore, the experiment revealed the need for more data for network training, which led to the introduction of the *FPP-extended* dataset. New training rounds performed on this pre-processed dataset improved accuracy levels, surpassing the 90% accuracy barrier in most of the trained models. Notably, the Longformer + CNN model achieved 94% accuracy with 100% precision, and other models, such as Word2Vec + CNN with 93.55%, Keras Embedding + CNN, and Keras Embedding + LSTM with 93% accuracy, also performed well with fewer computational resources. The CNN architecture showed a consistent performance and correlated with the best results obtained in the research.

The pre-training of Word2Vec weights was found to be a fundamental step in achieving good model performance. Significant gains in text understanding were observed compared to Keras Embedding models. However, as mentioned in Subsect. 5.2, these gains were only evident when pre-training was performed with more data, either in the *FPP-base* classification or the *FPP-extended* classification. This finding aligns with the recent trend of using more extensive pre-training data in large models to maximize results [6].

To determine the optimal model for utilization with the MCTI, it is essential to thoroughly analyze the top-performing models and understand their limitations. The model that achieved the best results was the Longformer in conjunction with the CNN network, boasting a remarkable accuracy of 94%. However, as indicated in Subsect. 4.5, Longformer necessitates substantial RAM to operate, and its model weight exceeds the maximum allowed by Github or Hugging Face repositories.

The second best model identified is Word2Vec + CNN, which does not exhibit any performance limitations and has a modest size of only 85.2 Mb. Because of that, this model was chosen for the classification prototype of the related project.

Another crucial aspect to consider is the opportunities that the model misclassified. Upon closer examination, it was found that some texts lack sufficient information for making an accurate decision, indicating a limitation related to the quality of data provided to the network.

Those *inconclusive* texts account for less than 5% of the *FPP-extended* dataset. Therefore, while the models may still benefit from additional training data and optimizations, their performance will be constrained by the quality of input data.

Another critical aspect to consider is the precision value obtained in some of the models. Although the results are positive, it cannot be guaranteed that the system has 100% accuracy, primarily due to the data quality issue highlighted earlier and the limited amount of data used for validation.

## 7 Final Considerations

Text classification is a well-established problem in NLP, and it has gained significant attention and advancements in recent years due to the growing popularity of artificial intelligence. The increasing interest of the MCTI in automating processes related to its portfolio of financial products (Subsect. 4.1) has been fueled by these advancements. Developing a system that can automatically identify and recommend funding opportunities within the context of the FPP requires an effective model capable of filtering these opportunities based on their eligibility for Brazilian projects. Therefore, this work proposes a solution to the challenge of classifying opportunity texts, which are often long, unstructured, non-uniform, and scarce. Through reviewing related work and the state-of-the-art, deep learning techniques, contextual representation of texts, and transfer learning were identified as motivating factors for developing the models used in this research. Word Embeddings with Keras and Word2Vec were used as textual representation mechanisms, along with Longformer for document-long embeddings. SNN, DNN, CNN, and LSTM models were employed as neural networks for classification. Experiments were conducted through cross-analysis of techniques and models to validate performance and determine the most suitable approach for the application.

The data used for training these models (Sect. 3) was obtained through data scraping scripts applied on online platforms of various financial institutions. The data was noisy and had text encoding issues, necessitating pre-processing. Experiments were designed (Table 5) to identify the best techniques to be applied, resulting in a significant reduction in the maximum sentence size, a decrease in the number of unique tokens by more than half, and an increase in classification performance by at least 2%.

Word2Vec and Longformer transfer learning models were employed during the research and development process. However, due to computational power limitations, it was impossible to perform pre-training or fine-tuning of the Longformer weights. Nonetheless, the model demonstrated high efficiency in classification, achieving 94% accuracy in the architecture that utilized CNN. Word2Vec also exhibited strong performance, with 93.55% accuracy in the architecture with CNN. Pre-training using unsupervised learning with texts of unlabeled opportunities resulted in up to 3.5% accuracy gain compared to weights trained solely with the *FPP-extended* dataset (Subsect. 5.2).

Despite the Longformer + CNN architecture producing the best classification results, it required significant computational power. Therefore, the Word2Vec + CNN architecture was selected for the prototype of the application to be used by the MCTI, highlighting the importance of the incremental approach followed in the project's methodology, which facilitated the identification of a suitable model that met the implementation requirements. The research achieved a new level of results in classifying project financing opportunities associated with the FPP, surpassing the work of [28]. These results led to the publication of a paper [25] and strengthened the collaboration between the government and the university.

For future work, it is suggested:

1. In order to enhance information quality and system performance, improvements in data acquisition and processing may be implemented, such as introducing an "inconclusive" class to accommodate entries where the eligibility criteria are unclear;



2. Comparison of the Word2Vec pre-training strategy with CNN for the classification of other datasets in the literature;
3. With the availability of greater computational power, fine-tuning the Longformer model and developing more optimizations to enable its use by the MCTI.

**Acknowledgments.** The Brazilian Ministry of Science, Technology, and Innovation (MCTI) has provided partial support for this project. We sincerely thank Dr. Joao Gabriel Souza, who led the efforts in constructing the dataset and graciously shared the data for this study.

## References






1. Ainslie, J., et al.: ETC: encoding long and structured inputs in transformers (2020). <https://doi.org/10.48550/ARXIV.2004.08483>, <https://arxiv.org/abs/2004.08483>
2. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: the long-document transformer (2020). <https://doi.org/10.48550/arXiv.2004.05150>, <https://arxiv.org/abs/2004.05150>
3. Brasil: Ministério de ciência, tecnologia e inovações. portfólio de produtos financeiros (2019). <https://ppf.mcti.gov.br/>
4. Brown, T., et al.: Language models are few-shot learners. *Adv. Neural. Inf. Process. Syst.* **33**, 1877–1901 (2020)
5. van den Bulk, L.M., Bouzembrak, Y., Gavai, A., Liu, N., van den Heuvel, L.J., Marvin, H.J.: Automatic classification of literature in systematic reviews on food safety using machine learning. *Curr. Res. Food Sci.* **5**, 84–95 (2022)
6. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.: Big self-supervised models are strong semi-supervised learners (2020)
7. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR abs/1810.04805* (2018). <http://arxiv.org/abs/1810.04805>
8. van Dinter, R., Catal, C., Tekinerdogan, B.: A decision support system for automating document retrieval and citation screening. *Expert Syst. Appl.* **182**, 115261 (2021)
9. Do, C.B., Ng, A.Y.: Transfer learning for text classification. In: Weiss, Y., Schölkopf, B., Platt, J. (eds.) *Advances in Neural Information Processing Systems*, vol. 18. MIT Press (2005). <https://proceedings.neurips.cc/paper/2005/file/bf2fb7d1825a1df3ca308ad0bf48591e-Paper.pdf>
10. Fei-Fei, L., Fergus, R., Perona, P.: A Bayesian approach to unsupervised one-shot learning of object categories. In: *Proceedings ninth IEEE International Conference on Computer Vision*, pp. 1134–1141. IEEE (2003)
11. Gron, A.: *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 1st edn. O’Reilly Media Inc, Sebastopol (2017)
12. Kim, Y.: Convolutional neural networks for sentence classification. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751. Association for Computational Linguistics, Doha, Qatar, October 2014. <https://doi.org/10.3115/v1/D14-1181>, <https://aclanthology.org/D14-1181>
13. Kontonatsios, G., Spencer, S., Matthew, P., Korkontzelos, I.: Using a neural network-based feature extraction method to facilitate citation screening for systematic reviews. *Expert Syst. Appl.* **X 6**, 100030 (2020)
14. Li, J., et al.: Multi-label text classification via hierarchical transformer-CNN. In: *2022 14th International Conference on Machine Learning and Computing (ICMLC)*. ICMLC 2022, pp. 120–125. Association for Computing Machinery, New York, NY, USA (2022). <https://doi.org/10.1145/3529836.3529912>

15. McCann, B., Bradbury, J., Xiong, C., Socher, R.: Learned in translation: Contextualized word vectors. In: NIPS (2017)
16. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems. NIPS'13, vol. 2, pp. 3111–3119. Curran Associates Inc., Red Hook, USA (2013)
17. Miller, E.G., Matsakis, N.E., Viola, P.A.: Learning from one example through shared densities on transforms. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662), vol. 1, pp. 464–471 (2000)
18. Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J.: Deep learning-based text classification: a comprehensive review. *ACM Comput. Surv.* **54**(3) (2021). <https://doi.org/10.1145/3439726>
19. Pan, S.J., Tsang, I.W.H., Kwok, J.T.Y., Yang, Q.: Domain adaptation via transfer component analysis. *IEEE Trans. Neural Netw.* **22**, 199–210 (2011)
20. Pan, S.J., Yang, Q.: A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**(10), 1345–1359 (2010). <https://doi.org/10.1109/TKDE.2009.191>
21. Pennington, J., Socher, R., Manning, C.: GloVe: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar, October 2014. <https://doi.org/10.3115/v1/D14-1162>, <https://aclanthology.org/D14-1162>
22. Peters, M.E., et al.: Deep contextualized word representations (2018). <https://doi.org/10.48550/ARXIV.1802.05365>, <https://arxiv.org/abs/1802.05365>
23. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
24. Raina, R., Battle, A., Lee, H., Packer, B., Ng, A.Y.: Self-taught learning: transfer learning from unlabeled data. In: Proceedings of the 24th International Conference on Machine Learning. ICML '07, pp. 759–766. Association for Computing Machinery, New York, USA (2007). <https://doi.org/10.1145/1273496.1273592>
25. Rocha, C.A.A., et al.: Using transfer learning to classify long unstructured texts with small amounts of labeled data. In: Proceedings of the 18th International Conference on Web Information Systems and Technologies - WEBIST, pp. 201–213. INSTICC, SciTePress (2022). <https://doi.org/10.5220/0011527700003318>
26. Ruder, S., Peters, M.E., Swayamdipta, S., Wolf, T.: Transfer learning in natural language processing. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials, pp. 15–18. Association for Computational Linguistics, Minneapolis, USA, June 2019. <https://doi.org/10.18653/v1/N19-5004>, <https://aclanthology.org/N19-5004>
27. Semberecki, P., Maciejewski, H.: Deep learning methods for subject text classification of articles, pp. 357–360, September 2017. <https://doi.org/10.15439/2017F414>
28. Silva, B., Alves, J., Rebeschini, J., Querol, D., Pereira, E., Celestino, V.: Data science applied to financial products portfolio. In: Annals of Meeting of National Association of Post-graduation and Research in Administration (2021)
29. Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune BERT for text classification? In: Sun, M., Huang, X., Ji, H., Liu, Z., Liu, Y. (eds.) CCL 2019. LNCS (LNAI), vol. 11856, pp. 194–206. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-32381-3\\_16](https://doi.org/10.1007/978-3-030-32381-3_16)
30. Thompson, N.C., Greenewald, K., Lee, K., Manso, G.F.: The computational limits of deep learning (2020). <https://doi.org/10.48550/ARXIV.2007.05558>, <https://arxiv.org/abs/2007.05558>
31. Wang, J., Wang, Z., Zhang, D., Yan, J.: Combining knowledge with deep convolutional neural networks for short text classification, pp. 2915–2921, August 2017. <https://doi.org/10.24963/ijcai.2017/406>

32. Weigang, L.: A study of parallel self-organizing map. arXiv preprint quant-ph/9808025 (1998)
33. Weigang, L., da Silva, N.C.: A study of parallel neural networks. In: IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339), vol. 2, pp. 1113–1116. IEEE (1999)
34. Xiao, L., Wang, G., Zuo, Y.: Research on patent text classification based on word2vec and LSTM. In: 2018 11th International Symposium on Computational Intelligence and Design (ISCID), vol. 01, pp. 71–74 (2018)
35. Zellers, R., et al.: Defending against neural fake news (2019). <https://doi.org/10.48550/ARXIV.1905.12616>, <https://arxiv.org/abs/1905.12616>
36. Zhou, H.: Research of text classification based on TF-IDF and CNN-LSTM. J. Phys. Conf. Ser. J. Phys. Conf. Ser. **2171**, 012021 (2022). <https://doi.org/10.1088/1742-6596/2171/1/012021>
37. Zhu, Y., et al.: Aligning books and movies: towards story-like visual explanations by watching movies and reading books (2015). <https://doi.org/10.48550/ARXIV.1506.06724>, <https://arxiv.org/abs/1506.06724>



# Integrating Linguistic and Citation Information with Transformer for Predicting Top-Cited Papers

Masanao Ochi<sup>1</sup>, Masanori Shiro<sup>2</sup>, Jun'ichiro Mori<sup>1</sup>,  
and Ichiro Sakata<sup>1</sup>

<sup>1</sup> Department of Technology Management for Innovation, Graduate School of Engineering, The University of Tokyo, Hongo 7-3-1, Bunkyo, Tokyo, Japan  
[masanao.oochi@gmail.com](mailto:masanao.oochi@gmail.com)

<sup>2</sup> Human Informatics and Interaction Research Institute, National Institute of Advanced Industrial Science and Technology, Umezono 1-1-1 Central2, Tsukuba, Ibaraki, Japan

**Abstract.** Academic literature contains many different types of data, including language, citations, figures, tables, etc. Released in 2017 for natural language processing, the Transformer model is widely used in fields as diverse as image processing and network science. Transformer models trained on large datasets have been shown to perform well on different tasks when trained on additional small amounts of new data. Most studies of classification and regression in the academic literature have designed individually customized features for specific tasks without fully considering data interactions. Customizing features for each task is costly and complex. This paper addresses this issue by proposing a basic framework that can be consistently applied to various tasks from the diverse information in academic literature. Specifically, we propose an end-to-end fusion method that combines linguistic and citation information of academic literature data utilizing two Transformer models. The experiments were conducted using a dataset on 67 disciplines extracted from the Web of Science, one of the largest databases about academic literature, and classified papers with the top 20% of citations five years after publication. The results show that the proposed method improves the F-measure by 0.028 compared to using only citation or linguistic information on average. Repeated experiments on 67 data sets also showed that the proposed model has the smallest standard deviation of F values. In other words, our proposed method shows **the best average performance and is stable with a small variance of the F-value**. We also conducted a comparative analysis between the dataset's characteristics and the proposed method's performance. The results show that our proposed method correlates poorly with the dataset's characteristics. In other words, our proposed method is highly versatile. Based on the above, our proposed method is superior regarding F-value, learning stability, and generality.

**Keywords:** Citation analysis · Scientific impact · Graph neural network · BERT · Transformer

## 1 Introduction

Recently, much of the scholarly literature has been digitized and research topics have been subdivided. However, to use limited resources efficiently, it is important to discover promising research topics worth investing in early and efficiently. In other words, there is now a need to develop technology to predict future research trends automatically. Previous studies on impact prediction of scholarly literature have used features specifically designed for each indicator [1, 3, 4, 6, 13, 23–25] or a link prediction using custom networks [20, 28, 30]. However, recent advances in deep learning technology have facilitated integrating different individual models and constructing more general-purpose models, such as the Transformer model [26]. The Transformer model, released in 2017, was initially used in natural language processing [8] but has since been widely used in various fields, including image processing [9] and network science [31]. This model has several advantages, including publishing trained models with large datasets and fine-tuning by applying new data to individual tasks. Traditionally, the impact of the academic literature has been evaluated either by creating individual features or by formulating the problem as a network link prediction problem. However, the rise of general-purpose models such as the Transformer may change this situation.

The actual scholarly literature contains various data, including language, citations, and images of figures and tables. Therefore, several studies have pointed out that network information, rather than linguistic information, may be necessary for predicting the impact of scholarly literature [17, 23]. In particular, Ochi et al. report that citation networks may be more biased than linguistic information in the embedding space of papers with future high citations [17]. This result indicates the need to develop a more advanced model than the BERT model using only linguistic information in the academic literature, such as the SPECTOR model [7], with the top-cited papers as teacher data.

Using the Transformer model in this study, we propose a new end2end fusion method of linguistic and citation information in scholarly literature data. Using a dataset extracted from the Web of Science, we evaluated the proposed method for classifying papers with the top 20% of citations five years after publication. We found that the proposed method improved the F-value by 0.028 compared to using only individual information. This method makes it possible to connect diverse data from the scholarly literature into end2end. Experimental results show the possibility of efficiently improving the accuracy of various classifications and predictions in actuality.

Our proposed method contributes to four points.

- We developed an end2end model that fuses linguistic features and a citation network of scholarly literature data.
- The proposed model automatically selects when citation network information is valid and when linguistic information is valid.
- The proposed model improves the classification accuracy of the papers with the highest number of citations after five years.
- The proposed model shows high learning stability and generality independent of the characteristics of the data set.

The remainder of this article will first introduce the related work in Sect. 2. In the section, we describe the context of the prediction of scholarly impact and clarify the needs of the end2end model, which connects linguistic and citation information. We describe our proposed model, including its architecture, in Sect. 3. Then we explain the experiment in detail in Sect. 4. We show the experimental results in Sect. 5, and discussions of the results are in Sect. 6. Finally, in Sect. 7, we show the scientific contribution of our work and note several challenges we can address in the future.

This paper extends and completes previous working papers [17, 18]. We have added more datasets and experiments compared to the previous paper to clarify the effectiveness and limitations of the proposed method. We also discuss the results in more detail and show that the proposed method may be effective on larger datasets.

## 2 Related Work

In this section, we categorize and summarize recent reports on studies of the Transformer model that are relevant to this study. We contextualize the Transformer model, describe its application and extension to scholarly literature data, and then describe the research conducted on the index, the influence of scholarly literature, its predictions, and challenges, and clarify the position of this study.

### 2.1 Transformer Model for Scholarly Data

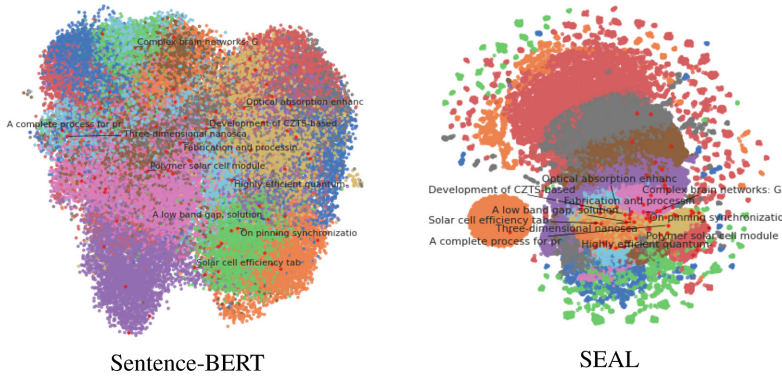
The Transformer model [26], one of the Encoder-Decoder models using Attention, is capable of large-scale learning due to its slight computational complexity and parallel computing capability. The Transformer model was quickly put to use when the BERT model [8] showed the highest accuracy on the GLUE dataset [27], a multi-task accuracy competition for natural language processing. Since then, its use has expanded in diverse fields, such as image processing [9], seismology [15], cardiology [29] and network science [31].

The application of the Transformer model to scholarly literature data is also underway. The first is the SciBERT model [5], which is based on the BERT model and trained on text data from academic literature. SciBERT focuses on generic Embedding acquisition for academic literature at the word level. However, the SPECTER model [7] attempts to obtain Embedding at the paper level rather than at the word level. The SPECTER model acquires Embedding at the paper level by making the papers with a citation relationship with each other a pair of positive examples.

### 2.2 The Influence of Scholarly Literature

However, scholarly literature contains not only text but also various types of information such as citations, figures, tables, authors, and institutional affiliations. Researchers used this information to index the influence of academic literature, for

example, the number of citations,  $h$ -index for authors [11], Journal Impact Factor (JIF) for journals [10], and Nature Index (NI) for research institutions. Many studies have predicted future  $h$ -index values [1, 3, 13, 24]. Acuna et al. calculated an equation for predicting the  $h$ -index. They showed that five main parameters are fundamentally crucial for prediction [1]: the number of publications, the current  $h$ -index value, the number of years since the first publication, the number of types of journals published to date, and the number of papers in top journals.



**Fig. 1.** Visualization results of the acquired distributed representation [17]. Color coding is the result of the K-means method.

There are some studies to predict the number of future citations of papers [4, 6, 23, 25]. Among them, Stegehuis et al. and Cao et al. predict the number of citations in the far future, considering the number of citations during 1–3 years after publication. In contrast, Sasaki et al. predict the number of citations after three years from publication directly [23]. The task evaluated in this study also predicts the number of citations five years after publication, just as Sasaki et al. did. Previous efforts to predict indicators have created various features and used them as input to the model.

There are attempts to predict the impact of scholarly literature more directly as a link prediction problem by creating a custom network. Yan et al. evaluated the impact of academic literature by creating a co-author network of countries, institutions, and authors and predicting their link relationships [28]. They showed that predicting author coauthorship was more difficult than predicting country or institution coauthorship. Park et al. created a citation network of patent information between the two fields and developed a model to predict future trends in the number of citations across fields [20]. They used it to predict increasing trends in linkages between the biotechnology field and the information technology field, showing that technological convergence is underway. Yi et al. constructed a bipartite graph, author, and keywords from the scholarly literature data [30]. With this, they developed a model to predict future changes in author interest. By evaluating the model as a link prediction problem between

authors and keywords, they show that it can predict future changes in each author’s interest based on past trends in authors’ keywords. Thus, the direct use of network information effectively predicts the influence of academic literature.

However, studies that used each data separately or combined the extracted features for classification or regression did not adequately consider the interactions among the data. It is also a challenge to make more active use of citation information rather than simply using it as teacher data, as in the SPECTER model. In particular, it is vital to building end2end models that fuse various academic literature data to build more general-purpose models. As a first step, this paper proposes an end2end fusion method of linguistic and citation information in academic literature data using the Transformer model.

### 3 Proposed Method

We built a model that can learn end2end by fusing linguistic and citation information among the diverse data possessed by the academic literature. However, is it necessary to fuse multiple pieces of information to predict the impact of academic literature? If a model can fully understand the text of a paper, is it sufficient to predict the impact of that paper? This section shows citation information may be more important than a paper’s content in predicting its impact. That is, we require a model that actively incorporates citation information. Hence, we propose a model that can be trained end2end by fusing linguistic and citation information.

#### 3.1 Linguistic or Citation Information?

Is it necessary to fuse language and citation information to predict the impact of academic literature? Is predicting the scholarly literature’s impact impossible if the model accurately understands the linguistic content? Several studies have provided rebuttal evidence to this question. Sasaki et al. constructed a linear model to predict the number of citations and reported that the features associated with the citation network are important [23]. Ochi et al. used a network embedding and a language model to examine how to place the top-cited papers in the embedding space [17]. The results are so impressive that we show them in Fig. 1.

In Fig. 1, the color coding indicates the result of clustering. The red plots sparsely shown with the titles of the papers are the top-cited papers. Comparing the visualization results of a language model (Sentence-BERT [22]) and a network embedding (SEAL [21]), we can observe that the top-cited papers are more concentrated in a network embedding model. The entropy of the top-cited papers is 2.900 for the Sentence-BERT model, while it is 1.742 for SEAL. In other words, the top-cited papers are more biased at the SEAL model than the Sentence-BERT model by 1.1 points in terms of the number of papers with the highest citations.

Thus, several studies have reported that, in some cases, citation information is more effective than linguistic information in predicting the impact of academic



literature. In other words, the model for predicting the impact of academic literature requires the active use of citation information.

### 3.2 Fusion Transformer Model of Linguistic and Citation Information

This study constructs a model to learn end2end by fusing linguistic and citation information from various academic literature data. Therefore, as shown in Fig. 2, we propose the method. The method uses a multilayer perceptron layer (MLP) to fuse the network and the Transformer model for language processing to learn future top-cited papers classification problems. We use Graph-BERT [31] as the Transformer for citation network information and Sci-BERT [5] as the Transformer for linguistic information. In the previous Sect. 3.1, we found a significant bias between the BERT model based on linguistic information and the Embedding model based on networks about the distribution of the papers with the highest number of future citations. Therefore, we considered that only one of the two types of information might be helpful for classification, so we used a parallel model for both rather than a multilayered model in which the output of SciBERT is input to Graph-BERT. Depending on the classification problem, we expect to affect the information via SciBERT is more critical when the linguistic information is valid, and the information via Graph-BERT is more important when the network information is valid. By fusing citation information, we can apply our model even when not all nodes in the network have language information.

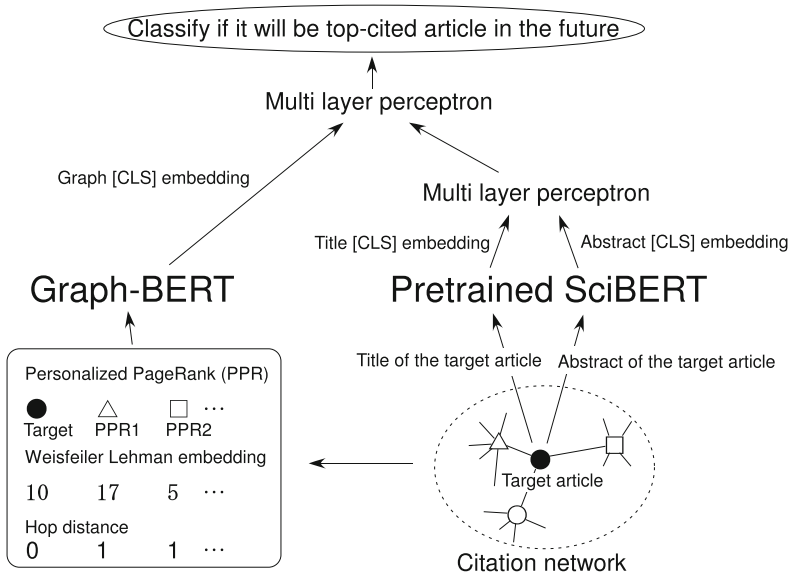


Fig. 2. Schematic diagram of proposed method

In Fig. 2, we first select a target paper. The target papers are randomly sampled nodes from the citation network. The proposed method learns and predicts a classification task to determine whether the target papers will likely be the top-cited papers in the future. First, we input three features into the Graph-BERT part. Personalized PageRank (PPR) [19], Weisfeiler-Lehman Embedding [16], and Hop Distance. Personalized PageRank is a personalized PageRank that computes the PageRank score customized for the target node for all nodes in the network. We order the nodes in decreasing order of PPR value, like a sequence of tokenized words in BERT. We compute Weisfeiler-Lehman Embeddings and input them as features for the aligned nodes. The Hop Distance is the shortest path length in the network from the target node and is input as a feature of the aligned nodes. Next, in Fig. 2, we input two pieces of information to Pre-trained SciBERT: the title and abstract of the target paper. We tokenized each and input them as a series of words, as in BERT.

We only use the [CLS] token, the classification token prefixed at the input of BERT, in the output of Graph-BERT and Pretrained SciBERT. This token allows for efficient training of the classification task. Finally, through the MLP layer, we combine the three [CLS] tokens to learn and predict the classification task of whether the target papers are probably the top-cited papers in the future.

## 4 Experiments

This section describes the experiments conducted to evaluate our proposed method. First, we describe the 67 small datasets of academic literature we have prepared for our experiments. Next, we train and evaluate our proposed model using a citation classification task. For this purpose, we describe the methods we compare and detail the learning and evaluation conditions.

### 4.1 Scientific Literature Dataset

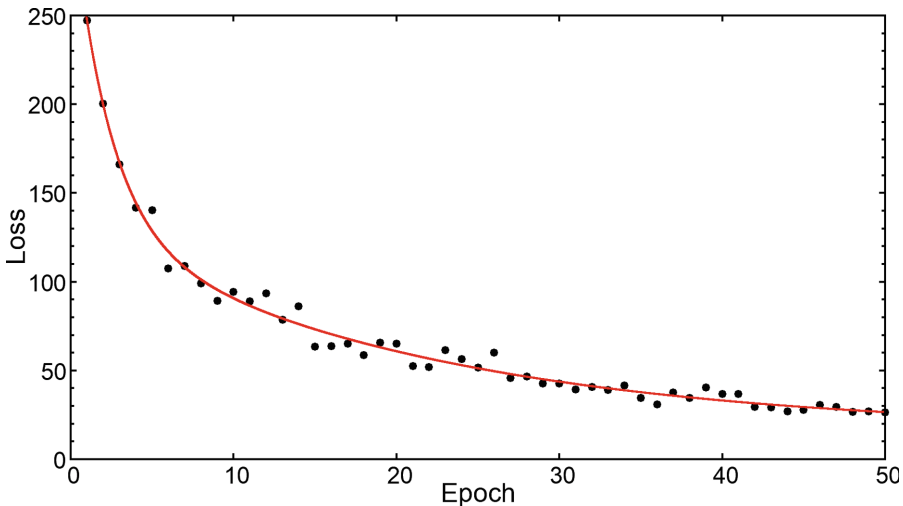
The data used are 67 small datasets extracted from the Web of Science<sup>1</sup>, which was one of the largest databases of academic literature. We search and extract each dataset from the Web of Science with specific queries. We carefully set up our queries to target relatively small research areas. We list in Tables 3, 4, 5 and 6 the queries we used to create the dataset in the Appendix A. We use the “Dataset ID” in tables, and this “Dataset ID” is common across all tables. All datasets were for articles published up to 2018. We present the features of each dataset we extracted in Table 7 in Appendix B. We targeted a relatively small research area and selected queries. As a result, we extracted datasets containing papers from 1,140 to 31,758. In the table, “Num. Articles” indicates the number of articles retrieved in Web of Science by the queries. Each dataset uses abstract information as linguistic information and citation information as network information. In the table, from “Num. Nodes” to “Gini coef. of Degree Dist.” represent the

<sup>1</sup> Web of Science <https://www.webofknowledge.com>.

characteristics of network information and from “Num. Abstracts” to “Word Perplexity” represent the characteristics of linguistic information. In particular, the “Num. Nodes” indicates the number of papers in the network, including papers appearing in the citation information. In contrast, the “Num. Abstracts” is small, indicating that abstract information does not exist in all nodes (papers) in the network.

## 4.2 Classification Problem Setup and Conditions

We consider positive cases as those papers published in 2013 from the dataset extracted in the previous section that is in the top 20% of citations after five years and negative cases as those that are not. We also randomly selected 70% of the papers in our dataset for training and the remainder for evaluation. To set the number of training epochs, we have observed reduced losses during training. The results are shown in Fig. 3. Although the loss decreases steadily, we selected 20 epochs because it is easy to overtrain if too many epochs are set. We also randomly sorted the data and ran the experiment 15 times for each dataset. We calculated Precision, Recall, and F-value as the classification results of the evaluated data in the trained model. We selected three comparison methods: Graph-BERT, SciBERT, and the proposed method. We chose only Graph-BERT, SciBERT, and the proposed method for comparison because Graph-BERT and SciBERT are elements of the proposed method, and the proposed method is a combined model of the two. We also used one MLP layer and softmax for the classification output. We used publicly available pretrained SciBERT models<sup>2</sup> and performed fine-tuning on each dataset.



**Fig. 3.** Loss reduction per epoch during training.

<sup>2</sup> SciBERT: <https://github.com/allenai/scibert>.

## 5 Results

The Precision, Recall, and F-value results of the classification averaging all datasets are shown in Table 1. “Graph-BERT”, “SciBERT” and “Proposed” in the table represent each method. The “**Bold Characters**” in the table represents the method with the best result for the evaluation index. First, we compare the results with the F-value. The F-value is the harmonic mean of the precision and recall, so the F-value measures the overall performance of the models. In the Table, the “Proposed” method shows 0.642, 0.614 for “Graph-BERT”, and 0.638 for “SciBERT” in F-value measurement. The higher the F-value, the better the result, so the proposed model showed the best results. Since this result is an average of all results, we performed a T-test on each result. The results show that the proposed model is statistically significantly superior at the  $p < 0.05$  significance level. In other words, the proposed method improves the classification results by 0.028 for Graph-BERT and 0.04 for SciBERT as the difference in F-values. This means our “Proposed” method shows the best result in the other two methods.

On the other hand, the results per dataset are diverse. We present our results per dataset in Table 8 in Appendix C. When we check the results for each dataset, we find that the proposed method performs the best of 19/67 datasets in the F-value measurement. The “Graph-BERT” method shows the best performance in 34/67 datasets. This indicates that Graph-BERT often shows better results than the proposed method. However, the proposed method performs better on average, meaning that Graph-BERT shows better results on some datasets and worse results on others. We focus on the standard deviation results in Table 1. In the Table, the “Graph-BERT” shows  $\pm 0.247$  standard deviation value in the F-value measurement, which is larger than the “Proposed” method standard deviation value  $\pm 0.120$ . The smaller the standard deviation of the results, the more stable the model training is. In other words, the proposed method shows the smallest standard deviation in F-values, indicating high learning stability.

These results show that the proposed model is the best average performance, indicating that the learning stability is high even when applied to diverse datasets.

**Table 1.** Classification Results Overall. \* $p < 0.05$

	Precision	Recall	F-value
Graph-BERT	<b>0.652*</b> $\pm 0.235$	0.618 $\pm 0.278$	0.614 $\pm 0.247$
SciBERT	0.636 $\pm 0.138$	0.685 $\pm 0.202$	0.638 $\pm 0.138$
Proposed	0.591 $\pm 0.131$	<b>0.749*</b> $\pm 0.182$	<b>0.642*</b> $\pm 0.120$

## 6 Discussion

In this section, we discuss the generalizability of the proposed model from two perspectives. The first is the stability of learning, and the second is a comparative analysis of the characteristics of datasets.

### 6.1 The Stability of Learning

Learning instability in deep learning has been noted in diverse literature. For example, there are papers on the gradient loss problem that occurs during fine-tuning in BERT [14] and papers on the over-smoothing [12] and over-squashing [2] problems in GNNs. Table 1 shows that the proposed method has a minor standard deviation of results. This slight standard deviation indicates that the learning is stable regardless of the data set and initial values. Table 1 also shows that Graph-BERT has the most significant standard deviation, meaning that learning is unstable. However, our proposed method improves this standard deviation by 0.127 in the F-value measurement. This improvement is clearly due to the pre-trained model SciBERT, which shows stable results because SciBERT is a pre-trained language model already trained from information in a large body of academic literature. Our model can improve accuracy and stability by utilizing citation information with the pre-trained language model.

### 6.2 Comparative Analysis

What characteristics of the datasets influence the difference in the models that show promising results for each dataset? To clarify this point, we calculated the correlation coefficients between the classification results of the F-values of each dataset (Table 8) and the features of the dataset (Table 7). We show the results in Table 2. The “Feature” column in the table indicates features. Values in the table in “**Bold Characters**” indicate the **weakest correlations** with absolute correlation coefficients in all methods.

The results show that of the eight features, the proposed method has the lowest correlation coefficients for six and the remaining two features. This indicates that the proposed method performs independently of the various characteristics of each dataset. In other words, the proposed method is highly versatile. On the other hand, the “Graph-BERT” method showed a significant correlation of 0.753 for the “Num. Edges” row. Other network-related features also showed strong correlations. This result indicates that Graph-BERT is sensitive to the network nature of citation relations in the dataset, and shows its usefulness in some datasets in this classification task. In fact, the Graph-BERT method performs well on 34/67 datasets, suggesting that the difference between successful and unsuccessful learning may be significant.

In this section, we discussed two aspects of the study: the stability of learning and the comparative analysis of the features of each dataset. We showed that the proposed method has high learning stability and is versatile, independent of each dataset’s features.

**Table 2.** The results of comparing the correlation coefficients between the F value of the classification result and each feature for each dataset.

Feature	Method		
	Graph-BERT	SciBERT	Proposed
Num. Articles	0.698	0.569	<b>0.514</b>
Num. Nodes	0.666	0.517	<b>0.504</b>
Num. Edges	0.753	0.623	<b>0.583</b>
Network Density (%)	0.517	0.373	<b>0.350</b>
Avg. Degree	0.342	0.361	<b>0.329</b>
Gini Coeff. of Degree Dist.	<b>0.417</b>	0.462	0.442
Num. Abstracts	0.708	0.643	<b>0.582</b>
Word Perplexity	0.429	<b>0.260</b>	0.320

## 7 Conclusion

In this paper, we proposed a model that uses the Transformer model to fuse linguistic and citation information from academic literature. The proposed model was trained and evaluated on 67 datasets extracted from the Web of Science and showed an average improvement in F-values of 0.028 over the Graph-BERT model alone and 0.04 over the SciBERT model alone. However, some results for individual datasets showed that the single model performed better, indicating that, in many cases, the proposed method tends to produce results comparable to those of the single model that performed better. In other words, the proposed method showed the best average performance and is stable with a small variance of the F-value. Comparative analysis of the results with the dataset’s features showed that the proposed method has the lowest correlation coefficient with the dataset’s features. This indicates that the proposed method produces good results regardless of the dataset’s characteristics. In other words, the proposed method is the most versatile. In any case, our proposed model improves the classification accuracy of the papers with the highest number of citations after five years. Therefore, the proposed model automatically selects when citation network information is valid and when linguistic information is valid. We conclude that we developed an end2end model that fuses linguistic features and a citation network of scholarly literature data.

## Limitation

Our proposed method has some limitations. The dataset applied in this study is relatively small. Also, since our model is a combination of the Transformer model, scalability is expected. We wanted to test the effectiveness of the proposed method on a larger dataset. Since the proposed model is an end2end model, we can quickly increase the number of tasks. We want to test the effectiveness of

the proposed method not only in the citation count classification but also for multiple tasks. We would also like to evaluate the integration of methods such as ViT [9] since there is information on figures and tables in the academic literature data.

The memory size of the GPU limits our method. Therefore, it is necessary to make adjustments such as devising keywords and dividing the data into disciplines small enough to be used by GraphBERT, or limiting the number of citations to papers that have been cited frequently in the past. Of course, if the data is made smaller, it will be sparse concerning the parameter space and may not yield appropriate learning results. However, this problem will be solved by the use of new GPUs.

**Acknowledgement.** This article is based on results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO) and supported by JSPS KAKENHI Grant Number JP21K17860 and JP21K12068. The funders had no role in the study design, data collection and analysis, decision to publish, or manuscript preparation.

## A Queries from Web of Science

We list in Tables 3, 4, 5 and 6 the queries we used to create the dataset from the Web of Science. We use a “Dataset ID” in tables, and this “Dataset ID” is common across all tables. We have multiple queries set up in each dataset. We registered articles matching any one of these queries in the dataset.

## B Network and Linguistic Features for Each Dataset

We present the features of each dataset we extracted in Table 7. We targeted a relatively small research area and selected queries. As a result, we extracted datasets containing papers from 1,140 to 31,758.

Two types of features are available: one for citation information and the other for language information. The feature on citation information is mainly a network-related indicator. These are the first six (Number of Articles, Number of Nodes, Number of Edges, Network Density, Average Degree, and Gini Coefficient of Degree Distribution). Next, the features related to linguistic information are mainly indicators related to language. These are the last two (Number of Abstracts, Word Perplexity). The Number of Articles represents the number of papers in each dataset. The Number of Nodes represents the number of nodes used in the network, including papers cited in the dataset. The Number of Edges represents the number of edges in the network. The Network Density

Table 3. Queries from Web of Science Dataset 1–20.

Dataset ID	Queries
01	2019-ncov,ace2,anxiety,coronavirus,coronavirus disease 2019,covid-19,epidemiology,hydroxychloroquine, infection,infection control,mental health,mers,mers-cov,mortality,outbreak,pandemic,pandemics,pneumonia, public health,sars,sars-cov,sars-cov-2,severe acute respiratory syndrome
02	acoustic analogy,acoustics,aeroacoustics,aerodynamic noise,beamforming,cavity,cavity flow,cfd, compressible flow,computational aeroacoustics,computational fluid dynamics,flow control, helmholtz resonator,jet,jet noise,large eddy simulation,les,muffler,noise,noise control,noise reduction
03	agriculture,australia,brazil,britain,canada,china,colonialism,economic growth,economic history,education, empire,england,enlightenment,france,gender,globalization,historiography,history,india,liberalism,migration
04	anisotropy,ballast,clay,constitutive model,constitutive relations,critical state,cyclic loading,deformation, dem,dilatancy,discrete element method,finite element method,granular material,granular materials, hypoplasticity,laboratory tests,liquefaction,numerical simulation,particle breakage,particle crushing, particle shape,plasticity,sand,sands,shear band,shear strength,soil mechanics,strain localization
05	adsorption,asp flooding,chemical flooding,contact angle,crude oil,enhanced oil recovery, enhanced oil recovery (eor),eor,foam,formation damage,heavy oil,interfacial tension,microemulsion, nanoparticles,numerical simulation,oil displacement efficiency,oil recovery,polyacrylamide,polymer, polymer flooding,porous media,rheology,surfactant,surfactant flooding,surfactants,viscoelasticity,viscosity
06	ahp,analytic hierarchy process,analytical hierarchy process,choquet integral,decision analysis, decision making,decision-making,evaluation,fuzzy ahp,fuzzy logic,fuzzy number,fuzzy numbers,fuzzy sets, fuzzy topsis,gis,goal programming,group decision making,group decision-making,intuitionistic fuzzy set, intuitionistic fuzzy sets,mcdm,multi-criteria decision making,multi-criteria decision-making
07	ab initio calculations,band structure,crystal structure,density functional theory,dft,elastic constants, elastic properties,electronic properties,electronic structure,equation of state,first principles, first-principles,first-principles calculation,first-principles calculations,fp-lapw,hardness, high pressure,mechanical properties,molecular dynamics,optical properties,phase diagram,phase transition
08	boundary layer,boundary layer flow,brownian motion,casson fluid,chemical reaction,entropy generation, free convection,heat and mass transfer,heat transfer,homotopy analysis method,joule heating, magnetic field,magnetohydrodynamics,mass transfer,mhd,mhd flow,micropolar fluid,mixed convection,nanofluid, nanofluids,natural convection,numerical solution,porous media,porous medium,radiation,stretching sheet
09	assessment,computerized adaptive testing,confirmatory factor analysis,differential item functioning, effect size,factor analysis,flynn effect,intelligence,iq,irt,item response theory,measurement, measurement invariance,meta-analysis,missing data,psychometrics,rasch model,reliability
10	austempering,austenite,bainite,dilatometry,dual phase steel,dual-phase steel,ebsd,ferrite,hardness, heat treatment,high strength steel,intercritical annealing,kinetics,martensite,mechanical properties, mechanical property,microstructure,phase transformation,phase transformations,precipitation
11	acoustic emission,anisotropy,brazilian test,confining pressure,constitutive model,crack,crack propagation, cyclic loading,damage,damage evolution,discrete element method,energy dissipation,failure mode,fracture, fracture mechanics,fracture toughness,granite,heterogeneity,mechanical properties,numerical simulation, rock,rock mechanics,sandstone,shpb,size effect,strain rate,strength,stress intensity factor,tensile strength
12	acute toxicity,analgesic,anti-inflammatory,anti-inflammatory activity,antibacterial, antibacterial activity,antidiabetic,antimicrobial,antimicrobial activity,antioxidant,antioxidant activity, cytotoxicity,diabetes,diabetes mellitus,ethnobotany,ethnomedicine,ethnopharmacology,flavonoids,inflammation, medicinal plant,medicinal plants,oxidative stress,phytochemical,phytochemicals,rats,streptozotocin,toxicity
13	accelerometer,attitude estimation,calibration,data association,data fusion,estimation, extended kalman filter,filtering,gps,inertial navigation,inertial navigation system,information fusion, initial alignment,integrated navigation,kalman filter,kalman filtering,maneuvering target tracking, multi-target tracking,navigation,nonlinear filtering,particle filter,particle filtering,sensor fusion,sins
14	algebra,assessment,cognitive load,cognitive load theory,comprehension,e-learning,education,feedback, fractions,instructional design,intelligent tutoring systems,learning,learning style,learning styles, mathematics,mathematics education,memory,metacognition,motivation,multimedia,multimedia learning, problem solving,proof,reading,reading comprehension,self-regulated learning,teaching,technology
15	amphetamine,associative learning,attention,choice,classical conditioning,concurrent schedules, conditioned taste aversion,conditioning,context,extinction,humans,key peck,latent inhibition,learning, lever press,morphine,multiple schedules,pavlovian conditioning,pigeon,pigeons,rat,rats,renewal,schizophrenia
16	atmospheric boundary layer,boundary layer,climatology,cloud microphysics,clouds,complex terrain,convection, convective storms,data assimilation,gpm,large-eddy simulation,model evaluation/performance,nowcasting, numerical weather prediction,numerical weather prediction/forecasting,precipitation,radar, radars/radar observations,rainfall,remote sensing,satellite,satellite observations,stable boundary layer
17	compressive strength,concrete,cracking,damage,ductility,durability,fatigue,fiber reinforced concrete, fiber-reinforced concrete,fibre reinforced concrete,flexural strength,fracture,fracture energy, fracture mechanics,fracture toughness,impact,mechanical properties,microstructure,numerical simulation, reactive powder concrete,reinforced concrete,self-compacting concrete,size effect,steel fiber
18	action,action observation,aging,attention,coordination,eeg,embodiment,fmri,grasping,human,imitation,kinematics, learning,mirror neurons,monkey,motor control,motor cortex,motor imagery,motor learning,movement,perception, proprioception,reaching,rehabilitation,stroke,touch,transcranial magnetic stimulation,virtual reality
19	air injection,bitumen,carbon dioxide,co,co flooding,co injection,combustion,crude oil,diffusion coefficient, enhanced oil recovery,eor,foamy oil,gas injection,heavy oil,heavy oil recovery,heavy oil reservoir, horizontal well,in situ combustion,in-situ combustion,interfacial tension,minimum miscibility pressure, numerical simulation,porous media,sagd,simulation,steam injection,thermal recovery,viscosity
20	congestion management,facts,facts devices,genetic algorithm,load shedding,optimal power flow,optimization, particle swarm optimization,phasor measurement unit,pmu,power flow,power system,power system stability, power system stabilizer,power systems,pss,reactive power,smart grid,stability,statcom,state estimation,svc,



**Table 4.** Queries from Web of Science Dataset 21–40.

Dataset ID	Queries
21	continental shelf,cretaceous,debris flow,delta,east china sea,faeces analysis,fluvial,geomorphology, grain size,holocene,late quaternary,miocene,pleistocene,provenance,quaternary,sea level,sediment, sediment transport,sedimentary facies,sedimentation,sedimentology,seismic stratigraphy,seismites, sequence stratigraphy,south china sea,stratigraphy,turbidite,turbidites,turbidity current
22	adaptation,amblyopia,attention,binocular vision,color,color vision,contrast,contrast sensitivity, depth perception,eye movements,human,monkey,motion,motion perception,optic flow,perception, perceptual learning,psychophysics,saccade,saccades,smooth pursuit,spatial frequency,spatial vision
23	boiling,boiling heat transfer,bubble,bubble dynamics,chf,condensation,correlation,critical heat flux, evaporation,flow boiling,flow pattern,heat exchanger,heat transfer,heat transfer coefficient, heat transfer enhancement,microchannel,microchannels,microgravity,nucleate boiling,numerical simulation, pool boiling,pressure drop,r134a,refrigerant,subcooled boiling,subcooled flow boiling,two-phase flow
24	16s rdna,16s rrna,16s rrna gene,actinobacteria,actinomycetes,alpha-proteobacteria,antarctica,archaea,bacillus, bacteria,bacteroidetes,chemotaxonomy,culturomics,flavobacteriaceae.geba,genome,marine bacteria,new species, new taxa,nov,novel species,phylogenetic analysis,phylogeny,polyphasic taxonomy,proteobacteria,seawater,soil
25	ads-cft correspondence,black hole,black hole thermodynamics,black holes,black holes in string theory, classical theories of gravity,dirac equation,exact solutions,extra dimensions,general relativity, hawking radiation,large extra dimensions,lovelock gravity,models of quantum gravity, nonlinear electrodynamics,phase transition,phase transitions,quasinormal modes,spacetime singularities
26	barium titanate,batio,ceramics,dielectric,dielectric constant,dielectric properties,dielectric property, electrical properties,ferroelectric,ferroelectric properties,ferroelectricity,ferroelectrics,lead-free, lead-free ceramics,lead-free piezoelectric ceramics,microstructure,morphotrophic phase boundary, perovskite,perovskites,phase transition,piezoelectric,piezoelectric ceramics,piezoelectric properties
27	amplitude ratios,diffusion,eigenvalue approach,elasticity,finite element method,fractional calculus, generalized thermoelasticity,gravity,initial stress,laplace transform,laplace transforms,magnetic field, magneto-thermoelasticity,micropolar,microstretch,normal mode analysis,phase velocity,rayleigh waves, reflection,relaxation time,rotation,thermal shock,thermoelastic,thermoelasticity,transversely isotropic
28	austenite,austenite stability,austenitic stainless steel,austenitic steel,cold rolling,ebsd,fatigue, grain size,hadfield steel,high manganese steel,martensite,martensitic transformation, mechanical properties,mechanical property,microstructure,nitrogen,phase transformation,retained austenite, shape memory effect,stacking fault energy,stainless steel,strain hardening,texture,trip,trip effect
29	a posteriori error estimate,a posteriori error estimates,a posteriori error estimation,adaptiveity, convergence,discontinuous galerkin,discontinuous galerkin method,discontinuous galerkin methods, domain decomposition,error estimate,error estimates,error estimation,finite element,finite element method, finite element methods,finite elements,finite volume method,fluid-structure interaction, immersed boundary method,incompressible flow,maxwell's equations,mixed finite element, mixed finite elements,navier-stokes,navier-stokes equations,optimal control,stability,stokes equations
30	cascading failure,cascading failures,centrality,clustering coefficient,complex network,complex networks, complex systems,degree distribution,econophysics,graph theory,network,network analysis,networks, opinion dynamics,phase transition,power law,preferential attachment,random graphs,resilience,robustness, scale-free,scale-free network,scale-free networks,simulation,social network,social network analysis
31	analytical solution,bearing capacity,composite foundation,dynamic response,finite element, finite element analysis,finite element method,foundations,interface,lateral load,liquefaction,model test, negative skin friction,numerical analysis,numerical simulation,offshore wind turbine,pile,pile foundation, pile foundations,pile group,pile groups,pile-soil interaction,piled raft,piles,sand,seismic response
32	1,25-dihydroxyvitamin d,25(oh)d,25-hydroxyvitamin d,breast cancer,calcitriol,calcium,cancer,children, cholecalciferol,deficiency,epidemiology,hypovitaminosis d,inflammation,insulin resistance,meta-analysis, multiple sclerosis,obesity,osteoporosis,parathyroid hormone,polymorphism,pregnancy,rickets,supplementation
33	australia,authenticity,china,climate change,cultural tourism,culture,destination,destination image, development,economic growth,ecotourism,heritage,heritage tourism,identity,malaysia,marketing,motivation, rural tourism,satisfaction,south africa,spain,sustainability,sustainable development,sustainable tourism
34	accretive operator,asymptotically nonexpansive mapping,banach space,banach spaces,cat(0) space, common fixed point,equilibrium problem,extragradient method,fixed point,fixed point problem,fixed points, hilbert space,hilbert spaces,iterative algorithm,maximal monotone operator,monotone mapping, nonexpansive mapping,nonexpansive mappings,projection,proximal point algorithm,split feasibility problem, strong convergence,variational inequalities,variational inequality,variational inequality problem
35	academic achievement,academic performance,achievement,achievement goals,adolescence,adolescents,anxiety, education,engagement,gender,higher education,intelligence,intrinsic motivation,learning,learning strategies, mathematics,motivation,personality,physical activity,physical education,self-concept,self-determination
36	brexit,china,climate change,constructivism,democracy,eu,eu law,europe,european commission, european integration,european union,europeization,foreign policy,germany,global governance,globalization, governance,human rights,international law,international organizations,international relations,legitimacy
37	bifeo,bismuth ferrite,ceramics,crystal structure,dielectric,dielectric constant,dielectric properties, electrical properties,ferroelectric,ferroelectricity,ferroelectrics,ferromagnetism,magnetic properties, magnetization,magnetolectric,magnetolectric coupling,magnetolectric effect,magnetostriction
38	beach,beach erosion,beach nourishment,breaking waves,climate change,coastal erosion,coastal management, erosion,gis,morphodynamics,numerical model,numerical modeling,remote sensing,sar,scatterometer, sediment transport,shoreline change,significant wave height,storm surge,surf zone,swan,turbulence,wave
39	abts,anti-inflammatory,antimicrobial,antimicrobial activity,antioxidant,antioxidant activity, antioxidant capacity,antioxidants,ascorbic acid,bioactive compounds,cytotoxicity,dpph,extraction,flavonoid, flavonoids,frap,free radicals,hplc,lipid peroxidation,oxidative stress,phenolic acids,phenolic compounds
40	china,climate,climate change,climate models,climate variability,cmp5,downscaling,drought,evapotranspiration, extreme events,floods,global warming,hydrology,mann-kendall test,precipitation,rainfall, regional climate model,rnoff,spi,statistical downscaling,streamflow,temperature,trend,trend analysis,trends,

Table 5. Queries from Web of Science Dataset 41–60.

Dataset ID	Queries
41	balanced truncation,control theory,controllability,decentralized control,descriptor systems, discrete-time systems,eigenstructure assignment,h control,iterative method,large-scale systems, linear systems,lyapunov equation,matrix equation,model order reduction,model reduction,nonlinear systems, optimal control,optimization,output feedback,riccati equation,robust control,robust stability,robustness
42	aging,attention,attentional blink,attentional capture,awareness,binocular rivalry,consciousness, decision making,egg,eye movement,eye movements,fmri,human,inhibition,memory,perception,prefrontal cortex, priming,reaction time,saccade,saccades,selective attention,spatial attention,vision,visual attention
43	arenavirus,bioterrorism,ebola,ebola virus,ebola virus disease,ebolavirus,emerging infectious diseases, epidemic,epidemiology,filovirus,glycoprotein,hemorrhagic fever,hendra virus,infectious diseases,jumin virus, lassa fever,lassa virus,marburg virus,nipah virus,one health,outbreak,public health,sierra leone
44	a. optical materials,crystal structure,d. luminescence,energy transfer,eu,europium,led,luminescence, luminescence properties,optical materials,optical properties,persistent luminescence,phosphor,phosphors, photoluminescence,rare earth,rare earths,red phosphor,sol-gel,thermal stability,thermoluminescence
45	accessibility,architecture,built environment,children,design,environment,gis,green infrastructure, green space,health,housing,landscape,mental health,nature,parks,perception,physical activity,place attachment, public space,quality of life,recreation,sense of place,space syntax,sustainability,urban design
46	accessibility,beijing,built environment,china,commuting,gis,hedonic model,high-speed rail,housing, housing market,housing price,housing prices,land use,mobility,planning,public transport,real estate, spatial analysis,sustainability,sustainable development,transit-oriented development,transport
47	calibration,climate change,dem,digital elevation model,evapotranspiration,flood,flood forecasting,floods,gis, hydrological model,hydrological modeling,hydrological modelling,hydrology,land use,land use change, modeling,modeling,rainfall,remote sensing,runoff,sensitivity analysis,streamflow,swat,swat model,uncertainty
48	boundary layer,cfd,channel flow,coherent structures,computational fluid dynamics, direct numerical simulation,dns,drag reduction,flow control,heat transfer,large eddy simulation, large-eddy simulation,les,numerical simulation,piv,rans,separation,turbulence,turbulence model, turbulence modeling,turbulence modelling,turbulence simulation,turbulent boundary layer
49	“parkinsons disease”,articulation,bilingualism,children,development,duration,dysarthria,english,french, intelligibility,intonation,japanese,language,language acquisition,language development,mismatch negativity, perception,phonetics,phonology,prosody,spanish,speech,speech perception,speech production
50	corrosion,corrosion fatigue,crack closure,crack growth,crack initiation,crack propagation,fatigue, fatigue crack growth,fatigue crack propagation,fatigue damage,fatigue life,fatigue life prediction, fatigue limit,fatigue strength,finite element analysis,finite element method,fractography, fracture mechanics,high cycle fatigue,life prediction,low cycle fatigue,microstructure,multiaxial fatigue, reliability,residual stress,s-n curve,stress intensity factor,variable amplitude loading
51	apatite,archean,china,crystal structure,epithermal,fluid inclusion,fluid inclusions,geochemistry,gold, gold deposit,granite,granitic pegmatite,hydrothermal alteration,iran,la-icp-ms,magnetite,mineral chemistry, mineralization,mineralogy,ore-forming fluid,pegmatite,pyrite,quartz,raman spectroscopy,rare earth elements
52	bayes estimation,bayes estimator,bayesian estimation,censored data,characterization,em algorithm,entropy, estimation,exponential distribution,failure rate,gamma distribution,generalized order statistics, hazard function,hazard rate,maximum likelihood,maximum likelihood estimation, maximum likelihood estimator,mean residual life,moments,monte carlo simulation,order statistics, pareto distribution,quantile function,record values,reliability,simulation,stochastic ordering
53	fano resonance,fano resonances,fdtd,finite element method,integrated optics,metamaterials,nanoantenna, nanoantennas,nanoparticles,nanophotonics,nonlinear optics,plasmon,plasmonic,plasmonics,plasmons, surface plasmon,surface plasmon polariton,surface plasmon polaritons,surface plasmon resonance
54	active distribution network,distributed generation,distributed generation (dg),distribution network, distribution networks,distribution system,distribution systems,genetic algorithm,microgrid, monte carlo simulation,network reconfiguration,optimization,particle swarm optimization,power quality, power system,power system reliability,power system restoration,reconfiguration,reliability
55	acheulean,atapuerca,biochronology,cut marks,early pleistocene,europe,france,holocene,iberian peninsula,italy, late pleistocene,lithic technology,magdalenian,mammalia,mammals,middle palaeolithic,middle pleistocene, moustertian,neanderthals,palaeoecology,palaeoenvironment,palaeolithic,paleoecology,pleistocene,pliocene
56	ann,arima,artificial neural network,artificial neural networks,data mining,deep learning,forecast, forecasting,genetic algorithm,global solar radiation,load forecasting,machine learning,neural network, neural networks,photovoltaic,prediction,renewable energy,short-term load forecasting,solar energy, solar radiation,support vector machine,time series,weibull distribution,wind energy,wind power
57	19th century,anatomy,aristotle,charles darwin,darwin,descartes,education,enlightenment,epistemology,ethics, eugenics,evolution,france,historiography,history,history of medicine,history of science,leibniz,mathematics, medicine,national socialism,natural history,newton,nineteenth century,psychiatry,psychology,religion,science
58	c-s-h,calcium aluminate cement,calcium silicate hydrate,calcium sulfoaluminate cement,cement, cement hydration,cement paste,clinker,compressive strength,concrete,durability,ettringite,fly ash,hydration, hydration (a),hydration products,kinetics,limestone,mechanical properties,microstructure,nanoindentation
59	bond,bond strength,carbonation,cathodic protection,cement,cement paste,chloride,chloride diffusion, chloride ion,chlorides,compressive strength,concrete,corrosion,corrosion rate,cracking,diffusion,durability, electrical resistivity,fly ash,microstructure,mortar,permeability,pore structure,porosity
60	asymmetric information,coordination,deteriorating items,deterioration,dynamic pricing,dynamic programming, eq,forecasting,game theory,genetic algorithm,heuristics,inventory,inventory control,inventory management, lot sizing,lot-sizing,optimization,pricing,production,production planning,revenue management,simulation, spare parts,stackelberg game,stochastic demand,supply chain,supply chain coordination,

**Table 6.** Queries from Web of Science Dataset 61–70.

Dataset ID	Queries
61	cogging torque,condition monitoring,efficiency,electric machines,electric vehicle,fault detection, fault diagnosis,fem,finite element analysis,finite element method,induction machine,induction motor, induction motors,modeling,optimization,permanent magnet,permanent magnet machines,permanent magnet motor
62	algae,anaerobic digestion,biodiesel,biofuel,biofuels,biogas,biomass,biorefinery,botryococcus braunii, carbon dioxide,chlorella,chlorella vulgaris,cyanobacteria,extraction,fatty acid,fatty acids,flocculation, growth,harvesting,lipid,lipid extraction,lipids,microalgae,nutrient removal,photobioreactor,photosynthesis
63	bank erosion,bed load,bedload,bedload transport,channel morphology,dams,erosion,flood,floodplain,floods, fluvial geomorphology,geomorphology,gis,holocene,human impact,hydraulic geometry,lidar,numerical simulation, remote sensing,riparian vegetation,river,river restoration,rivers,sediment,sediment budget
64	asymptotic behavior,bifurcation,critical exponent,critical groups,critical growth,critical point, critical point theory,critical sobolev exponent,elliptic system,elliptic systems,existence, fractional laplacian,ground state solution,morse index,morse theory,mountain pass theorem, multiple solutions,nehari manifold,p-laplacian,positive solution,positive solutions,resonance, schrödinger equation,semilinear elliptic equation,semilinear elliptic equations,uniqueness
65	drainage,drip irrigation,evaporation,evapotranspiration,groundwater,groundwater recharge, hydraulic conductivity,hydrology,infiltration,irrigation,modeling,preferential flow,recharge,runoff,salinity, saturated hydraulic conductivity,soil,soil moisture,soil water,soil water content,soils,solute transport
66	austenitic stainless steel,c. hydrogen embrittlement,carbon steel,cathodic protection,corrosion, corrosion fatigue,delayed fracture,diffusion,dislocation,fatigue,fractography,fracture,high strength steel, hydrogen,hydrogen diffusion,hydrogen embrittlement,hydrogen induced cracking,hydrogen permeation, hydrogen trapping,hydrogen-induced cracking,mechanical properties,microstructure,pipeline steel,sec
67	ant colony optimization,column generation,combinatorial optimization,dynamic programming, genetic algorithm,genetic algorithms,grasp,heuristic,heuristics,integer programming,local search,logistics, metaheuristic,metaheuristics,multi-objective optimization,optimization,routing,scheduling, simulated annealing,simulation,tabu search,time windows,transportation,traveling salesman problem,tsp,

represents the network density, where  $N$  is the number of nodes and  $E$  is the number of edges; the network density  $D$  calculates as  $D = \frac{2E}{N(N-1)}$ . The Average Degree represents the average degree, calculated as  $\frac{E}{N}$ , and the average number of edges per node. The Gini Coefficient of Degree Distribution represents the degree distribution’s Gini coefficient and indicates whether the edges are biased toward a particular node. The Number of Abstracts represents the number of abstracts, which is the number of abstract information included in the data set. The language model may not be fully utilized if this value is low. The Word Perplexity represents the perplexity of words, an indicator of lexical diversity. If the vocabulary in a dataset  $D$  is  $V$ , and the proportion of occurrences of a word  $w_i$  is denoted by  $P(w_i)$ , then the perplexity  $PP(D)$  is calculated by Equation  $PP(D) = \prod_{i=0}^V P(w_i)^{-P(w_i)}$ .

## C Result for Each Dataset

We present our results per dataset in Table 8. When we check the results for each dataset, we find that the proposed method only performs the best of 19/67 datasets in the F-value measurement. The “Graph-BERT” method shows the best performance in 34/67 datasets.

**Table 7.** Network and linguistic features for each dataset.

Dataset ID	Num. Articles	Num. Nodes	Num. Edges	Network Dens. (%)	Avg. Degree	Gini Coeff. of Degree Dist	Num. Abstracts	Word Perplexity
01	2,395	50,948	71,713	0.005526	2.815	0.616	1,657	752.286
02	2,297	42,117	52,704	0.005942	2.503	0.574	1,785	806.093
03	12,079	230,672	230,161	0.000865	1.996	0.492	3,455	1010.945
04	2,144	45,051	59,470	0.005860	2.640	0.593	1,695	1099.948
05	7,097	145,348	215,607	0.002041	2.967	0.624	5,493	1663.292
06	2,541	47,070	78,145	0.007054	3.320	0.661	2,455	637.565
07	4,912	112,399	183,794	0.002910	3.270	0.653	4,787	1197.106
08	5,621	70,680	172,923	0.006923	4.893	0.726	5,394	254.619
09	3,374	81,870	91,784	0.002739	2.242	0.536	2,542	2485.374
10	17,528	210,997	351,898	0.001581	3.336	0.648	13,230	1021.305
11	5,336	96,416	126,961	0.002732	2.634	0.586	4,427	1481.089
12	13,298	321,404	453,688	0.000878	2.823	0.613	10,373	2666.764
13	8,697	118,802	157,105	0.002226	2.645	0.584	7,846	970.382
14	13,926	286,351	344,278	0.000840	2.405	0.565	10,165	3425.753
15	21,226	358,035	613,214	0.000957	3.425	0.655	11,448	2370.453
16	1,902	36,636	60,180	0.008968	3.285	0.648	1,613	1178.198
17	20,017	309,201	436,950	0.000914	2.826	0.606	16,936	2088.358
18	8,314	171,619	264,224	0.001794	3.079	0.637	6,036	2109.605
19	26,105	521,474	650,068	0.000478	2.493	0.569	22,721	3979.790
20	5,871	80,734	112,392	0.003449	2.784	0.599	5,429	927.696
21	5,834	206,204	280,155	0.001318	2.717	0.609	4,215	3426.185
22	7,342	114,557	205,503	0.003132	3.588	0.672	4,238	1893.942
23	6,741	90,982	169,726	0.004101	3.731	0.684	6,113	766.738
24	4,067	98,658	160,213	0.003292	3.248	0.654	3,666	1891.281
25	1,140	25,211	50,084	0.015760	3.973	0.699	1,097	287.971
26	18,978	165,277	430,777	0.003154	5.213	0.715	17,190	914.252
27	2,184	29,776	45,203	0.010197	3.036	0.628	1,547	264.154
28	4,310	64,827	106,555	0.005071	3.287	0.642	4,021	671.604
29	11,022	159,570	266,151	0.002091	3.336	0.655	9,416	279.527
30	23,585	463,865	696,140	0.000647	3.001	0.638	20,602	4324.730
31	4,088	61,478	77,427	0.004097	2.519	0.573	3,330	1285.666
32	23,141	283,583	633,654	0.001576	4.469	0.711	14,331	1456.106
33	5,887	150,614	175,801	0.001550	2.334	0.556	4,797	1826.010
34	3,144	32,137	74,688	0.014464	4.648	0.720	3,009	173.619
35	16,553	388,511	524,262	0.000695	2.699	0.607	12,748	3006.321
36	31,758	628,486	682,262	0.000345	2.171	0.528	14,110	640.960
37	28,592	319,727	686,086	0.001342	4.292	0.686	26,990	1123.632
38	7,299	132,117	188,618	0.002161	2.855	0.613	5,222	2854.658
39	11,202	227,444	381,894	0.001476	3.358	0.658	9,654	1932.958
40	18,278	407,361	670,188	0.000808	3.290	0.661	14,984	3546.054
41	3,752	46,948	79,053	0.007173	3.368	0.641	3,436	286.443
42	8,856	160,656	329,190	0.002551	4.098	0.706	6,206	1378.180
43	5,338	85,439	163,025	0.004467	3.816	0.698	3,716	901.600
44	20,993	257,723	555,967	0.001674	4.314	0.697	18,716	762.037
45	16,796	359,276	467,947	0.000725	2.605	0.592	10,606	3032.806
46	6,020	123,734	141,664	0.001851	2.290	0.544	4,456	1831.336
47	6,298	154,960	226,555	0.001887	2.924	0.628	5,680	2598.111
48	10,602	180,006	274,883	0.001697	3.054	0.630	9,074	783.051
49	11,725	257,144	367,024	0.001110	2.855	0.622	7,105	1498.251
50	9,534	130,985	195,307	0.002277	2.982	0.621	7,889	994.092
51	5,123	146,753	250,203	0.002324	3.410	0.676	4,003	1427.195
52	2,801	50,818	58,442	0.004526	2.300	0.538	2,378	943.279
53	17,971	227,421	534,558	0.002067	4.701	0.722	16,452	869.963
54	4,661	67,817	96,925	0.004215	2.858	0.609	4,397	1048.016
55	3,565	140,363	184,053	0.001868	2.623	0.601	2,617	1491.436
56	27,426	453,321	639,433	0.000622	2.821	0.614	24,434	2984.775
57	7,768	172,239	169,844	0.001145	1.972	0.489	1,513	1715.224
58	12,346	194,977	289,775	0.001524	2.972	0.624	10,615	1264.390
59	20,509	300,367	442,577	0.000981	2.947	0.620	16,420	1956.335
60	3,309	54,124	75,852	0.005179	2.803	0.603	2,976	1268.338
61	3,877	49,798	62,969	0.005079	2.529	0.565	3,608	1023.036
62	14,120	281,405	476,360	0.001203	3.386	0.664	10,514	2366.466
63	8,705	236,930	323,512	0.001153	2.731	0.609	7,135	4412.041
64	2,717	29,989	53,384	0.011872	3.560	0.646	2,479	114.824
65	9,936	209,095	306,397	0.001402	2.931	0.623	7,797	3154.676
66	27,360	361,442	617,026	0.000945	3.414	0.654	22,808	1562.402
67	10,072	158,517	237,351	0.001889	2.995	0.630	9,489	1950.027

**Table 8.** Classification Results for each dataset.

Dataset ID	Graph-BERT			SciBERT			Proposed Method		
	Precision	Recall	F-value	Precision	Recall	F-value	Precision	Recall	F-value
01	<b>0.526</b> ±0.191	0.274±0.146	0.327±0.124	0.515±0.104	0.481±0.126	0.475±0.070	0.503±0.060	<b>0.489</b> ±0.124	<b>0.482</b> ±0.059
02	0.224±0.219	0.267±0.285	0.225±0.216	<b>0.644</b> ±0.124	0.775±0.096	<b>0.689</b> ±0.062	0.519±0.099	<b>0.783</b> ±0.072	0.617±0.067
03	0.143±0.156	0.210±0.334	0.136±0.261	<b>0.296</b> ±0.118	0.267±0.154	0.271±0.127	0.236±0.048	<b>0.390</b> ±0.187	<b>0.286</b> ±0.083
04	0.154±0.140	0.329±0.400	0.198±0.201	<b>0.387</b> ±0.058	0.738±0.067	<b>0.502</b> ±0.037	0.355±0.047	<b>0.843</b> ±0.047	0.497±0.041
05	<b>0.792</b> ±0.064	<b>0.776</b> ±0.222	<b>0.757</b> ±0.123	0.730±0.090	0.602±0.193	0.627±0.125	0.703±0.086	0.505±0.172	0.560±0.128
06	0.132±0.104	0.200±0.204	0.152±0.129	0.436±0.057	<b>0.846</b> ±0.116	<b>0.568</b> ±0.036	<b>0.639</b> ±0.101	0.779±0.068	0.553±0.073
07	0.611±0.040	0.397±0.109	0.470±0.088	<b>0.620</b> ±0.049	0.642±0.152	0.616±0.061	0.604±0.048	<b>0.765</b> ±0.141	<b>0.684</b> ±0.043
08	0.722±0.052	0.509±0.115	0.587±0.082	<b>0.741</b> ±0.077	0.640±0.152	0.668±0.078	0.607±0.056	<b>0.860</b> ±0.078	<b>0.707</b> ±0.027
09	<b>0.426</b> ±0.274	0.415±0.305	0.311±0.142	0.392±0.108	0.385±0.127	<b>0.358</b> ±0.067	0.254±0.118	<b>0.511</b> ±0.222	0.309±0.079
10	<b>0.792</b> ±0.064	<b>0.776</b> ±0.222	<b>0.757</b> ±0.123	0.730±0.090	0.602±0.193	0.627±0.125	0.703±0.086	0.505±0.172	0.560±0.128
11	0.577±0.087	0.361±0.080	0.434±0.066	<b>0.628</b> ±0.060	0.525±0.227	0.536±0.118	0.548±0.039	<b>0.731</b> ±0.123	<b>0.619</b> ±0.039
12	<b>0.827</b> ±0.050	<b>0.860</b> ±0.116	<b>0.835</b> ±0.056	0.713±0.047	0.705±0.096	0.702±0.031	0.746±0.057	0.698±0.122	0.711±0.067
13	<b>0.732</b> ±0.096	0.651±0.252	0.642±0.158	0.648±0.068	0.702±0.128	0.661±0.047	0.555±0.035	<b>0.876</b> ±0.046	<b>0.678</b> ±0.025
14	<b>0.793</b> ±0.043	0.830±0.077	<b>0.808</b> ±0.027	0.736±0.041	0.768±0.146	0.741±0.075	0.716±0.082	<b>0.844</b> ±0.093	0.766±0.035
15	<b>0.808</b> ±0.034	<b>0.937</b> ±0.035	<b>0.866</b> ±0.009	0.631±0.030	0.747±0.125	0.677±0.047	0.600±0.040	0.737±0.080	0.642±0.088
16	0.391±0.149	0.433±0.298	0.361±0.154	0.458±0.049	0.511±0.189	0.472±0.114	<b>0.478</b> ±0.066	<b>0.763</b> ±0.159	<b>0.576</b> ±0.051
17	<b>0.792</b> ±0.039	<b>0.887</b> ±0.085	<b>0.832</b> ±0.029	0.766±0.044	0.649±0.160	0.687±0.094	0.659±0.045	<b>0.725</b> ±0.139	<b>0.679</b> ±0.048
18	<b>0.757</b> ±0.063	<b>0.753</b> ±0.145	<b>0.741</b> ±0.068	0.743±0.057	0.727±0.090	0.729±0.040	0.668±0.050	0.696±0.104	0.674±0.043
19	<b>0.856</b> ±0.037	<b>0.811</b> ±0.087	<b>0.829</b> ±0.041	0.703±0.034	0.724±0.078	0.710±0.032	0.721±0.047	0.760±0.094	0.734±0.032
20	<b>0.642</b> ±0.206	0.413±0.206	0.452±0.158	0.617±0.102	0.422±0.143	0.478±0.092	0.593±0.068	<b>0.481</b> ±0.240	<b>0.493</b> ±0.142
21	<b>0.740</b> ±0.055	0.589±0.167	0.638±0.120	0.640±0.041	0.632±0.117	0.628±0.059	0.605±0.032	<b>0.743</b> ±0.152	<b>0.659</b> ±0.071
22	<b>0.787</b> ±0.033	<b>0.720</b> ±0.107	<b>0.746</b> ±0.054	0.685±0.050	0.616±0.135	0.635±0.071	0.668±0.042	0.565±0.140	0.600±0.075
23	<b>0.791</b> ±0.067	0.681±0.191	<b>0.707</b> ±0.109	0.763±0.072	0.630±0.168	0.668±0.080	0.651±0.043	<b>0.762</b> ±0.120	0.695±0.052
24	0.573±0.067	0.478±0.145	0.510±0.096	<b>0.658</b> ±0.060	0.536±0.111	0.581±0.067	0.583±0.049	<b>0.683</b> ±0.096	<b>0.624</b> ±0.043
25	<b>0.309</b> ±0.123	0.381±0.234	0.319±0.141	<b>0.550</b> ±0.117	0.376±0.185	<b>0.406</b> ±0.109	0.316±0.094	<b>0.414</b> ±0.163	0.391±0.074
26	<b>0.795</b> ±0.029	<b>0.817</b> ±0.102	<b>0.801</b> ±0.041	0.705±0.056	0.742±0.154	0.708±0.058	0.656±0.051	0.675±0.143	0.652±0.047
27	0.385±0.080	0.336±0.159	0.331±0.094	<b>0.458</b> ±0.046	0.706±0.214	0.539±0.082	0.454±0.045	<b>0.736</b> ±0.134	<b>0.554</b> ±0.040
28	<b>0.622</b> ±0.064	0.471±0.147	0.517±0.096	0.549±0.034	0.884±0.086	<b>0.673</b> ±0.020	0.504±0.011	<b>0.931</b> ±0.042	0.654±0.008
29	<b>0.830</b> ±0.062	0.799±0.114	<b>0.805</b> ±0.042	0.603±0.052	<b>0.803</b> ±0.148	0.676±0.044	0.570±0.032	0.802±0.164	0.654±0.056
30	<b>0.845</b> ±0.056	0.826±0.078	<b>0.830</b> ±0.021	0.678±0.062	0.807±0.096	0.758±0.020	0.701±0.064	<b>0.859</b> ±0.080	<b>0.766</b> ±0.019
31	0.463±0.040	0.379±0.097	0.411±0.064	<b>0.607</b> ±0.084	0.696±0.160	0.629±0.055	0.559±0.054	<b>0.808</b> ±0.088	<b>0.655</b> ±0.026
32	<b>0.814</b> ±0.047	<b>0.889</b> ±0.071	<b>0.846</b> ±0.015	0.739±0.032	0.750±0.064	0.741±0.023	0.684±0.061	<b>0.832</b> ±0.107	0.742±0.025
33	0.402±0.116	0.391±0.226	0.369±0.148	<b>0.622</b> ±0.100	0.835±0.078	<b>0.704</b> ±0.053	0.502±0.050	<b>0.968</b> ±0.034	0.659±0.039
34	<b>0.833</b> ±0.072	0.618±0.113	<b>0.699</b> ±0.072	0.627±0.061	0.724±0.107	0.664±0.036	0.554±0.034	<b>0.896</b> ±0.076	0.681±0.019
35	<b>0.802</b> ±0.031	0.821±0.092	<b>0.807</b> ±0.033	0.754±0.053	0.769±0.091	0.754±0.024	0.683±0.036	<b>0.873</b> ±0.059	0.764±0.014
36	<b>0.857</b> ±0.031	<b>0.855</b> ±0.091	<b>0.852</b> ±0.039	0.720±0.046	0.753±0.069	0.732±0.014	0.724±0.043	0.698±0.072	0.706±0.021
37	<b>0.792</b> ±0.020	<b>0.871</b> ±0.048	<b>0.829</b> ±0.014	0.749±0.047	0.794±0.114	0.763±0.038	0.744±0.050	0.680±0.121	0.701±0.049
38	<b>0.808</b> ±0.040	<b>0.826</b> ±0.159	<b>0.804</b> ±0.085	0.755±0.036	0.421±0.120	0.531±0.098	0.687±0.048	0.655±0.141	0.658±0.084
39	<b>0.810</b> ±0.057	0.612±0.141	0.682±0.084	0.739±0.074	<b>0.827</b> ±0.069	<b>0.774</b> ±0.020	0.659±0.074	0.817±0.103	0.720±0.028
40	0.792±0.043	0.717±0.132	0.742±0.064	<b>0.874</b> ±0.047	0.657±0.035	0.668±0.106	0.740±0.056	0.736±0.055	<b>0.835</b> ±0.090
41	0.388±0.200	0.277±0.183	0.309±0.175	<b>0.514</b> ±0.061	0.874±0.152	<b>0.634</b> ±0.040	0.465±0.015	<b>0.931</b> ±0.086	0.619±0.028
42	<b>0.777</b> ±0.035	0.781±0.153	<b>0.768</b> ±0.075	0.695±0.059	0.772±0.120	0.721±0.032	0.669±0.065	<b>0.806</b> ±0.094	0.723±0.023
43	<b>0.794</b> ±0.085	0.622±0.136	0.687±0.093	0.708±0.054	0.758±0.106	0.727±0.057	0.688±0.038	<b>0.840</b> ±0.064	<b>0.754</b> ±0.025
44	<b>0.771</b> ±0.031	0.820±0.103	<b>0.790</b> ±0.036	0.710±0.049	0.823±0.078	0.758±0.026	0.621±0.042	<b>0.836</b> ±0.092	0.708±0.021
45	<b>0.803</b> ±0.035	0.831±0.073	<b>0.813</b> ±0.029	0.689±0.060	0.792±0.084	0.731±0.022	0.640±0.025	<b>0.873</b> ±0.067	0.722±0.023
46	0.493±0.039	0.449±0.095	0.465±0.060	0.538±0.062	0.646±0.213	0.563±0.080	<b>0.561</b> ±0.100	<b>0.695</b> ±0.204	<b>0.595</b> ±0.068
47	<b>0.764</b> ±0.052	0.656±0.186	0.686±0.095	0.679±0.054	0.778±0.094	<b>0.719</b> ±0.026	0.580±0.028	<b>0.893</b> ±0.085	0.700±0.023
48	<b>0.788</b> ±0.027	0.696±0.115	<b>0.733</b> ±0.062	0.677±0.070	0.693±0.132	0.671±0.036	0.622±0.039	<b>0.810</b> ±0.082	0.699±0.026
49	<b>0.764</b> ±0.031	0.752±0.111	<b>0.752</b> ±0.041	0.629±0.030	<b>0.819</b> ±0.092	0.707±0.028	0.642±0.027	0.724±0.134	0.672±0.055
50	<b>0.778</b> ±0.029	0.771±0.154	<b>0.763</b> ±0.090	0.658±0.049	0.689±0.097	0.666±0.035	0.625±0.030	<b>0.781</b> ±0.106	0.689±0.037
51	<b>0.837</b> ±0.067	<b>0.732</b> ±0.160	<b>0.764</b> ±0.072	0.755±0.051	0.522±0.153	0.601±0.115	0.686±0.052	0.637±0.114	0.653±0.065
52	0.229±0.154	0.278±0.297	0.216±0.173	0.339±0.041	<b>0.529</b> ±0.132	<b>0.407</b> ±0.059	<b>0.368</b> ±0.061	0.478±0.134	0.403±0.053
53	<b>0.889</b> ±0.034	0.819±0.091	<b>0.848</b> ±0.044	0.657±0.044	<b>0.899</b> ±0.068	0.755±0.015	0.691±0.057	0.746±0.164	0.700±0.069
54	<b>0.521</b> ±0.098	0.516±0.184	0.500±0.129	0.482±0.040	0.580±0.240	0.499±0.102	0.474±0.034	<b>0.739</b> ±0.129	<b>0.572</b> ±0.045
55	0.545±0.174	0.497±0.266	0.499±0.211	<b>0.683</b> ±0.094	0.574±0.183	0.594±0.100	0.659±0.059	<b>0.811</b> ±0.153	<b>0.601</b> ±0.078
56	<b>0.836</b> ±0.023	0.932±0.047	<b>0.880</b> ±0.013	0.675±0.044	<b>0.950</b> ±0.031	0.788±0.022	0.662±0.057	0.920±0.040	0.767±0.027
57	0.212±0.108	0.262±0.254	0.212±0.134	0.295±0.162	0.243±0.203	0.236±0.137	<b>0.356</b> ±0.106	<b>0.562</b> ±0.267	<b>0.418</b> ±0.137
58	<b>0.757</b> ±0.033	0.849±0.113	<b>0.794</b> ±0.048	0.688±0.075	0.857±0.128	0.751±0.038	0.757±0.062	<b>0.921</b> ±0.078	<b>0.719</b> ±0.038
59	<b>0.815</b> ±0.021	0.799±0.141	<b>0.797</b> ±0.082	0.711±0.055	<b>0.834</b> ±0.116	0.759±0.034	0.736±0.040	0.680±0.162	0.691±0.093
60	0.411±0.171	0.378±0.228	0.369±0.185	<b>0.486</b> ±0.044	0.790±0.166	<b>0.590</b> ±0.052	0.417±0.019	<b>0.949</b> ±0.056	0.578±0.021
61	0.401±0.181	0.319±0.299	0.279±0.146	<b>0.434</b> ±0.046	0.815±0.153	<b>0.559</b> ±0.057	0.381±0.027	<b>0.896</b> ±0.116	0.533±0.041
62	<b>0.774</b> ±0.057	<b>0.831</b> ±0.076	<b>0.797</b> ±0.019	0.669±0.063	0.763±0.126	0.701±0.039	0.688±0.049	0.695±0.111	0.683±0.032
63	<b>0.817</b> ±0.048	<b>0.914</b> ±0.087	<b>0.858</b> ±0.036	0.709±0.061	0.723±0.127	0.704±0.049	0.681±0.057	0.759±0.131	0.707±0.038
64	0.548±0.084	0.433±0.158	0.456±0.096	<b>0.670</b> ±0.105	0.493±0.132	0.551±0.095	0.575±0.060	<b>0.826</b> ±0.145	<b>0.666</b> ±0.046
65	<b>0.771</b> ±0.034	<b>0.740</b> ±0.088	<b>0.751</b> ±0.031	0.667±0.050	0.572±0.167	0.596±0.087	0.624±0.039	0.686±0.173	0.631±0.073
66	<b>0.770</b> ±0.022	<b>0.858</b> ±0.052	<b>0.810</b> ±0.014	0.714±0.039	0.810±0.096	0.753±0.031	0.727±0.053	0.781±0.107	0.746±0.045
67	<b>0.820</b> ±0.029	0.749±0.104	0.777±0.050	0.782±0.036	0.823±0.077	<b>0.798</b> ±0.027	0.694±0.031	<b>0.894</b> ±0.030	0.781±0.013

## References

1. Acuna, D.E., Allesina, S., Kording, K.P.: Predicting scientific success. *Nature* **489**(7415), 201–202 (2012). <https://doi.org/10.1038/489201a>
2. Alon, U., Yahav, E.: On the bottleneck of graph neural networks and its practical implications. In: International Conference on Learning Representations (2021). <https://openreview.net/forum?id=i80OPhOCVH2>
3. Ayaz, S., Masood, N., Islam, M.A.: Predicting scientific impact based on *h*-index. *Scientometrics* **114**(3), 993–1010 (2017). <https://doi.org/10.1007/s11192-017-2618-1>
4. Bai, X., Zhang, F., Lee, I.: Predicting the citations of scholarly paper. *J. Informetrics* **13**(1), 407–418 (2019). <https://doi.org/10.1016/j.joi.2019.01.010>, <http://www.sciencedirect.com/science/article/pii/S1751157718301767>
5. Beltagy, I., Lo, K., Cohan, A.: SciBERT: pretrained language model for scientific text. In: EMNLP (2019)
6. Cao, X., Chen, Y., Liu, K.R.: A data analytic approach to quantifying scientific impact. *J. Informetrics* **10**(2), 471–484 (2016). <https://doi.org/10.1016/j.joi.2016.02.006>, <http://www.sciencedirect.com/science/article/pii/S1751157715301346>
7. Cohan, A., Feldman, S., Beltagy, I., Downey, D., Weld, D.S.: Specter: document-level representation learning using citation-informed transformers. In: ACL (2020)
8. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics (2019). <https://doi.org/10.18653/v1/n19-1423>
9. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, 3–7 May 2021. OpenReview.net (2021). <https://openreview.net/forum?id=YicbFdNTTy>
10. Garfield, E., Sher, I.H.: New factors in the evaluation of scientific literature through citation indexing. *Am. Doc.* **14**(3), 195–201 (1963). <https://doi.org/10.1002/asi.5090140304>, <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.5090140304>
11. Hirsch, J.E.: An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences* **102**(46), 16569–16572 (2005). <https://doi.org/10.1073/pnas.0507655102>, <https://www.pnas.org/content/102/46/16569>
12. Li, G., Müller, M., Thabet, A., Ghanem, B.: DeepGCNs: can GCNs go as deep as CNNs? (2019)
13. Miró, Ò., et al.: Analysis of h-index and other bibliometric markers of productivity and repercussion of a selected sample of worldwide emergency medicine researchers. *Emergency Med. J.* **34**(3), 175–181 (2017). <https://doi.org/10.1136/emered-2016-205893>, <https://emj.bmj.com/content/34/3/175>
14. Mosbach, M., Andriushchenko, M., Klakow, D.: On the stability of fine-tuning BERT: misconceptions, explanations, and strong baselines (2021)
15. Mousavi, S.M., Ellsworth, W.L., Zhu, W., Chuang, L.Y., Beroza, G.C.: Earthquake transformer-an attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nat. Commun.* **11**(1), 1–12 (2020)

16. Niepert, M., Ahmed, M., Kutzkov, K.: Learning convolutional neural networks for graphs. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 2014–2023. PMLR, New York, New York, USA, 20–22 June 2016. <https://proceedings.mlr.press/v48/niepert16.html>
17. Ochi, M., Shiro, M., Mori, J., Sakata, I.: Which is more helpful in finding scientific papers to be top-cited in the future: Content or citations? Case analysis in the field of solar cells 2009. In: Mayo, F.J.D., Marchiori, M., Filipe, J. (eds.) Proceedings of the 17th International Conference on Web Information Systems and Technologies, WEBIST 2021, 26–28 October 2021, pp. 360–364. SCITEPRESS (2021). <https://doi.org/10.5220/0010689100003058>
18. Ochi, M., Shiro, M., Mori, J., Sakata, I.: Classification of the top-cited literature by fusing linguistic and citation information with the transformer model. In: Decker, S., Mayo, F.J.D., Marchiori, M., Filipe, J. (eds.) Proceedings of the 18th International Conference on Web Information Systems and Technologies, WEBIST 2022, Valletta, Malta, 25–27 October 2022, pp. 286–293. SCITEPRESS (2022). <https://doi.org/10.5220/0011542200003318>
19. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. Technical report 1999-66, Stanford InfoLab, November 1999. <http://ilpubs.stanford.edu:8090/422/>, previous number = SIDL-WP-1999-0120
20. Park, I., Yoon, B.: Technological opportunity discovery for technological convergence based on the prediction of technology knowledge flow in a citation network. *J. Informetrics* **12**(4), 1199–1222 (2018). <https://doi.org/10.1016/j.joi.2018.09.007>, <https://www.sciencedirect.com/science/article/pii/S1751157718300907>
21. Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics (2015)
22. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, November 2019. <https://arxiv.org/abs/1908.10084>
23. Sasaki, H., Hara, T., Sakata, I.: Identifying emerging research related to solar cells field using a machine learning approach. *J. Sustain. Dev. Energy Water Environ. Syst.* **4**, 418–429 (2016). <https://doi.org/10.13044/j.sdewes.2016.04.0032>
24. Schreiber, M.: How relevant is the predictive power of the h-index? A case study of the time-dependent hirsch index. *J. Informetrics* **7**(2), 325–329 (2013). <https://doi.org/10.1016/j.joi.2013.01.001>, <http://www.sciencedirect.com/science/article/pii/S1751157713000035>
25. Stegehuis, C., Litvak, N., Waltman, L.: Predicting the long-term citation impact of recent publications. *J. Informetrics* **9** (2015). <https://doi.org/10.1016/j.joi.2015.06.005>
26. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017). <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>

27. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.: GLUE: a multi-task benchmark and analysis platform for natural language understanding. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pp. 353–355. Association for Computational Linguistics, Brussels, Belgium, November 2018. <https://doi.org/10.18653/v1/W18-5446>, <https://aclanthology.org/W18-5446>
28. Yan, E., Guns, R.: Predicting and recommending collaborations: an author-, institution-, and country-level analysis. *J. Informetrics* **8**(2), 295–309 (2014). <https://doi.org/10.1016/j.joi.2014.01.008>, <https://www.sciencedirect.com/science/article/pii/S1751157714000091>
29. Yan, G., Liang, S., Zhang, Y., Liu, F.: Fusing transformer model with temporal features for ECG heartbeat classification. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 898–905. IEEE (2019)
30. Yi, Z., Ximeng, W., Guangquan, Z., Jie, L.: Predicting the dynamics of scientific activities: a diffusion-based network analytic methodology. *Proceedings of the Association for Information Science and Technology*, vol. 55, no. 1, pp. 598–607 (2018). <https://doi.org/10.1002/pra2.2018.14505501065>, <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/pra2.2018.14505501065>
31. Zhang, J., Zhang, H., Xia, C., Sun, L.: Graph-BERT: only attention is needed for learning graph representations. *CoRR abs/2001.05140* (2020). <https://arxiv.org/abs/2001.05140>





# Soft Web Intelligence with the J-CO Framework

Paolo Fosci  and Giuseppe Psaila 

University of Bergamo - DIGIP, Viale Marconi 5, 24044 Dalmine, BG, Italy  
{paolo.fosci, giuseppe.psaila}@unibg.it

**Abstract.** In the last two decades a plethora of approaches have been proposed to perform Web Intelligence to discover useful knowledge over the World-Wide Web. However, variety and vastness of the Web are still making this task a hard challenge.

Nonetheless, the Web is evolving. An example is the advent of the *JSON* format as the practical standard for exchanging data over the Internet.

In our previous work, we proposed the concept of Soft Web Intelligence: it is a modern interpretation of Web Intelligence based on the current technological panorama, in which *JSON* data sets can be gathered and stored within *JSON* document stores and processed by means of Soft Computing so as to Soft Querying them. Soft Web Intelligence is enabled by the *J-CO* Framework, a software tool that is natively able to manage, soft-query and transform collections of *JSON* documents, located either in *NoSQL* repositories or over the Internet. The paper illustrates our vision by presenting a plausible case study based on a weekly-updated data set that reports COVID-19 cases in European Countries.

**Keywords:** Soft web intelligence · J-CO framework · Continued acquisition of *JSON* data sets from Web sources · Soft querying *JSON* data sets

## 1 Introduction

Web Intelligence [22] has been introduced two decades ago with the aim of developing tools for collecting and analyzing data from the World-Wide Web. Indeed, the original definition (proposed in [22]) said: “*Web Intelligence (WI) exploits Artificial Intelligence (AI) and advanced Information Technology (IT) on the Web and Internet*”.

Nonetheless, many works are still published on the topic, because the goal was (and still is) very ambitious and wide. Furthermore, the Web is evolving constantly on directions that often do not help the implementation of Web Intelligence tasks.

In spite of this general trend, a major event has occurred that could help the development of Web Intelligence solutions. This event is the advent of *JSON* (JavaScript Object Notation) as the practical standard for sharing data sets over the Internet, in place of XML. The consequent evolution and acceptance of “*JSON* document stores” (*NoSQL* databases that manage *JSON* data sets in a native way), has dramatically changed the technological panorama since [22]; thus, it is reasonable thinking about novel interpretations of the idea of Web Intelligence.

Fuzzy Logic and Soft Computing are undoubtedly considered methods and techniques for Artificial Intelligence. So, why not to use them for Web-Intelligence scopes,

for processing data sets represented as collections of *JSON* documents that are acquired from the Web and are stored within *JSON* document stores?

The idea came out while working on the *J-CO* Framework, a tool under development at University of Bergamo (Italy), specifically designed to perform complex integration and querying on *JSON* data sets; its query language, named *J-CO-QL<sup>+</sup>*, currently provides powerful soft-querying capabilities: soft conditions can be exploited, so as to express imprecise and vague selection conditions on *JSON* data sets.

Based on these considerations, a novel interpretation of Web Intelligence, named *Soft Web Intelligence*, was conceived in [11]. In this paper, we consolidate the vision and present a novel case study, by means of which we show how *Soft Web Intelligence* can be practically performed by means of a modern and stand-alone tool (the *J-CO* Framework) that opens the way to further developments for the future.

The paper is organized as follows. Section 2 provides the background from which the work has originated. Section 3 provides a consolidated vision of the concept of *Soft Web Intelligence*. Section 4 presents the current organization of the *J-CO* Framework, highlighting those components and capabilities that enable *Soft Web Intelligence*. Section 5 extensively presents a practical case study, in which we show how to conceive, set up and perform tasks in a scope of *Soft Web Intelligence*. Finally, Sect. 6 draws conclusions and future work.

## 2 Background

In this section, we present the background from which this work has originated.

The concept of Web Intelligence was introduced in [22]. Clearly, it originates from the concept of “Business Intelligence”, which encompasses methods and tools for reporting, making multi-dimensional analysis and, more in general, discovering knowledge from data possibly gathered in data warehouses and data marts [16]. In some sense, we can figure out that Web Intelligence is the evolution of Business Intelligence towards the World Wide Web.

The work [22] talks about Artificial Intelligence as the key tool for performing Web Intelligence, but what is Artificial Intelligence?

Even though many people think about Neural Networks and Deep Learning, when hearing the term Artificial Intelligence, also techniques for Data Mining are very popular techniques that belong to the area of Artificial Intelligence. Consequently, their application to Web Intelligence has been immediately recognized [12], together with the consciousness that peculiarities of the Web constituted (and still constitute) a hard challenge, if compared to their application on flat data. The work [28] provides an interesting application of Web Intelligence to Digital Libraries, being mainly focused on link analysis and Web-log analysis.

Fuzzy Logic and Soft Computing are considered as a side part of Artificial Intelligence. Indeed, Zadeh, the inventor of Fuzzy-Set Theory [23], provided his own interpretation for the concept of Web Intelligence in [24, 25]. Specifically, Zadeh thought about the concept of “WebIQ” or “WIQ”, as the evolution of the idea of “Machine IQ”: machines become more and more intelligent as far as their capability to answer vague or imprecise questions is concerned; Fuzzy-Set Theory provides the formal tools towards

WebIQ. Another remarkable interpretation of the concept of Web Intelligence is “Computational Web Intelligence” [26], i.e., the adoption of “Computational Intelligence” to Web Intelligence scopes. We also mention “Brain Informatics” [27], whose goal is to foster Web Intelligence through techniques that come out from studying human brain.

It is worth noticing that, at the best of our knowledge, in the literature there are very few papers that are explicitly focused on the adoption of fuzzy logic and soft computing in Web-Intelligence scopes. We found the paper [13], in which authors exploit soft computing in a group decision-making system to express preferences, while Web Intelligence is used to gather knowledge to be exploited to make decisions. Another related paper is [17], in which the authors propose to use FUZZYALGOL, a fuzzy procedural programming language [21], to perform soft querying on Web sources.

As far as data formats are concerned, in the two passed decades since [22], the major event has been the advent of *JSON* as *de-facto* standard format for information interchange over the Internet, in place of XML. As a direct consequence, in the area of *NoSQL* databases, “*JSON* document stores” such as *MongoDB* have been proposed and have become very popular. These consideration motivated the development of the *J-CO* Framework, i.e., developing a general-purpose tool for performing complex manipulation and querying of *JSON* data sets [2]. The original capabilities reported in [2, 19] have been extended with soft-querying capabilities [7–11]. Furthermore, the *J-CO* Framework has been exploited as an execution engine for a domain-specific language for soft querying spatial features in *GeoJSON* documents [6, 20].

Currently, the *J-CO* Framework is a unique tool in the panorama of the few proposals for exploiting Fuzzy-Set Theory in the world of *JSON* data management. Specifically, [1, 15] proposed *fMQL*, an extension of *MQL* (the *MongoDB* query language). The extension is based on *JSON* documents previously tagged with “fuzzy labels”. In contrast, [14] proposed an extension of the *MongoDB* data model to support fuzzy values at the level of single document field. Definitely, they propose a fuzzy *JSON* document store, which was implemented on top of *MongoDB*.

At the best of our knowledge, no other proposals are somehow related to the *J-CO* Framework.

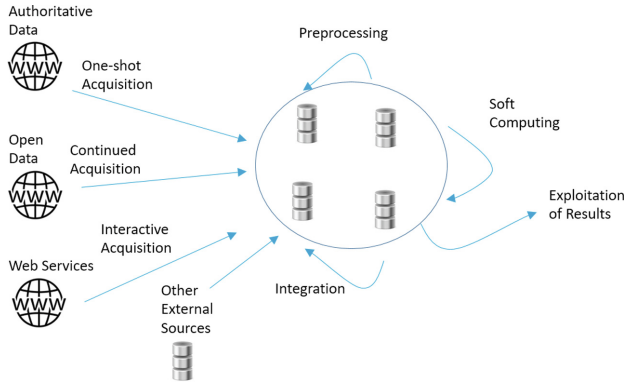
Consequently, we can say that our vision of *Soft Web Intelligence*, which will be introduced in Sect. 3, can be considered as a modern interpretation of the original idea of Web Intelligence, on the basis of the current scientific and technological panorama.

### 3 Vision: Soft Web Intelligence

Is there a uniform vision of Web Intelligence? Based on the literature (that we summarized in Sect. 2), the answer is clearly “no, there is not a uniform vision of *Web Intelligence*”. Probably this is due to the fact that different problems ask for different solutions and conduct to different visions.

In [11], we proposed the notion of *Soft Web Intelligence*, thinking about scopes in which it is necessary to collect *JSON* data sets from Web sources, so as to query them in a vague or imprecise way and obtain useful knowledge.

Our vision characterizes these scopes in a general way, regardless of the specific data sets to acquire and process, in which soft computing and soft querying can signif-



**Fig. 1.** Vision of Soft Web Intelligence.

icantly help analysts to find what they are looking for from web sources. The vision is depicted in Fig. 1; hereafter, we present it.

– **Web Sources.** In *Soft Web Intelligence*, data come from many Web sources. In order to understand how to deal with these sources, we provide a classification of them.

- **Authoritative Data.** Public Administrations and Authorities often publish data that can be considered “immutable” or “slowly mutable”. As an example, the reader can think about borders of municipalities, regions, and so on: they do not change or, if they change, this happens rarely and is due to the approval of acts by governments and parliaments.

Practically, this means that authoritative data sets can be downloaded only once and stored within databases. As far as their format is concerned, *JSON* is gaining more and more popularity. Considering again the case of administrative borders, they are often provided as *GeoJSON* data sets, which is a specific *JSON* format for representing geographical information layers [5].

- **Open-Data Portals.** Open-Data portals are a channel commonly used by public administrations to distribute data sets that concern the administered territory, so as to achieve the requirement of “transparency” (this is why they are said “open”).

On Open-Data portals, it is possible to find data sets that change continuously and it could be of interest for the analyst not to lose the previous versions; in such a case, it is necessary to perform a “continued acquisition” of these data sets, to accumulate all versions that change in time.

As far as the format of data is concerned, among the others *JSON* is becoming more and more adopted.

- **Web Services.** Often, useful data are provided by Web Services in a way that it is not possible to obtain them with one single call. For example, this is the case of “geo-coding”: provided latitude and longitude of a point on the Earth surface, a web service for geo-coding returns the address of the point; clearly, this kind of service asks for as many calls as the number of points to geo-code.

As far as the format of data provided by web services, we can say that substantially all of them provide data as *JSON* documents.

- **Data Storage.** In our vision, the *Data Storage Area* (the blue circle in Fig. 1), encompasses systems for storing data sets to integrate and process. Since we are considering scopes in which data sets are provided as *JSON* documents, *NoSQL* databases must be exploited; in particular, “*JSON* document stores” (or, simply, “*JSON* stores”) have imposed on the panorama of DBMSs (DataBase Management Systems), due to their ability to store and query *JSON* documents in a native way.
- **External Data.** Any kind of external data (i.e., not directly acquired from Web sources) can provide valuable information. Clearly, since we are considering a scope where data are represented as *JSON* data sets and stored in *JSON* stores, external data should be represented as *JSON* data sets as well.
- **Processing Tasks.** Many processing tasks could be figured out. In Fig. 1, we report the types that we consider most relevant (without excluding any other processing tasks, in principle).
  - **Pre-processing.** Once downloaded, a data set usually needs to be pre-processed, so as to clean it and change the format, to make it suitable for further processing. If data to be pre-processed are obtained through continued acquisition, their format is known and pre-processing could be automated.
  - **Integration.** Once data sets are collected, they could be integrated. Integration could be either an independent task or a part of pre-processing; this latter situation occurs when it must be performed on data sets obtained through continued acquisition, whose structure is known in advance.
  - **Soft Computing.** Once data sets to analyze have been built, it is possible to perform tasks of “soft computing”. Here, a plethora of techniques can be used in principle, such as Neural Networks; indeed, many techniques that fall into the area of Artificial Intelligence (AI) are also considered Soft-Computing techniques. A specific category of these techniques is devoted to “Soft Querying”: by relying on Fuzzy-Set Theory, they allow for specifying “Soft Conditions”, i.e., conditions whose evaluation does not provide a crisp “true” or “false” value, but a “satisfaction degree” in the  $[0, 1]$  range. This way, items that do not exactly match conditions can still emerge, possibly providing analysts with unexpected results. In this paper, we focus on soft querying, but the vision is open to any kind of soft-computing technique.

The next section presents the *J-CO* Framework, which currently provides the technical capabilities to build a scope of *Soft Web Intelligence*. Indeed, we conceived the concept of *Soft Web Intelligence* while introducing novel capabilities into the *J-CO* Framework based on the needs of our colleagues working on Geography (see [3,4]).

## 4 The J-CO Framework

In this section, we provide a synthetic presentation of the *J-CO* Framework, by highlighting the features that were specifically introduced in [11] to support the vision of *Soft Web Intelligence*.

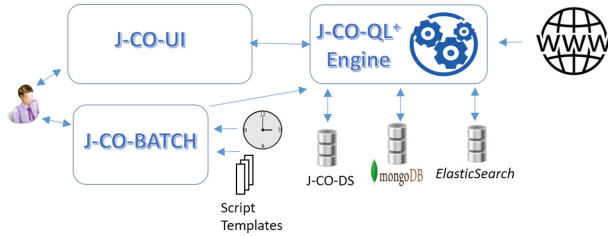


Fig. 2. Organization of the *J-CO* Framework.

#### 4.1 Organization of the Framework

The *J-CO* Framework is designed to provide data engineers and data analysts with a powerful support for gathering, integrating and querying possibly-large collections of *JSON* data sets. The framework is built around *J-CO-QL<sup>+</sup>*, a query language for specifying high-level transformations of collections of *JSON* documents in a declarative way. The language will be presented in action in Sect. 5.

Figure 2 depicts the tools that constitute the framework. We briefly introduce them hereafter.

- *J-CO-QL<sup>+</sup> Engine*. This component actually executes *J-CO-QL<sup>+</sup>* scripts (i.e., queries). It is able to retrieve data from external document databases (for example, managed by *MongoDB*) and save results into them; it also can send HTTP requests to web services and portals to get *JSON* data sets directly from Web sources.
- *J-CO-DS*. This component is a novel document store specifically designed to store large or very large single documents (such as many *GeoJSON* documents), so as to overcome limitations of other document databases (such as *MongoDB*). Section 4.2 is devoted to present this component.
- *J-CO-BATCH* is an off-line executor of *J-CO-QL<sup>+</sup>* scripts. It has been introduced in [11] to enable scheduled and/or parameterized execution of scripts. Section 4.3 is devoted to present this component.
- *J-CO-UI*. This is the user interface, by means of which analysts can write *J-CO-QL<sup>+</sup>* scripts, submit them to the *J-CO-QL<sup>+</sup> Engine* and inspect results.

In the remainder of this section, we present the *J-CO-DS* and the *J-CO-BATCH* components in details.

#### 4.2 *J-CO-DS*

As anticipated, *J-CO-DS* is a *JSON* store which is able to store very large documents [18]. It was specifically introduced in the *J-CO* Framework to overcome practical limitations provided by other *JSON* stores, such as *MongoDB*: indeed, *MongoDB* is not able to store large documents whose internal representation (named *BSON*) is greater than 16 MByte; furthermore, *MongoDB* adds an extra field named `_id` to documents, which alter *JSON* documents from the original ones.

*J-CO-DS* does not provide any computational capabilities, such as a query language, because it is an integral part of the *J-CO* Framework, so, queries can be performed by means of *J-CO-QL*<sup>+</sup> and its engine.

Although *J-CO-DS* organizes a database as a set of “collections” of *JSON* documents, as *MongoDB*, it provides three different types of collection.

**Static collections** are the classical collections, i.e., created and possibly updated either by the user or by a *J-CO-QL*<sup>+</sup> script.

**Dynamic collections** are thought for continued acquisition (they have been introduced in [11]); they are created specifying a pool of Web sources; the content of the collection is periodically updated with novel documents.

Let us define a dynamic collection in more details.

- A “dynamic-source descriptor” is a tuple

$$dsd = \langle name, URL, frequency, mode \rangle$$

in which many properties are defined. The *name* member is the name provided by the user to characterize the source. The *URL* member is the URL to contact to get documents. The *frequency* member is the frequency with which gathering must be repeated; admitted values can be expressed in hours (e.g. *2H*), days (e.g. *4D*) or weeks (e.g. *5W*). Finally, the *mode* member denotes the way to update the content of the collection; its value can be either ‘*overwrite*’, i.e., only the last image of the Web source is kept, or ‘*append*’, i.e., a new image of the Web source is appended to those already present in the collection.

- Thus, a dynamic collection *dc* is described by the tuple

$$dc = \langle name, sources \rangle$$

where *name* uniquely identifies the collection within the database, while *sources* is a non-empty array of dynamic-source descriptors.

Notice that different sources could be accessed with different frequencies, depending on the frequency they provide novel data. Furthermore, for each source a different update mode can be specified. To understand this choice, we briefly illustrate how the dynamic collection works.

- Data collected from each source are managed in a dedicated storage area, so as to be managed independently of data coming from the other sources.
- A Web source usually provides either a single *JSON* document, or a *JSON* array (usually an array of documents, but not necessarily).

For each single request sent to the Web source that returns data (either a single document or an array), a *JSON* document is created, with the following fields: *source* is the name of the source (the *name* field in the *dsd* source descriptor); *url* is the contacted URL; *timestamp* is the acquisition time; *data* actually contains the returned data (either a single *JSON* document or an array).

This new document is inserted into the data space of the web source.

- In the case the ‘‘`overwrite`’’ update mode is selected, the previous content of the data space is cleaned and the new document is added. This means that only one document at a time is present in the data space.  
In the case the ‘‘`append`’’ update mode is selected, the new document is added to the previous ones possibly present in the data space.

The rationale behind the two update modes is the following. The ‘‘`append`’’ mode is suitable when the goal is to keep all images of the Web Source. The usefulness of the ‘‘`overwrite`’’ mode is to decouple the  $J\text{-CO-QL}^+$  batches and scripts from the Web source and the intermediate network connection that provides the data to process: this way, data can be easily processed, independent of the availability of the Web source and of the Internet connection.

**Virtual collections** provide a unified abstraction to many Web sources (they have been introduced in [11]). They are defined by specifying a pool of URLs, which provide collections of *JSON* documents. When a user or a  $J\text{-CO-QL}^+$  script gets the virtual collection, the actual content is obtained by calling all associated services on-the-fly. In virtual collections, a source descriptor is a tuple

$$vsd = \langle name, URL \rangle$$

where only the *name* and *URL* members are present (it does not make sense to think about update mode or update frequency, because Web-source images are acquired on the fly).

A virtual collection is, in turn, defined by the tuple

$$vc = \langle name, sources \rangle$$

where *name* uniquely identifies the collection in the database, while *sources* is a non-empty array of virtual-source descriptors *vsd*.

Dynamic and Virtual collections can be altered only through the managing interface of  $J\text{-CO-DS}$ ; they cannot be updated by  $J\text{-CO-QL}^+$  scripts.

### 4.3 J-CO-BATCH

$J\text{-CO-BATCH}$  has been introduced in [11]. Its goal is to provide a ‘‘batch execution’’ of  $J\text{-CO-QL}^+$  scripts. Hereafter, we present the functionality it provides.

- **Template Execution.** A *Template* is a  $J\text{-CO-QL}^+$  script that is provided with ‘‘macro-parameter’s’’. Their lexical structure is `##parameterName##`, since  $J\text{-CO-QL}^+$  has no similar lexical elements.

In order to execute a template *t*, a context *C* must be provided; *C* is a key/value map, by means of which a macro-substitution of all macro-parameters  $mp \in t$ , with the corresponding value  $C(mp)$ , is performed; if all macro-parameters in *t* have a corresponding value in *C*, the *t* template is actually transformed into an executable  $J\text{-CO-QL}^+$  script.

Only constant values (either numbers or strings enclosed within single or double quotes), identifiers and path-expressions are valid values for macro-parameters in *C*.



- **Batch Execution.** Users can launch the execution of templates in batch mode, by providing the  $C$  context as a “property file”, as illustrated in Listing 5.
- **Scheduled Batch Execution.** Batch execution of template can be scheduled, either as a “one shot” execution, in which the template is executed at a given predefined instant, or as a “repeated” execution, to execute a template multiple times.

## 5 Soft Web Intelligence: A Practical Case Study

The ultimate goal of this paper is to show a practical case study that is in the scope of *Soft Web Intelligence*. By means of it, it will be clear our perspective.

### 5.1 Case Study

The recent COVID-19 pandemic provides a very interesting application of the idea of *Soft Web Intelligence*. Hereafter, we present the application context.

- The European Union (EU) has asked its Member States to weekly provide data about the number of official COVID-19 cases and the number of deaths. Currently (April 2023) these data are still communicated to the European Union.
- On the Open-Data portal of the European Centre for Disease Prevention and Control (ECDC<sup>1</sup>, a specific data set with the latest version of the data set is published, as a *JSON* array. This array contains all weekly updates for each country; consequently, it would be possible for an analyst to perform an historical study of the pandemic for European countries.
- Each item in this data set refers to a specific country by means of its country code. If it were necessary to add an extra piece of information, such as the name of the country capital, it would be necessary to look up for it on external web services that provide such a kind of information. *GeoNames*<sup>2</sup>, a geographical database accessible from the Internet, provides a pool of web services for this purpose.
- There are several formats for country codes. ISO<sup>3</sup>, the International Organization for Standardization, defines two standard formats, named *alpha-2* and *alpha-3*, which use, respectively, two and three letters to identify a country. Thus, an analyst must be aware that ECDC identifies countries by means of an ISO *alpha-3* code, while *GeoNames* uses an ISO *alpha-2* code.

We foresee that the analyst would like to address the problem presented hereafter.

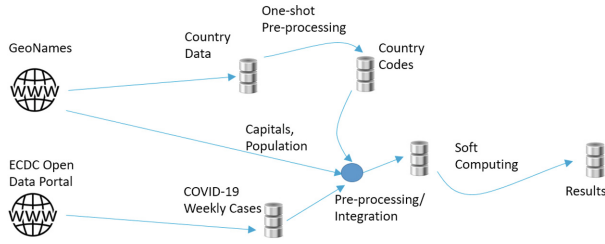
**Problem 1.** *Based on the data set providing the weekly counts for COVID-19 cases, look for those EU countries that, in the last week, could show an alert because the cases are too high in relation with the country population; for each of those countries, report also the name of its capital.*

Referring to the vision depicted in Fig. 1, Problem 1 is clearly a problem of *Soft Web Intelligence*, as illustrated in Fig. 3. It can be effectively managed by the *J-CO* Framework, as illustrated hereafter.

<sup>1</sup> ECDC web site: <https://www.ecdc.europa.eu/en>).

<sup>2</sup> GeoNames web site: <https://www.geonames.org/>.

<sup>3</sup> ISO, country code web site: <https://www.iso.org/iso-3166-country-codes.html>.



**Fig. 3.** Depicting the case study.

- A data set containing country data is provided by a *GeoNames* web service as a collection of *JSON* documents; this collection is stored in a *JSON* document store for later exploitation. However, in order to obtain a mapping between *ISO alpha-2* and *ISO alpha-3* country codes (and remove all unnecessary data) a one-shot pre-processing activity must be performed, so as to obtain only country codes. This task can be easily done by means of *J-CO-QL<sup>+</sup>*.
- Weekly updates of the data sets describing COVID-19 cases are made from the ECDC Open-Data portal. A dynamic collection managed by *J-CO-DS* is automatically filled in with the last image of the data set; this way, subsequent steps can be performed in an off-line manner, with respect to the Open-Data portal.
- The data set describing COVID-19 weekly cases can be automatically pre-processed so as to integrate it with country codes and with further data concerning each single country, such as the name of its capital and its population, if this latter information is missing. Again, a web service exposed by *GeoNames* provides this piece of information.
- Finally, the soft query is performed, by running a *J-CO-QL<sup>+</sup>* template through *J-CO-BATCH*. The final collection contains the data of interest and can be later accessed for knowledge exploitation and dissemination.

In the following Sections, we present each single step in details.

## 5.2 Defining a Dynamic Collection Gathering Data About COVID-19 Pandemic

In order to collect the latest version of the ECDC data about COVID-19 pandemic, a database managed by *J-CO-DS* is created. In this database, a dynamic collection (see Sect. 4.2) is created, so that it acquires the latest version of the data set locally.

Listing 1 shows the commands issued to *J-CO-DS* to create the database and the virtual collection.

- Line 1 in Listing 1 creates the database `softIntelligenceDb`.

```

{
  "country"           : "Luxembourg",
  "countryCodeIsoAlpha2" : "LU",
  "countryCodeIsoAlpha3" : "LUX"
}

```

**Fig. 4.** Example of document in the static `countryCodes` collection.

- Line 2 creates the dynamic collection, with name `euOfficialCovidData`; notice the `--url` parameter, which specifies the URL to contact to get the collection on the ECDC Open-Data portal.
- To complete the definition of the dynamic collection, Line 3 sets the update frequency for the dynamic collection to one week (denoted as `1W`), since the frequency default value is one day. Then, Line 4 sets the update mode of the dynamic collection as `type 1`, that means “overwrite” mode, i.e., the new image of the Web source replaces the previous one (the default update mode is `type 0`, i.e., “append” mode).

This way, *J-CO-DS* starts getting the last state of the data set each week. The reader can notice that this way it is possible to have a clear list of data sets that are acquired from external sources. As an effect, the World Wide Web is actually viewed as a giant data store.

The `softIntelligenceDb` database holds also a static collection, named `countryCodes`; a document in this collection reports, for each European country, both the ISO *alpha-2* and the ISO *alpha-3* codes. Figure 4 shows an example document related to Luxembourg. The `countryCodes` collection is obtained by a previous acquisition activity: these codes are made available by a GeoNames web service<sup>4</sup>, together with other information; a *J-CO-QL*<sup>+</sup> script, that we do not show for the sake of space, was executed to transform the data in the format shown in Fig. 4. Since country codes are static, they can be stored within a static collection, which is ready for further processing activities.

---

**Listing 1.** *J-CO-DS* commands: getting EU data.

---

```

1. create-database --name softIntelligenceDb
2. create-dynamic-collection --database softIntelligenceDb
   --collection euOfficialCovidData
   --url https://opendata.ecdc.europa.eu/covid19/nationalcasedeath/json/
3. set-frequency --database softIntelligenceDb --collection euOfficialCovidData
   --index 0 --frequency 1W
4. set-update-type --database softIntelligenceDb
   --collection euOfficialCovidData --index 0 --type 1

```

---

<sup>4</sup> GeoNames web service to get country codes:

<http://api.geonames.org/countryInfoJSON?formatted=true&username=paolofosci>.

### 5.3 Repeated Scheduled Pre-Processing

---

#### Listing 2. *J-CO-QL*<sup>+</sup> script: preprocessing EU COVID-19 data - Part 1.

---

```

1. USE DB softIntelligenceDb
   ON SERVER jcods 'http://127.0.0.1:17017';

2. GET COLLECTION euOfficialCovidData@softIntelligenceDb;

3. EXPAND
   UNPACK WITH .data
   ARRAY .data TO .countryData;

4. SAVE AS euCovidDataUnpacked;

5. GROUP
   PARTITION WITH .countryData.item.year_week,
                 .countryData.item.country
   BY .countryData.item.country, .countryData.item.indicator
   INTO .array
   GENERATE
   BUILD {
     .country : .countryData.item.country,
     .indicator : .countryData.item.indicator,
     .year_week : MAX_ARRAY (.array, STRING,
                             .countryData.item.year_week)
   };

6. JOIN OF COLLECTIONS temporary AS c1,
   euCovidDataUnpacked AS c2
   CASE
   WHERE
     .c1.country = .c2.countryData.item.country AND
     .c1.indicator = .c2.countryData.item.indicator AND
     .c1.year_week = .c2.countryData.item.year_week AND
     .c1.indicator = "cases"
   GENERATE
   BUILD {
     .country : .c1.country,
     .indicator : .c1.indicator,
     .yearWeek : .c1.year_week,
     .countryCode : .c2.countryData.item.country_code,
     .population : .c2.countryData.item.population,
     .cases : .c2.countryData.item.weekly_count
   };

```

---

In order to pre-process the data about COVID-19 pandemic, so as to extract and complete *JSON* documents concerning the last week, a *J-CO-QL*<sup>+</sup> script is run by *J-CO-BATCH*, which is the engine that executes *J-CO-QL*<sup>+</sup> scripts in a scheduled way (see Sect. 4.3). The script is reported in Listings 2 and 3, due to its length. Hereafter, we present it in details: this way, we also present the *J-CO-QL*<sup>+</sup> language in action.

*Selecting Data about the Last Week.* The first part of the script (reported in Listing 2) actually extracts, from the `euOfficialCovidData` dynamic collection in the `softIntelligenceDb` database, the *JSON* collection that actually describes COVID-19 cases in the last communicated week. A series of steps is necessary to achieve this (apparently simple) goal.

- The `USE DB` instruction on Line 1 actually connects the execution process to the database; this way, in the remainder of the script, the `softIntelligenceDb` database will be visible to instructions.

```

{
  "timestamp" : "2023-04-07T03:59:25.855",
  "url"       : "https://opendata.ecdc.europa.eu/covid19/.../",
  "data"      : [
    {
      "country"       : "Austria",
      "country_code"  : "AUT",
      "continent"     : "Europe",
      "population"    : 8978929,
      "indicator"     : "deaths",
      "year_week"     : "2020-01",
      "source"        : "TESSy COVID-19",
      "note"          : ""
    },
    ..., // other country documents
    {
      "country"       : "Luxembourg",
      "country_code"  : "LUX",
      "continent"     : "Europe",
      "population"    : 645397,
      "indicator"     : "cases",
      "weekly_count"  : 1428,
      "year_week"     : "2023-08",
      "rate_14_day"   : 358.2291,
      "cumulative_count" : 373263,
      "source"        : "TESSy COVID-19",
      "note"          : ""
    },
    ..., //other country documents
  ]
}

```

**Fig. 5.** Fragment of COVID-19 collection retrieved on line 2 in Listing 2.

```

{
  "timestamp" : "2023-04-07T03:59:25.855",
  "url"       : "https://opendata.ecdc.europa.eu/covid19/.../"
  "countryData" : {
    "position" : 6930,
    "item" : {
      "continent" : "Europe",
      "country"   : "Luxembourg",
      "country_code" : "LUX",
      "cumulative_count" : 290244,
      "indicator" : "cases",
      "note" : "",
      "population" : 645397,
      "rate_14_day" : 1745.2824,
      "source" : "TESSy COVID-19",
      "weekly_count" : 5981,
      "year_week" : "2022-25"
    }
  }
}

```

**Fig. 6.** Example of document in the temporary collection generated by Line 3 in Listing 2, extracted from the document in Fig. 5.

- The GET COLLECTION instruction on Line 2 actually acquires the content of the euOfficialCovidData dynamic collection in the softIntelligenceDb database.

Figure 5 depicts a fragment of this collection. Notice that dynamic collections in *J-CO-DS* contain one document for each monitored URL; since we associated only one URL and for it only the last version is kept (see Listing 1), only one single doc-

ument is stored and thus retrieved by the `GET COLLECTION` instruction. Specifically, notice that the actual collection acquired from the ECDC Open-Data portal is contained within the `data` array field. In Fig. 5 we report two documents, one concerned with Austria and one concerned with Luxembourg. Notice that the fields contained in the two documents are not exactly the same, i.e., documents can be heterogeneous.

The collection produced by the instruction becomes the new *temporary collection* of the process, ready to become the input to the next instruction (see [19]).

- From Line 3 a pool of instructions extracts the most recent *JSON* documents from the temporary collection generated by Line 2. The `EXPAND` instruction on Line 3 unnests all the *JSON* documents in the `data` field of the lonely document in the temporary collection depicted in Fig. 5; Fig. 6 depicts one unnested document, specifically the one depicted in Fig. 5 related to Luxembourg.

Specifically, in Fig. 6, notice that in place of the source `data` array, there is a `countryData` field that contains a complex sub-document with two fields: the `position` field reports the position formerly occupied by the document in the source array; the `item` field actually contains the unnested document. Outside the `data` field, all fields previously present in the source global document are reported.

- Line 4 temporarily saves the current temporary collection into the *Intermediate-Results Database* (or *IRDB*): this is a private database for the query process under execution, within which it is possible to save temporary results to be exploited later in the same query process. Although the temporary collection is saved into the database, it is not affected and it survives as temporary collection of the query process.

In Line 4, notice that the *IRDB* is implicitly referred, in that the collection name is not followed by `@<dbname>` (as in Line 2).

- The goal of Line 5 is to discover, for each country, the most recent week for which data are reported. In Fig. 6, within the `item` field, notice the `year_week` field; its value is a string with the format `yyyy-wk`, where `yyyy` denotes the year, while `wk` is the number of week in the year.

To obtain documents like the one reported in Fig. 7 for Luxembourg, the `GROUP` instruction on Line 5 behaves as follows. (i) The `PARTITION` clause selects all documents in the input temporary collection having the necessary fields. (ii) The `BY` clause specifies the “grouping fields”, specifically the country name and the indicator. For each group of documents having the same values for the grouping fields, a unique document is created, having all the grouping fields and an array field whose name is specified by the `INTO` clause (in this case, the name is `array`), containing all the input grouped documents.

The `GENERATE` clause specifies a kind of “post-processing” of the generated documents, so as to prepare them to be inserted into the output temporary collection. On Line 5, the `BUILD` block actually projects the documents on the grouping field and adds the `year_week` field, whose value is obtained by the built-in `MAX_ARRAY` function, which is used to extract the maximum week identifier from within the documents in the `array` field. An example of final document for Luxembourg related to the indicator named “cases” is depicted in Fig. 7.

- Finally (for Listing 2), the temporary collection generated by Line 5 is exploited to select only the documents of interests from within the collection temporarily saved in the *IRDB*, which contains all the documents unnested from the official data set reported in Fig. 5.

Specifically, the `JOIN OF COLLECTIONS` instruction on Line 6 joins documents in the current temporary collection (see Fig. 6) and in the novel collection named `euCovidDataUnpacked`, which is saved within the *IRDB* (see Fig. 7); the former is aliased as `c1`, while the latter is aliased as `c2`. All pairs of documents  $d_1 \in c1$  and  $d_2 \in c2$  are considered, building, for each  $(d_1, d_2)$  pair, a novel *od* document with two fields, named `c1` and `c2`;  $d_1$  is the value of the `c1` field, while  $d_2$  is the value of the `c2` field.

Each *od* document is processed by the `CASE` clause, in which the `WHERE` selection condition selects those documents that are actually of interest; specifically, the condition selects those *od* documents that are obtained by pairing a document in `c1`, denoting the last week identifier for a country and an indicator, with the corresponding document in `c2`. This way, only the most recent documents for each country are actually obtained as `c2` fields.

For all these *od* documents, the `GENERATE` clause actually restructures the output documents, so as to flatten them, as reported in Fig. 8.

```
{
  "country"      : "Luxembourg",
  "indicator"    : "cases",
  "year_week"   : "2023-13"
}
```

**Fig. 7.** Example of document after Line 5 in Listing 2.

```
{
  "cases"       : 1011,
  "country"     : "Luxembourg",
  "countryCode" : "LUX",
  "indicator"   : "cases",
  "population"  : 645397,
  "yearWeek"   : "2023-13"
}
```

**Fig. 8.** Example of document in the temporary collection generated by Line 6 in Listing 2.

*Completing Documents with External Data.* Listing 3 reports the second part of the scheduled *J-CO-QL*<sup>+</sup> script that pre-processes data about COVID-19 pandemic.

**Listing 3.** *J-CO-QL<sup>+</sup>* script: preprocessing EU COVID-19 data - Part 2.

```

7. JOIN OF COLLECTIONS temporary AS c1,
    countryCodes@softIntelligenceDb AS c2
CASE
  WHERE WITH .c1.countryCode AND
    .c1.countryCode = .c2.countryCodeIsoAlpha3
  GENERATE
    BUILD {
      .country           : .c1.country,
      .indicator         : .c1.indicator,
      .yearWeek         : .c1.yearWeek,
      .population        : .c1.population,
      .cases             : .c1.cases,
      .countryCode       : .c1.countryCode,
      .countryCodeIsoAlpha2 : .c2.countryCodeIsoAlpha2
    };

8. LOOKUP FROM WEB
FOR EACH WITH .countryCode
  CALL "http://api.geonames.org/countryInfoJSON"
    + "?formatted=true&lang=en"
    + "&country=" + .countryCodeIsoAlpha2
    + "&username=paolofosci&style=full";

9. EXPAND
UNPACK WITH .data.geonames
  ARRAY .data.geonames TO .geonames;

10. FILTER
CASE
  WHERE WITH .geonames, .source AND
    WITH .source.population
  GENERATE
    BUILD {
      .country           : .source.country,
      .indicator         : .source.indicator,
      .yearWeek         : .source.yearWeek,
      .population        : .source.population,
      .cases             : .source.cases,
      .countryCode       : .source.countryCode,
      .countryCodeIsoAlpha2 : .source.countryCodeIsoAlpha2,
      .capital           : .geonames.item.capital
    }
  WHERE WITH .geonames, .source AND
    WITHOUT .source.population
  GENERATE
    BUILD {
      .country           : .source.country,
      .indicator         : .source.indicator,
      .yearWeek         : .source.yearWeek,
      .population        : TO_INT(.geonames.item.population),
      .cases             : .source.cases,
      .countryCode       : .source.countryCode,
      .countryCodeIsoAlpha2 : .source.countryCodeIsoAlpha2,
      .capital           : .geonames.item.capital
    };

11. SAVE AS euMostRecentCovidData@softIntelligenceDb;

```

Specifically, the documents that were obtained by Line 6 in Listing 2 must be completed with the name of the capital of each country and, if missing, with the population. The script is presented hereafter.

```

{
  "cases"           : 1011,
  "country"         : "Luxembourg",
  "countryCode"     : "LUX",
  "countryCodeIsoAlpha2" : "LU",
  "indicator"       : "cases",
  "population"      : 645397,
  "yearWeek"        : "2023-13"
}

```

**Fig. 9.** Example of document in the temporary collection generated by Line 7 in Listing 3.



```

{
  "timestamp" : "2023-04-08T02:57:29.138",
  "url"       :... "http://api.geonames.org/countryInfoJSON?...&country=LU&..."
  "source"   : {
    "cases"      : 1011,
    "country"    : "Luxembourg",
    "countryCode" : "LUX",
    "countryCodeIsoAlpha2" : "LU",
    "indicator"  : "cases",
    "population" : 645397,
    "yearWeek"  : "2023-13"
  }
  "data"     : {
    "geonames" : [
      {
        "areaInSqKm"   : "2586.0",
        "capital"      : "Luxembourg",
        "continent"    : "EU",
        "continentName" : "Europe",
        "countryCode"  : "LU",
        "isoAlpha3"    : "LUX",
        "countryName"  : "Luxembourg",
        "currencyCode" : "EUR",
        "population"   : "607728",
        ... // other less interesting fields
      }
    ]
  }
}

```

**Fig. 10.** Example of document generated by Line 8 in Listing 3.

- The `JOIN OF COLLECTIONS` instruction on Line 7 actually extends documents with the most recent indicators for each country with the country codes in the ISO *alpha-2* and ISO *alpha-3* formats. Remember that they are described in the `countryCodes` collection already present in the `softIntelligenceDb` database (see Sect. 5.2 and Fig. 4). We do not explain the instruction in details; the resulting collection contains exactly the same document in the input temporary collection, but extended with the `countryCodeIsoAlpha2` field (ISO *alpha-2* code), as illustrated in Fig. 9, where we report the same document depicted in Fig. 8 extended with the country codes.
- In order to add the name of the capital and, if missing, the population of each country, it is necessary to contact an external web service. This is done by the `LOOKUP FROM WEB` instruction on Line 8.

This statement has been introduced in [11]: for each document in the input temporary collection that satisfies the `FOR EACH` clause, the URL composed in the `CALL` clause is called. Specifically, a web service provided by GeoNames is called, providing the value of the `countryCodeIsoAlpha2` field to the `country` parameter of the URL (which requires an ISO *alpha-2* country-code to select a country).

The corresponding output document is structured as illustrated in Fig. 10: the `timestamp` field denotes the time the web services was contacted, while the `url` field denotes the full URL; the `source` field contains the source document for which the call was done, while the `data` field contains the *JSON* documents received from the web service.

```

{
  "timestamp" : "2023-04-08T02:57:29.138",
  "url"       : "http://api.geonames.org/countryInfoJSON? ...&country=LU&...",
  "data"      : {},
  "source"   : {
    "cases"           : 1011,
    "country"         : "Luxembourg",
    "countryCode"     : "LUX",
    "countryCodeIsoAlpha2" : "LU",
    "indicator"       : "cases",
    "population"      : 645397,
    "yearWeek"        : "2023-13"
  },
  "geonames" : {
    "position" : 1
    "item"     : {
      "areaInSqKm" : "2586.0",
      "capital"    : "Luxembourg",
      "continent"  : "EU",
      "continentName" : "Europe",
      "countryCode" : "LU",
      "isoAlpha3"   : "LUX",
      "countryName" : "Luxembourg",
      "currencyCode" : "EUR",
      "population"  : "607728",
      ..., ... // other less interesting fields
    }
  }
}

```

**Fig. 11.** Example of document generated by Line 9 in Listing 3.

In the specific case, notice that this document contains a single `geonames` field, which is an array of one single document that describes the country (GeoNames returns always an array of documents).

- Consequently, it is necessary to flatten the documents: this is done in Line 9 by the `EXPAND` instruction, that unnests the only document contained in the `geonames` array. Figure 11 reports the document that is obtained from the document reported in Fig. 10.
- The process is completed by the `FILTER` instruction on Line 9. It contains two `WHERE` branches: the first one specifies what to do if the `population` field was present in the original document; the second branch specifies what to do if the `population` field was missing in the source document.

The final result is the same and, as illustrated in Fig. 12, documents have been enriched with the name of the capital and, if it was missing, with the population of the country.

- The final collection of the scheduled pre-processing script is saved by Line 10 into the `euMostRecentCovidData` collection within the `softIntelligenceDb` database, ready to be queried by analysts.

## 5.4 Template for Soft Querying Data

Without soft querying, it is not possible to talk about *Soft Web Intelligence*. Indeed, Problem 1 asks for exploiting soft querying capabilities, because through them it is possible to evaluate conditions in an imprecise way and rank documents.

```

    {
      "capital"           : "Luxembourg",
      "cases"            : 1011,
      "country"          : "Luxembourg",
      "countryCode"      : "LUX",
      "countryCodeIsoAlpha2" : "LU",
      "indicator"        : "cases",
      "population"       : 645397,
      "yearWeek"         : "2023-13"
    }

```

**Fig. 12.** Example of document in the collection generated by Listings 2 and 3.

Listing 4 reports the  $J\text{-CO-QL}^+$  template developed for this goal. It is executed by the  $J\text{-CO-BATCH}$  tool in the  $J\text{-CO}$  Framework. Since a template is parametric, analysts could use it with different values of its parameters, so as to cope with different needs while using the same  $J\text{-CO-QL}^+$  code. Hereafter, we present it in details.

---

**Listing 4.**  $J\text{-CO-QL}^+$  script: soft querying EU COVID-19 data.

---

```

1. CREATE FUZZY OPERATOR covidAlertFO
  PARAMETERS
    population TYPE INTEGER,
    cases      TYPE INTERGE
  PRECONDITION population > 0 AND
    cases > 0
  EVALUATE 1000 * 1000 * cases / population
  POLYLINE
    [ ( 0, 0.00), ( 30, 0.00), ( 100, 0.01),
      ( 200, 0.15), ( 750, 0.50), (1500, 0.80),
      (2500, 0.90), (4000, 1.00) ];

2. USE DB softIntelligenceDb
   ON SERVER jcodes 'http://127.0.0.1:17017';

3. GET COLLECTION euMostRecentCovidData@softIntelligenceDb;

4. FILTER
  CASE WHERE WITH .population, .cases
  GENERATE
    CHECK FOR FUZZY SET covidAlert
    USING covidAlertFO (.population, .cases)
    ALPHACUT ##alphaCutThreshold## ON covidAlert
    BUILD {
      .capital           : .capital,
      .cases             : .cases,
      .country           : .country,
      .indicator         : .indicator,
      .countryCode       : .countryCode,
      .countryCodeIsoAlpha2 : .countryCodeIsoAlpha2,
      .population        : .population,
      .yearWeek          : .yearWeek,
      .casesPerMillion   : TO_INT (1000*1000*.cases/.population),
      .covidRisk         : MEMBERSHIP_TO (covidAlert)
    }
  DEFUZZIFY;

5. SAVE AS euCountryCovidRisk@softIntelligenceDb;

```

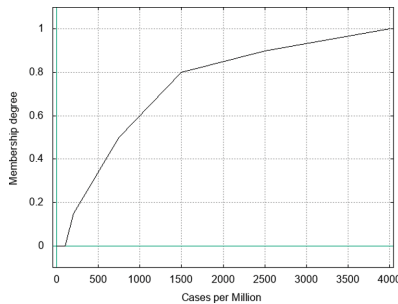
---

*Defining a Fuzzy Operator.* Line 1 of the script in Listing 4 creates a “fuzzy operator”, i.e., an operator that can be exploited to evaluate the degree of membership of a document to a fuzzy set [23].

Very shortly, a “fuzzy set”  $A$  in a universe  $U$  is a mapping  $A : U \rightarrow [0, 1]$ : given an item  $x \in U$ ,  $A(x) \in [0, 1]$  is the membership degree of  $x$  to  $A$ ; in other words, if  $A(x) = 1$ , then  $x$  fully belongs to  $A$ ; if  $A(x) = 0$ , then  $x$  does not belong at all to  $A$ ; if  $0 < A(x) < 1$ , then  $x$  partially belongs to  $A$  (the greater the value, the greater the way  $x$  belongs to  $A$ ).

For example, given a car  $a$ , we could see that it is in the set of *Fast\_Cars* if its maximum speed is greater than 200 km/h, i.e.,  $Fast\_Cars(a) = 1$ . But what happens if the maximum speed is 190 km/h? The membership degree could be  $Fast\_Cars(a) = 0.80$ , meaning that the car is not so fast.

The instruction on Line 1 of the script in Listing 4 creates a fuzzy operator named `covidAlertFO`. It receives two formal parameters named `population` and `cases` (as integer numbers), whose values must be both greater than 0 (as stated by the `PRECONDITION` clause). If the precondition is true, the `EVALUATE` clause computes the number of “cases per million”; the result is used as  $x$ -axis value against the membership function defined by the `POLYLINE` clause; the function is depicted in Fig. 13: if the  $x$ -axis value is less than 0 cases per million, the corresponding 0 value is returned as membership degree; if the  $x$ -axis value is greater than 4000 cases per million, the corresponding value 1 is returned as membership degree. Notice that the shape of the curve enhances the membership when cases per million are greater than 30.



**Fig. 13.** Membership function for the `covidAlertFO` fuzzy operator.

*Soft Querying.* The remaining lines in the template reported in Listing 4 actually perform the soft query, based on the values of the macro parameter named `##alphaCutThreshold##`, which allows for specifying a minimum threshold for membership degrees, as illustrated hereafter.

Line 2 in Listing 4 connects to the `softIntelligenceDb` database, from which Line 3 acquires the content of the `euMostRecentCovidData` collection produced by the scheduled pre-processing script illustrated in Sect. 5.3. Remember that the collection contains the documents describing the last communicated cases of COVID-19 by each European country (see Fig. 12).

Line 4 is the only instruction necessary to perform the soft query, provided that the fuzzy operator has been defined (on Line 1). Hereafter, we explain each single clause.

- The `WHERE` condition selects all documents having both the `population` and the `cases` fields; by construction, all documents should have them both.
- Within the `GENERATE` section, the `CHECK FOR FUZZY SET` clause evaluates the membership degree of each selected document to a fuzzy set named

`covidAlert`. The membership is evaluated by means of the “soft condition” specified after the `USING` keyword.

Specifically, the soft condition exploits the `covidAlertFO` fuzzy operator (defined on Line 1): the operator is called passing the values of the `population` and of the `cases` fields. The membership degree provided by the fuzzy operator becomes the membership degree of the document to the `covidAlert` fuzzy set. As an example of this intermediate state, Fig. 14 depicts the document that derives from the one presented in Fig. 12: after the evaluation of the `CHECK FOR FUZZY SET` clause. A special `~fuzzysets` field is now present, which contains the `covidAlert` field whose value is the membership degree of the document to the homonym fuzzy set. In the *J-CO-QL*<sup>+</sup> data model, the `~fuzzysets` field behaves as a key/value map, where the key is the name of a fuzzy set, while the value is the membership degree; this way, the membership degrees of the document to multiple fuzzy sets can be represented.

- The `ALPHACUT` clause specifies a minimum threshold for the membership to the `covidAlert` fuzzy set: only documents having a membership degree that is no less than the specified threshold are actually inserted into the output temporary collection.

In the template reported in Listing 4, the value is not specified: in its place, we find a “macro parameter” named `alphaCutThreshold` (reported in red color, for the sake of clarity). The value for the macro parameter is specified at execution time, as depicted in Listing 5 (a configuration file is submitted to *J-CO-BATCH* to provide macro parameters when a template is executed). In this case, the threshold of 0.5 is chosen, meaning that all documents having a membership degree to the `covidAlert` fuzzy set less than 0.5 are discarded.

---

**Listing 5.** Example of property file used for macro parameters in Listing 4.

---

```
alphaCutThreshold = 0.5
```

---

- The `BUILD` block of Line 4 in Listing 4 generates the final structure of documents that passes the `ALPHACUT` threshold. In particular, notice the `covidRisk` field, whose value is the membership degree to the `covidAlert` fuzzy set (it is provided by the `MEMBERSHIP_TO` built-in function). This field will denote the level of risk in the last (or more recent) week for the specified country.
- Finally, the `DEFUZZIFY` option “defuzzifies” the documents, i.e., it removes the `~fuzzysets` field from documents. As an example, consider the document depicted in Fig. 15, which is the final shape of the document reported in Fig. 14.

The template reported in Listing 4 ends with Line 5: the `SAVE AS` instruction saves the temporary collection generated by the instruction on Line 4 into the `euCountryCovidRisk` within the `softIntelligenceDb` database, for further exploitation by the analyst.

We can conclude that now the *J-CO* Framework can be actually exploited to perform Web Intelligence tasks directly working on *JSON* documents returned by Web sources. Furthermore, it is possible to exploit the support to soft querying provided by *J-CO-QL*<sup>+</sup>: this feature is unique in the panorama, in particular because it is provided by a stand-alone tool.

```

{
  "capital"           : "Lussemburgo",
  "cases"            : 1011,
  "casesPerMillion"  : 1566,
  "country"          : "Luxembourg",
  "countryCode"      : "LUX",
  "countryCodeIsoAlpha2" : "LU",
  "covidRisk"        : 0.806647778281193,
  "indicator"        : "cases",
  "population"       : 645397,
  "yearWeek"         : "2023-13",
  "~fuzzysets"       : {
    "covidAlert" : 0.806647778281193
  }
}

```

**Fig. 14.** Example of document generated after the CHECK FOR FUZZY SET clause of Line 4 in Listing 4.

```

{
  "capital"           : "Lussemburgo",
  "cases"            : 1011,
  "casesPerMillion"  : 1566,
  "country"          : "Luxembourg",
  "countryCode"      : "LUX",
  "countryCodeIsoAlpha2" : "LU",
  "covidRisk"        : 0.806647778281193,
  "indicator"        : "cases",
  "population"       : 645397,
  "yearWeek"         : "2023-13"
}

```

**Fig. 15.** Example of document generated by the template in Listing 4.

## 6 Conclusions

Moving from the initial idea, introduced over two decades ago to exploit data that can be found on the Word-Wide Web, to discover useful knowledge, in this paper we propose a modern interpretation of Web Intelligence, based on the current technological panorama in which data can be gathered from the Internet as *JSON* collections, possibly stored in *NoSQL* repositories and then processed by means of Soft Computing. This vision, that we named Soft Web Intelligence, has been enabled by the *J-CO* Framework, a software tool under development at the University of Bergamo (Italy), which is natively able to perform Soft Computing over data gathered from either different web sources or *JSON* document stores.

After defining our proposal for Soft Web Intelligence in details, and introducing the *J-CO* Framework, we presented a plausible case-study, to show that our vision is practically applicable. The case-study exploited real data about COVID-19 cases, made available by the Open-Data portal of ECDC, a European-Union Agency, and geographical information accessible through the web-services provided by *GeoNames*, a geographical database available on the Internet, to perform a Soft Web Intelligence activity in order to define real-time COVID-19 alert thresholds in European Countries.

The case-study demonstrates the feasibility of the concept of *Soft Web Intelligence*; nonetheless, this is due to the effectiveness of the *J-CO* Framework that, at the best of our knowledge, currently has no competitor platforms or tools that are able to enable (non-programmer) analysts to perform similar tasks.

While developing the *J-CO* Framework, we also carry on a continued research activity in order to improve performance and introduce new features. Currently, we are pursuing a line of evolution concerning the extension toward multi-grade fuzzy analysis, allowing users to define their own multi-grade fuzzy-set models and operators; furthermore, we are working on a general approach for defining soft aggregations.

## References

1. Abir, B.K., Amel, G.T.: Towards fuzzy querying of NOSQL document-oriented databases. *DBKDA* **2015**, 163 (2015)
2. Bordogna, G., Capelli, S., Ciriello, D.E., Psaila, G.: A cross-analysis framework for multi-source volunteered, crowdsourced, and authoritative geographic information: the case study of volunteered personal traces analysis against transport network data. *Geo-spat. Inf. Sci.* **21**(3), 257–271 (2018)
3. Burini, F., Cortesi, N., Gotti, K., Psaila, G.: The urban nexus approach for analyzing mobility in the smart city: towards the identification of city users networking. *Mob. Inf. Syst.* **2018**, 6294872 (2018)
4. Burini, F., Cortesi, N., Psaila, G.: From data to rhizomes: applying a geographical concept to understand the mobility of tourists from geo-located tweets. In: *Informatics*, vol. 8(1), p. 1. Multidisciplinary Digital Publishing Institute (2021)
5. Butler, H., Daly, M., Doyle, A., Gillies, S., Hagen, S., Schaub, T., et al.: The geojson format. Internet Engineering Task Force (IETF) (2016)
6. Fosci, P., Marrara, S., Psaila, G.: Geosoft: a language for soft querying features within geojson information layers. In: Marchiori, M., Domínguez Mayo, F.J., Filipe, J. (eds.) *Web Information Systems and Technologies. WEBIST WEBIST 2020 2021*. LNBP, vol. 469, pp. 196–219. Springer International Publishing, Cham (2023). [https://doi.org/10.1007/978-3-031-24197-0\\_11](https://doi.org/10.1007/978-3-031-24197-0_11)
7. Fosci, P., Psaila, G.: *J-CO*, a framework for fuzzy querying collections of *JSON* documents (Demo). In: Andreassen, T., De Tré, G., Kacprzyk, J., Legind Larsen, H., Bordogna, G., Zadrozny, S. (eds.) *FQAS 2021*. LNCS (LNAI), vol. 12871, pp. 142–153. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-86967-0\\_11](https://doi.org/10.1007/978-3-030-86967-0_11)
8. Fosci, P., Psaila, G.: Towards flexible retrieval, integration and analysis of Json data sets through fuzzy sets: a case study. *Information* **12**(7), 258 (2021)
9. Fosci, P., Psaila, G.: Intuitionistic fuzzy sets in J-CO-QL +? In: García Bringas, P., (et al). 17th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2022). LNNS, vol. 531, pp. 134–145. Springer, Cham. [https://doi.org/10.1007/978-3-031-18050-7\\_13](https://doi.org/10.1007/978-3-031-18050-7_13)
10. Fosci, P., Psaila, G.: Soft integration of geo-tagged data sets IN J-CO-QL+. *ISPRS Int. J. Geo Inf.* **11**(9), 484 (2022)
11. Fosci, P., Psaila, G., et al.: Towards soft web intelligence by collecting and processing json data sets from web sources. In: 18th International Conference on Web Information Systems and Technologies, pp. 302–313. No. 302, SCIPRESS (2022)
12. Han, J., Chang, K.C.: Data mining for web intelligence. *Computer* **35**(11), 64–70 (2002)
13. Kacprzyk, J., Zadrozny, S.: Soft computing and web intelligence for supporting consensus reaching. *Soft. Comput.* **14**(8), 833–846 (2010)

14. Medina, J.M., Blanco, I.J., Pons, O.: A fuzzy database engine for MongoDB. *Int. J. Intell. Syst. Online library* **37**, 5691–5764 (2022)
15. Mehrab, F., Harounabadi, A.: Apply uncertainty in document-oriented database (MongoDB) using F-xml. *J. Adv. Comput. Res.* **9**(3), 87–101 (2018)
16. Negash, S., Gray, P.: Business intelligence. In: Negash, S., Gray, P. (eds.) *Handbook On Decision Support Systems 2*, pp. 175–193. Springer, Heidelberg (2008). [https://doi.org/10.1007/978-3-540-48716-6\\_9](https://doi.org/10.1007/978-3-540-48716-6_9)
17. Poli, V.S.R.: Fuzzy data mining and web intelligence. In: *International Conference on Fuzzy Theory and Its Applications (iFUZZY)*, pp. 74–79. IEEE (2015)
18. Psaila, G., Fosci, P.: Toward an analyst-oriented polystore framework for processing JSON geo-data. In: *International Conference on Applied Computing 2018, Budapest; Hungary, 21–23 October 2018*, pp. 213–222. IADIS (2018)
19. Psaila, G., Fosci, P.: J-CO: a platform-independent framework for managing geo-referenced JSON data sets. *Electronics* **10**(5), 621 (2021)
20. Psaila, G., Marrara, S., Fosci, P.: Soft querying GeoJSON documents within the j-co framework. In: *WEBIST*, pp. 253–265 (2020)
21. Reddy, P.V.S.: FUZZYALGOL: fuzzy algorithmic language for designing fuzzy algorithms. *J. Comput. Sci. Eng.* **2**(2), 21–24 (2010)
22. Yao, Y.Y., Zhong, N., Liu, J., Ohsuga, S.: Web Intelligence (WI) research challenges and trends in the new information age. In: Zhong, N., Yao, Y., Liu, J., Ohsuga, S. (eds.) *WI 2001. LNCS (LNAI)*, vol. 2198, pp. 1–17. Springer, Heidelberg (2001). [https://doi.org/10.1007/3-540-45490-X\\_1](https://doi.org/10.1007/3-540-45490-X_1)
23. Zadeh, L.A.: Fuzzy sets. *Inf. Control* **8**(3), 338–353 (1965)
24. Zadeh, L.A.: A note on web intelligence, world knowledge and fuzzy logic. *Data Knowl. Eng.* **50**(3), 291–304 (2004)
25. Zadeh, L.A.: Web intelligence, world knowledge and fuzzy logic – the concept of web IQ (WIQ). In: Negoita, M.G., Howlett, R.J., Jain, L.C. (eds.) *KES 2004. LNCS (LNAI)*, vol. 3213, pp. 1–5. Springer, Heidelberg (2004). [https://doi.org/10.1007/978-3-540-30132-5\\_1](https://doi.org/10.1007/978-3-540-30132-5_1)
26. Zhang, Y.Q., Lin, T.Y.: Computational web intelligence (CWI): synergy of computational intelligence and web technology. In: *World Congress on Computational Intelligence*, vol. 2, pp. 1104–1107. IEEE (2002)
27. Zhong, N., et al.: Web intelligence meets brain informatics. In: Zhong, N., Liu, J., Yao, Y., Wu, J., Lu, S., Li, K. (eds.) *WImBI 2006. LNCS (LNAI)*, vol. 4845, pp. 1–31. Springer, Heidelberg (2007). [https://doi.org/10.1007/978-3-540-77028-2\\_1](https://doi.org/10.1007/978-3-540-77028-2_1)
28. Zuccala, A., Thelwall, M., Oppenheim, C., Dhiensa, R.: Web intelligence analyses of digital libraries: a case study of the national electronic library for health (NELH). *J. Doc.* **63**, 558–589 (2007)





# An NLP Approach to Understand the Top Ranked Higher Education Institutions' Social Media Communication Strategy

Alvaro Figueira<sup>1</sup> and Lirielly Nascimento<sup>2</sup>(✉)

<sup>1</sup> CRACS/INESCTEC and FCUP, University of Porto, Porto, Portugal  
arfiguei@fc.up.pt

<sup>2</sup> Department of Computer Science, FCUP, University of Porto, Porto, Portugal  
LVITORUGO@gmail.com

**Abstract.** In this paper we examine the use of social media as a marketing channel by Higher Education Institutions (HEI) and its impact on the institution's brand, attracting top professionals and students. HEIs are annually evaluated globally based on various success parameters to be published in rankings. Specifically, we analyze the Twitter publishing strategies of the selected HEIs, and we compare the results with their overall ranking positions. Our study shows that there are no significant differences between the well-known university rankings based on Kendall  $\tau$  and RBO metrics. However, our data retrieval indicates a tendency for the top-ranked universities to adopt specific strategies, which are further emphasized when analyzing emotions and topics. Conversely, some universities have less prominent strategies that do not align with their ranking positions. This study provides insights into the role of social media in the marketing strategies of HEIs and its impact on their global rankings.

**Keywords:** Higher education institutions · Social media communication · Twitter · Ranking analysis · Publishing strategies

## 1 Introduction

Over time, the number of available rankings has increased, enabling individuals to make more informed decisions. This trend is also observed in Higher Education Institutions (HEI), with university rankings becoming increasingly prevalent and diverse. The objective of creating these rankings is to measure and evaluate success in various areas or criteria. The metrics used to evaluate HEI are continuously improving, as are the methods to determine them more accurately. These institutions are commonly evaluated based on criteria such as student success, research volume, funding and awards, internationalization, employment, and connections to industry, among others.

Today, there are several leading indexes for HEI. Probably the best known and most widely used are the CWUR<sup>1</sup>, QS<sup>2</sup>, Leiden<sup>3</sup>, ARWU<sup>4</sup> (also known as the Shanghai

<sup>1</sup> <https://www.cwur.org/>.

<sup>2</sup> <https://www.topuniversities.com/>.

<sup>3</sup> <https://www.leidenranking.com/>.

<sup>4</sup> <https://www.shanghairanking.com/>.

ranking), and URAP<sup>5</sup>. It has been shown [16] that the correlation between these indices has been strong over the years. Therefore, despite some small variations in the indexes, explained by the difference in the evaluation criterion used for each of them, the overall picture given by one does not differ much from the others.

Several studies have investigated the comparison of university rankings and the challenges inherent in such comparisons, including works such as [2, 14] and [19]. In contrast, in this article we take a different approach by comparing HEI rankings with their Twitter posting strategies. Our goal is to analyze the extent to which the external communication of HEIs differs from one another, rather than discussing the ranking itself. The motivation for this analysis is the critical role that external communication plays in recruiting new students, distinguished researchers, and funding [9], particularly as the image projected by HEIs becomes increasingly important. Given that Twitter (and now Facebook and LinkedIn) is one of the most widely used networks in academia, we believe it is essential to review the performance and strategies of HEIs in this network. Ultimately, we seek to understand whether rankings reflect differences in how HEIs convey their messages.

Prior research has delved into the social media publications of HEIs, including works such as [5], and have developed methods for analyzing their postings [8], as well as inspecting the publication strategy in top-ranked HEIs [4]. In this study, we adopt a longitudinal perspective by analyzing and comparing a more extensive range of HEIs, rather than solely focusing on those ranked near the top. Our objective is to identify and compare variations in external communication among HEIs, as we vary their ranking position significantly.

This paper is structured as follows: In Sect. 2, we provide an overview of our analysis and the criteria for selecting a particular ranking and sampling HEIs. In Sect. 3, we conduct an analysis of the data collected. In Sect. 4, we use a vector space model to compare all HEIs and analyze the results. Finally, in Sect. 5, we summarize the research process and draw our final conclusions.

## 2 Data Retrieval

In this study, we selected four of the most commonly used ranking systems (CWUR, Shanghai, US News, and QS) to investigate potential small variations between them, despite recognizing the results from [16]. We utilized the Kendall distance and Kendall correlation coefficient (“Kendall’s  $\tau$ ”) metrics [6, 12] to compare the rankings. The Kendall distance is 0 for identical top-k lists and 1 for completely different ones. The Kendall  $\tau$  is a measure of correspondence between two rankings, with values close to 1 indicating strong agreement and values close to  $-1$  indicating strong disagreement. Another frequently used metric in comparing ranked lists is Rank Biased Overlap (“RBO”), where 1 means an identical ranking, and 0 means disjoint lists. The RBO is more robust in dealing with top weightiness [20].

Our goal was to determine if one ranking had significant variations compared to the others. The results for Kendall distance were zero for all combinations of ranking

---

<sup>5</sup> <https://urapcenter.org/>.

comparisons. For Kendall  $\tau$  (and RBO), the results were 0.64 (0.95) for CWUR vs. Shanghai, 0.63 (1.00) for CWUR vs. US News, and 0.47 (0.05) for CWUR vs. QS. Although there was less strong similarity between CWUR and QS, the general conclusion was that there were no significant variations in the rankings. Consequently, we only considered the CWUR ranking for further analysis.

We collected posts from HEIs ranked from 1 to 10 and in positions 100, 200, 300, 400, and 500. This wide range of rankings provides a perspective on top performing HEIs as well as any differences across a broad range of the ranking list. Table 1 shows the positions and their respective rankings across the four indexes. As can be seen, the ranking differences for the selected HEIs were not significant for the purposes of this paper.

**Table 1.** The position of the HEIs on the four rankings chosen. Updated from [7].

High Education Institution	CWUR	Shanghai	USNews	QS
Harvard University	1	1	1	5
Massachusetts Institute of Technology	2	3	2	1
Stanford University	3	2	3	3
University of Cambridge	4	4	8	3
University of Oxford	5	7	5	2
Princeton University	6	6	16	20
University of Chicago	7	10	15	10
Columbia University	8	8	6	19
University of Pennsylvania	9	15	13	13
California Institute of Technology	10	9	9	6
Boston University	99	101–150	65	112
University of Lisbon	200	201–300	197	356
University at Buffalo	300	301–400	280	338
University of Porto	308	201–300	255	295
University of Oklahoma, Norman	400	501–600	425	651–700
Federal University of Minas Gerais	500	401–500	256	651–700

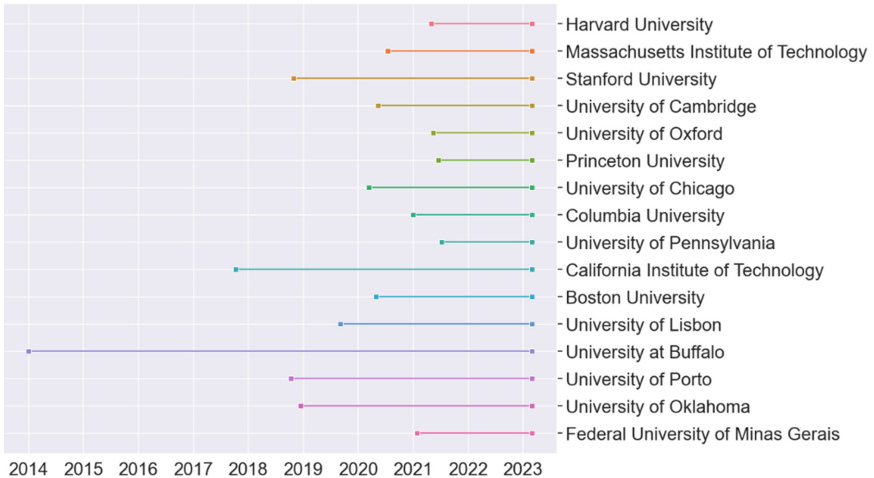
We had to make some changes to the list of HEIs used for our analysis. Instead of position 100, we chose position 99, as Keio University (position 100) had stopped tweeting after April 2020. We also included the University of Porto, out of curiosity, as it is the university of the authors.

To retrieve the most recent 2500 tweets for each HEI, we built an in-house tweet collector and set the last possible post at 31 July 2022. Tweets were extracted on two occasions - on the 5th and 17th of August 2022. However, the Twitter API did not return all 2500 tweets for the University of Lisbon (only 1583) and the University of Buffalo (only 1235). We excluded any retweets during this retrieval process as these two HEIs

had not yet posted 2500 tweets. To increase our sample size, we collected the recent tweets published after the initial period of analysis until February 2023.

Due to different posting frequencies among HEIs, the time span for the retrieved tweets varies for each institution. In Fig. 1, we depict the common period for tweet posts across all HEIs. As shown in the figure, the biggest common period is between July 2021 and February 2023.

In the next section, we inspect the retrieved data and perform a more detailed analysis of publishing time and content.

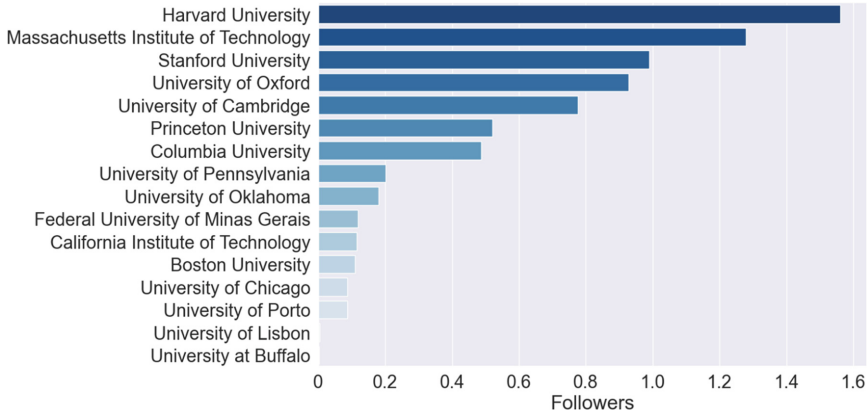


**Fig. 1.** Data collected period for each HEI. Updated from [7].

### 3 Data Analysis

Previous studies have explored the analysis of social media publications in HEIs, including works such as [8], and have used machine learning methods to analyze the publication strategy in top ranked HEIs [4]. Our approach differs in that we take a longitudinal perspective and analyze a larger set of HEIs, not solely those ranked at the top, as we anticipate observing changes as we move further down the ranking list. We begin our analysis by examining the number of followers for each HEI using Fig. 2 Number of followers as of February 2023.

Looking at Fig. 2 below, we can see that Harvard has the greatest number of followers with more than 1.56 million, followed by MIT with more than 1.27 million, Stanford and Oxford with more than 928K, Cambridge with more than 776K, Princeton and Columbia with 520K and 486K respectively, than Pennsylvania and Oklahoma with more than 201K and 179K respectively, Federal University of Minas Gerais, California Institute of Technology and Boston, each one with more than 119K, 114K, 109K respectively, Porto and Chicago in the sequence with more than 86K, and with less than 11K are Lisbon



**Fig. 2.** Number of followers, in millions, as of February 2023. Updated from [7].

and Buffalo, in this sequence. Table 2, below, depicts the mean and maximum number of posts for the daily tweet frequency for all the High Education Institutions.

We observed that the California Institute of Technology has the smallest standard deviation in posting frequency, while Lisbon has the highest. This suggests that the posting strategy at the California Institute of Technology is more consolidated, with approximately two posts per day. Massachusetts Institute of Technology, Boston, Chicago, Oklahoma-Norman, Stanford, Porto, Buffalo, and California Institute of Technology all publish between two to three posts per day. We noted the high number of posts (355) from the University of Lisbon on September 25th, 2021, which was due to the university's response to new students being approved.

Figure 3 shows the box-plot graph for each university's daily tweet frequency. It is apparent that the University of Pennsylvania's daily tweet frequency has a normal distribution with a median of around eight tweets per day and no outliers. Similarly, Massachusetts Institute of Technology and Harvard University show normal distribution with only two and three outliers respectively, all above the upper limit. There is a common pattern among Chicago, Oklahoma, Federal University of Minas Gerais, Porto, Boston, Princeton, and the California Institute of Technology, in which there is almost a normal distribution with some outliers causing the shape to be slightly squeezed (Fig. 4). This indicates that these universities do not have a consistent number of tweets, and it may vary slightly.

Another similar pattern can be seen in the plots of Oxford, Stanford, Cambridge, Lisbon, and Buffalo, where the mean is clearly visible above one post, indicating that these universities have some consistency in their daily tweets. However, in the case of Pennsylvania, Massachusetts Institute of Technology, and Harvard, we still see that pattern, but at a smaller level, presenting an unbalanced Gaussian distribution.

To analyze the posting frequency across different days and times, we constructed a tweet frequency table for all HEIs in the intersection period, crossing the weekday with the posting hour. This resulted in the heat map presented in Fig. 5. From the heat map, we observed that posts from the Universities of Pennsylvania and Oklahoma are concentrated between 2 PM to 9 PM on weekdays. Harvard, Princeton, Chicago, and

**Table 2.** Posting daily frequency (decreasing order). Updated from [7].

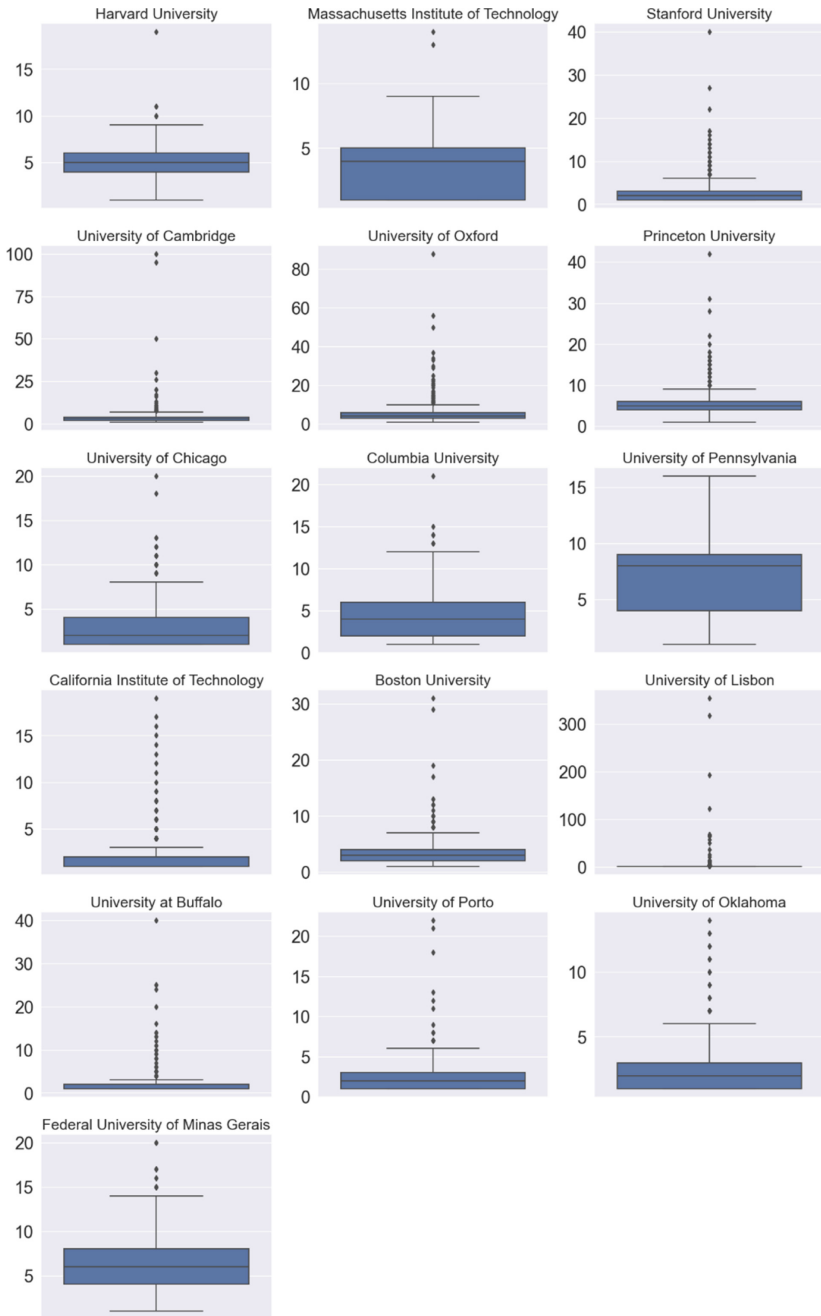
Rank	High Education Institution	Mean	Max.	Std.
9	University of Pennsylvania	7.07	16	2.91
500	Federal University of Minas Gerais	6.05	20	3.21
5	University of Oxford	5.81	88	6.21
6	Princeton University	5.66	42	3.63
1	Harvard University	5.36	19	1.72
8	Columbia University	4.19	21	2.89
200	University of Lisbon	3.95	355	23.36
4	University of Cambridge	3.50	100	5.29
2	Massachusetts Institute of Technology	3.31	14	1.78
99	Boston University	3.16	31	2.28
7	University of Chicago	2.86	20	2.14
400	University of Oklahoma - Norman	2.66	14	1.73
3	Stanford University	2.55	40	2.55
308	University of Porto	2.25	22	1.65
300	University at Buffalo	2.12	40	3.28
10	California Institute of Technology	2.05	19	1.62

Boston, post throughout the week, though they only post during working hours. However, we also observed that high frequency posting from MIT, Pennsylvania, Oklahoma, and Federal University of Minas Gerais is condensed into a short period of time and weekdays. This suggests a regular and systematic approach to external communication that may be considered an editorial approach.

We also created a set of word clouds for each HEI, based on all retrieved posts and the common posting period. Figure 6 shows the word clouds for each HEI using all available retrieved posts. We observed that most HEIs prioritize projecting their image, with their name being the most used term. However, Columbia, Boston, Lisbon, and Oklahoma differ from this pattern. We also noticed that the terms ‘student’ and ‘research’ are common across almost all HEIs, highlighting their focus on these topics and specific segments of readers.

Our analysis of the word clouds for each HEI reveals some interesting patterns. For instance, most HEI prioritize projecting their own name, as it appears as the most used term. However, Columbia, Boston, Lisbon, and Oklahoma stand out for using other terms more frequently. In addition, terms such as “student” and “research” are common across most HEI, indicating their focus on these topics and specific reader segments.

It is worth noting that University of Lisbon does not emphasize these terms to a high degree, while University of Porto and Federal University of Minas Gerais use their Portuguese counterparts, “estudante” and “pesquisa”, respectively. Engagement actions directed towards newcomers are also common in most HEI, often by congratulating them



**Fig. 3.** Boxplots of daily posting for each HEI. Updated from [7].

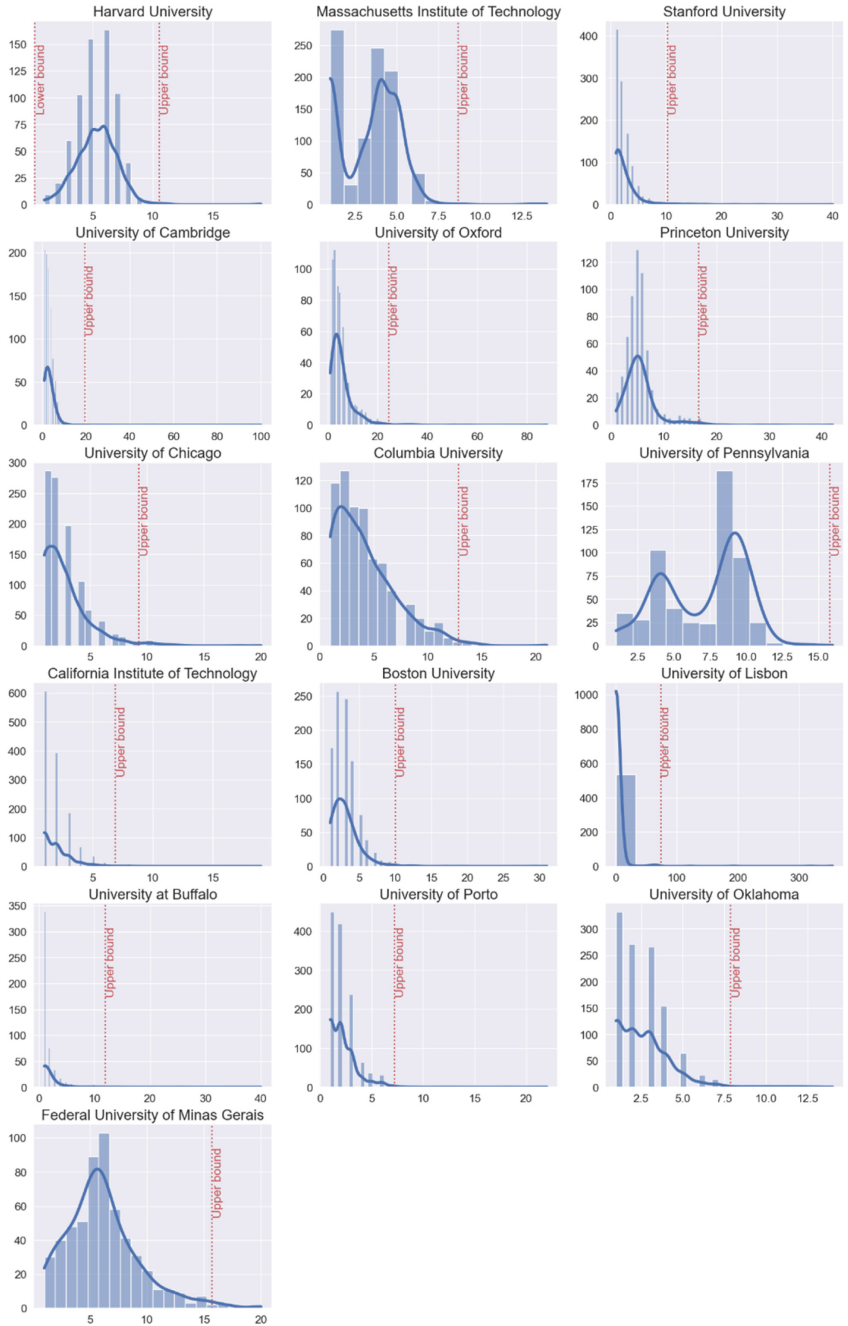


Fig. 4. Distribution of posting frequencies. Updated from [7].



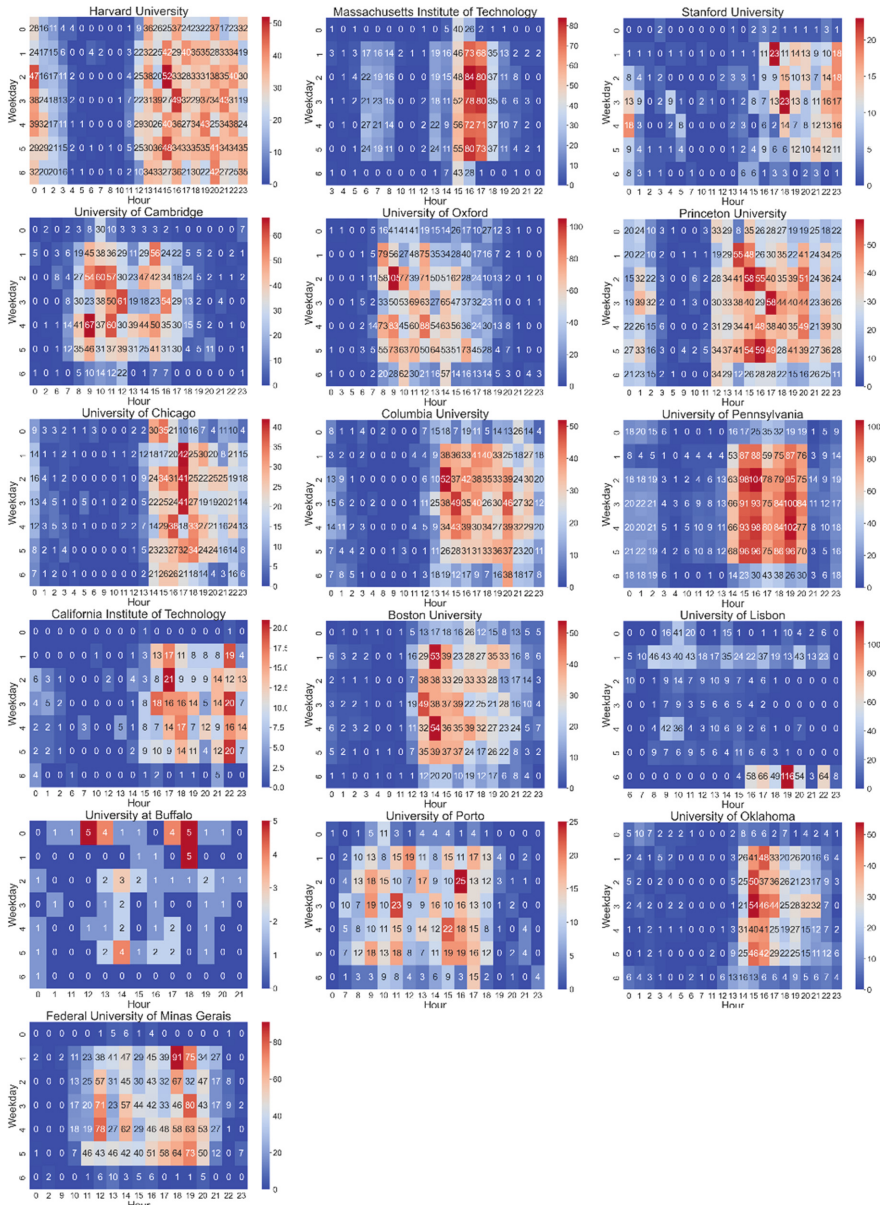


Fig. 5. Publication weekday and time. Updated from [7].

using terms like “first”, “year”, and “new”. Finally, the terms “pandemic” and “vaccine” are still prevalent in posts from Harvard and Oxford, but not in other HEI, suggesting an important editorial difference. When we focus on the common publishing period (Fig. 7), we only observe two minor changes: (a) an increase in engagement actions in

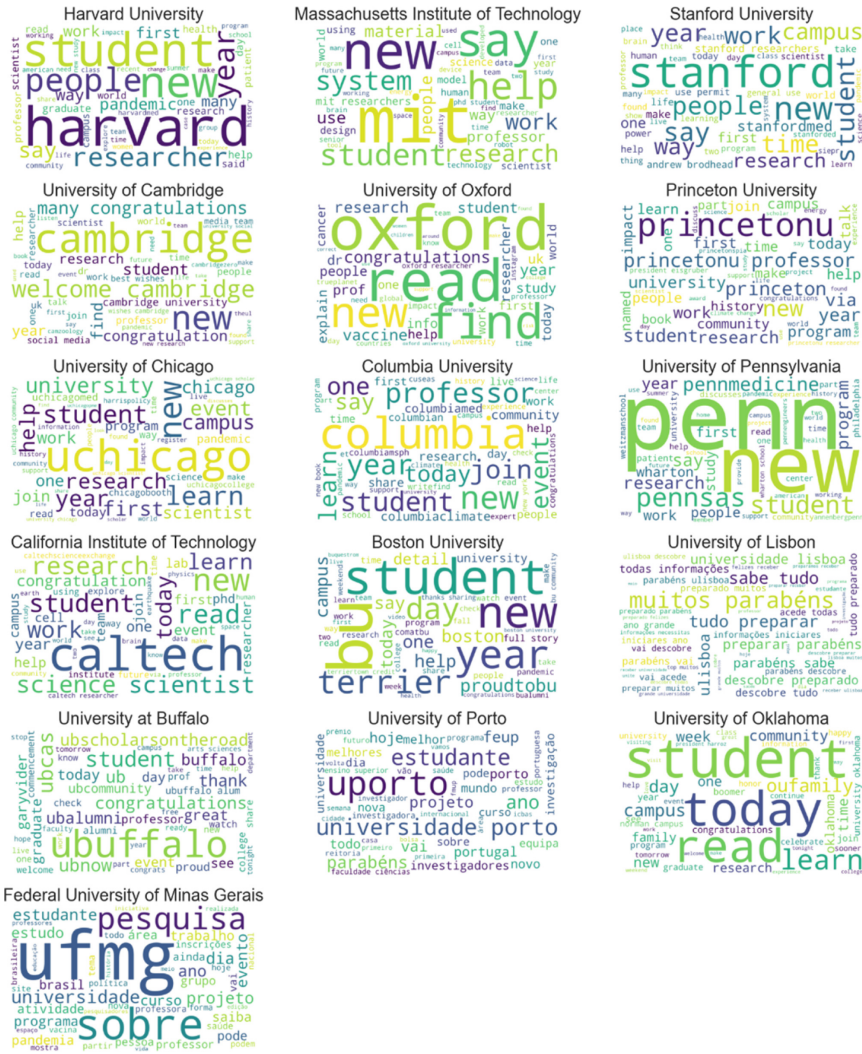


Fig. 6. Word cloud for each HEI considering all retrieved posts. Updated from [7].

Columbia compared to other terms and (b) a reduction in the importance of branding and projecting the institutional image at University of Porto.

Overall, our analysis suggests that despite different HEI publishing at varying periods and frequencies, they employ very similar strategies in terms of textual content. Thus, it remains unclear whether there is a general mapping between the ranking lists and the strategies and publishing patterns adopted by each HEI.

To gain further insights into the content of the tweets, we also analyzed the sentiment of each post for all HEI. To accomplish this, we used the TextBlob library (0.16.0) in a Python implementation, which categorizes each post as positive, neutral, or negative. The sentiment value returned corresponds to the analysis of the text.

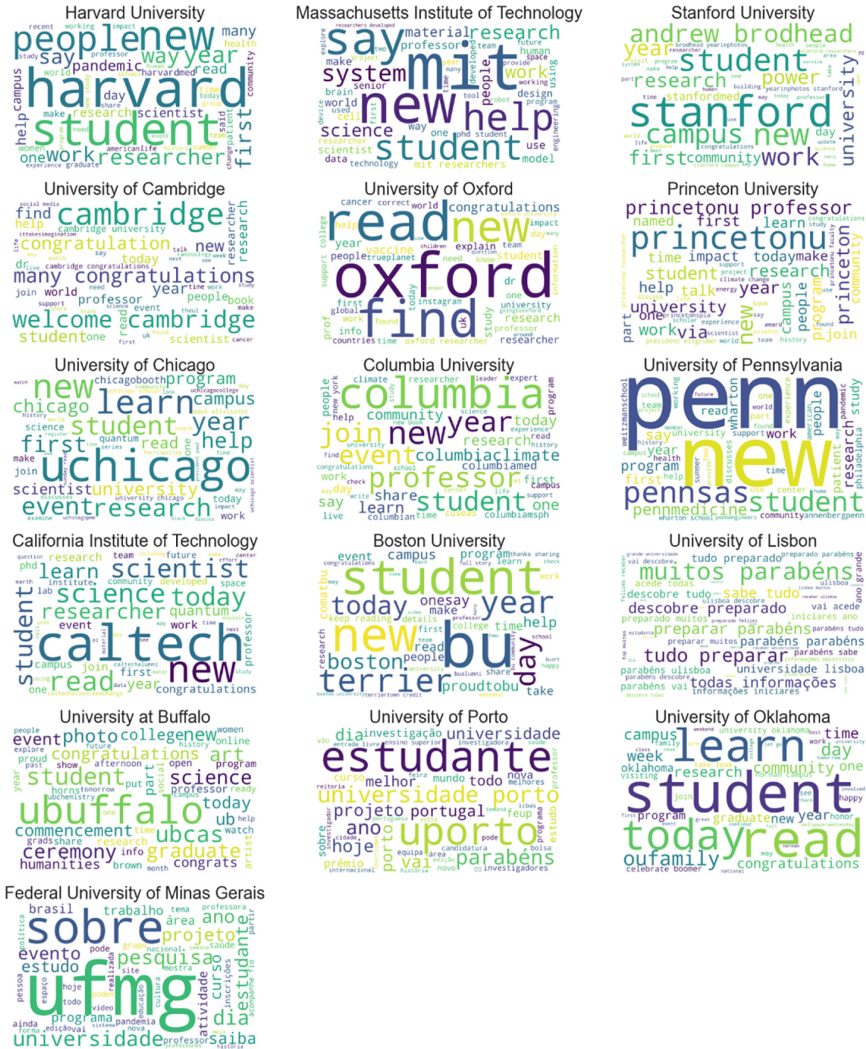
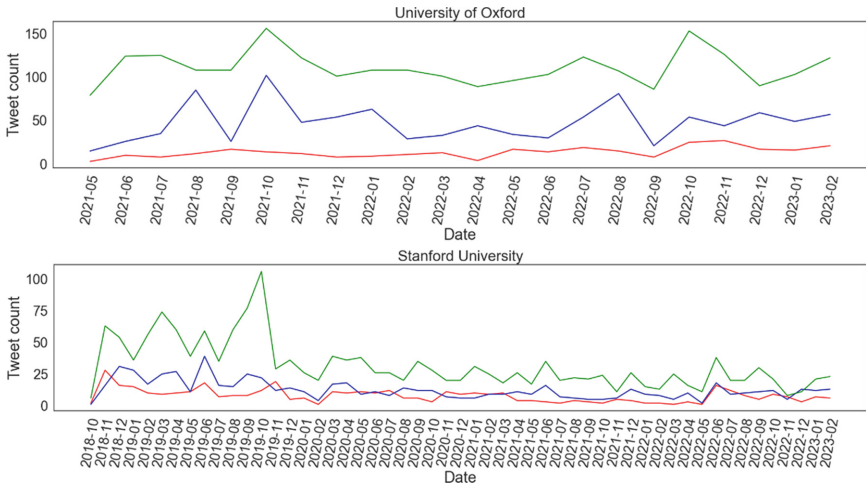


Fig. 7. Word cloud for each HEI considering the common period. Updated from [7].

As an example, we present the sentiment analysis of Harvard’s tweets over time in Fig. 8, where we group tweets by month.

For the sake of saving space, we do not present the graphs for all HEIs in this section. However, we will use the computed values to compare HEIs in the next section.

Another approach to visualizing the sentiment data is presented in Fig. 9, where we show the cumulative sentiment frequency on a daily basis. To produce this graph, we used the following strategy: for each HEI, we started with a countable variable set to zero. For each day of analysis, we added to the countable variable if the tweets had a positive sentiment, subtracted if they had a negative sentiment, and left the countable variable unchanged if the sentiment was neutral.



**Fig. 8.** Monthly evolution of sentiment for Oxford and Stanford posts. Negative sentiment in red, neutral in blue and positive in green. Updated from [7].

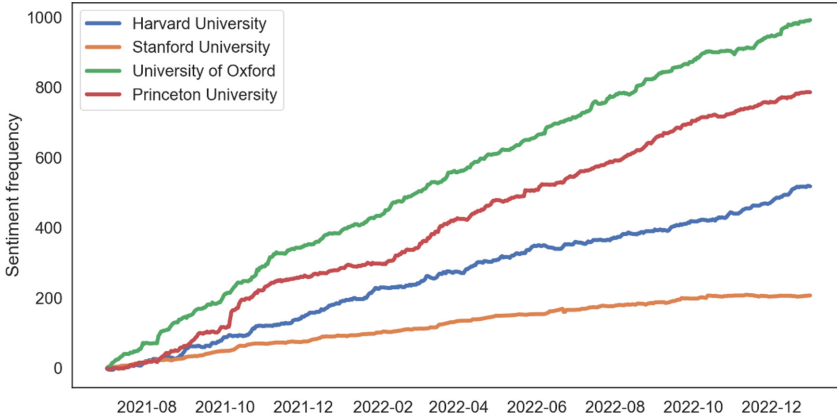
Figure 9 provides a cumulative view of the sentiment frequency for the four universities. It is notable that the sentiment of daily tweets for all four universities appears to be increasingly positive, as evidenced by the linearly increasing behavior of the lines over time. Moreover, the growth rate of the sentiment lines for Oxford, Princeton, and Harvard appears to be similar, with only slight differences in slope, whereas Stanford’s sentiment line remains relatively constant. This difference can be attributed to the fact that the three former institutions have a larger number of positive tweets compared to negative ones, whereas Stanford appears to have published almost an equal number of positive and negative tweets, as indicated by Fig. 8. While we present sentiment graphs only for these four universities, we will use the computed sentiment values to compare all HEIs in the following section.

To gain insights into the emotions conveyed by the general tweets of each HEI, we employed the NRCLex library to predict the sentiments and emotions in the text. The library identifies various emotional affects, such as anger, anticipation, disgust, fear, joy, negative, positive, surprise, sadness, and trust. To facilitate visualization of the emotional categories, we assigned colors: green to positive emotions, red to negative emotions, and orange to neutral emotions.

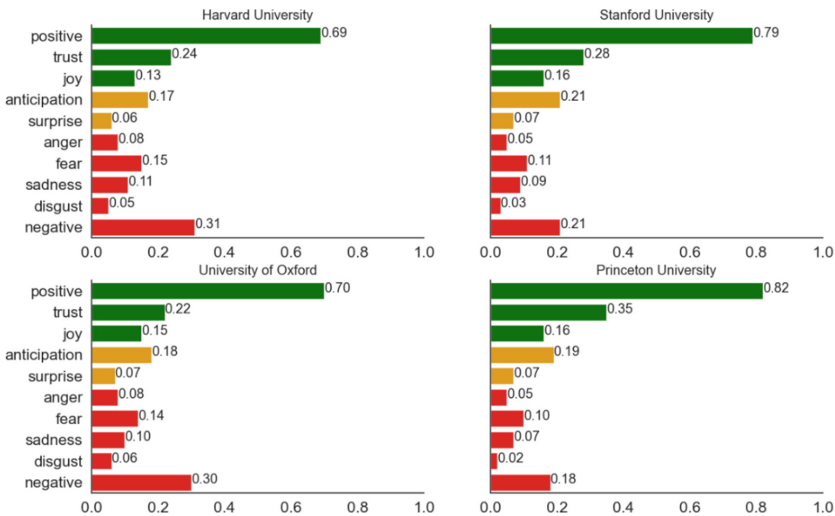
Figure 10 shows that the overall emotion proportions for the top four universities are quite similar. It is noteworthy that the sum of positive and negative sentiments is equal to 1, and the same applies to the sum of trust, joy, anticipation, surprise, anger, fear, sadness, and disgust. Our aim was to detect any unusual patterns, but none were found.

We also conducted a topic modeling analysis to explore the themes covered in the general tweets of each HEI during the analysis period. To perform the analysis, we employed the Gensim library for topic modeling and the Topic Coherence Metric [17] to identify the optimal number of topics in the text corpus.

Before applying the Latent Dirichlet Allocation (LDA) [3, 11, 21] model for topic modeling, we preprocessed the text by removing all punctuations, special characters,



**Fig. 9.** Cumulative sentiment frequency daily for four of the top 10 HEIs.



**Fig. 10.** Proportion of emotions and sentiments found in the overall tweets for each HEI.

numbers, links, and capitalization. Next, we removed stop words that are not relevant to topic modeling and applied a lemmatization function to convert each word to its base or dictionary form. Table 3 displays the topics identified for each HEI.

The topics discovered through our topic modeling analysis were labeled with short names and are presented in Table 3. The table displays each topic found per HEI, with a coherence score for the number of topics found and a Jaccard distance between topics. Coherence measures the quality of topics in terms of human interpretability, while Jaccard distance indicates the distance between topics, with a range from 0 (close) to 1 (distant).

The twelve topics discovered were Research and Development (R&D), Beginning of the Academic Year (BAY), University Event (UE), Education (E), Health Care and

**Table 3.** Topics found by the LDA model for each HEI.

Harvard	MIT	Stanford	Cambridge	Oxford	Princeton	Chicago	Columbia	Pennsylvania	Caltech	Boston	Porto	Oklahoma	UFMG
0.34	0.32	0.37	0.37	0.43	0.35	0.31	0.26	0.30	0.25	0.34	0.31	0.34	0.33
0.83	0.69	0.15	0.81	0.64	0.76	0.81	0.82	0.78	0.15	0.72	0.66	0.74	0.82
BAY	BAY	BAY			BAY			BAY	BAY	BAY	BAY	BAY	BAY
R&D	R&D	R&D	R&D	R&D		R&D		R&D	R&D				R&D
			UE	UE		UE	UE			UE		UE	UE
E					E	E	E	E			E		
HCM				HCM		HCM							
			C				C					C	
			T						T				
							AP						AP
								S					
					PD								
						DI							
			CA										

Medicine (HCM), Community (C), Technology (T), Academic Programs (AP), Sports (S), Professional Development (PD), Diversity and Inclusion (DI), and Collaboration and Accessibility (CA). The maximum number of topics found was 5 for Chicago and Cambridge, while the minimum number was 2 for MIT, Stanford, Boston, and Porto.

Upon analysis of Table 3, we observed that HEIs generally follow a similar publication strategy, with four main topics found across a large number of institutions. Beginning of the Academic Year was found in 10 of the 16 HEIs, Research and Development in 9 HEIs, University Event in 7 HEIs, and Education in 5.

## 4 Grouping the Strategies

To conduct a more in-depth analysis, we sought to quantitatively compare the publication strategies of HEIs. Since we will be using numerical quantities, we can compare all HEIs at once. Our goal is to perform an unsupervised classification that will group the HEIs according to the metrics we will use.

Our analysis will focus solely on the publication patterns, and we will not consider factors such as employment, student success, or research funding. We aim to use metrics obtained from the analysis of the retrieved tweets to group the HEIs and compare the results with the rankings.

### 4.1 The Feature Space Vector Model

To reflect most of the analysis we have done previously, we choose 10 features to represent the publishing behavior of each HEI. Those are:

- Mean daily posting frequency
- Max daily posting frequency
- Ratio of publishing in weekends (Saturday + Sunday)
- Ratio of publishing during night period (9pm to 7am)
- Mean positive sentiment
- Mean neutral sentiment

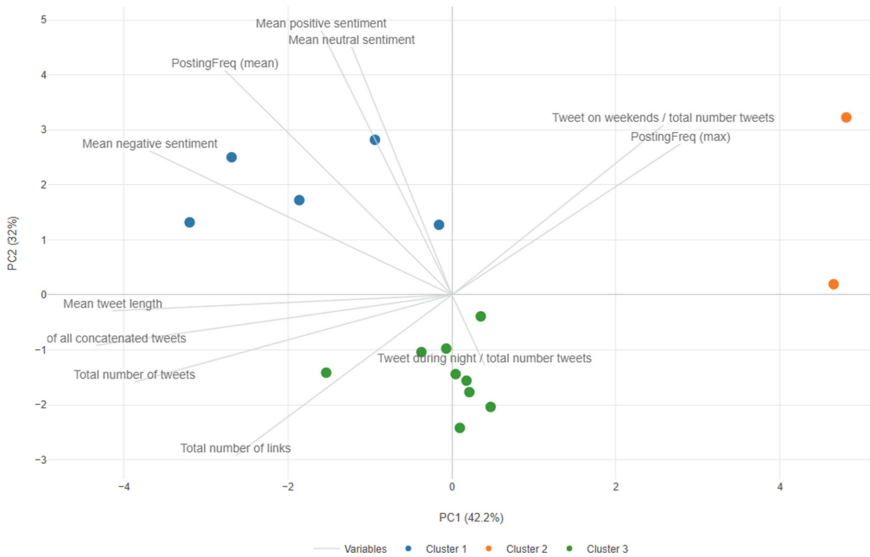
- Mean negative sentiment
- Mean tweet length (text)
- Length of all concatenated tweets (text)
- Total number of links used in the text

These features represent most of the analysis described previously and now are used together to represent a signature of each HEI posting behavior.

### 4.2 Clustering the HEI

We represent each HEI as a vector in a 10-dimensional vector space model and compute the distances between HEIs to determine which ones are closer to each other. We use the standard k-means algorithm to group HEIs based on these metrics. We experimented with generic k-means [15] using both the Floyd algorithm [13] and the Hartigan-Wong [10] algorithms, but the results were almost identical. To minimize inter-cluster distances, we tried different numbers of clusters and compared them using the “elbow method”. Eventually, we decided to use three clusters to group the HEIs.

To visualize the grouping results, we present a mapping of each HEI colored according to its assigned cluster in Fig. 11. We use a PCA transformation [1] to represent 10-dimensional points in two dimensions. In this visualization we are presenting information retrieved from the HEI in the period of July 2021 to February 2023 [7].



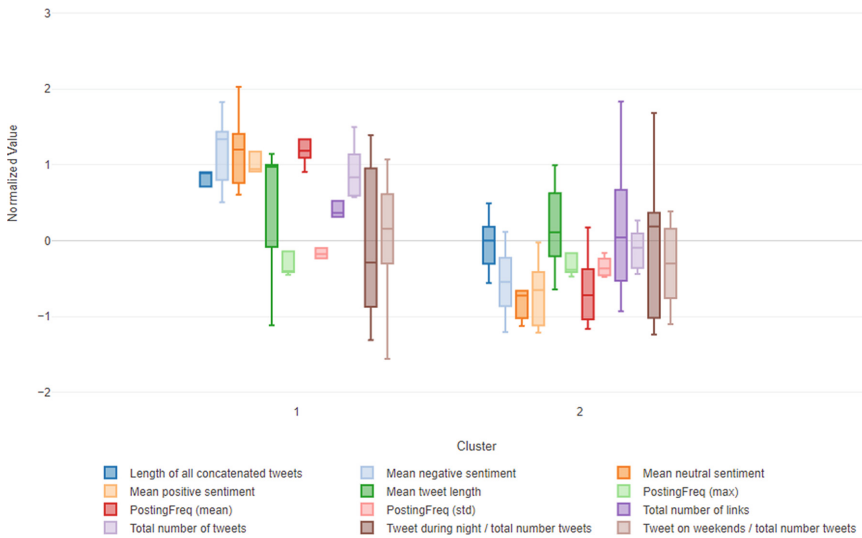
**Fig. 11.** Positioning and clustering the HEIs in a 2D projection of the feature space [7].

We can confirm this clustering makes sense because there is a clear distinction of the 3 groups: HEI in blue in the second quadrant (cluster 1), HEI in orange in the first quadrant (cluster 2), and HEI in green (mostly around the separation between the third and the fourth quadrants (cluster 3).

To complete the analysis, we checked the distribution of the normalized values of the 10 features in each cluster Fig. 12, below using boxplots.

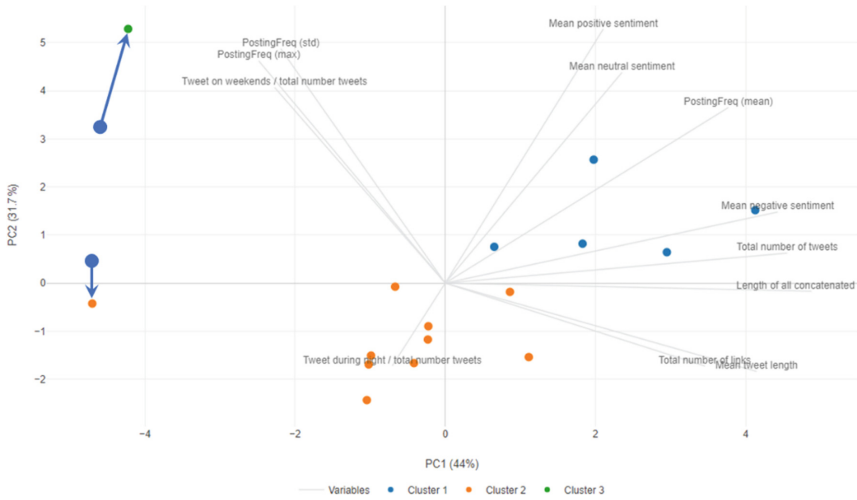
The boxplot in Fig. 12 only shows the results for clusters 1 and 2, as cluster 3 has only one HEI and therefore it makes no sense to present its result. However, we can see that cluster 1 and 2 may suggest a well-established strategy between the HEI present in the two clusters. We suggest that, as the compact distribution of variables in both clusters show us the concentration of these variables on the positive side for cluster 1 and on the negative side for cluster 2, note that all data were normalized before this analysis.

We have expanded our previous dataset [7] by collecting new data from July 2021 to February 2023 in order to better understand the clustering of HEIs and gain deeper insights on how to group them more cohesively. The major differences that we have found include a) posting frequency mean, b) mean positive sentiment, and c) length of all concatenated tweets. With this updated data, we have re-examined the biplot.



**Fig. 12.** Distribution of each variable in each cluster [7].





**Fig. 13.** Distribution of each variable in each cluster highlighting the major changes.

In Fig. 13, the points with arrows were, in the previous experience, part of the same cluster. However, in the updated dataset, these points moved away from each other, leading to a different clustering result. Of those two, the one in orange was assigned to cluster 2, while the green one became the only one in cluster 3. These two were respectively University of Buffalo and University of Lisbon. The other points also moved slightly but did not affect the clustering significantly.

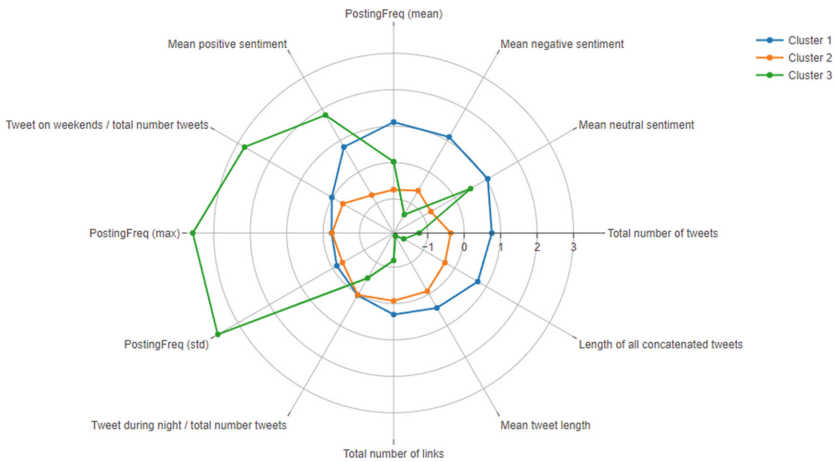
### 4.3 Analysis of the Results

In order to compare these results with the rankings, we use Table 4, where we include the cluster assignment (second column) together with the four ranking lists. We ordered the table with respect to column “cluster”, hence grouping HEIs that belong to the same cluster.

The clustering results show that the HEIs in cluster 1, except for University of Minas Gerais, are all ranked in the top 10 positions of the ranking lists. On the other hand, the HEIs in cluster 2 span a wide range of positions in the rankings. Cluster 3 contains only one HEI, University of Lisbon (UL), which is located in the middle of the ranking list (position 200 in CWUR). The comparison between UL and University of Buffalo (UB) reveals that their publishing strategies have diverged, with UB’s strategy becoming more consolidated and UL’s becoming more chaotic with less clear objectives. Overall, HEIs in cluster 1 tend to be ranked higher, while those in cluster 2 can’t be properly characterized by the features used. We can see this in the radar plot of Fig. 14.

**Table 4.** Cluster assignment. Updated from [7].

High Education Institution	Cluster	CWUR	Shanghai	USNews	QS
Columbia University	1	8	8	6	19
Federal University of Minas Gerais	1	500	401–500	456	651–701
Harvard University	1	1	1	1	5
Princeton University	1	6	6	16	20
University of Oxford	1	5	7	5	2
University of Pennsylvania	1	9	15	13	13
Boston University	2	99	101–150	65	112
California Institute of Technology	2	10	9	9	6
Massachusetts Institute of Technology	2	2	3	2	1
Stanford University	2	3	2	3	3
University at Buffalo	2	300	301–400	280	338
University of Cambridge	2	4	4	8	3
University of Chicago	2	7	10	15	10
University of Oklahoma - Norman	2	400	501–600	425	651–700
University of Porto	2	308	201–300	255	295
University of Lisbon	3	200	201–300	197	356



**Fig. 14.** Distribution of each variable in each cluster highlighting the major changes.

## 5 Conclusions

In this paper, we have explored the relationship between publishing strategies and the ranking of Higher Education Institutions (HEIs). An extended dataset from [7] was gathered and analyzed in terms of frequency, date, and content. We represented HEIs as vectors within a 10-dimensional space and grouped them utilizing the k-means clustering algorithm.

Our results indicate a correlation between top-ranked HEIs and the consolidation of publication strategies. The variables examined effectively differentiate non-consolidated strategies; however, additional, or alternative features are required to better segment cluster 2. Furthermore, we observed that higher-ranked HEIs exhibit more expressive sentiments, greater tweet lengths, and increased posting frequencies.

Additionally, we examined the emotions and topics present in the overall content of the tweets for each HEI. Our analysis revealed that the highest-ranked HEIs share similar publication strategies, characterized by the presence of four primary topics across a majority of the institutions.

As a future direction, it would be valuable for researchers to develop a predictive model employing Long Short-Term Memory (LSTM) techniques, merging the collected retrieval information to predict emotions and topics in forthcoming posts from each HEI. Additionally, the ability to analyze images present in posts should also be included in the analysis.




## References

1. Abdi, H., Williams, L.J.: Principal component analysis. *Wiley Interdiscipl. Rev. Comput. Statist.* **2**(4), 433–459 (2010)
2. Aguillo, I., Bar-Ilan, J., Levene, M., Ortega, J.: Comparing university rankings. *Scientometrics* **85**(1), 243–256 (2010)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**(Jan), 993–1022 (2003)
4. Coelho, T., Figueira, Á.: Analysis of top-ranked HEI publications' strategy on Twitter. In: 2021 IEEE International Conference on Big Data (Big Data), pp. 5875–5877. IEEE (2021)
5. Dumpit, D.Z., Fernandez, C.J.: Analysis of the use of social media in Higher Education Institutions (HEIs) using the technology acceptance model. *Int. J. Educ. Technol. High. Educ.* **14**(1), 1–16 (2017)
6. Field, A.P.: Kendall's coefficient of concordance. *Encycl. Statist. Behav. Sci.* **2**, 1010–1011 (2005)
7. Figueira, A., Nascimento, L.: Do top Higher Education Institutions' social media communication differ depending on their rank? In: Proceedings of the 18th International Conference on Web Information Systems and Technologies, ISBN 978-989-758-613-2, ISSN 2184-3252, pp. 355–362 (2022)
8. Figueira, Á.: Uncovering social media content strategies for worldwide top-ranked universities. *Proc. Comput. Sci.* **138**, 663–670 (2018)
9. Gajić, J.: Importance of marketing mix in higher education institutions. *Eur. J. Appl. Econ.* **9**(1), 29–41 (2012)
10. Hartigan, J.A., Wong, M.A.: Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, **28**(1), 100–108 (1979)

11. Hong, L., Davison, B.D.: Empirical study of topic modeling in twitter. In: Proceedings of the First Workshop on Social Media Analytics, pp. 80–88 (2010)
12. Kendall, M.G.: A new measure of rank correlation. *Biometrika* **30**(1/2), 81–93 (1938)
13. Linde, Y., Buzo, A., Gray, R.: An algorithm for vector quantizer design. *IEEE Trans. Commun.* **28**(1), 84–95 (1980)
14. Liu, N.C.: The story of academic ranking of world universities. *Int. Higher Educ.* 54 (2009)
15. MacQueen, J.: Classification and analysis of multivariate observations. In: 5th Berkeley Symposium Mathematics Statistics Probability, pp. 281–297 (1967)
16. Olcay, G.A., Bulu, M.: Is measuring the knowledge creation of universities possible?: A review of university rankings. *Technol. Forecast. Soc. Chang.* **123**, 153–160 (2017)
17. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 399–408 (2015)
18. Scott, A.J., Symons, M.J.: Clustering methods based on likelihood ratio criteria. *Biometrics* **27**(2), 387 (1971). <https://doi.org/10.2307/2529003>
19. Van Raan, A.F.: Challenges in ranking of universities. In: Invited Paper for the First International Conference on World Class Universities, Shanghai Jiao Tong University, Shanghai, pp. 133–143 (2005)
20. Webber, W., Moffat, A., Zobel, J.: A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst. (TOIS)* **28**(4), 1–38 (2010)
21. Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitter rank: finding topic-sensitive influential twitterers. In: Proceedings of the third ACM International Conference on Web Search and Data Mining, pp. 261–270 (2010)



# Influence of Demographic Variables and Usage Behaviour on the Perceived User Experience

Jessica Kollmorgen<sup>1</sup> , Martin Schrepp<sup>2</sup> , and Jörg Thomaschewski<sup>3</sup> 

<sup>1</sup> University of Applied Sciences Emden/Leer, Emden, Germany  
jessica.kollmorgen@ux-researchgroup.com

<sup>2</sup> SAP SE, Walldorf, Germany  
martin.schrepp@sap.com

<sup>3</sup> University of Applied Sciences Emden/Leer, Emden, Germany  
joerg.thomaschewski@hs-emden-leer.de

**Abstract.** Users form an overall impression concerning the user experience (UX) based on their perception of special UX qualities. Therefore, measuring users' perceptions of these particular UX aspects is essential for determining the UX of a product. The measured hedonic qualities, e.g. stimulation or aesthetics, and pragmatic qualities, e.g. efficiency or learnability, form a suitable overall impression of the perceived user experience of the product. In practice, the measurement of such qualities is often carried out with the help of standardized questionnaires such as the SUS, UMUX, or UEQ. However, the same product sometimes shows large differences in the ratings of different users. It is conceivable that other factors, for example, demographics, usage frequency, or experience with a product, can influence UX ratings. In a previous study (Kollmorgen, Schrepp & Thomaschewski, 2022), the four products Netflix, Microsoft PowerPoint, BigBlueButton, and Zoom were examined for differences in the UX ratings according to such factors. In the present paper, the data set was extended by two additional products of different product categories to deepen and broaden the investigation of the influences of external factors on the perceived UX of products, with a specific focus on their impact on pragmatic and hedonic qualities.

**Keywords:** User experience · Usability · UEQ-Short · UMUX-LITE · SUS · Pragmatic quality · Hedonic quality · Product knowledge · Frequency of use

## 1 Introduction

To evaluate how well products meet the requirements of their users, questionnaires are often used. Standard questionnaires like the User Experience Questionnaire (UEQ) (Laugwitz, Schrepp & Held, 2008), the Usability Metric for User Experience (UMUX) (Finstad, 2010), or the System Usability Scale (SUS) (Brooke, 1996) can be used to measure the usability and user experience (UX) of products. This makes it possible to align the needs of users as closely as possible with the products (Schrepp, 2021).

To gain an appropriate overall impression of the measured products, it is important to distinguish between hedonic and pragmatic factors (Hassenzahl, Diefenbach & Göritz,

2010). Pragmatic qualities (PQ) are associated with a product's ability to assist users in achieving specific goals, while hedonic qualities (HQ) are geared towards fulfilling psychological needs that go beyond the sole purpose of task completion, such as stimulation or aesthetics (Hassenzahl, 2008; Winter et al., 2017).

However, results often show that different users do not perceive the user experience or usability of the same product in the same way. This could be attributed to several factors. On the one hand, studies have already shown that the importance of hedonic and pragmatic UX factors depends on the product category (Winter et al., 2017; Kollmorgen et al., 2021; Meiners et al., 2021; Schrepp et al. 2023). In one study, for example, it became clear that for the product category of online banking, pragmatic UX factors such as trust or quality of content were rated as important, in contrast to hedonic factors, such as stimulation or aesthetics. On the other hand, also a different usage behaviour can have an impact on the perceived usability and user experience of a product. E.g., people who use a product more frequently typically know it better, have adjusted their usage behaviour to avoid typical UX problems of the product, and therefore perceive the user experience differently. Conversely, a product is presumably only used more frequently if it offers a good user experience.

This led to the first research question, *RQ 1: Are there external factors besides the classic UX factors that influence the perceived user experience of a product and assist in explaining the differences in UX ratings?*

However, the way the product is used can have varying effects on pragmatic and hedonic factors. While having a high level of expertise with a product is likely to lead to higher ratings for pragmatic quality, it is uncertain if this same effect applies to hedonic qualities.

Based on this research question, the study by Kollmorgen, Schrepp and Thomaschewski (2022a) investigated which impacts external factors can have on the pragmatic and hedonic qualities of well-known products. For this purpose, four products from three different product categories were selected, which have been heavily used in recent years. The streaming platform Netflix, the video conferencing tools Zoom and BigBlueButton, and the presentation software Microsoft PowerPoint. These products support leisure activities at home as well as remote working and thus display a quite heterogeneous set of use cases and user experience factors.

Building on that, this paper extends and deepens the findings of this first study from Kollmorgen, Schrepp and Thomaschewski (2022a) by collecting data on two other products that are also heavily used and well-known: the social network platform TikTok and the online banking software PayPal. TikTok, just like Netflix, is mainly used for leisure and thus should have a stronger focus on hedonic qualities such as fun and visual aesthetics. PayPal, on the other hand, has more pragmatic purposes and focuses mainly on the efficient fulfilment of working tasks. This product selection ensures that the influences on both hedonic and pragmatic UX factors are reviewed and deepened with two additional product categories, resulting in six products of five product categories overall.

This led to the overarching second research question, *RQ2: To what extent are the pragmatic as well as the hedonic quality of products influenced by the external factors*

*mentioned above?* Does the impact of these factors influence pragmatic and hedonic qualities differently?

This paper is structured as follows: After a presentation of the UX questionnaires used in Sect. 2, the interindividual differences in the perception of UX are explained in Sect. 3, which serve as the basis for answering the research questions. The methodology of the two studies developed on this basis is then explained in Sect. 4. The results of these are presented in Sect. 5 and form the basis for answering the two research questions in Sect. 6, concluding the article in Sect. 7 with a summary and outlook.

## 2 UX Questionnaires

The goal of this research is to investigate the influence of demographic factors and differences in the usage experience or usage frequency on the subjective impression of persons concerning UX. The standard method to measure such subjective UX impressions are questionnaires. But user experience itself is a quite heterogeneous concept that contains many facets. There are many established standard UX questionnaires (see Schrepp, 2021a for an overview) available that measure aspects of UX, but they deviate to some extent from the specific UX aspects they consider. For that reason, the studies use three quite common UX standard questionnaires that will be shortly introduced in this section.

### 2.1 System Usability Scale (SUS)

The SUS (Brooke, 1996, 2013) is a short questionnaire that focuses on the measurement of classical usability aspects, for example, usefulness, consistency, or ease of learning. The original publication announced the SUS as a “quick and dirty usability scale”. But despite this modest description, the SUS is currently still one of the most used usability questionnaires and there is a huge number of papers that investigate the psychometric properties of the SUS (see Lewis, 2018 for an overview).

The 10 items of the SUS are short statements that describe aspects of usability:

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought that the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

The SUS contains a single scale and produces an overall score between 0 and 100. Each item can be rated on a 5-point agreement scale with the endpoints Strongly Disagree (left) and Strongly Agree (right).

For items 1, 3, 5, 7, and 9 agreement represents a positive evaluation, and these items are scored as 0 to 4 from left to right. For items 2, 4, 6, 8, and 10 agreement represents a negative evaluation, and these items are scored as 0 to 4 from right to left, thus in the opposite direction. Thus, a 4 represents the most positive evaluation, and a 0 the most negative evaluation. The scores for the 10 questions are added up to a participant score between 0 and 40, which is then multiplied by 2.5 to scale it between 0 and 100 (the argumentation for this rescaling is that a score between 0 and 100 is easier to communicate). The SUS score for a product is then simply the average over all participant scores.

## 2.2 Short Form of Usability Metric for User Experience (UMUX-LITE)

The UMUX-LITE (Finstad, 2010) is a short usability questionnaire that contains the two items:

- This system’s capabilities meet my requirements.
- This system is easy to use.

The measurement concept of the UMUX-LITE is related to the Technology Acceptance Model (Davis, 1986). This concept assumes that user acceptance of a new technology is based on its perceived usefulness (first item of the UMUX-LITE) and perceived ease of use (second item of the UMUX-LITE).

Participants can rate these items on a 7-point response scale with the endpoints *Strongly disagree* (left) and *Strongly agree* (right). The responses are scored as 0 to 6 from disagreement to agreement, therefore 0 is the most negative, and 6 the most positive evaluation. Just like in the SUS, the item scores are added up to a participant score between 0 and 12. This score is then rescaled to 0 to 100 by dividing it by 12 and multiplying it by 100. The UMUX-LITE score for a product is then the average over all participant scores.

The UMUX-LITE thus provides a high-level measurement of UX related to the concept underlying the technology acceptance model.

## 2.3 Short Form of the User Experience Questionnaire (UEQ-S)

The original User Experience Questionnaire (UEQ) (Laugwitz, Schrepp & Held, 2008) measures UX by six pragmatic and hedonic UX aspects (*Attractiveness, Efficiency, Perspicuity, Dependability, Stimulation, Novelty*). The UEQ contains 26 items in the form of a semantic differential.

A short form with just 8 items was developed (Schrepp, Hinderks & Thomaschewski, 2017) to support use cases that require short completion times. This short version (UEQ-S) contains only two scales for pragmatic (task-related UX qualities) and hedonic (non-task-related UX qualities).

The items of the UEQ-S are:



obstructive	0 0 0 0 0 0	supportive
complicated	0 0 0 0 0 0	easy
inefficient	0 0 0 0 0 0	efficient
confusing	0 0 0 0 0 0	clear
boring	0 0 0 0 0 0	exciting
not interesting	0 0 0 0 0 0	interesting
conventional	0 0 0 0 0 0	inventive
usual	0 0 0 0 0 0	leading edge

The first 4 items form the scale for pragmatic quality and the second four items the scale for hedonic quality. An overall value is determined by the mean over all 8 items, it represents to overall impression concerning UX. The items are scored from -3 (negative term) to +3 (positive term). The scale scores are simply the mean over all items in the corresponding scale and all participants in a study.

The UEQ-S questionnaire and all supporting material (handbook, translations in more than 30 languages, and an Excel-based data analysis tool) are available free of charge at <https://www.ueq-online.org/>.

### 2.4 Differences Between the Three Questionnaires

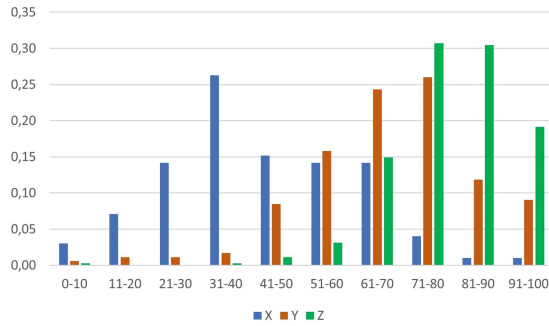
When analyzing the items of SUS, UMUX-LITE, and UEQ-S semantically, it is clear that all three questionnaires measure distinct concepts of UX. The SUS measures only classical usability criteria, for example, ease of learning, consistency, efficiency, or controllability, thus setting the focus on aspects that support or hinder users to work on their tasks. This aspect is considered also in the UMUX-LITE (measured here in a single question *This system is easy to use*) and in the UEQ-S (in the four items of the scale for pragmatic quality). The UMUX-LITE covers in addition to usability the usefulness of a product. The UEQ-S covers in addition to pragmatic quality also the hedonic quality or fun of the use of a product (by the four items of the scale hedonic quality). For the research questions of this study, this is a relevant aspect, as demographic factors or usage experience and frequency may only impact specific UX aspects and not all of them.

## 3 Interindividual Differences in the Perception of UX

Of course, different persons have different perceptions of the UX of a product. A nice example to demonstrate this can be found in Rummel & Schrepp (2018). Figure 1 (presentation taken from Schrepp, 2021b based on data described in Rummel & Schrepp, 2018) shows the distribution of ratings, grouped into intervals of length 10, obtained with the System Usability Scale (SUS) for three products.

Product Z is clearly rated much better than Product X. But it is interesting to note that the observed ratings for both products span the entire range. Thus, even for the on average poorly rated Product X, there are users that give quite high ratings. And for the on average good-rated Product, Z there are some strongly dissatisfied users. What are the reasons for such strong interindividual differences in the perception of UX?

Of course, demographic factors, for example, gender, age, or the cultural background of a user, can influence the UX perception of a product. But a majority of studies



**Fig. 1.** Relative frequencies of individual SUS scores for three different example products X, Y, and Z (Rummel & Schrepp, 2018 and Schrepp, 2021b).

concerning the SUS (see Lewis, 2018 for a summary) found no effect of age and gender on SUS ratings. For the UEQ (Laugwitz et al., 2008), a recent study (Aufderhaar et al., 2019) found no substantial differences between the ratings of men and women for some websites.

Personality traits, usually conceptualized based on the Five-Factor Model of personality (John & Srivastava, 1999; McCrae & John, 1992), can also influence ratings of UX questionnaires (for example Kortum & Oswald, 2018; Liapis et al., 2019; Devaraj et al., 2008; or Braun et al., 2019).

If users rate the UX of a product, they have to recall usage episodes from past interactions. Assume, for example, that a UX questionnaire asks about the speed of a system’s response to user inputs or commands. If users do not remember any long waiting times, they will rate this aspect as positive. If users remember a lot of situations where the system responds too slowly, the rating will be negative. Thus, the interaction history of a user with a product will of course have an impact on the perceived UX of that product.

The level of experience or the frequency of use may also impact the UX perception. A study by McLellan et al. (2012) found, for example, that experienced users tend to provide more positive UX ratings.

It is important to note that the impact of variables like age, gender, experience, or usage frequency on UX scores of standardized questionnaires depends on the concrete product. For general rules like, for example, “Gender has for all possible products no impact on S” (where S is a scale from a UX questionnaire) it will always be possible to find counterexamples. Assuming, for example, a website that is highly optimized for a purely female target group: will gender have an impact on UX ratings? If the design target is reached, then the answer will most likely be “Yes”. Therefore, products that are used for different usage scenarios are investigated in the studies.

## 4 Methodology

As already explained, the perceived user experience of a product can, on the one hand, depend on demographic factors or usage behaviour. On the other hand, these external factors can exert varying degrees of influence on pragmatic and hedonic qualities. For

this reason, a total of six popular products from five product categories were considered (see Table 1). The assignment of products to specific product categories as well as their analogies in the importance of UX factors are formed below based on Meiners et al. (2021).

**Table 1.** Examined products with UX focus and product category.

Focus	Product	Product category
Pragmatic quality	Microsoft PowerPoint	Presentation
	PayPal	Online banking
Pragmatic and hedonic quality	BigBlueButton	Video conferencing
	Zoom	Video conferencing
Hedonic quality	Netflix	Video streaming
	TikTok	Social network

#### 4.1 First Study: Netflix, PPT, BBB and Zoom

The first study was published at the 19<sup>th</sup> International Conference on Web Information Systems and Technologies (WEBIST) 2022 by Kollmorgen, Schrepp & Thomaschewski (2022a, 2022b). In this data collection, participants were recruited from various universities and through a panel and were compensated monetarily for their participation in the study. The target groups were provided with either German or English questionnaires between September and December 2021.

An online survey was conducted to gather data on the external factors influencing the four products. The survey begins with a brief set of instructions, followed by the collection of demographic information and usage behaviour details from the participants. Specifically, the following information was requested:

- *Age*
- *Gender*: Male (M), Female (F), Divers (D)
- *Usage frequency* (How often do you use < product name >?): Not very frequent, Several times a month, Several times a week, On a daily basis
- *Knowledge* (How good is your knowledge of < product name >?): Low, Medium, Strong, Excellent
- *Duration of use* (How long have you been using < product name >?): Less than a week, Since more than a week, Since more than 6 months, Since more than a year, Since more than 5 years

Participants were not required to answer all the questions in the survey, which is why an additional “No answer” category was included. Following the section with demographic and behavioural questions, the survey included the two items from the UMUX-LITE, eight items from the UEQ-S, and ten items from the SUS as described

above. At the end of the survey, participants were given the opportunity to provide free-form comments on the strengths and weaknesses of the product.

Completing this overall questionnaire, consisting of questions on demographic and usage data as well as on the items of the three questionnaires, took the respondents on average about 3 to 4 min. This shows that by using the short versions of the questionnaires, respondents were able to answer all questions and still spend very little time.

In this first study, the products Netflix, Microsoft PowerPoint (PPT), BigBlueButton (BBB) and Zoom were surveyed on demographics, usage patterns, and their perceived user experience.

We expected that the four products cover the range from pragmatic to hedonic quality. Netflix, belonging to the video streaming product category, is primarily used for private purposes and thus is expected to focus more on hedonic quality. In such cases of products that are primarily intended for private use, such as *Netflix*, hedonic factors like fun or beauty should not be neglected (Hassenzahl, 2001). On the other hand, PPT, from the product category presentation software, is used to complete work tasks, thus it is expected to have a strong focus on pragmatic quality. BBB and Zoom, both belonging to the video conferencing product category, are used for both private and professional purposes, which is why they should take into account both hedonic and pragmatic needs.

The obtained data sets were cleaned to enhance their quality. Any data records with a processing time that was too brief or too few clicks, or that had an incorrect response to the quality assurance question, were removed, resulting in the elimination of 97 records and leaving 338 records in total. For the four online surveys, the following numbers of responses were collected: Netflix ( $N = 97$ ), BBB ( $N = 76$ ), Zoom ( $N = 76$ ) and Microsoft PowerPoint ( $N = 89$ ). The participants had an average age of roughly 28 years, and more detailed information is available in the research protocol (Kollmorgen, Schrepp & Thomaschewski, 2022b).

## 4.2 Second Study: TikTok and PayPal

Building on the results of the initial data collection, the potential for deepening the answers to research questions was identified. There should be a stronger focus on assessing influences on hedonic and pragmatic quality. BBB and Zoom from the first study are both expected to focus equally on both pragmatic and hedonic quality, while Netflix (HQ focus) and PPT (PQ focus) only should depict one of the two. For this reason, the products TikTok (HQ focus) and PayPal (PQ focus) were selected as additions for the second data collection, as they both are expected to focus mainly on one quality each. TikTok, from the product category social network, is expected to focus more on hedonic quality due to the nature of the product, since, for example, UX factors such as aesthetics and novelty are considered important for this product category. PayPal, the online banking product category, is expected to place a correspondingly stronger focus on pragmatic quality, therefore pragmatic UX factors such as dependability and efficiency are more important here (Kollmorgen, Meiners, Schrepp & Thomaschewski, 2021).

Thus, in the first study, there were already two products that should equally focus on both pragmatic and hedonic quality, but only one product each that should focus more on PQ and HQ, respectively. The aim of adding these two products was therefore to

deepen the statements on the influence on hedonic and pragmatic quality, in particular, since the entire range is then evenly covered.

The second study is therefore a replication study based on the first study from 2022. English questionnaires with the same structure were used, in which only the product names were changed to TikTok and PayPal. The surveys were again conducted via the panel and the respondents received monetary compensation.

After the corresponding data collection, the data were cleaned analogously to the first study. 27 records that did not meet the criteria were removed, resulting in a remaining set of 114 records for TikTok and 111 for PayPal. Detailed results can be found in the protocol (Kollmorgen, Schrepp & Thomaschewski, 2023).

## 5 Results

In the following, the data from both studies are considered and compared together in order to obtain more meaningful results. On the one hand, this is possible because it is a replication study, which means that the data from both surveys can be interpreted in the same way. On the other hand, sufficient data is also available for both new products so that analyses and interpretations can be carried out.

To ensure a meaningful interpretation of the influence of the demographic factors and usage behaviour on UX metrics, a lower limit for the number of participants in a category had to be defined. For example, it would not be meaningful to say that users who have only been using a product for a short time rate a product significantly good/bad if only 5 out of 100 respondents placed themselves in this category of usage frequency. As a lower limit, it was determined that a category under consideration must have at least  $N = 10$  records. This is based on the fact that an average of 94 data records were collected per survey and a quantity threshold of 10% was set, which is established in statistical research. If the lower limit was not exceeded, the results were not interpreted and are shown in italics in the corresponding tables and diagonally patterned in the bars of the corresponding figures.

In the following, the results of the ratings are presented first, followed by the analyses of the demographic factor of gender as well as by the external factors usage frequency, knowledge, and duration of use. The results serve as the basis for answering the first research question *RQ 1: Are there external factors besides the classic UX factors that influence the perceived user experience of a product and assist in explaining the differences in UX ratings?*

### 5.1 Rating of the Products

First, the ratings of the products, in general, are to be discussed. For this purpose, the overall UX ratings of the individual questionnaires must be compared. To facilitate the comparison with the SUS and UMUX-LITE scales (from 0 to 100), the UEQ-S ratings (from  $-3$  to  $+3$ ) were converted into percentages. This involved using a simple percentage calculation by scaling the values to 0–6, then multiplying by 100 and dividing by 6. The corresponding scaled scores are shown in Table 2.

**Table 2.** Scale values. Range 0–100. The UEQ-S scores were converted for better comparability. The UEQ-S measures the PQ and HQ in their own scales, which are shown separately here in the two lower lines.

	Netflix	PPT	Zoom	BBB	PayPal	TikTok
UMUX-LITE	80.67	72.28	77.85	67.54	83.86	76.90
SUS	82.89	70.67	76.81	70.36	78.25	75.68
UEQ-S	67.00	54.17	64.00	56.67	65.50	70.83
UEQ-S PQ Scale	70.17	66.33	75.17	68.17	76.83	69.67
UEQ-S HQ Scale	63.67	42.00	52.83	45.17	54.17	72.00

### Pragmatic Quality

Studies have shown with the help of high correlations that SUS, UMUX-LITE, and UEQ-S PQ Scale all measure a similar concept (Schrepp, Kollmorgen & Thomaschewski, 2023). This is also visible in Table 2 since the ratings show only minor differences. These results become clear in a summary ordering of the product ratings between the three questionnaires. In descending order were evaluated (see Table 2):

- UMUX-LITE: PayPal, Netflix, Zoom, TikTok, PPT, BBB
- SUS: Netflix, PayPal, Zoom, TikTok, PPT, BBB
- UEQ-S PQ Scale: PayPal, Zoom, Netflix, TikTok, BBB, PPT

That is why a comparison of the products can be made here first.

It is visible from Table 2 that PayPal is rated highest in terms of pragmatic quality for UMUX-LITE and UEQ-S. This may be related to the product's very strong focus on pragmatic quality in the online banking category (see also Meiners et al., 2021).

PPT and BBB, on the other hand, are rated the worst. It appears that Microsoft PowerPoint is perceived as too complex to effectively achieve goals. This is indicated by 21 out of the 37 open responses to the survey on PPT, which noted that the software's many different functions are overly extensive, complicated, or illogical. For instance, creating customized slide designs was mentioned as a particularly challenging aspect of using the software. Concerning BigBlueButton, it is often specified for use in the work/education environment, which reinforces the pragmatic focus, making users more critical in this regard. As a result, 8 out of 19 open responses to the BBB survey cited the absence of certain functions, such as the ability to control user volume, as the reason for their dissatisfaction.

### Hedonic Quality

However, if we look at the UEQ-S HQ Scale, we get a different picture:

- UEQ-S PQ Scale: TikTok, Netflix, PayPal, Zoom, BBB, PPT

Looking at the hedonic quality (UEQ-S HQ Scale, Table 2), it is clear that TikTok is rated by far the best, followed by Netflix. As explained, both products are expected to have a hedonic focus, since they are used voluntarily in leisure time and are rarely prescribed by other people such as employers.

This is also a common outcome observed when evaluating HQ. The UEQ-S assesses the level of enjoyment and novelty associated with a product. However, since the products except for TikTok examined in the study have been available on the market for some time, they are considered less novel. As a result, design revisions are frequently employed in practice. This is one of the reasons why the HQ scores are notably lower in direct comparison to the PQ scores. This trend is also discussed in Sect. 6.

Even if one must not overinterpret these results, the influence of hedonic quality is evident in the corresponding products.

## 5.2 Impact of Gender

The initial analysis examines whether gender influences the ratings of SUS, UMUX-LITE, or UEQ-S scales for the six products. However, it is worth noting that there were overall only six self-identified diverse participants and three respondents who chose the “No answer” option, resulting in insufficient data to produce meaningful results for these categories. As a result, the focus will be on comparing the ratings between male and female participants. Table 3 displays the percentage proportions of male and female participants for all six product evaluations.

**Table 3.** Distribution of male and female participants.

Gender	Netflix N = 97	PPT N = 89	Zoom N = 76	BBB N = 76	PayPal N = 111	TikTok N = 114
Male	55%	74%	50%	54%	46%	47%
Female	43%	26%	47%	45%	51%	49%

Table 4 presents the values of the three UX questionnaires categorized by gender. Regarding the UEQ-S, the overall value is being used, which means the pragmatic and hedonic qualities are not considered separately at the moment.

The variation in ratings between Zoom and BBB is intriguing. Even though both products are in the same category and cater to similar use cases, there is a significant difference in the way females and males rate them across all three UX scales. Females rate Zoom much higher than males (see Table 4), whereas no such trend can be observed for BBB. It is possible that this is because BBB is predominantly used in an educational setting, whereas Zoom is a more versatile video conferencing tool that is employed for both personal and professional communication.

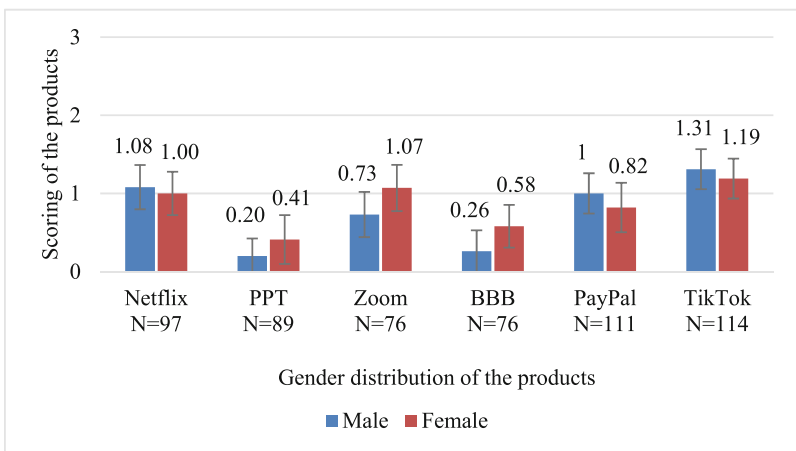
The gender of the participants had a statistically significant influence (ANOVA,  $p < .05$ ) for Zoom on all three questionnaires. Female participants tended to rate Zoom better than male participants. For the other five investigated products Netflix, PPT, BBB, PayPal, and TikTok an analysis of variance (ANOVA,  $p < .05$ ) showed that there is no statistically significant influence of gender on the scores. Detailed results can be found in the protocol (Kollmorgen, Schrepp & Thomaschewski, 2023).

Figure 2 illustrates the UEQ-S scores (from Table 4) categorized by gender. As can be seen, there are only small differences between the gender ratings. This difference is

**Table 4.** Impact of gender on the 3 UX scales. Range 0–100 for UMUX-LITE and SUS, from –3 to +3 for UEQ-S.

Questionnaire	Gen-der	Netflix N = 97	PPT N = 89	Zoom N = 76	BBB N = 76	PayPal N = 111	TikTok N = 114
UMUX-LITE	M	81.90	72.22	75.66	66.87	85.74	78.09
	F	80.20	72.46	82.64	68.38	82.12	75.15
SUS	M	84.40	69.62	73.36	69.82	79.13	78.19
	F	81.90	73.70	82.64	70.81	77.75	73.09
UEQ-S	M	1.08	0.20	0.73	0.26	1.00	1.31
	F	1.00	0.41	1.07	0.58	0.82	1.19

only significant for *Zoom*, but there is a slight tendency that female participants give higher ratings, except for *Netflix* (this is true for all three questionnaires). Therefore, due to the medium sample sizes, it cannot be ruled out that there is no effect of gender on the ratings, but in each case, the effect is quite small.

**Fig. 2.** Influence of gender on the UEQ-S scores. Range from –3 to +3.

### 5.3 Impact of Usage Frequency

As already explained, the perception and evaluation of UX can be influenced by the frequency of usage. When users actively engage with the product being evaluated more often, they are more likely to identify its features, advantages, and disadvantages. In addition, users may also adapt their behaviour to avoid known usability issues, which could be overlooked during their product evaluation, so that frequent users may rate the product better than non-frequent users.



Table 5 displays the percentage distribution of usage frequency across products, where the percentage distribution for Zoom, for instance, is already established by its product type: it targets students who have predetermined times during the course of modules in which they use the product.

**Table 5.** Distribution of usage frequency.

Usage frequency	Netflix N = 97	PPT N = 89	Zoom N = 76	BBB N = 76	PayPal N = 111	TikTok N = 114
Not very freq.	9%	58%	25%	41%	13%	19%
Sev. times a month	36%	31%	33%	29%	58%	11%
Sev. times a week	38%	8%	32%	28%	26%	16%
Daily basis	16%	2%	4%	3%	3%	48%

Table 6 further examines usage frequency and displays the values for the three questionnaires. Usage frequencies with fewer than 10 participants are shown in italics.

As observed, the more frequently a product in these categories is used, the better the UX score in the questionnaires is. This correlation is not surprising, as good UX tends to result in increased usage frequency, and over time, users with more frequent product usage are likely to have a better impression.

An ANOVA ( $p < .05$ ) showed that the frequency of usage had a significant impact on the SUS scores for Netflix, Zoom, and TikTok. In addition, a significant impact on the UMUX-LITE scores for Zoom, BBB, and TikTok as well as for the UEQ-S scores for Netflix and TikTok could be found.

Figure 3 depicts the SUS scores (from Table 6) for the six products investigated, in relation to self-reported usage frequency. Usage frequencies with fewer than 10 participants are shown diagonally patterned. Many of the differences in scores are relatively high i.e., the impact on usage frequency on the scale scores also leads to meaningful differences. It is noteworthy that Netflix, Zoom, and PayPal are consistently rated higher than PPT, BBB, and TikTok in all usage frequency categories.

## 5.4 Impact of Knowledge

It is also possible that experience with the range of products being evaluated could impact the evaluation process. Similar to increased usage frequency, greater knowledge of the products may lead to a clearer identification of their advantages and disadvantages.

Table 7 displays the percentage distribution of self-reported knowledge among the participants.

Consistent with previous observations, it became evident that Netflix, Zoom, and PayPal receive better ratings overall. However, TikTok clearly stands out in terms of this external factor, scoring the best overall, except for users who have little knowledge of the product. PPT and BBB again achieve the worst UX ratings.

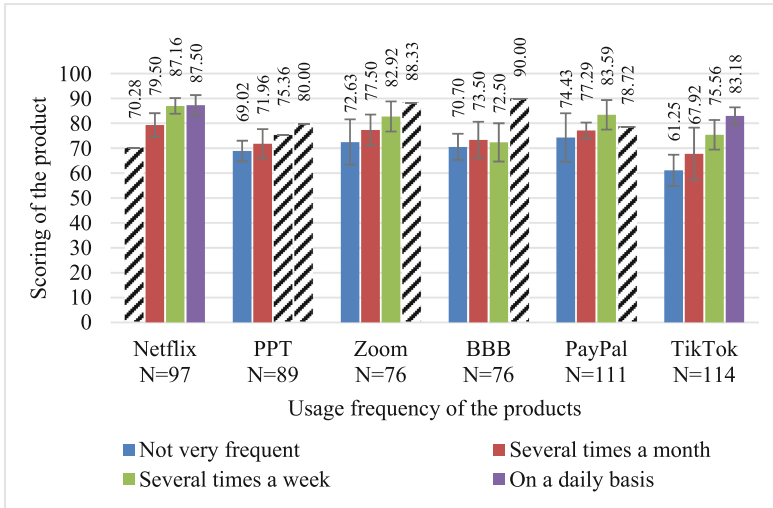
**Table 6.** Impact of usage frequency. Range 0–100 for UMUX-LITE, and SUS; from –3 to + 3 for UEQ-S. Usage frequencies with fewer than N = 10 participants are shown in italics.

Questionnaire	Frequency	Netflix N = 97	PPT N = 89	Zoom N = 76	BBB N = 76	PayPal N = 111	TikTok N = 114
UMUX-LITE	Not very freq.	74.07	70.07	75.44	66.15	77.65	50.00
	Sev. tim. month	79.29	72.02	78.67	73.75	81.25	68.06
	Sev. tim. week	83.78	<i>82.14</i>	83.68	75.69	91.15	82.87
	Daily basis	80.21	<i>83.34</i>	<i>88.89</i>	<i>83.33</i>	<i>84.57</i>	89.55
SUS	Not very freq.	70.28	69.02	72.63	70.70	74.43	61.25
	Sev. tim. month	79.50	71.96	77.50	73.50	77.29	67.92
	Sev. tim. week	87.16	<i>75.36</i>	82.92	72.50	83.59	75.56
	Daily basis	87.50	<i>80.00</i>	<i>88.33</i>	<i>90.00</i>	<i>78.72</i>	83.18
UEQ-S	Not very freq.	<i>-1.88</i>	0.18	0.68	0.35	0.7	0.25
	Sev. tim. month	0.95	0.26	1.08	0.62	0.69	1.27
	Sev. tim. week	1.20	<i>0.54</i>	0.82	0.69	1.38	1.22
	Daily basis	1.38	<i>0.81</i>	<i>1.54</i>	<i>1.00</i>	<i>0.99</i>	1.76
UEQ-S	Not very freq.	<i>0.31</i>	0.89	1.25	1.04	1.10	0.26
	Sev. tim. month	1.08	0.97	1.70	1.19	1.29	1.17
PQ scale	Sev. tim. week	1.43	<i>1.39</i>	1.66	1.40	2.17	1.01
	Daily basis	1.53	<i>2.00</i>	<i>2.42</i>	<i>2.00</i>	<i>1.76</i>	1.63
UEQ-S	Not very freq.	<i>-0.56</i>	<i>-0.53</i>	0.12	<i>-0.34</i>	0.41	0.24
	Sev. tim. month	0.83	<i>-0.46</i>	0.47	0.05	0.08	1.38
HQ scale	Sev. tim. week	0.97	<i>-0.32</i>	<i>-0.02</i>	<i>-0.02</i>	0.59	1.42
	Daily basis	1.22	<i>-0.38</i>	<i>0.67</i>	<i>0.00</i>	<i>0.22</i>	1.89

Table 8 displays the UX ratings for the three questionnaires based on the reported knowledge of the products, which was the basis for the calculations. Knowledge with fewer than 10 participants is shown in italics.

Additionally, the assumption that greater experience with the products can lead to better evaluations is further supported. On average, participants rated the products more positively when they reported greater knowledge of them. This trend is clearly visible across all three questionnaires. Figure 4 also displays a graphical representation of this trend for UEQS (from Table 8). Usage frequencies with fewer than 10 participants are shown diagonally patterned.

With the help of various ANOVA analyses ( $p < .05$ ), statistically significant influences were also found for the external factor knowledge. For the UMUX-LITE there is, except for Netflix and PayPal, a significant impact of knowledge on the scores. For SUS the impact is significant, except for PPT and PayPal. For the UEQ-S there is only a significant impact of knowledge observed for TikTok.



**Fig. 3.** Influence of usage frequency on the SUS scores. Range 0–100. Usage frequencies with fewer than 10 participants are shown diagonally patterned.

**Table 7.** Distribution of self-reported product knowledge.

Knowledge	Netflix N = 97	PPT N = 89	Zoom N = 76	BBB N = 76	PayPal N = 111	TikTok N = 114
Low	7%	9%	20%	25%	12%	20%
Medium	22%	51%	41%	45%	59%	33%
High	54%	35%	34%	20%	23%	33%
Excellent	19%	6%	5%	1%	5%	13%

**5.5 Impact of Duration of Use**

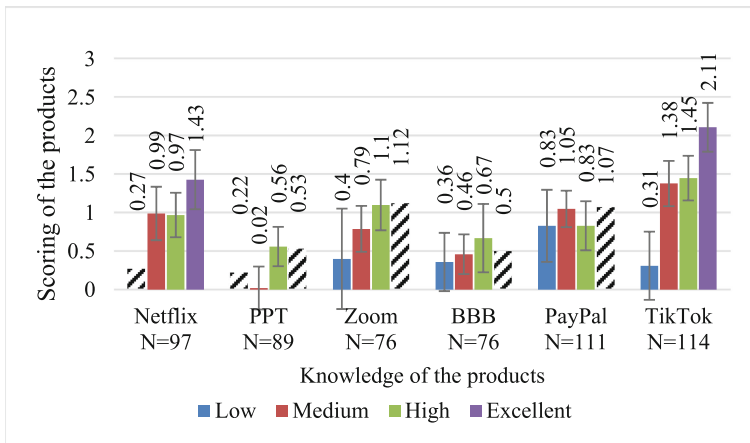
It is reasonable to assume that users who have been using a product in these categories for an extended period may have a better understanding of it. This does not necessarily imply that they know all of the product’s functions and can operate it flawlessly, but rather that they can navigate it based on their needs. Conversely, users who have only used a product for a short time may struggle to achieve their goals. It is necessary to investigate the impact of usage duration on the ratings.

For the duration of use (see Table 9) there is for most products one category that clearly dominates. Thus, it is not really a surprise that there is for most products no significant impact of this variable on the scores. An exception is TikTok, where the duration of use indeed significantly influenced the questionnaire scores according to an ANOVA ( $p < .05$ ).

The distribution of responses in terms of duration of use is shown in Table 9. It is visible that most categories contain fewer than 10 participants. Therefore, no further

**Table 8.** Distribution of self-reported product knowledge. Range 0–100 for UMUX-LITE, and SUS; from –3 to +3 for UEQ-S. Knowledge with fewer than N = 10 participants is shown in italics.

Questionnaire	Know-ledge	Netflix N = 97	PPT N = 89	Zoom N = 76	BBB N = 76	PayPal N = 111	TikTok N = 114
UMUX-LITE	Low	75.00	66.63	67.78	61.84	82.25	50.00
	Medium	78.97	68.15	76.61	73.04	82.64	79.95
	High	79.65	77.69	83.33	76.67	82.62	84.68
	Excellent	87.50	86.67	95.84	83.33	91.11	98.33
SUS	Low	77.08	63.12	68.33	66.32	74.89	62.07
	Medium	77.74	68.56	74.84	73.01	79.24	79.66
	High	82.36	74.03	82.98	77.67	79.18	76.28
	Excellent	92.36	81.00	91.25	80.00	77.98	89.17
UEQ-S	Low	0.27	0.22	0.40	0.36	0.83	0.31
	Medium	0.99	0.02	0.79	0.46	1.05	1.38
	High	0.97	0.56	1.10	0.67	0.83	1.45
	Excellent	1.43	0.53	1.12	0.50	1.07	2.11



**Fig. 4.** Influence of knowledge on the UEQ-S scores. Range from –3 to +3. Knowledge with fewer than 10 participants is shown diagonally patterned.

statements about the data are made. The complete data can be found in the protocol (Kollmorgen, Schrepp & Thomaschewski, 2023).

**Table 9.** Distribution of duration of use.

Duration of use	Netflix N = 97	PPT N = 89	Zoom N = 76	BBB N = 76	PayPal N = 111	TikTok N = 114
Shorter	2%	1%	8%	39%	3%	18%
More than a year	64%	10%	88%	55%	51%	66%
More than 5 years	33%	88%	3%	0%	46%	6%

## 6 Discussion

In the following, the results are discussed in order to answer the two research questions.

- *RQ1: Are there external factors besides the classic UX factors that influence the perceived user experience of a product and assist in explaining the differences in UX ratings?*
- *RQ 2: To what extent are the pragmatic as well as the hedonic quality of products influenced by the external factors mentioned above?*

### 6.1 External Influencing Factors

To answer the first research question, the results of the two studies must first be considered in terms of the influence of external factors. Starting with **gender**, the results did not show a significant impact on the UX scale scores except for Zoom. Nevertheless, women tended to rate the products better in all surveys.

In contrast, significant influences of the **usage frequency** on the perceived UX could be demonstrated. Especially for the products Netflix, Zoom, and TikTok, influences by the usage frequency were found in different ways with the three questionnaires. This shows the affirmation that the more often a product is used, the better the perceived user experience is, and vice versa.

Influences were also shown regarding experience with the product (**knowledge**). Here, however, differences were more pronounced. Thus, all three questionnaires for TikTok found significant influence by the knowledge on the UX ratings. This may be related to the fact that TikTok is the youngest of the six products surveyed and focuses on innovation, which means that new functions are regularly provided by the social media platform. Accordingly, with a better experience with TikTok, more benefits can be perceived in the UX. For PayPal, in contrast, no significant influence of knowledge was shown in the scores of the questionnaires. This may be related to the nature of the product. As an online payment service, PayPal only offers limited and intuitive functionalities. Thus, there is not much to learn, and an increasing experience may not cause better usability impressions. No clear trends emerged for the other products. For Zoom and BBB, significant influences by knowledge were found on ratings in the usability-focused questionnaires SUS and UMUX-LITE. For Netflix, significant results were found only in the SUS, and for PPT only in the UMUX-LITE. This may be due to the sample size as well as the selected target group (students).

No meaningful results could be obtained for the *duration of use* in this study. This is due to the varying extent of data records per duration category. The distribution of responses to the duration of use question corresponds to the maturity of the product. Microsoft PowerPoint and PayPal have been on the market the longest, which is why they show more data sets in the “More than 5 years” category. In contrast, BBB and Zoom have only gained popularity in the last few years, mainly due to the COVID-19 pandemic and the shift of work as well as social life to digital. The social network TikTok was also released only a few years ago and the study, therefore, shows only a few respondents who have been using the product since almost the beginning.

As anticipated, however, our study’s findings show that usage frequency and knowledge are likely to have an impact on ratings, but that this impact depends on the concrete product. Although the majority of cases did not show significant effects for the other two factors, this could be attributed to the limited sample size and uneven distribution of participants in different categories. However, there was a discernible trend in the data.

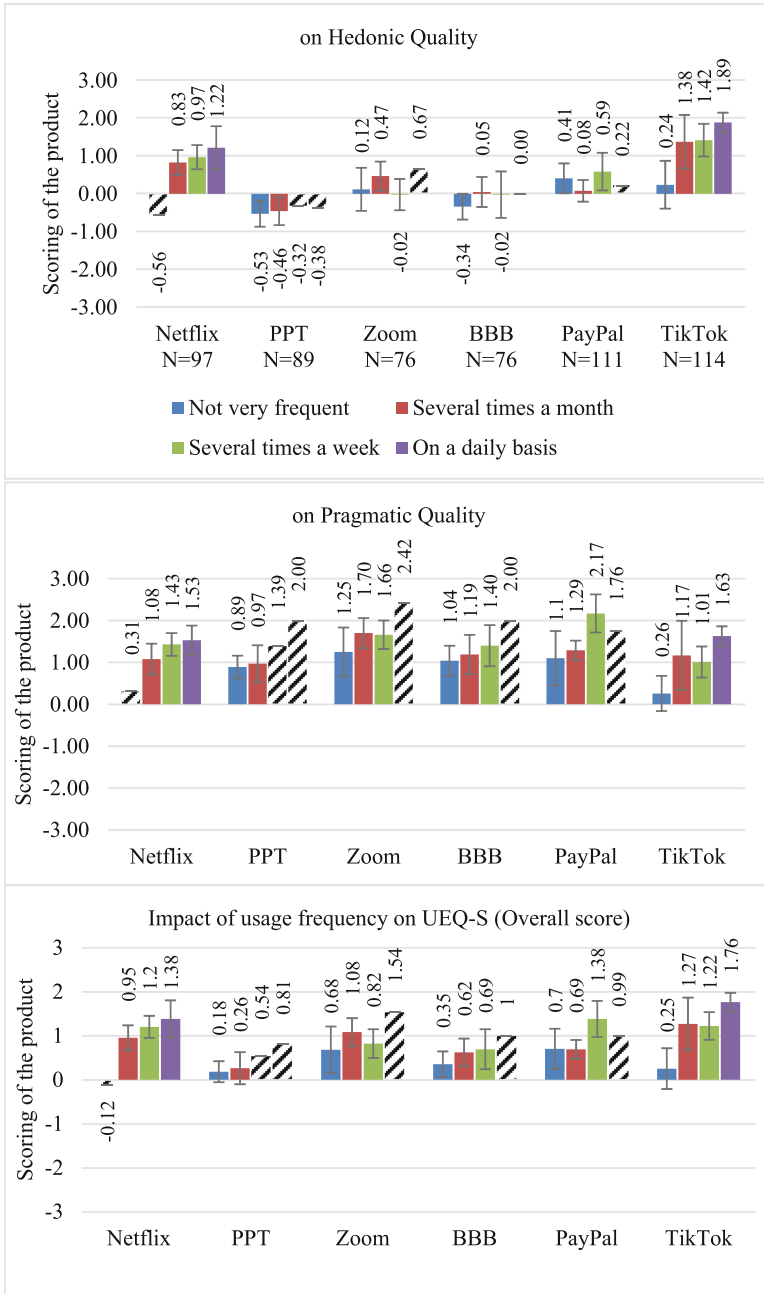
In answer to the first research question, it can therefore be stated that there are definitely external factors that influence the perceived UX of products. Usage frequency could be proven as such an influencing factor, and trends were also visible for the experience with the product. Through these findings, another explanation for the differences in UX ratings of the same products can be confirmed.

## 6.2 Influences on Pragmatic and Hedonic Qualities

Finally, the identified significant influences of the external factors with respect to both the pragmatic and hedonic quality of the products are considered in particular to answer the second research question *RQ2: To what extent are the pragmatic as well as the hedonic quality of products influenced by the external factors mentioned above?*

Section 5.3 presented evidence that *usage frequency* significantly impacts product ratings, which is once again summarized in Fig. 5 ( $N < 10$  patterned). Consequently, this external factor is examined once more in the context of pragmatic and hedonic quality, which is the focus of Fig. 5 (according to Table 6). The figure reveals a trend that is consistently visible with the overall UEQ-S ratings, particularly for pragmatic quality. This trend was also significantly observed concerning the PQ for Netflix, Zoom, and TikTok and the HQ for Netflix, BBB, and TikTok in the ANOVA tests ( $p < .05$ ). The detailed information can be found in the Research Protocol (Kollmorgen, Schrepp & Thomaschewski, 2023).

However, some discussions of specific impacts on the pragmatic and hedonic quality can also be made with respect to the external factor knowledge. Significant impacts were found for the hedonic-focused product TikTok. The reasons for this are, as explained, that TikTok, as a young product, relies on users gradually becoming familiar with the innovative methods. If looked more closely at the results of the ANOVA, this is also reflected in the PQ/HQ scales of the UEQ-S. For TikTok, significant influences were found for both the PQ and the HQ. This speaks for the innovativeness of the product. As users become more familiar with TikTok, they learn the functions they need to fulfil (recreationally designed) their goals. At the same time, using the social network appeals to them more when they can better understand the functions. On the other hand, the novel functions of the platform seem overloaded for inexperienced users. Thus, people



**Fig. 5.** Influence of usage frequency on the UEQ-S overall, UEQ-S PQ and UEQ-S HQ scores. Usage frequencies with fewer than 10 participants are shown diagonally patterned.

gave the worst ratings of perceived UX for TikTok when they rated their knowledge as poor.

For the products BBB and Zoom, which both cover hedonic needs in addition to pragmatic needs similar to TikTok, significant influences were found in two of three questionnaires. Both had a significant influence of the knowledge on the PQ, which can be justified by the prescriptiveness of the use in the professional environment and thus the more pragmatic focus of the persons surveyed.

For the pragmatically focused PayPal, on the other hand, no influences were found. As explained, one reason for this could be that PayPal cannot fulfil all the criteria for an online banking tool. For the pragmatically focused product PowerPoint, too, a significant influence was only found in one of the three questionnaires.

Overall, in answer to the second research question, this chapter demonstrated the extent to which the external factors can influence the pragmatic and hedonic quality of the products. It became clear that the external factors only influence or can influence specific UX aspects, but not all of them.

## 7 Summary and Future Work

In this paper, it is argued that there are differences in the UX ratings of certain products that cannot be explained solely by their membership in different product categories. The resulting research question (*RQ 1: Are there external factors besides the classic UX factors that influence the perceived user experience of a product and assist in explaining the differences in UX ratings?*) was answered by conducting studies on a total of six products of different product categories using three short questionnaires. In these, questions on demographic and usage behavioural factors were asked to be able to examine correlations with the aid of analyses. The products examined in the studies, Netflix (HQ focus), TikTok (HQ focus), BigBlueButton (PQ/HQ focus), Zoom (PQ/HQ focus), Microsoft PowerPoint (PQ focus), and PayPal (PQ focus), cover different usage scenarios (see Table 1) and the importance of pragmatic and hedonic qualities, therefore, differs among the products. This creates the possibility to determine the influence of external factors on both scales in particular. This formed the basis for answering the second research question (*RQ2: To what extent are the pragmatic as well as the hedonic quality of products influenced by the external factors mentioned above?*).

With regard to the first research question, significant influences on the UX ratings could be determined for the usage behavioural factors usage frequency and knowledge. No conclusions could be drawn for the duration of use, as the distribution of responses was too heavily skewed towards one category in each case. In relation to the demographic factor gender, only an influence for the product Zoom was detected, so that at least a trend and possible influence of gender on specific products is visible. This represents a possible starting point for further correlation studies.

The external factors were then assessed to answer the second research question on the influence of pragmatic and hedonic quality. It was found that hedonic-focused products (TikTok, Netflix, BBB/Zoom) were rated better on average when users had a higher usage frequency. This is understandable since the products are used for leisure and thus are not prescribed by anyone. Conversely, they were rated worse by those who used the



products less. However, the situation is different for the pragmatically focused products (PPT, PayPal). While PPT is very complex and extensive, PayPal does not fulfil all functions of the product category online banking, which means that for both products no significant influence by the usage frequency was found. The same applies to the external factor knowledge. For hedonically focused products (TikTok, partially BBB/Zoom), significant influences on the perceived UX could be detected. For pragmatically focused products (PayPal, PPT), on the other hand, no correlation was found.

The study can therefore be seen on the one hand as a recommendation to consider not only the purely product-specific UX factors but also the usage behaviour in the success concept of products. It should also be noted that the PQ should not be neglected for hedonically focused products and the HQ for pragmatically focused products.

The research has indicated that the choice of the measuring instrument is crucial in drawing accurate conclusions from the results. For instance, if a product's hedonic quality is a critical success factor, then it is imperative to use a dedicated scale to measure it. When a usability-focused method like SUS or UMUX-LITE is utilized, it may not be possible to identify variations in hedonic quality within the results.

Finally, an outlook on future work can be given. Firstly, the number of respondents available for our study was due to the division into the respective categories of the influencing factors relatively low. This is particularly problematic because the respondents were not evenly distributed across all categories of the influencing factors investigated. Therefore, some of the results are based on a small number of respondents, and they need to be corroborated with a more extensive range of products. Therefore, larger sample sizes will be relevant for future work.

Also, an extension of the range of products considered can shed even more light on the influence of external factors on the perceived UX of certain products and product categories as well as on the PQ and HQ in particular. Here, products of the same product categories (e.g., Amazon Prime as an alternative to Netflix, Instagram as an alternative to TikTok) would be conceivable.

## References

1. Aufderhaar, K., Schrepp, M., Thomaschewski, J.: Do women and men perceive user experience differently? *Int. J Interact. Multimedia Artif. Intell.* **5**, 63–67 (2019)
2. Braun, M., Chadowitz, R., Alt, F.: User experience of driver state visualizations: a look at demographics and personalities. In: *IFIP Conference on Human-Computer Interaction*, pp. 158–176. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-29390-1\\_9](https://doi.org/10.1007/978-3-030-29390-1_9)
3. Brooke, J.: SUS-a quick and dirty usability scale. In: Jordan, P., Thomas, B. (eds.) *Usability Evaluation in Industry 189(194)*, pp. 4–7. Taylor & Francis, London (1996)
4. Brooke, J.: SUS - A Retrospective. *J. Usability Stud.* **8**(2), 29–40 (2013)
5. Davis, F.: A technology acceptance model for empirically testing new end-user information systems - Theory and results. PhD Thesis, Massachusetts Inst. of Technology (1986)
6. Devaraj, S., Easley, R.F., Crant, J.M.: Research note - How does personality matter? Relating the five-factor model to technology acceptance and use. *Inf. Syst. Res.* **19**(1), 93–105 (2008)
7. Finstad, K.: The Usability Metric for User Experience. *Interact. Comput.* **22**(5), 323–327 (2010). <https://doi.org/10.1016/j.intcom.2010.04.004>

8. Hassenzahl, M.: The effect of perceived hedonic quality on product appealingness. *Int. J. Hum.-Comput. Interact.* **13**(4), 481–499 (2001). [https://doi.org/10.1207/S15327590IJHC1304\\_07](https://doi.org/10.1207/S15327590IJHC1304_07)
9. Hassenzahl, M.: Towards an experiential perspective on product quality. In: Proceedings of the 20th International Conference of the Association Francophone d'Interaction Homme-Machine on - IHM '08, Metz, France: Association for Computing Machinery, pp. 11–15 (2008). <https://doi.org/10.1145/1512714.1512717>
10. Hassenzahl, M., Diefenbach, S., Göritz, A.: Needs, affect, and interactive products - Facets of user experience. *Interact. Comput.* **22**(5), 353–362 (2010). <https://doi.org/10.1016/j.intcom.2010.04.002>
11. John, O.P., Srivastava, S.: The big five trait taxonomy: history, measurement, and theoretical perspectives. In: Pervin, L.A., John, O.P. (eds.) *Handbook of personality: Theory and research*, pp. 102–138. Guilford Press (1999)
12. Kollmorgen, J., Meiners, A.-L., Schrepp, M., Thomaschewski, J.: Ermittlung relevanter UX-Faktoren je Produktkategorie für den UEQ+. In Wienrich, C., Wintersberger, P. and Weyers, B. (Ed.). *Mensch und Computer 2021 – Workshopband*, Bonn: Gesellschaft für Informatik e.V (2021). <https://doi.org/10.18420/muc2021-mci-ws01-362>
13. Kollmorgen, J., Schrepp, M., Thomaschewski, J.: impact of usage behavior on the user experience of netflix, microsoft powerpoint, bigbluebutton and zoom. In: Proceedings of the 18th International Conference on Web Information Systems and Technologies – WEBIST, Valletta, Malta, pp. 397–406, 2022 (2022a). <https://doi.org/10.5220/0011380100003318>
14. Kollmorgen, J., Schrepp, M., Thomaschewski, J.: Protocol for a comparison of three short user experience questionnaires (2022b). <https://doi.org/10.13140/RG.2.2.32773.01760>
15. Kollmorgen, J., Schrepp, M., Thomaschewski, J.: Protocol for the Influence of demographic variables and usage behavior on the perceived user experience protocol - version 0.1/2023 (2023). <https://doi.org/10.13140/RG.2.2.17138.38087>
16. Kortum, P., Oswald, F.L.: The impact of personality on the subjective assessment of usability. *Int. J. Hum.-Comput. Interact.* **34**(2), 177–186 (2018)
17. Laugwitz, B., Held, T., Schrepp, M.: Construction and evaluation of a user experience questionnaire. In: Holzinger, A. (ed.) *HCI and Usability for Education and Work*, pp. 63–76. Springer Berlin Heidelberg, Berlin, Heidelberg (2008). [https://doi.org/10.1007/978-3-540-89350-9\\_6](https://doi.org/10.1007/978-3-540-89350-9_6)
18. Lewis, J.R.: The system usability scale: past, present, and future. *Int. J. Hum.-Comput. Interact.* **34**(7), 577–590 (2018). <https://doi.org/10.1080/10447318.2018.1455307>
19. Liapis, A., Katsanos, C., Xenos, M., Orphanoudakis, T.: Effect of personality traits on UX evaluation metrics: a study on usability issues, valence-arousal and skin conductance. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–6 (2019)
20. McCrae, R.R., John, O.P.: An introduction to the five-factor model and its applications. *J. Pers.* **60**(2), 175–215 (1992)
21. McLellan, S., Muddimer, A., Peres, S.C.: The effect of experience on system usability scale ratings. *J. Usability Stud.* **7**(2), 56–67 (2012)
22. Meiners, A.-L., Kollmorgen, J., Schrepp, M., Thomaschewski, J.: Which UX aspects are important for a software product? In Schneegass, S., Pfleging, B., Kern, D. (Eds.). *Mensch und Computer 2021. MuC '21: Mensch und Computer 2021*. Ingolstadt Germany, 05 09 2021 08 09 2021. New York, NY, USA: ACM. pp. 136–139 (2021). <https://doi.org/10.1145/3473856.3473997>
23. Rummel, B., Schrepp, M.: UX-Fragebögen: was steckt in der Varianz?. In Dachselt, R., Weber, G. (Eds.). *Mensch und Computer 2018 – Workshopband*, Bonn: Gesellschaft für Informatik e.V (2018)

24. Schrepp, M., Hinderks, A., Thomaschewski, J.: Design and evaluation of a short version of the user experience questionnaire (UEQ-S). *Int. J. Interact. Multimedia Artif. Intell.* **4**(6), 103–108 (2017). <https://doi.org/10.9781/ijimai.2017.09.001>
25. Schrepp, M. (2021a). *User Experience Questionnaires: How to use questionnaires to measure the user experience of your products? KDP*. ISBN-13: 979–8736459766
26. Schrepp, M. (2021b). Measuring User Experience with Modular Questionnaires. In *International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Jakarta, Indonesia. DOI: <https://doi.org/10.1109/ICACSIS53237.2021.9631321>
27. Schrepp, M., Kollmorgen, J., Thomaschewski, J.: A Comparison of SUS, UMUX-LITE, and UEQ-S. In *Journal of User Experience* **18**(2), pp. 86–104. DOI
28. Schrepp, M., et al.: On the Importance of UX Quality Aspects for Different Product Categories. *Int. J. Interact Multimedia Artif. Intell.* (2023, (in press)). <https://doi.org/10.9781/ijimai.2023.03.001>
29. Winter, D., Hinderks, A., Schrepp, M., Thomaschewski, J.: Welche UX-Faktoren sind für mein Produkt wichtig? In Hess, S., Fischer, H. (Ed.). *Mensch und Computer 2017 – Usability Professionals*, Regensburg: Gesellschaft für Informatik e.V., pp. 191–200 (2017). <https://doi.org/10.18420/muc2017-up-0002>



# Categorizing UX Aspects for Voice User Interfaces Using the Kano Model

Kristina Kölln<sup>1,2</sup> , Andreas M. Klein<sup>1,2</sup>  , Jana Deutschländer<sup>1</sup> ,  
Dominique Winter<sup>3</sup> , and Maria Rauschenberger<sup>1</sup> 

<sup>1</sup> Faculty of Technology, University of Applied Sciences  
Emden/Leer, Emden, Germany

{kristina.koelln, maria.rauschenberger}@hs-emden-leer.de

<sup>2</sup> Department of Computer Languages and Systems,  
University of Seville, Seville, Spain

{andreas.klein, jana.deutschlaender}@ux-researchgroup.com

<sup>3</sup> University of Siegen, Siegen, Germany  
dominique.winter@designik.de

**Abstract.** Although voice user interfaces (VUIs) are widely available, they currently face challenges such as low adoption rates and user concerns. Users assess products through user experience (UX) aspects. Thus, knowing UX aspects for VUIs and their prioritization will improve UX and reduce challenges. In this study, we use a user-centered mixed-methods approach to identify and prioritize UX aspects of VUIs. Thereby, we identified 32 VUI UX aspects from the perspective of intensive users. We then applied the Kano model to categorize these UX aspects for which we analyzed  $N = 195$  VUI users. One thing we found was that 21 VUI UX aspects are distinctively prioritized, such as privacy, data security, and ad-free as *must-be*, and simplicity, comprehension, and error-free as *one-dimensional*. These findings can help VUI developers to prioritize specific UX aspects according to their target group's needs, enabling them to create VUIs that benefit and excite their users.

**Keywords:** Voice user interface · User Experience · Kano · Voice assistants · UX · Prioritization · Mixed methods · Human-centered design

## 1 Introduction

Considering user experience (UX) [11] can help product designers to overcome challenges such as low adoption rates or user concerns by achieving *acceptance by design* [33]. The human-centered design (HCD) framework has become widely accepted as a means to develop products with a positive UX. HCD is a holistic approach that focuses on the user to design products with a UX that fits the target group [11]. For the UX assessment of a product, many different UX aspects play a role (*e.g.*, perspicuity, efficiency, or dependability). The prioritization of

the UX aspects will vary based on the product and the individual user [34]. To develop products that, say, excite users, it is important to know which UX aspects are relevant for each product. For example, in our research, we focus on Voice User Interfaces (VUIs).

We define VUIs as any kind of software and device combination controlled by the user's spoken input [20]. VUIs have become increasingly popular in recent years, as many devices now come with built-in voice control (*e.g.*, smartphones), and users appreciate the comfort that comes with their use [20]. However, although many people own a VUI, the adoption rate is still rather low [39]. There are various possible reasons for this. For example, some users are concerned about which data is collected and how, while others mention the need for better understanding of commands [18,33]. To overcome such challenges, we need to know which UX aspects are relevant for VUI users and how to prioritize them. Recent research has included several attempts to define important UX aspects of VUIs using an expert-driven process [8,16,19]. We plan to identify relevant VUI UX aspects from the users' point of view, as the HCD framework suggests. For the prioritization of product qualities such as UX aspects, the Kano model has become established in the UX research community (*e.g.*, [5,24,31,38]). The Kano model is utilized to enable the target group to categorize individual aspects and determine their prioritization [12].

In this article, we present the identified UX aspects using a user-centered mixed-methods approach [11,27] (*i.e.*, a combination of qualitative and quantitative user studies). We concentrate on intensive users for the identification of the UX aspects because they can offer profound insights due to their extensive usage. We then use the Kano model to prioritize these UX aspects based on the VUI users' opinions of our survey with participants ( $N = 195$ ). Thus, we can provide VUI developers with a detailed understanding of the users' needs and wants, thereby enabling them to develop VUIs that fit the context and users.

This article is structured as follows: Sect. 2 introduces the background and related work of VUI UX research, mixed methods, and the Kano model. Section 3 explains the mixed-methods process for identifying the UX aspects as well as the process for categorizing the identified UX aspects into the Kano model. Section 4 presents our results. Section 5 discusses the findings and provides a critical assessment of the limitations of our research. Section 6 concludes our research and provides an outlook on future work.

## 2 Background and Related Work

Speech intelligibility, correct command execution, data security, and privacy are among the current challenges when it comes to using VUIs [18,33,39]. In order to meet users' needs and overcome existing barriers and reservations towards the use of VUIs, evaluation is required [14]. UX assessment that considers specific UX aspects for VUIs is an essential evaluation method to achieve this. In the following, we briefly introduce UX, how to identify UX aspects for VUIs, and VUI assessment approaches and methods. We then present the categorization of Kano [12] as a methodology for prioritizing.

## 2.1 UX of VUIs

The concept of UX is a holistic one. It takes into account emotions, cognition, and physical actions before, during, and after using a product [11]. UX has a set of distinct quality criteria, including pragmatic (*i.e.*, classical usability criteria such as *efficiency*) and hedonic (*i.e.*, non-goal criteria such as *stimulation*) [30]. These UX quality criteria, also called UX aspects, can be identified and evaluated by empirical studies. Focusing on relevant UX aspects enables efficient product development and evaluation (*e.g.*, using the most suitable questionnaires) [42].

Still, there has not yet been a consensus on UX measurement specifically for VUIs [14,37]. Various methods are available for VUI evaluation, but they do not necessarily focus on UX [14]. For example, one study analyzed six questionnaires commonly used for VUI evaluation and assessed their suitability for various UX dimensions [19]. Its authors recommend combining questionnaires to cover UX more comprehensively or measuring a distinct UX dimension in detail. Another VUI evaluation method entails using heuristics, such as guidelines for design and evaluation. However, they tend to focus on usability and disregard certain UX aspects [23,41].

Another applicable approach for measuring different UX aspects for VUIs is the modular questionnaire concept UEQ+ [36]. Due to its modularity, this approach is very flexible, as researchers can, for example, combine three speech quality scales with three of the twenty UEQ+ scales currently available. That way, researchers can create a questionnaire related to their particular research question for product-specific UX aspect evaluation [17]. Examples of other UEQ+ scales are *attractiveness*, *novelty*, and *efficiency*. The voice quality scales are constructed with human-computer interaction (HCI) and the *VUI design process* in mind [16]. The user, system, and context significantly influence HCI [7]. Therefore, improving the *VUI design process* requires a deep understanding of the context, user, and application to define relevant evaluation criteria [3].

## 2.2 Mixed-Methods Approach

A mixed-methods study [27] combines detailed insights from qualitative research with broader insights from quantitative analysis. Thus, a mixed-methods study's results provide insights into the depth and breadth of a research question. In recent studies, mixed-methods approaches have become increasingly popular [27], as they provide certain advantages. For example, mixed methods can be applied in single questionnaire experiments if a questionnaire combines standardized and open questions [1]. A further example is the comprehensive study design [9], where the combined use of standardized questionnaires and semi-structured interviews allows researchers to cover broader aspects while gaining in-depth information.

## 2.3 Kano Categorization

The Kano model [12] is a concept in quality management theory that was developed in the 1980s. The concept describes the theory that there are objective and

subjective qualities, from which a model was derived for categorizing product qualities. Quality characteristics differ in their influence on user (dis)satisfaction. For example, some can increase user satisfaction, while others mainly prevent user dissatisfaction [43]. In practice, the model is especially used for requirements engineering. It is even part of the basic knowledge for certifications in requirements engineering [10].

Particularly relevant categories for quality characteristics are *attractive features* (i.e., not expected, but exciting when present), *one-dimensional features* (i.e., the more prevalent, the more satisfied the user), and *must-be features* (i.e., no influence on satisfaction when present, but frustrating when absent). In addition, there are *indifferent features* (i.e., neither their presence nor absence is significant) and *reverse features* (i.e., they frustrate the user when present). These two features are to be avoided to achieve a positive UX.

The categorization of the individual quality characteristic into the Kano model is done by answering two questions. In the first question, the participants are asked how they would find it if the VUI fulfilled a certain characteristic (functional question). In the second question, the participants are asked how they would find it if the characteristic were not fulfilled (dysfunctional question). The participants have five options to answer each question: “*I like it that way,*” “*It must be that way,*” “*I am neutral,*” “*I can live with it that way,*” and “*I dislike it that way.*” These options do not represent a scale; instead, they reflect different feelings of users. A feature is categorized as *questionable* if the answers to the questions are contradictory.

The Kano model is a way to categorize the qualities of a product. UX is a quality to be assigned to the product. For instance, the Kano model has been used to develop targeted UX design for older user groups [38]. In a case study, Kano categories were used for UX grading of use cases, features, and requirements within product development [31]. Furthermore, the model has been utilized to rate requirements for the interactive design of mobile applications and investigate the UX of live-stream sales [5,24].

### 3 Methodology

With a mixed-methods approach, we aim to identify the missing UX aspects that users consider when using VUIs and categorize them according to the Kano categories. We want to enable VUI developers to focus on the elements that are currently of importance to VUI users. Therefore, our aim is to answer the following research questions (RQ):

**RQ1:** What are intensive users’ UX aspects for VUIs?

**RQ2:** Which UX aspects are required by the users?

**RQ3:** Which UX aspects have the potential to excite the users?

First, we identify UX aspects for VUIs (RQ1) that are relevant for users and their VUI evaluation. For this, we concentrate on the target group of intensive users because of their extensive VUI knowledge and insights. Intensive VUI users

use VUIs regularly (*i.e.*, from daily to several times a week) in private or professional environments [18,20]. We utilize the Kano model to obtain user feedback regarding the prioritization of UX aspects as must-have (RQ2) or attractive (RQ3). For this, we do not filter for intensive users because all VUI users and non-users have their opinions when given a certain statement. Participants do not need more in-depth experience to answer the questions. VUI developers can use this knowledge to determine which UX aspects should be prioritized for their development goal or context.

### 3.1 Identification of UX Aspects for VUIs

To identify intensive users' UX aspects for VUIs, we conducted a mixed-methods study and analyzed the collected data with qualitative content analysis [20,26].

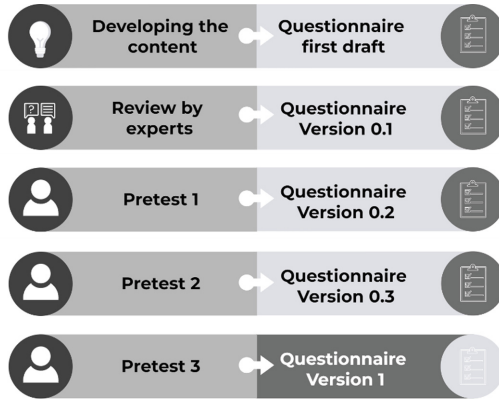
**Mixed-Methods Study.** In order to answer RQ1, we chose a user-centered mixed-methods study. We used a combination of semi-structured interviews and an online survey with a larger number of users to ask intensive users about their experiences.

For the interviews, we applied the semi-structured expert interview methodology [2]. We constructed our interview guidelines to answer RQ1. Thus, it consists of 19 qualitative questions about the participants' positive and negative expectations and experiences regarding VUIs as well as their contexts of use. The interview guidelines were constructed in German and English versions as well as an additional version to interview users whose children also use a VUI. The full interview guidelines are available in the research protocol of the mixed-methods study [21].

The interviews ( $N = 10$ ) were conducted and recorded from April to May 2021 via online video sessions using Microsoft Teams or, in one case, a phone call. Afterward, they were transcribed with a simple scientific transcript [4] and made anonymous. In two cases, we had to document the interviews with a memory log because the recording failed. All transcripts and memory logs are available in the research protocol of the mixed-methods study in their original language [21].

For the survey, we developed a questionnaire for German-, English- and Spanish-speaking participants using *Google Forms*. The questionnaire aims to verify the detailed results of the interviews and compare them with a broader sample of participants. We designed it as follows (Fig. 1): First, we developed the content by combining quantitative and qualitative questions with the aim of answering the research questions. This first draft was then presented to four UX experts (with at least three years of experience) for feedback. Based on their remarks, we made changes (*e.g.*, to the informative texts and order of questions) to develop Version 0.1 of the questionnaire. With Version 0.1, we conducted a pretest in which a tester completed the survey under the supervision of one of the authors. Based on the pretest observations, further changes were made to the questionnaire. These were mainly limited to changes in the wording in





**Fig. 1.** The development process of the questionnaire [20].

order to clarify the questions. These changes resulted in a new version of the questionnaire, which was also pretested. In this way, we performed three consecutive pretests, which resulted in the final questionnaire: Version 1. The complete questionnaire can be found in the research protocol in English, German, and Spanish versions [21]. The questionnaire ( $N = 76$ ) was conducted from April to June 2021. We distributed it via the social media platforms *LinkedIn*, *Facebook*, and *Twitter*, as well as through the personal networks of the authors.

**Study Participants.** We chose intensive VUI users to be our study participants for the identification of UX aspects for VUIs because they have in-depth experience with VUIs and can provide comprehensive insights into their use. The participants for the qualitative and quantitative parts of the study were acquired separately. For the qualitative part, we interviewed mostly male participants with heterogeneous backgrounds ( $N = 10$ , Table 1).

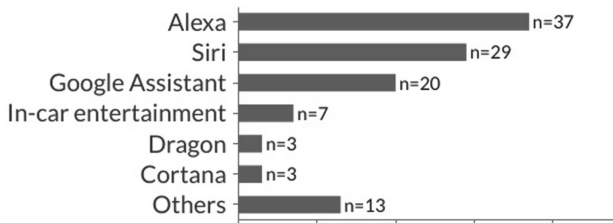
**Table 1.** Participants’ duration of use, devices, and applications [20].

Participants	Duration of use	Devices	Application
P1	3 years	Alexa	Accessibility (visual), smart home control
P2	> 5 years	Alexa, Siri	Accessibility (visual), librarianship
P3	> 10 years	Alexa*, Siri	Accessibility (visual), smartphone control
P4	3 years	Alexa, Siri	Accessibility (visual), search queries
P5	> 10 years	Dragon, Siri	Accessibility (visual), working tool, smartphone control
P6	> 10 years	Alexa, Dragon, Siri	Accessibility (motor), working tool, smartphone control
P7	> 5 years	Alexa, Siri	VUI development
P8	> 5 years	Alexa, in-car entertainment	Smart home control
P9	1 year	Google Assistant	Timer, search queries
P10	> 5 years	Alexa, smartphone**	Radio substitute, (fun) search queries

(\*stopped using Alexa, \*\*unknown smartphone brand)

Another inclusion criterion was *at least one year of use* so that the participants could demonstrate corresponding experience values and have experienced various situations with VUIs. We included participant 7 (P7) even though he did not call himself an intensive user. Although P7 does not use VUIs in a personal context, he works in VUI development and is very familiar with VUIs; thus, we consider him to qualify as an intensive user.

We acquired international survey participants for the quantitative part of the study ( $N = 76$ ). We then excluded participants ( $n = 24$ ) due to the following reasons:  $n = 1$  duplicate,  $n = 5$  records had fewer than three questions answered, and  $n = 18$  participants did not meet the target group requirements of a high frequency of use. Thus, we analyzed the data of intensive users ( $N = 52$ ).



**Fig. 2.** The devices used by the survey participants ( $N = 52$ ) [20].

Most of these survey participants reported using Alexa ( $n = 37$ ) or Siri ( $n = 29$ ) (Fig. 2). The participants were allowed to name more than one VUI.

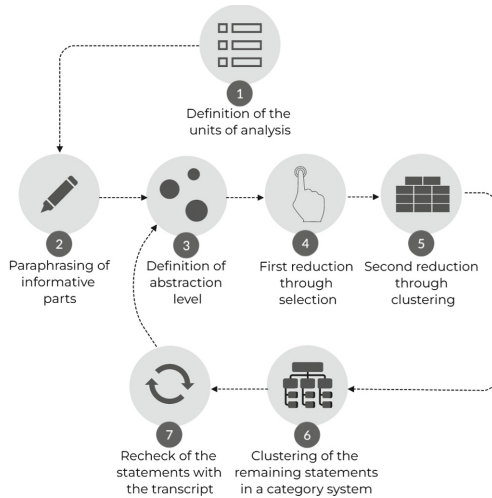
The usage scenarios of the participants in our study are wide-ranging. While six interview participants stated that they use VUIs because of a disability (Table 1), 48 survey participants primarily cited comfort as a reason for VUI use (Fig. 3). In addition, four of our interview participants have some kind of work relationship with VUIs, and 12 of our survey participants use a VUI as a tool at work. Thus, our group of study participants considered both private and professional contexts for their VUI evaluations.



**Fig. 3.** The survey participants' ( $N = 52$ ) reasons for VUI use [20].

Overall, the participants in this mixed-methods study form a heterogeneous and international group with a wide variety of usage scenarios and reasons for use.

**Qualitative Content Analysis.** To analyze the qualitative data from the interviews and surveys, we use the summarizing content analysis of Mayring [26] (Fig. 4) and an adjusted summarizing content analysis to ensure the comparability of the results [20]. The process starts with the definition of the units of analysis (Fig. 4 step 1). In this case, these are the transcripts of the interviews. Then, we mark and paraphrase the informative parts of the interviews to summarize the content into its key messages (Fig. 4 step 2). While doing so, we identified the abstraction level for the key messages (Fig. 4 step 3). These key messages are named *codes* in the following. Afterward, we removed all codes not related to RQ1 (Fig. 4 step 4). The remaining codes were clustered when they had similar key messages (Fig. 4 step 5). Following this, we applied a category system to the clusters, building our list of UX aspects (Fig. 4 step 6). After these enhancements to the code system, we conducted a final recheck with a second round of coding to ensure the code system would still fit the coded interview parts (Fig. 4 step 7).



**Fig. 4.** Content analysis [26] of the interviews based on [20].

To compare the data from our interviews to the data from the survey, we used the content analysis [26] with the following differences: (1) We used the interview code system as a starting point for the survey analysis instead of starting from scratch. (2) The abstraction level is the same as the abstraction level of the interview code system, instead of starting from scratch. (3) We started with the

interview category system (UX aspects) for the survey analysis and added new UX aspects if needed.

With the software tool MAXQDA Standard 2000 (Release 20.4.1), two authors alternately coded the interview and survey data. For the interview analysis, the authors took turns as follows: Author A coded P1, then author B coded P2, and so on until P10 was reached. In a second round, the authors exchanged the participants' transcripts, so author A coded P2, then author B coded P1, and so on. One author analyzed all qualitative data for the survey analysis, followed by the second author. If the coding led to too many changes, a third round would have been necessary to ensure consensus. However, it was not required in this case because only minor changes were made, and the content analysis of the interviews and survey concluded after two runs each.

### 3.2 Kano Categorization of UX Aspects for VUIs

Aspects and components of products can be assigned to the expectations of users according to the aforementioned categorization described by Kano [12]. The UX aspects for VUIs describe the quality of experience of a VUI. Therefore, we can categorize the UX aspects of VUIs according to Kano.

**Categorization Process.** The categorization into the Kano model is done via discrete analysis. For each participant, we determined which Kano category each UX aspect fell into. For this purpose, the answers to the functional and dysfunctional questions are used to make a categorization (Table 2). For example, if a participant answers the functional question with “*It must be that way*” and the dysfunctional question with “*I dislike it that way,*” the feature is identified as a must-be feature. If a participant answers both questions with contradictory answers, the feature is deemed questionable (*i.e.*, not reasonably categorizable for this participant).

**Table 2.** Example of the process of categorizing items into the Kano model based on [12].

Functional Question	+	Dysfunctional Question	<i>rightarrow</i>	Kano Category
It must be that way	+	I dislike it that way	→	Must-Be
I like it that way	+	I dislike it that way	→	One-Dimensional
I like it that way	+	I am neutral	→	Attractive
I am neutral	+	I am neutral	→	Indifferent
I dislike it that way	+	It must be that way	→	Reverse

The functional and dysfunctional questions are based on the description of each UX aspect. For example, the UX aspect “comprehension” was described

by the participants as “*The VUI understands the user correctly, even if they do not speak very clearly.*” The functional question about this UX aspect is, “*How do you feel if a VUI understands you correctly, even if you do not speak very clearly?*” The dysfunctional question is, “*How do you feel if a VUI DOES NOT understand you correctly when you do not speak too clearly?*” The full questionnaire is available in the following research protocol of the Kano categorization [22].

**Survey Participants.** We recruited our survey participants via *Prolific* [32], which is a crowd-working platform that provides a subject pool for research [28] with high-quality data [29]. We provided the study’s title and description as well as the external study link (LimeSurvey questionnaire), including the study completion code, to the participants and pre-screened them to select only those who were fluent in English.

A total of 219 records were then processed through data cleaning. The following exclusion criteria were used:

- more than three questionable features ( $n = 9$ )
- more than 28 identical categorizations ( $n = 11$ )
- less than two seconds per item ( $n = 6$ )
- age under 18 or over 85 ( $n = 2$  and  $n = 2$  with corrupted data in this field)

Participants could be excluded due to one or more of these exclusion criteria. Our final data set consists of  $N = 195$  participants.

## 4 Results

First, we present the UX aspects found from the user-centered mixed-methods study. Then, we show the results of the categorization of the UX aspects according to Kano.

### 4.1 UX Aspects for VUIs

We identified 32 UX aspects for intensive users of VUIs (see Table 3). Each UX aspect is defined by the key messages given by the study participants. This way, we make sure to interpret the UX aspects as intended by the participants and not by personal associations of, say, a practitioner working with them. The original statements are available in the mixed-methods research protocol [21].

For each interview and survey participant, we consider only one single mention of a UX aspect. Thereby, we ensure that no misplaced weighting results from multiple entries by a single participant. Due to the small number of participants, we decided not to set a minimum number of participants who must have named a UX aspect. Instead, we decided to use the categorization of Kano to determine whether a UX aspect is of importance to VUI users in general.

**Table 3.** The identified aspects the target group named for evaluating UX of VUIs [20].

Index	Aspect	Interpretation of the participants	Int. (N = 10)*	Sur. (N = 52)*
1	Comprehension	The VUI understands the user correctly, even if they do not speak very clearly.	10	37
2	Error-free	Both the result and the operation do not give errors, wrong answers or misunderstandings.	6	34
3	Aesthetic	The hardware of the VUI is supposed to be minimalistic. Visual feedback about the status (listening, processing, disabled, etc.) is positively received as long as it is discreet.	3	35
4	Range of functions	The VUI has as many functions and application possibilities as possible.	4	29
5	Simplicity	The operation is easy to perform and contains as few steps as possible.	8	23
6	Effectivity	The user reaches their goal.	9	17
7	Support of the user	The VUI helps users to achieve their goals.	3	22
8	Humanity	The user has the feeling of talking to a human being. They can conduct a normal dialogue, the VUI persona responds with humor and empathy, and the voice sounds natural.	4	21
9	Personal fulfillment	The VUI allows the user to live out their personality. They can speak in their dialect and do not have to alter their voice in order to be better understood.	3	21
10	Context sensitivity	The VUI knows its user, understands the current situation, and can remember the context of the conversation.	4	20
11	Efficiency	The user reaches their goal without detours.	7	17
12	Privacy	The VUI should not permanently listen in on, interrupt, or even record private conversations.	6	16
13	Data security	If personal information must be provided, it can be trusted that it will not be shared and will be handled ethically.	7	15
14	Time-saving	The user does not need to fetch a device or press a button; they can immediately start the usage process. They receive the results immediately after the request.	6	14
15	Politeness	The VUI does not insult the user; it allows them to finish their sentences and does not activate without being asked.	0	19
16	Linking with third-party products	Many third-party products should be compatible. The VUI can easily be connected with them, and there are no errors in communication.	8	11
17	Safety	The VUI gives the user physical and privacy security. For example, security is given by enabling operation in the car without removing the hands from the steering wheel or by protecting the data from external access.	2	16
18	Capability to learn	The VUI can learn new commands, learn the personality of its user, and exercise appropriate reactions. Incorrectly learned commands can be deleted.	2	15
19	Intuitiveness	The user does not need to learn special vocabulary, but can immediately communicate with the VUI using their everyday language. Setting up and learning how to use the VUI is possible without additional help.	5	12
20	Practicality	The VUI helps the user with everyday challenges.	7	10
21	Reliability	The VUI responds only when it is addressed, without false activation. The results are correct and verified. The quality of the interaction should be consistently high.	0	15
22	Help with errors	If an error occurs, a way to fix it is shown. There is a help function.	3	11
23	Convenience	The user can use the VUI in any situation without having to make an effort. For example, they can use it from the sofa, bed, or desk.	7	7
24	Fun	The VUI is fun to use and its humor is appropriate.	3	8
25	Customizability	The persona of the VUI can be set by the user according to their preferences (gender, language, humor, voice, etc.)	0	8
26	Flexibility	The VUI can adapt to different users and situations.	4	4
27	Voice	The voice of the VUI is pleasant and clearly understandable.	0	7
28	Responsiveness	The VUI responds as soon as it is addressed, but only when it is addressed.	0	6
29	Independency	The user does not need any assistance in using the VUI. It allows additional independence for users who would have problems operating a GUI (for example, those with visual or motor impairment, dyslexics, and children).	4	2
30	Innovation	The VUI has new, modern, and unique features.	2	1
31	Ad-free	Advertising is not played or can be turned off.	0	2
32	Longevity	The VUI can be used for a long time, does not break quickly, and does not need to be repeatedly replaced with the latest model.	0	2

The aspects are sorted by the total number of mentions. \*Number of mentions by the interview (Int.) and survey (Sur.) participants

## 4.2 Categorization of UX Aspects

It is determined which percentage of the overall sample assigns a UX aspect to each Kano category. The overall assignment to the Kano categories is then determined by the majority of the study participants. In order to verify that

the assignment into a category does not result from a random distribution, we performed the test proposed by Fong [6]:

$$|a - b| < 1.65 \cdot \sqrt{\frac{(a + b)(2n - a - b)}{2n}} \tag{1}$$

As shown in the formula,  $a$  denotes the frequency in the category with the most mentions (*i.e.*, the category to which the UX aspect was assigned), and  $b$  is the frequency in the category with the second most mentions. The sum of the ratings considered in the evaluation is represented by  $n$ . The result of the Fong test depends primarily on how large the difference is between the assignment to the first category and the assignment to the second category. According to Fong, the assignment of the UX aspect can be considered significant if the statement of inequality is not true [25].

Following the discrete data analysis and the Fong test, we identified a list of 21 UX aspects that were unambiguously categorized into one of the Kano categories with a confirmed result of the Fong test (Table 4).

**Table 4.** Distinct results of the discrete analysis to categorize the UX aspects into the Kano model (N = 195).

UX Aspect	Category	Must-be	One-dimensional	Attractive	Indifferent	Reverse	Questionable
Data security	Must-Be	124 (63.5%)	59 (30.2%)	2 (1%)	4 (2%)	3 (1.5%)	3 (1.5%)
Privacy	Must-Be	93 (47.6%)	64 (32.8%)	2 (1%)	5 (2.5%)	7 (3.5%)	24 (12.3%)
Ad-free	Must-Be	91 (46.6%)	68 (34.8%)	20 (10.2%)	15 (7.6%)	0 (0%)	1 (0.5%)
Simplicity	One-Dimensional	67 (34.3%)	109 (55.8%)	9 (4.6%)	9 (4.6%)	1 (0.5%)	0 (0%)
Comprehension	One-Dimensional	31 (15.8%)	99 (50.7%)	40 (20.5%)	21 (10.7%)	3 (1.5%)	1 (0.5%)
Error-free	One-Dimensional	66 (33.8%)	97 (49.7%)	21 (10.7%)	8 (4.1%)	0 (0%)	3 (1.5%)
Voice	One-Dimensional	71 (36.4%)	96 (49.2%)	15 (7.6%)	13 (6.6%)	0 (0%)	0 (0%)
Practicality	One-Dimensional	22 (11.2%)	91 (46.6%)	44 (22.5%)	34 (17.4%)	4 (2%)	0 (0%)
Help with errors	One-Dimensional	57 (29.2%)	87 (44.6%)	30 (15.3%)	20 (10.2%)	0 (0%)	1 (0.5%)
Convenience	One-Dimensional	45 (23%)	86 (44.1%)	33 (16.9%)	28 (14.3%)	3 (1.5%)	0 (0%)
Time-saving	One-Dimensional	36 (18.4%)	86 (44.1%)	34 (17.4%)	33 (16.9%)	4 (2%)	2 (1%)
Support of the user	One-Dimensional	36 (18.4%)	85 (43.5%)	42 (21.5%)	29 (14.8%)	2 (1%)	1 (0.5%)
Flexibility	One-Dimensional	31 (15.8%)	84 (43%)	41 (21%)	36 (18.4%)	3 (1.5%)	0 (0%)
Personal fulfillment	One-Dimensional	49 (25.1%)	80 (41%)	31 (15.8%)	32 (16.4%)	2 (1%)	1 (0.5%)
Effectivity	One-Dimensional	41 (21%)	74 (37.9%)	42 (21.5%)	35 (17.9%)	2 (1%)	1 (0.5%)
Fun	One-Dimensional	35 (17.9%)	73 (37.4%)	45 (23%)	37 (18.9%)	4 (2%)	1 (0.5%)
Efficiency	One-Dimensional	34 (17.4%)	70 (35.8%)	53 (27.1%)	34 (17.4%)	1 (0.5%)	3 (1.5%)
Capability to learn	One-Dimensional	44 (22.5%)	65 (33.3%)	26 (13.3%)	41 (21%)	15 (7.6%)	4 (2%)
Context sensitivity	One-Dimensional	35 (17.9%)	57 (29.2%)	31 (15.8%)	41 (21%)	26 (13.3%)	5 (2.5%)
Humanity	Indifferent	14 (7.1%)	20 (10.2%)	44 (22.5%)	88 (45.1%)	26 (13.3%)	3 (1.5%)
Aesthetic	Indifferent	43 (22%)	26 (13.3%)	35 (17.9%)	67 (34.3%)	11 (5.6%)	13 (6.6%)

The 21 UX aspects are distinctively categorized and are distributed as follows:

Three UX aspects are distinctively categorized as **must-be**: *privacy*, *data security*, and *ad-free*.

16 UX aspects are distinctively categorized as **one-dimensional**: *comprehension*, *error-free*, *simplicity*, *effectivity*, *support of the user*, *personal fulfillment*, *context sensitivity*, *efficiency*, *time-saving*, *capability to learn*, *practicality*, *help with errors*, *convenience*, *fun*, *flexibility*, and *voice*.

Two UX aspects are distinctively categorized as **indifferent**: *aesthetic* and *humanity*.

No UX aspects are distinctively categorized as **attractive**, **reverse**, or **questionable**. However, we have 11 UX aspects that are ambiguous and can be categorized as at least two different Kano categories (Table 5).

**Table 5.** UX aspects that are not distinct for only one Kano category (N = 195).

UX Aspect	Must-be	One-dimensional	Attractive	Indifferent	Reverse	Questionable
Customizability	17 (8.7%)	41 (21%)	64 (32.8%)	66 (33.8%)	5 (2.5%)	2 (1%)
Independency	51 (26.1%)	63 (32.3%)	46 (23.5%)	34 (17.4%)	1 (0.5%)	0 (0%)
Innovation	25 (12.8%)	62 (31.7%)	48 (24.6%)	58 (29.7%)	2 (1%)	0 (0%)
Intuitiveness	79 (40.5%)	80 (41%)	15 (7.6%)	20 (10.2%)	0 (0%)	1 (0.5%)
Linking with third-party products	37 (18.9%)	55 (28.2%)	35 (17.9%)	41 (21%)	20 (10.2%)	7 (3.5%)
Longevity	87 (44.6%)	93 (47.6%)	5 (2.5%)	9 (4.6%)	1 (0.5%)	0 (0%)
Politeness	83 (42.5%)	83 (42.5%)	12 (6.1%)	13 (6.6%)	3 (1.5%)	1 (0.5%)
Range of functions	27 (13.8%)	57 (29.2%)	38 (19.4%)	62 (31.7%)	8 (4.1%)	3 (1.5%)
Reliability	93 (47.6%)	85 (43.5%)	9 (4.6%)	5 (2.5%)	0 (0%)	3 (1.5%)
Responsiveness	80 (41%)	85 (43.5%)	15 (7.6%)	15 (7.6%)	0 (0%)	0 (0%)
Safety	66 (33.8%)	79 (40.5%)	17 (8.7%)	25 (12.8%)	4 (2%)	4 (2%)

The 11 UX aspects that are ambiguous are distributed as follows: Six UX aspects were mainly categorized as both **must-be and one-dimensional**: *politeness* (43% each one-dimensional and must-be), *intuitiveness* (41% each one-dimensional and must-be), *longevity* (48% one-dimensional, 45% must-be), *responsiveness* (44% one-dimensional, 41% must-be), *safety* (41% one-dimensional, 34% must-be), and *reliability* (48% must-be, 44% one-dimensional).

One UX aspect could be categorized as both **one-dimensional and indifferent**: *range of functions* (32% indifferent, 29% one-dimensional).

Another could be categorized as both **attractive and indifferent**: *customizability* (34% indifferent, 33% attractive).

Three UX aspects could not be assigned to any Kano category because the variance was too large: *linking with third-party products*, *independency*, and *innovation*.

It is noteworthy that the UX aspects *humanity* and *context sensitivity* each received 13% as a **reverse** feature and that *privacy* has 12% **questionable** answers.



## 5 Discussion

The identification of UX aspects for certain products is not a new research field. However, instead of following a theoretical approach on what a UX aspect should mean, we followed a user-centered approach to identify 32 UX aspects for VUIs (**RQ1**). These 32 UX aspects represent what intensive users think about when evaluating the UX of VUIs. Established literature had already defined a few of the UX aspects that our participants named, such as *efficiency* and *effectivity* [11] and *aesthetic* [35]. However, these aspects are not specified for VUIs. Additionally, some of our UX aspects can be found as part of other known VUI UX aspects (e.g., *simplicity* and *politeness*), which may be part of the UX factor *likeability* of the *Subjective Assessment of Speech System Interfaces* (SASSI) [8], but they have not yet been explicitly considered. Lastly, several UX aspects for VUIs named by our participants are new and unique to our research, such as *independency* and *context sensitivity*.

From the survey participants ( $N = 52$ ), the majority reported using Alexa ( $n = 37$ ) or Siri ( $n = 29$ ), with participants being allowed to name more than one VUI. In contrast to our results, a representative German study ( $N = 3184$ ) found that Google Assistant (12%) and Alexa (9%) are the most commonly used VUIs [39]. However, since we did not look for brand-specific evaluations, we believe our results are still applicable to general VUI use.

Following the identification of the UX aspects, we categorized them in the Kano model with a larger set of general VUI users ( $N = 195$ ). This ensured that the UX aspects are relevant not only for intensive users but for all VUI users, even potential ones. In addition, this categorization into the Kano categories provides VUI developers a better understanding of the prioritization of individual UX aspects (**RQ2** and **RQ3**).

The UX aspects that were categorized as **must-be** are *privacy*, *data security*, and *ad-free*. These are qualities that VUI developers should always consider when developing VUIs. Studies have already shown that privacy and data security are huge concerns for VUI users [15], which is in line with this categorization. However, our study also revealed *ad-free* to be a must-be for our study participants. Since advertising is a popular way to provide free services, this topic should be researched further.

**One-dimensional** UX aspects include *comprehension*, *error-free*, *simplicity*, *effectivity*, *support of the user*, *personal fulfillment*, *context sensitivity*, *efficiency*, *time-saving*, *capability to learn*, *practicality*, *help with errors*, *convenience*, *fun*, *flexibility* and *voice*. One-dimensional features should also always be considered. Even though they are not rated as important as must-be features, their absence has a negative impact. Furthermore, one-dimensional features improve the UX of a VUI the more they are fulfilled. Lastly, we identified two UX aspects that were categorized as **indifferent** by the VUI users: *aesthetic* and *humanity*. These UX aspects could be neglected under certain circumstances since their absence does not seem to have a notable impact.

Besides the UX aspects that were distinctively categorized as one specific Kano category, we identified a few UX aspects that are ambiguous and could be

sorted into at least two categories. This does not necessarily imply that these UX aspects are of less importance. For example, six UX aspects are mostly categorized as **must-be** or **one-dimensional**: *politeness*, *safety*, *intuitiveness*, *reliability*, *responsiveness*, and *longevity*. Since must-be and one-dimensional are both important feature categories for the users, these UX aspects should still be prioritized. It becomes more complicated with UX aspects that can be either a positive or an indifferent feature. For example, *range of functions* is both one-dimensional and indifferent, and *customizability* is both attractive and indifferent. To ensure the correct prioritization of these specific UX aspects, the target group should be defined more clearly for the to-be-developed product. This also applies to the UX aspects *linking with third-party products*, *independency*, and *innovation* since these are too diffuse to be sorted into any kind of category.

Developers concerned with the humanity and context sensitivity of VUIs should pay particular attention to their target group and development goal. These UX aspects were assigned as a reverse feature by 13% of the respondents, which means that these aspects can negatively impact the UX of a VUI if implemented. For example, an experiment identified two groups among those that use smart home applications. One user group likes more effective voice commands, while the other likes more natural speech [40].

We did not identify any reverse feature. This could be because the UX aspects were positively formulated from the earlier data collection. Our intention was to show how the users wish their VUIs to be. So, if a user was disappointed that a VUI did not understand them, we interpreted that as meaning a good understanding would be appreciated. Thus, the results of the Kano categories can be interpreted as confirmation of the UX aspects. If they had been sorted into the reverse category, they would have had a negative influence on the UX if implemented.

We could also not identify a distinct attractive feature. This is particularly interesting because these features would enable developers to excite the users, not just satisfy their basic needs. This could be the result of the live circle of Kano features [12], as attractive features become one-dimensional and, eventually, basic. An American consumer report showed that US adults' VUI use stagnated. The VUI adoption rate seems to have peaked at 50–60% in the US population, meaning that users and non-users would probably need something exciting to encourage them to use new voice technology [13].

Although we did not identify a questionable feature, privacy has 12% questionable answers. This could be due to the importance of the topic for the participants. The pre-tests found that the question was answered vehemently with “absolutely not” for both the functional and the dysfunctional questions. The functional question was: “How would you feel if the VUI does not permanently listen in on, interrupt, or record private conversations?” The dysfunctional question was: “How would you feel if the VUI permanently listens in on, interrupts, or even records private conversations?” After the pretest, the “not” was more emphasized in the layout. This might not have been enough to prevent hasty answers.

The results have not been collected for a specific product, as we asked our mixed-methods study participants to rate VUIs in general. Therefore, they have to be narrowed down more precisely for the individual cases. Matzler and Hinterhuber [25] suggest that, when evaluating a specific product, the relative importance and degree of fulfillment should be queried in addition to the functional and dysfunctional questions. Since we did not evaluate a specific product, this question was deliberately omitted, but it can be helpful in the categorization for further development projects.

One limitation of our study is that our study participants are mostly from WEIRD (Western, Educated, Industrialized, Rich, and Democratic) countries. For example, the participants from the Kano categorization survey were only citizens of the UK and the USA. While this sample was convenient and provided valuable data for our research question, it may not be representative of other populations. For example, non-WEIRD populations may have different values, problems, and behaviors that could impact the generalization of our findings. Future research should include a more diverse sample of participants, especially participants with a non-WEIRD background, to determine whether our results are applicable to a broader population.

## 6 Conclusion and Future Work

We explored the experiences of intensive users regarding VUIs with a user-centered mixed-methods approach. Thereby, we were able to identify a list of 32 potential UX aspects for VUIs. Although some of these UX aspects are already present in the literature, we gathered the UX aspects from the user's point of view. This led to several new UX aspects for VUIs, such as *independency* and *context sensitivity*. The definition of each UX aspect is based on what intensive users expect from a VUI regarding it.

With a subsequent categorization according to Kano, we narrowed down the relevance of the individual UX aspects for VUIs more precisely. We confirmed 19 UX aspects as distinct and relevant for VUI users and 11 as ambiguous but at least partly relevant. Two were categorized as mostly indifferent. By determining the degree of distinctiveness of the individual UX aspects in the Kano categories “*must-be*,” “*one-dimensional*,” “*attractive*,” “*indifferent*,” and “*reverse*,” we provide VUI developers with guidelines to generate a better UX, which can increase the acceptance of VUIs.

VUI developers can now use these UX aspects as a starting point to determine their development focus. Thus, by narrowing down their own target group, they can develop a precise UX vision for their specific product. These aspects can then also serve as a basis for selecting the right evaluation method for the planned product. For example, they could select the VUI questionnaire from the available ones (*e.g.*, UEQ+ Voice Scales, SUS, and SASSI) that best fits the relevant UX aspects. Future research should investigate in greater detail whether there are indeed relevant differences between the various UX aspects in terms of importance for specific target groups (*e.g.*, various usage contexts or backgrounds of the users).

**Acknowledgements.** This is an extension of the previously published conference paper from WEBIST'22: Proceedings of the 18th International Conference on Web Information Systems and Technologies. First, we would like to thank all participants for their time and the insights they provided. We also want to thank Anna Weigand for the tremendous support in finalizing the paper.

## References

1. Biermann, M., Schweiger, E., Jentsch, M.: Talking to Stupid?!? Improving voice user interfaces. In: Fischer, H., Hess, S. (eds.) *Mensch und Computer 2019 - Usability Professionals*. pp. 53–61. Gesellschaft für Informatik e.V. Und German UPA e.V., Bonn (2019). <https://doi.org/10.18420/muc2019-up-0253>
2. Bogner, A., Littig, B., Menz, W.: Interviews mit Experten: eine praxisorientierte Einführung (Interviewing experts: a practice-oriented introduction). Springer Fachmedien Wiesbaden, Wiesbaden (2014). <https://doi.org/10.1007/978-3-531-19416-5>
3. Cohen, M.H., Giangola, J.P., Balogh, J.: *Voice User Interface Design*. Addison-Wesley, Boston (2004)
4. Dresing, T., Pehl, T.: *Praxisbuch interview, transkription & analyse, audiotranskription. Anleitungen und Regelsysteme für qualitativ Forschende. (Manual on interviewing, transcription and analysis, audio transcription. Software Guides and Practical Hints for Qualitative Researchers.)*. Dr. Dresing und Pehl, Marburg (2018)
5. Feng, Y.L., Huang, C.H.: Study on user experience of live streaming sales based on ISM and kano quality model. *J. Phys.: Conf. Ser.* **1748**(4) (2021). <https://doi.org/10.1088/1742-6596/1748/4/042046>
6. Fong, D.: Using the self-stated importance questionnaire to interpret Kano questionnaire results. *Center Q. Manage. J.* **5**(3), 21–23 (1996)
7. Hassenzahl, M., Tractinsky, N.: User experience a research agenda. *Behav. Inf. Technol.* **25**(2), 91–97 (2006). <https://doi.org/10.1080/01449290500330331>
8. Hone, K.S., Graham, R.: Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natl. Lang. Eng.* **6**(3&4), 287–303 (2000). <https://doi.org/10.1017/S1351324900002497>
9. Iniesto, F., Coughlan, T., Lister, K.: Implementing an accessible conversational user interface: applying feedback from university students and disability support advisors. In: Vazquez, S.R., Drake, T., Ahmetovic, D., Yaneva, V. (eds.) *Proceedings of the 18th International Web for All Conference*, pp. 1–5. ACM, New York, NY, USA (2021). <https://doi.org/10.1145/3430263.3452431>
10. International Requirements Engineering Board (IREB): IREB certified professional for requirements engineering - foundation level - syllabus version 3.1.0. Tech. rep., <https://www.ireb.org/en/about/ireb/1> (2022). [https://www.ireb.org/content/downloads/2-cpre-foundation-level-syllabus-3-0/cpre-foundationlevel-syllabus\\_en\\_v.3.1.pdf](https://www.ireb.org/content/downloads/2-cpre-foundation-level-syllabus-3-0/cpre-foundationlevel-syllabus_en_v.3.1.pdf)
11. ISO 9241–210: Ergonomics of human-system interaction Part 210: humancentred design for interactive systems. Tech. rep., <https://www.iso.org/committee/53372.html> (2019). <https://www.iso.org/standard/77520.html>
12. Kano, N., Seraku, N., Takahashi, F., Tsuji, S.i.: (attractive quality and must-be quality). *J. Jpn. Soc. Q. Control* **31**(4), 147–156 (1984). <https://cir.nii.ac.jp/crid/1572261550744179968>

13. Kinsella, B.: Voice assistant adoption clustering around 50% of the population (2022), <https://voicebot.ai/2022/04/15/voice-assistant-adoption-clustering-around-50-of-the-population/>
14. Klein, A.M., Kölln, K., Deutschländer, J., Rauschenberger, M.: Design and evaluation of voice user interfaces: what should one consider? In: Design, Operation and Evaluation of Mobile Communications: 4th International Conference, MOBILE 2023, Held as Part of the 25th HCI International Conference, HCII 2023, Copenhagen, Denmark, 23–28 July 2023, Proceedings, pp. 167–190 (2023). [https://doi.org/10.1007/978-3-031-35921-7\\_12](https://doi.org/10.1007/978-3-031-35921-7_12)
15. Klein, A.M., Hinderks, A., Rauschenberger, M., Thomaschewski, J.: Exploring voice assistant risks and potential with technology-based users. In: Proceedings of the 16th International Conference on Web Information Systems and Technologies - Volume 1: WEBIST, pp. 147–154. INSTICC, SciTePress, Portugal (2020). <https://doi.org/10.5220/0010150101470154>
16. Klein, A.M., Hinderks, A., Schrepp, M., Thomaschewski, J.: Construction of UEQ+ scales for voice quality. In: Proceedings of the Conference on Mensch Und Computer, pp. 1–5. MuC '20, Association for Computing Machinery, New York, NY, USA (2020). <https://doi.org/10.1145/3404983.3410003>
17. Klein, A.M., Hinderks, A., Schrepp, M., Thomaschewski, J.: Measuring user experience quality of voice assistants. In: 2020 15th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1–4. IEEE, Seville, Spain (2020). <https://doi.org/10.23919/CISTI49556.2020.9140966>
18. Klein, A.M., Rauschenberger, M., Thomaschewski, J., Escalona, M.J.: Comparing voice assistant risks and potential with technology-based users: a study from Germany and Spain. *J. Web Eng.* **7**(16), 1991–2016 (2021). <https://doi.org/10.13052/jwe1540-9589.2071>
19. Kocaballi, A.B., Laranjo, L., Coiera, E.: Understanding and measuring user experience in conversational interfaces. *Interact. Comput.* **31**(2), 192–207 (2019). <https://doi.org/10.1093/iwc/iwz015>
20. Kölln, K., Deutschländer, J., Klein, A.M., Rauschenberger, M., Winter, D.: Identifying user experience aspects for voice user interfaces with intensive users. In: Proceedings of the 18th International Conference on Web Information Systems and Technologies, pp. 385–393. SCITEPRESS - Science and Technology Publications (2022). <https://doi.org/10.5220/0011383300003318>
21. Kölln, K., Deutschländer, J., Klein, A.M., Rauschenberger, M., Winter, D.: Protocol for identifying user experience aspects for voice user interfaces with intensive users (2022). <https://doi.org/10.13140/RG.2.2.26828.49287>
22. Kölln, K., Klein, A.M., Deutschländer, J., Winter, D., Rauschenberger, M.: Protocol for categorizing UX aspects for voice user interfaces using the kano model (2023). <https://doi.org/10.13140/RG.2.2.32565.55528>
23. Langevin, R., Lordon, R.J., Avrahami, T., Cowan, B.R., Hirsch, T., Hsieh, G.: Heuristic evaluation of conversational agents. In: Kitamura, Y., Quigley, A., Isbister, K., Igarashi, T., Bjørn, P., Drucker, S. (eds.) Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–15. ACM, New York, NY, USA (2021). <https://doi.org/10.1145/3411764.3445312>
24. Li, X., You, Y.: Kano model analysis required in app interactive design based on mobile user experience. *Int. J. Multimedia Ubiquitous Eng.* **11**(11), 247–258 (2016). <https://doi.org/10.14257/ijmue.2016.11.11.21>
25. Matzler, K., Hinterhuber, H.H.: How to make product development projects more successful by integrating kano's model of customer satisfaction into quality function deployment. *Technovation* **18**(1), 25–38 (1998). <https://doi.org/10.>

- 1016/S0166-4972(97)00072-2, <https://www.sciencedirect.com/science/article/pii/S0166497297000722>
26. Mayring, P.: Qualitative Content Analysis, vol. 1994. UVK Univ.-Verl, Konstanz (1994)
  27. McKim, C.A.: The value of mixed methods research: a mixed methods study. *J. Mixed Methods Res.* **11**(2), 202–222 (2017). <https://doi.org/10.1177/1558689815607096>
  28. Palan, S., Schitter, C.: Prolific.ac a subject pool for online experiments. *J. Behav. Exper. Finance* **17**, 22–27 (2018). <https://doi.org/10.1016/j.jbef.2017.12.004>, <https://linkinghub.elsevier.com/retrieve/pii/S2214635017300989>
  29. Peer, E., Rothschild, D., Gordon, A., Evernden, Z., Damer, E.: Data quality of platforms and panels for online behavioral research. *Behav. Res. Methods* **54**, 1643–1662 (2021). <https://doi.org/10.3758/s13428-021-01694-3>
  30. Preece, J., Rogers, Y., Sharp, H.: Interaction design: beyond human-computer interaction. In: *Interaction Design: Beyond Human-Computer Interaction*. J. Wiley & Sons, New York (2002)
  31. Primrose, M.C.: User experience grading via kano categories. In: *2010 18th IEEE International Requirements Engineering Conference*, pp. 331–336 (2010). <https://doi.org/10.1109/RE.2010.47>
  32. Prolific Academic Ltd.: <https://www.prolific.co>. Accessed 06 Mar 2023
  33. Rauschenberger, M.: Acceptance by Design: voice assistants. In: *1st AI-DEbate Workshop: workshop establishing An InterDisciplinary perspective on speech-BASed TEchnology*, p. 27.09.2021. OvGU, Magdeburg, Germany (2021). <https://doi.org/10.25673/38476>, <https://opendata.uni-halle.de//handle/1981185920/38717>
  34. Schrepp, M., et al.: On the importance of UX quality aspects for different product categories. *IJIMAI (International Journal of Interactive Multimedia and Artificial Intelligence)* (2022)
  35. Schrepp, M., Thomaschewski, J.: Construction and first validation of extension scales for the user experience questionnaire (UEQ) (2019). <https://doi.org/10.13140/RG.2.2.19260.08325>
  36. Schrepp, M., Thomaschewski, J.: Design and validation of a framework for the creation of user experience questionnaires. *Int. J. Interact. Multimedia Artif. Intell.* **5**(7), 88–95 (2019). <https://doi.org/10.9781/ijimai.2019.06.006>
  37. Seaborn, K., Urakami, J.: Measuring voice UX quantitatively. In: Kitamura, Y., Quigley, A., Isbister, K., Igarashi, T. (eds.) *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–8. ACM, New York, NY, USA (2021). <https://doi.org/10.1145/3411763.3451712>
  38. Song, X.: User experience design of elderly-oriented social apps based on kano model—the case of wechat. In: Duffy, V.G., Gao, Q., Zhou, J., Antona, M., Stephanidis, C. (eds.) *HCI International 2022 - Late Breaking Papers: HCI for Health, Well-being, Universal Access and Healthy Aging*, pp. 546–558. Springer Nature Switzerland, Cham (2022). [https://doi.org/10.1007/978-3-031-17902-0\\_39](https://doi.org/10.1007/978-3-031-17902-0_39)
  39. Tas, S., Hildebrandt, C., Arnold, R.: Voice assistants in Germany. *WIK Wissenschaftliches Institut für Infrastruktur und Kommunikationsdienste GmbH, Bad Honnef, Germany* (2019). <https://www.wik.org/en/publications/publication/no-441-voice-assistants-in-germany>, nr.441
  40. Vacher, M., Caffiau, S., Portet, F., Meillon, B., Roux, C., Elias, E., Lecouteux, B., Chahuara, P.: Evaluation of a context-aware voice interface for ambient assisted living: qualitative user study vs. quantitative system evaluation. *ACM Trans. Access. Comput.* **7**, 1–36 (5 2015). <https://doi.org/10.1145/2738047>

41. Wei, Z., Landay, J.A.: Evaluating speech-based smart devices using new usability heuristics. *IEEE Pervasive Comput.* **17**(2), 84–96 (2018). <https://doi.org/10.1109/MPRV.2018.022511249>
42. Winter, D., Hinderks, A., Schrepp, M., Thomaschewski, J.: Welche UX Faktoren sind für mein Produkt wichtig? (Which UX factors are important for my product?). In: Hess, S., Fischer, H. (eds.) *Mensch und Computer MuC 2017*. Gesellschaft für Informatik e. V. und die German UPA e.V. (2017). <https://doi.org/10.18420/muc2017-up-0002>
43. Witell, L., Löfgren, M., Dahlgaard, J.J.: Theory of attractive quality and the Kano methodology the past, the present, and the future. *Total Q. Manage. Bus. Excellence* **24**, 1241 – 1252 (2013). <https://doi.org/10.1080/14783363.2013.791117>



# How We Evaluate the Accessibility of an Infographic: A Pilot Study Through SUS Questionnaire

Alessio Caccamo<sup>(✉)</sup> 

Department of Planning, Design, Technology of Architecture, Sapienza – University of Rome,  
00193 Rome, RM, Italy

alessio.caccamo@uniroma1.it

**Abstract.** The present study aims to evaluate the accessibility of infographics by using the System Usability Scale (SUS) questionnaire, as well as to underline the necessity of new curricula on infographic literacy. The study was conducted on a sample of 200 participants [100 Visual Designer Graduate – 100 Other Disciplines Graduated]. The participants were given a set of infographics to evaluate based on their usability and understandability. The results showed that there were significant differences in the scores based on the level of education of the participants, with higher design education levels leading to better scores. The study also highlighted the importance of developing new curricula on infographic literacy, as the current educational system does not provide enough opportunities for students to learn about infographics and their proper use. This is especially important given the increasing prevalence of infographics in various fields, from journalism to science communication. Without proper education on infographic literacy, individuals may struggle to fully understand and utilize the information presented in infographics, leading to potential misinterpretations or misunderstandings. The findings of this study have important implications for educators and practitioners alike. Educators should prioritize the development of new curricula on infographic literacy to better prepare students for the increasing prevalence of infographics in various fields. Practitioners, on the other hand, should strive to make their infographics more accessible and user-friendly, especially for individuals with lower educational backgrounds.

**Keyword:** Infographic · Evaluation · Information Design

## 1 Introduction

Infographics are increasingly prevalent in our daily lives, appearing in various contexts ranging from news media to scientific publications. They are often used to convey complex information in a visually appealing and easy-to-understand manner. Nevertheless, despite their growing popularity, there is still much to be understood about how individuals perceive and understand infographics.

Previous studies [1] have focused on the usability and effectiveness of infographics, often using quantitative measures such as the System Usability Scale (SUS) to assess



their overall accessibility. While these studies have provided valuable insights into the usability of infographics, they have not fully explored the cognitive processes involved in perceiving and understanding these visual representations of information.

The present paper aims to enrich the scientific context surrounding the issue of perceiving and understanding infographics by providing a framework for understanding the cognitive processes involved. We argue that understanding the cognitive mechanisms involved in processing infographics is essential for improving their design and making them more accessible to a wider range of individuals. To achieve this goal, we draw on a range of literature from cognitive psychology and visual perception to provide a comprehensive framework for understanding the cognitive processes involved in perceiving and understanding infographics.

By examining these processes in the context of infographics, we aim to provide a better understanding of how individuals perceive, understand, and evaluate, these visual representations of information.

## **2 Infodemic and Data Visualization: The Context**

### **2.1 Living in a Visual Environment**

We live in a visual environment and in a data-driven world [2] whose information is ubiquitous and of crucial importance for understanding different fields of knowledge [3] because what we see is what we know [4]. However, in the current post-pandemic scenario from COVID-19, the practices of mass mis and disinformation have re-emerged and have originated new terms, such as ‘infodemic’ and ‘disinfodemic’ [5] i.e., the uncontrolled production and dissemination of information whose degeneration is caused by fake news [6]. Contemporary society is facing a continuous evolution of habits, processes, and tools in the field of communication and information, because of the advancement of technologies and the massive amount of interaction inputs [7]. The production of data increases year by year, intentionally or unintentionally to our actions [8], and the Orwellian society does not seem to be so far away.

In fact, we are in the condition where we have a higher availability of information than our mind can handle usefully: information overload. The same people, through their activities - smartphones, home automation and ICT systems - produce incredible amounts of Small and Big Data, the stratification of which - without a method of selection, organisation, and interpretation - is just a “disorderly collection of information” [9], which leads mankind into a condition of chaos caused by the constant and simultaneous use of all communication possibilities, whether out of haste or ignorance. On the one hand, due to the haste to instantly do something that others might do to our disadvantage and to immediately acquire control of a medium of communication in any way, and on the other hand, due to ignorance of all the possibilities that the rush does not give us a way to know [10]. However, this rapid spread is not surprising, given the vast number of people who use the Internet to communicate, socialise, consume, and share information [11].

The terms ‘disinformation’, ‘misinformation’ and ‘propaganda’ are sometimes used interchangeably, and their definitions are shifting and overlapping [12]. All three involve false or misleading messages spread in the form of information content, whether in

the form of mainstream communication, online messages, advertisements, or published articles. These forms can be generically grouped under the terminology of information disorder [13]. However, it is important to distinguish genuine from malicious messages and those that are designed, produced, or distributed by ‘agents’ who intend to cause harm from those that are not. Information disorder can appear in the form of text, image, infographics, video and audio, or a combination of these, and be both created or manipulated by humans - as is the case with ‘deepfakes’ - or synthetically generated by artificial intelligence-enabled tools [11].

In relation to the above scenario, the communicative artefacts derived from Data Journalism are not exempt from information disorder, and in fact are well suited to being manipulated and distorted for political purposes [14]. As Huff [15] states, a well-packaged statistic performs better than a big lie. A first issue in terms of errors that needs to be addressed is what Jones [11] refers to as the epistemological error. Data, in fact, never represent a fact in the objective sense [17] but rather a description of an event or phenomenon, thus forming a virtual model of reality that by its very nature is subject to errors and fallacies [18]. This is because scientific models of the visual-figurative type have always been virtual, and their novelty is to be found in the fact that they are the most real virtual models ever conceived. More real models in the sense of more formally, structurally, and functionally resembling - the objects depicted [15].

## 2.2 Encoding Visual Artifacts

In an infodemic, the citizen needs to process information quickly, using automated thinking parameters - bias and pattern - favouring interpretations that require minimal cognitive effort and primarily reflecting prior knowledge. Today’s population needs to analyse information that is interconnected with society and the environment and that is continuously transmitted, remixed, and shared [20].

This mix and match of multimedia content that makes use of text, images, and data in several formats, shapes the way we perceive reality through visual communication, feeding the reinforcement of biases due to a substantial illiteracy towards a conscious consumption and production of communicative artefacts. The key role within the process of information disorder is played by the receiver of communication, who can transform from consumer to producer through the tools offered by digital technologies [21].

Therefore, the correct coding and critical encoding of the communication artefact is the cornerstone in which the decision is made as to whether information is to be considered reliable. However, this process is particularly tricky.

In fact, the activity of understanding and interpreting is strongly influenced by a series of biases that blend social, cognitive, and perceptual components, opening the question of a critical literacy in the consumption and creation of communicative-infographic artefacts, as because of technological democratisation everyone can design graphics, but only a few know how [2].

But we understand graphs can lie, either intentionally or unintentionally [22]. Huff [15], Tufte [23], Cairo [18] and Jones [16] present several examples of published graphs that are designed in such a manner as to produce misleading interpretations of the data. Furthermore, the same “seductive language of data” [15] is often used to sensationalise, inflate, confuse, and oversimplify. In fact, Meyer, Shinar & Leiser [24] state that the

relative effectiveness of a visualisation may depend in part on the characteristics of the user population, while Carpenter & Shah [25] point out that individual differences in graphical knowledge may play as significant a role in the comprehension process as variation in the properties of the graph itself. Cleveland & McGill [26] provide a list of the most relevant perceptual features in reading graphs. These include, in order of precision: (i) Position along a common scale; (ii) Positions along unaligned scales; (iii) Length, direction, angle; (iv) Area; (v) Volume, curvature; (vi) Shading, colour saturation. While Freedman & Shah [27], identify three determining factors in the successful comprehension of visualisation, such as: (i) the visual properties of the representation; (ii) prior knowledge in reading graphs; and (iii) prior knowledge regarding the content of the visualisation. Therefore, the success or failure of comprehension could be due to an interrelationship of several factors, which, according to Glazer [28], Friel, Curcio & Bright [29], Shah & Hoeffner [30], would involve: (i) the domain of prior knowledge; (ii) the constant enjoyment of infographic content; (iii) the design of the graphic itself; (iv) the context of the artefact's appearance, and (v) cognitive and social models. It is from these premises that the cognitive, social mechanisms underlying the spread of information disorder that are inevitably influenced by perceptual bias should be studied.

### 3 The Accessibility Issue of Data Visualization

#### 3.1 The False Myth of the Universal [Visual] Language

We owe the first systematic theorisation of visual language to the Bauhaus. Lupton [31] argues that particularly in the writings of Kandinsky, Klee, Moholy-Nagy and others, information graphics - i.e., what would later take the denomination of Data Visualisation - served as a model for a new aesthetic between didactics and poetics. Scientific grids, graphs, and diagrams were seen as the basis of an anti-illusionistic but universally comprehensible visual script, a graphic language, which goes beyond the conventions of perspectival realism but is objectively related to material facts [31]. In fact, Kandinskij and Klee were not solely interested in the expressive possibilities of the graphic sign but focused their pedagogical efforts on normalising their knowledge through the definition of universal principles of visual forms [32]. Earlier attempts can be found in the works of Superville, Jones, Blanc and Crane [32]. Similarly, Neurath, through the *Buildstatistik*, committed himself to the construction, on the one hand of a code of unified science [33] i.e., a new hieroglyphic script, which contemplates the immediate comprehensibility of iconic images, with a set of rules for their textual composition.

Later, Moholy-Nagy [34] identifies the basic principles of visual representation based on compositional variables such as dynamism and stasis. Kepes [35] subdivides visual language - defined as the language of vision - into plastic organisation and visual representation. Bertin [36] constructs a narrative around the topic of visual language applied to statistics - offering a refined analysis of the visual variables of data processing, its organisation, value, and purpose - by reinterpreting it under the lens of semiotics. Dondis [4] describes in detail the primitive elements of visual language, the rules of syntax and perception, even introducing the term *Visual Literacy*.

Historically, it was Balchin and Coleman [37] who coined the term *Graphicacy* referring to the skills of orientation, understanding and use of cartography for educational

purposes, i.e., it represents a competence that combines mathematical, textual, media, technological and graphic skills. Graphicacy is, therefore, to be understood as the competence relating to infographic language skills. A graphically literate citizen therefore can read and write through the language of graphs, mastering its grammar and using it critically to form and shape. This interaction is also referred to as the 'language of design' by Schön [38]. In this regard, a study conducted by Culbertson & Powers [39] examined various types of graphs and tried to detect the effectiveness of correlations between Graphicacy and other skills, such as verbal skills. Therefore, can we refer to innate competence?

Several studies have investigated the population's difficulties in perceiving graphics, arguing that comprehension and aspects beyond the most obvious proportional relationships can cause extreme difficulties [40, 41, 42, 43]. This is because in order for a visualisation to be correctly processed, the receiver applies two evaluation dimensions [44]:

- **Technical Mapping.** Represents the methods by which the visualisation was created: (i) Direct, when the user can deduce the underlying data, (ii) and Indirect, when the user is unable to deduce the underlying data.
- **Data Focus.** Represents what is communicated by the graphic: (i) Intrinsic, if the image facilitates the intuition of data by cognitively effective means, or Extrinsic, if the image facilitates the communication of the meaning implied by the data.

This is because the definition of a visual language considers that there is not only an exchange between code and message but also a transformation from message to code. The use of signs requires an interpretation on the part of the performer, even if they are available in their semantic and cognitive form [45], as a polysemic nature is present in the diagram [36]. In fact, the difference between verbal and visual language lies in the arbitrariness of the verbal sign that has no natural relationship with the concept it represents [31] and that can be influenced by cultural differences and dictates missing in the understanding and interpretation of a photograph, symbol, or diagram [46].

### 3.2 The Aesthetic Bias

Messages resulting from communicative-infographic artefacts are sent and received on three levels [4]: representational, astrational and symbolic. The former refers to how we see and recognise elements from context and through experience. The latter, to kinaesthetic properties and the reduction of visual components into basic elements. The third level, to the codified sign system. The proper knowledge and understanding of these levels define a visually literate subject [4]. Despite this, the high level of visual wrapping of data into potentially false news reduces the assessment of information awareness and correctness to a mere aesthetic matter [47] as we seem disposed to suspend our critical judgement when looking at data visualisation [32]: if it appears visually pleasing, then it will probably also be trustworthy, this because in visual design the boundary between seducing and informing is not so strict [48]. Furthermore, even with the proliferation of studies on information clutter and the observation that we live in a 'visual' society, there are minimal studies on the relationship between visual images and information disorder [49]. According to Hemsley & Snyder [48] the credibility of a fact - and the knowledge it

generates - depends on experience, perception, and social norms. As Fontana [50] argues, when people are faced with news, they apply a way of thinking based on cognitive fusion and belief systems, in other words, they simplify information, leading to a partial and incomplete interpretation, which leads to complete misrepresentation [48].

The basic requirement for understanding communicative-infographic artefacts is that understanding processes should take place to build inner representations. With respect to the consumption of infographics, it is possible to see how the visualisation of complex phenomena - considering the understanding/knowledge value of the Latin term *video* [7] - is the result of actions of encoding and decoding - using a language - of the message by an emitter and a receiver [51].

They - being a communicative tool for conveying information - necessarily require a reception phase of the visual message and therefore a perception phase to be understood as a process through which the information gathered by the sensory organs is arranged into objects, events or scenes equipped with meaning for the subject. This process is anything but purely objective and direct, as it is not limited to transferring the distal stimulus - i.e., the communicative-infographic artefact - into a proximal stimulus - the image imprinted on our retina - and consequently into a percept - i.e., the mental processing of it - without any need for integration or elaboration by our intellect. These comprehension processes are influenced by individual characteristics, such as domain content knowledge or visual-spatial skills of Graphicacy [29, 36, 52, 53] and the features of the stimulus, i.e., graphic, purpose and contense. Bertin [36] himself, with reference to graphic comprehension, identifies three levels of interaction - or questioning - that have an impact on the level of reading comprehension and that tie in with Curcio's [52] theory of comprehension:

- the level of the graphic system: to be understood as the canvas from which information is extracted, which generates an elementary reading.
- the level of internal processing: to be considered as the process of reduction - influenced by Gestalt theory - of the elements of the composition that pushes the subject to read new information through an intermediate reading.
- the level of external processing: which is configured as a general reading of the information in the graphic system, and therefore, a global reading.

More specifically, regarding the phase of the perception process of communicative-infographic artefacts, we can try to apply the Bayesian model of perception - an evolution of von Helmholtz's approach and unconscious inference - which, in adding up past experiences and stimuli, introduces a probabilistic constant whereby we tend to process the stimuli and produce a specific result on the basis of a mathematical model that may lead us to consider the information as correct or 'plausibly' more accurate. All this can lead to cognitive errors, the reason for which, according to Gillies & Giorello [54], lies in the fact that when faced with a general question or situation, the 'correct' answer cannot be found within a short period of time. For this reason, we tend to use a limited set of basic concepts and a suboptimal inferential mechanism to obtain the solution we consider to be optimal with respect to a balance between answer and process time, but not in an absolute sense: a compromise solution. Furthermore, being able to consider the communicative-infographic artefacts of sensory representations, as they are processed

through sight, we should consider the resistance to informational errors - such as illusions of optimality - which persist despite recognising their illusory nature and which in Data Visualization risk being misleading. In fact, as Neurath [33] explains, diagrams, while being of undoubted explanatory value, even though they are immediately comprehensible by numbers, they generate a sense of partial strangeness in observers without special skills and foment a feeling of not fully understanding.

In addition, the features of the message, such as consistency and shapes of representation, can induce people's trust in the information visualised. The (potential) intuitiveness of any communicative-infographic artefact is to be found in the visual nature of its language - strongly emphasised using iconographic elements - which generates a perception of simplicity, effectiveness and credibility that is superior even to written text [47]. A misunderstanding of the relationship with Data Visualisation is to assume that it - being based on scientific rigour, data and numbers - has unquestionable credibility, mapping a phenomenon in an unambiguous manner. However, any graphic is a representation of reality and can reveal as much as it can conceal [18] and the passage from data to information is a succession of actions and processes [55], which each gate can correspond to mistranslations - voluntary or not - that affect the entire final product.

A good communicative-infographic artefact - aesthetically speaking - decreases the levels of guarding and critical attitude. The interpretation and perception of different types of data are strongly influenced by the language and composition used in the visualisation. In visual information elaboration, this process can also be influenced by pre-attentive processing - which takes place in the sensory memory - that processes visual attributes such as colour, positioning, and shapes almost instantaneously, without the intervention of awareness [56]. The appearance of an infographic can attract or repel the reader's attention [36], which can distort their interpretation by emphasising misleading content and generating a distorted sense of credibility [49]; if they are neutralised by this emotion, they will be more likely to pay attention to what is shown on the representation. The confirmation bias should be read in this sense. According to Nickerson [57], this bias can occur when a person consciously or unconsciously restricts the field of analysis to observations only, i.e., data and information that are favourable to the confirmation of his or her beliefs, hypotheses, and expectations, disregarding and not examining any other alternative information that might disprove the acquired position. This condition appears to be common in individuals with information disorder [58]. In addition, Rajsic, Wilson and Pratt [59] suggest that people, when faced with the need for a simple visual analysis of an artefact, tend to apply this form of bias, which can be identified in two types of bias: the first form - active - aims to seek confirmation - the second - passive - to evade contradictory information. In fact, a person can actively or passively select information from his or her digital environment [60]: when a user actively seeks to identify and process only information that confirms his or her idea, he or she can be said to suffer from the active form of confirmation bias. On the other hand, if a person is passive, information that contradicts his or her idea of the world will be rejected and therefore not confirmed. In this case, one can say that one suffers from the passive form of confirmation bias. Faced with the risk of information overload, humans will tend to choose the simplest solution and with respect to a novelty - such as a piece of information - they will be inclined to uncritically choose that which most satisfies their

prior knowledge and does not lead to any disruption of their general coherence with the fact [61]. In a double study conducted by the University of Washington, Wobbrock et al. [62] it was found that articles with an average number of images of around three to seven were more credible than articles with very few or very many images. A weighted use of images to support discourse is therefore considered an element of credibility, which the authors refer to as the Goldilocks zone, a zone of balance between graphic and textual elements.

Finally, context plays an important role. As expressed by Bertin [36], a graph is not ‘drawn’ once and for all; it is ‘constructed’ and reconstructed until it reveals all the relationships formed by the interaction of the data. To make a useful graphic, we need to know what has gone before and what will follow’. According to this approach, therefore, the understanding does not lie in exclusively decoding the representation itself as a static object, but in comprehending the social actions through which the graphic was originally constructed [63]. In fact, the reader confusion is to be expected if it is not clear why a certain type of cut on reality has been made (problem data, visual scenario, complex theory, etc.) [64]. It is only by understanding the contexts for in which infographics are to be designed and how the data were obtained that a more complete understanding of graphics will be achieved. Therefore, understand the process of making the visual representation. The question is not about the specific type or quality of data, but how it will be presented, which can introduce errors and lead to wrong conclusions [23].

## 4 Data Visualization and Accessibility: A Pilot Studio

### 4.1 Methods and Design of the Research

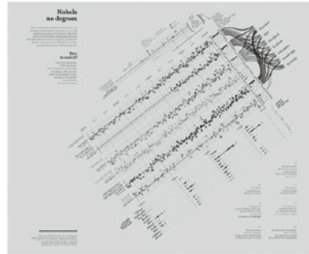
The study conducted examines the accessibility of information that is conveyed through infographics and specifically in five Data Journalism products. The perception and interpretation of users in the use of such content is analysed. It is therefore investigated whether the basic knowledge offered by educational curricula or mere previous experience is sufficient to obtain a good level of access to information. Therefore, the investigation focused on the following questions:

- Q1. Is infographic language understandable by all?
- Q2. Is there a correlation between the accessibility of infographics and the degree of representativeness?
- Q3. Is infographic literacy innate?
- Q4. Is training in design a determining factor in skill?

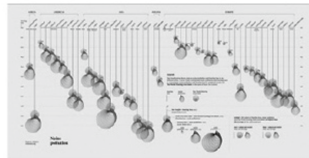
The infographics - selected according to the degree of iconicity of the representation, applying Anceschi’s [65] depiction scale (see Figs. 1 and 2) - were evaluated by a homogeneous sample of 200 graduates [ $M = 100 - F = 100$  - average age 22], divided into two groups according to the criterion of certified competence. Namely:

- Group A. Graduates in other disciplines.
- Group B. Graduates in Visual & Graphic Design and related disciplines.

**Infogr. 1 | Nobles, no degree.**  
G. Lupi in The Reading,  
Il Corriere della Sera  
**Iconicity GRADE: 7 Iv.**



**Infogr. 2 | Noise Pollution.**  
F. Fragapane in The Reading,  
Il Corriere della Sera  
**Iconicity GRADE: 6 Iv.**



**Fig. 1.** Selected Infographics and related Iconicity Grade. Disclaimer: Fair use of images for research purpose [1].

From a methodological point of view, the usability test was individually conducted by an on-line form, anonymously and guaranteeing all privacy and data protection regulations in compliance with GDPR. The study applied a three-variable correlation design: two independent variables (i) the System Usability Scale (SUS) and (ii) the degree of iconicity of the representation; and one di-pendant variable, namely the amount of information extracted from the infographic. The SUS scores were calculated based on Brooke's [66] scale and the Lewis-Sauro model [67]. For the ratings related to the amount of information extracted, the proportion of information each participant extracted was evaluated on a scale from 0 to 5. For degrees of iconicity, the 7-point scale remained the same.

## 4.2 Preliminary Results

Tables 1 and 2 present the mean, standard deviation and skewness values for the SUS questionnaire ratings, and the number of information extracted from all infographics, divided by the two samples. In general, both samples revealed a usability of the assessed infographics in terms of the degree of iconicity of their representation. In Group A, not a single infographic crosses the minimum threshold of 68 average points - indicative according to the SUS scale of an artefact at the limit of usability. In Group B, this threshold is instead overcome only by infographic No. 4 and No. 5.

The ratings of the two groups studied show a linear progressive trend as the submitted infographic becomes less and less iconic. In specific terms, Group B - Design graduates - score, on average, 38% higher than Group A - graduates from other disciplines, leading to the assumption that prior knowledge is crucial in terms of perceiving and comprehending the displayed information as argued by Kosslin [68] and Cairo [69]. Nevertheless, when analysing the results of the individual infographics in more detail, some interesting findings arise to answer the research questions at the start.





**Table 1.** Sus Questionnaire Results – Group A [1].

	<i>Mean</i>	<i>Std. D.</i>	<i>Asym.</i>	<i>Grade</i>	<i>Ico-Lv</i>
<i>Infogr. 1</i>	37,78	16,71	-0,39	F	7
<i>Extracted info</i>	1,71	1,35	0,39		
<i>Infogr.2</i>	45,17	21,33	1,47	F	6
<i>Extracted info</i>	2,72	1,02	-0,2		
<i>Infogr.3</i>	46,45	20,66	-0,28	F	5
<i>Extracted info</i>	2,43	1,38	0,12		
<i>Infogr.4</i>	55,68	20,8	1,59	D	4
<i>Extracted info</i>	2,61	-0,15	-0,43		
<i>Infogr.5</i>	58,02	23,97	-0,23	D	3
<i>Extracted info</i>	2,87	1,66	-0,13		

**Table 2.** Sus Questionnaire Results – Group B [1].

	<i>Mean</i>	<i>Std. D.</i>	<i>Asym.</i>	<i>Grade</i>	<i>Ico-Lv</i>
<i>Infogr. 1</i>	60,62	23,63	-0,52	D	7
<i>Extracted info</i>	2,95	1,61	-0,18		
<i>Infogr.2</i>	65,32	20,80	-1,22	C	6
<i>Extracted info</i>	4,33	1,38	-2,10		
<i>Infogr.3</i>	65,9	20,64	-1,12	C	5
<i>Extracted info</i>	3,76	1,33	-0,87		
<i>Infogr.4</i>	69,4	18,17	-0,77	C	4
<i>Extracted info</i>	3,63	1,41	-0,50		
<i>Infogr.5</i>	74,65	15,39	-1,09	B	3
<i>Extracted info</i>	3,75	1,32	-0,74		

Consider the iconicity grade 7 infographic and the iconicity grade 3 one (see Table 3). The infographic No. 1 achieves a SUS score in Group A of 37.78 (Grade F), in contrast to a 60.30 (Grade D) in Group B, showing a variation of 60.8% between the two results. In the infographic No. 2, the SUS value scored by the first sample is 58.02 (Grade D), compared to an average value of 74.65 (Grade C) in the second one, scoring a positive variance of 28.7%. The proportion of increase between the two sets of samples is a remarkable fact in that there is a progressive drop in the performance gap, the higher the degree of iconicity tends to value of less than 5. In fact, as can be seen in Table 3, the SUS values of Group B go from virtually sustained increases of + 60.5% (Infographic No. 1), to values of + 41.9% (Infographic No. 3), and the lowest increase value is + 24.7% (Infographic No. 4). Adding to this, Table 4 highlights that in both Group A and B, the

**Table 3.** Group A and B - Differences between SUS results compared [1].

	<i>Iconicity grade</i>	<i>SUS   Mean</i>	<i>SUS Grade</i>	
Infogr. 1	7			
<i>Group A</i>		37,78	F	–
<i>Group B</i>		60,92	D	+60,5%
Infogr. 2	6			
<i>Group A</i>		45,17	F	–
<i>Group B</i>		65,32	C	+44,6%
Infogr. 3	5			
<i>Group A</i>		46,45	F	–
<i>Group B</i>		65,90	C	+41,9%
Infogr. 4	4			
<i>Group A</i>		55,68	D	–
<i>Group B</i>		69,40	C	+24,7%
Infogr. 5	3			
<i>Group A</i>		58,02	D	–
<i>Group B</i>		74,65	B	+28,7%

average usability performance and the number of information extracted in the individual infographics follows a significant positive progression ( $r_{12}$ ) with values between 0.61 and 0.79.

In Table 5, two highly significant data emerge.

The first is the negative correlation between SUS and Iconicity ( $r_{13}$ ), a signal of an inversely proportional relationship between the abstraction of the representation and its ease of use. The second one, related to the first, demonstrates that the trend to retrieve information from the infographic tends to be favoured by its iconicity ( $r_{23}$ ). If in the former case we observe an extremely close correlation - with a value of  $-0.97$  in both groups - in the later, the range of values expands, moving between  $-0.77$  in Group A and  $0.28$  in Group B.

About question Q1, namely whether infographics are accessible, the results of the SUS show that the infographics submitted for the test - with the exclusion of the infographic No. 4 and No. 5 evaluated by Group B - do not pass, as an average score, the minimum levels of accessibility even though they are in potential to be very effective visual artefacts. Nevertheless, if we were to conduct a combined average of ratings between Group A and Group B - more likely to reflect the reality - none of the infographics would achieve the minimum rating of 68 (the highest value would be set at 66.3 of the infographic of iconicity grade 3). The individual ratings of the two groups in relation to the single infographic reveal how the accessibility of information is particularly variable within the same sample (see Fig. 3). Infographic No. 1 - iconicity grade 7 - scores above 68 pts by 47% of Group B and only 1% of Group A. Infographic No. 2, on the other hand,

**Table 4.** Correlation between Sus Questionnaire Results and extracted information [1].

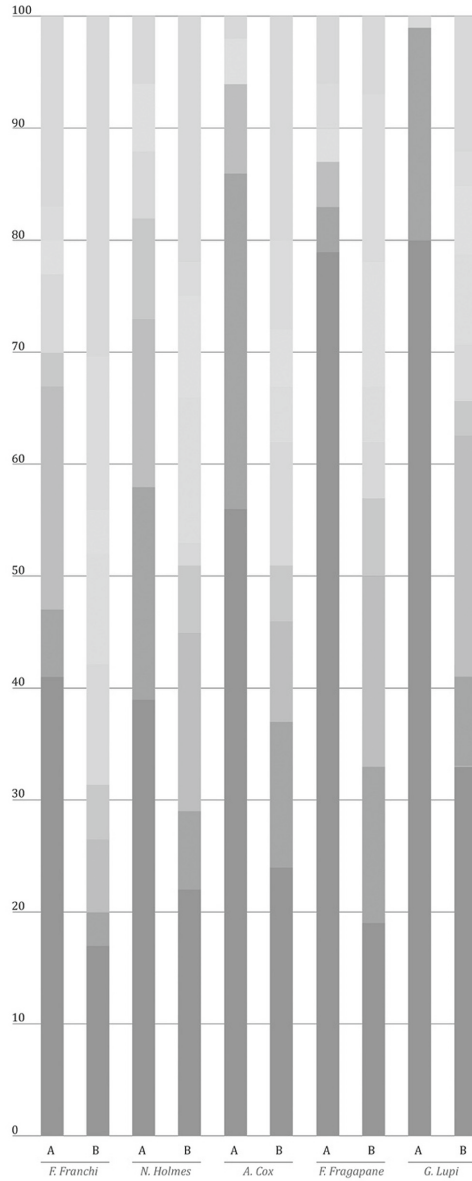
	<i>r 12   SUS and Extracted Information</i>
Infogr. 1	
<i>Group A</i>	0,66
<i>Group B</i>	0,73
Infogr. 2	
<i>Group A</i>	0,69
<i>Group B</i>	0,61
Infogr. 3	
<i>Group A</i>	0,64
<i>Group B</i>	0,64
Infogr. 4	
<i>Group A</i>	0,69
<i>Group B</i>	0,79
Infogr. 5	
<i>Group A</i>	0,70
<i>Group B</i>	0,63

**Table 5.** Correlation between Sus Questionnaire Results, extracted information and grade of iconicity [1].

	<i>r12</i>	<i>r13</i>	<i>r23</i>
<i>Group A + B</i>	0,87	-0,53	-0,29
<i>Group A</i>	0,82	-0,97	-0,77
<i>Group B</i>	0,27	0,97	-0,28

13% (A), 51% (B). Infographic No. 3, 10% (A), 59% (B). Infographic No. 4, 30% (A), 59% (B). Ultimately, infographic No. 5, 33% (A), 75% (B). This fluctuation suggests - answering question Q1 and Q3 - that the skill of reading visual artefacts cannot be considered an innate attribute and that infographics themselves are not so easily accessible in terms of information retrieval.

Reflecting on question Q2, namely whether iconicity of representation influences the usability of infographics, we can assume that in both samples, there is an increase in SUS ratings as the degree of iconicity leans towards the figurative as against the abstract. Group A increases by 53.6%, Group B by 23.2%. The greater growth in the first group may be because 'non-design-literate' subjects have greater problems in terms of processing abstract graphics, and in contrast, 'design-literate' subjects have less difficulty and for that reason, a more consistent performance. Keeping these values in mind,



**Fig. 3.** SUS scores distributed among all the two samples.

we can nevertheless consider the fact that the performance gap also tends to narrow according to the degree of iconicity, revealing a possible relationship between basic graphic competence and infographic reading ability. On the other hand, grade 6 and 7 infographics make use of a complete visual alphabet whose grammar is not intuitive, as is underlined by the large gap between the values of 60.5%. In contrast, grade 3

infographics, apart from being generally better evaluated by both samples, show a smaller delta between the different performances.

Last, respect to question Q4, prior knowledge in the discipline of Visual Design seems to favour greater reading skill than the performance obtained by Group A. Basic competence in Graphic Design seems to be a crucial factor in the encoding process specifically in visual representations that lead towards hypotheticals or pure abstraction (see No. 1 and 2 infographics). Nevertheless, it does not appear - now - to be a transversally acquired skill that is decisive in accessing the higher levels of information offered by data visualisation.

## 5 Conclusion

In summary, the contribution - starting from the evidence contained in the literature - focused on the issue of usability, i.e., the accessibility of information when represented through Information Design languages. In particular, the results obtained, and the correlations made confirm the trends found in the literature on the need for visual literacy for a correct decoding and perception of the information displayed. In general, almost all the infographics studied did not reach the minimum threshold of usability, thus opening the reflection to two questions, in terms of competence and design. The data at hand lead us to hypothesise that Graphicacy - which tends to be more developed in Designers, aided by the Designerly component - is decisive in achieving higher, though not excellent, levels of usability of communication-infographic artefacts. This points to the need for a democratisation of these skills not from a professionalising perspective but from a cultural and access perspective. Finally, the low level of usability achieved by the communicative-infographic artefacts raises questions in terms of design and the correct use of high levels of iconicity of data representation.

Today, people need to analyse information that is interconnected with society and the environment and that is continuously transmitted, remixed, and conditioned [20]. The visual translation of data into information makes use of a language with a specific grammar of signs and channels [36, 70]. However, reading images is far from intuitive as understanding the message can only take place if one is aware of the codes - such as the use of fonts, the iconographic choices, and the use of colour, as well as the arrangement of the pieces of a table, distilled over millennia of figurative and scriptural conventions [48]. If the correct encoding and decoding [18, 71] does not take place, communication fails [3]. The issue thus described fits into the international debate that has developed in recent years on the centrality of policy investment in digital literacy and digital skills to provide citizens with adequate cognitive tools to decode and encode information from data [72]. The difficulties are due - first - to a low level of what Balchin and Coleman [37] define as Graphicacy and which plays a key role in the cognitive learning process [73] and in Data Literacy.

From the preliminary analysis of the data, it emerges that the ability to read is necessary today and that studies on the skills necessary for correct decoding are more necessary than ever. Graphicacy and Basic Design alone do not appear to be so decisive in favouring this cognitive process. The test conducted raises issues in terms of both reading and production training. Finally, the low level of usability achieved by the

communicative-infographic artefacts raises questions in terms of Design and the correct use of high levels of iconicity in data representation. Even though the sample had an undergraduate level of education, the data and existing literature confirms that Graphicacy has been totally neglected in comparison to its ‘big brothers’ Literacy, Numeracy and Articulatory. This points to the need for a democratisation of these skills, not from a professional perspective but from a cultural and access perspective.

With the ever-increasing spread of information clutter, digital communication and the rise of data infrastructures, it is now more imperative than ever to incorporate teaching methodologies aimed at conscious production that takes into account the political, economic and social impacts - i.e., the economic and social impacts of the digital economy, economic and social impacts - i.e. an ethical dimension of the project - and, on the other hand, to an evaluation of the communicative-infographic artefacts with which we are confronted on a daily basis [74] by developing a critical attitude that avoids enthusiasm for uncritical data or dataisms [75] as it is necessary to learn how to read a graph before understanding it [18], since the veracity of the information contained in a data display is never absolute, but must be critically contextualised according to the objectives of those who want to use the initial data.

## References

1. Caccamo, A.: Data visualization, accessibility and graphicacy: a qualitative study of communicative artifacts through SUS questionnaire. In: Proceedings of the 18th International Conference on Web Information Systems and Technologies - WEBIST, SciTePress, pp. 422–430 (2022)
2. Wong, D.M.: *The Wall Street Journal Guide to Information Graphics: The Dos and Don'ts of Presenting Data, Facts, and Figures*. WW Norton, New York (2010)
3. Meirelles, I.: *Design for Information: An Introduction to the Histories, Theories, and Best Practices Behind Effective Information Visualizations*. Rockport Publishers, London (2013)
4. Dondis, D.A.: *Primer of Visual Literacy* (Rev. ed.). The MIT Press, Boston (1973)
5. Zarocostas, J.: How to fight an infodemic. *Lancet* **395**, 676 (2020)
6. WHO. <https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation>. Accessed 26 Mar 2023
7. Caccamo, A., Mariani, M.: Data Design: la Comunicazione progettata attraverso i dati. *Comunicazione Puntodoc*, pp. 138–148 (2020)
8. Jones, B., O'Donnell, K.: *Data Literacy Fundamentals: Understanding the Power & Value of Data*. Data Literacy Press, Boston (2020)
9. Marzocca, F.: *Il nuovo approccio scientifico verso la transdisciplinarietà*. Edizioni Mythos, Rome (2014)
10. Munari, B.: *Design e comunicazione visiva. Contributo a una metodologia didattica*. Laterza, Milan (2017)
11. Howard, P. N., Neudert, L., Prakash, N.: *Digital misinformation/disinformation and children*. United Nations Children's Fund (UNICEF) (2021)
12. Guess, A.M., Lyons, B.A.: Misinformation, disinformation, and online propaganda. *Social Media and Democracy: The State of the Field, Prospects for Reform*, pp. 10–33 (2020)
13. Wardle, C., Derakhshan, H.: *Information disorder: toward an interdisciplinary framework for research and policy making*. Council of Europe Report, 27. Council of Europe, Strasbourg (2019)

14. Thomson, T.J., Angus, D., Dootson, P., Hurcombe, E., Smith, A.: Visual mis/disinformation in journalism and public communications: current verification practices, challenges, and future opportunities. *Journalism Practice*, pp. 1–25 (2020)
15. Huff, D.: *Mentire con le statistiche*. Monti & Ambrosini, Milan (2007)
16. Jones, B.: *Avoiding Data Pitfalls: How to Steer Clear of Common Blunders When Working with Data and Presenting Analysis and Visualizations*. Wiley, Hoboken (2019)
17. Loukissas, Y.A.: *All Data Are Local: Thinking Critically in a Data-Driven Society*. The MIT Press, Boston (2019)
18. Cairo, A.: *Come i grafici mentono. Capire meglio le informazioni visive*. Cortina Raffaello, Milan (2020)
19. Maldonado, T.: *Reale e virtuale*. Feltrinelli, Milan (2005)
20. Manovich, L.: *Remixability and Modularity*. [http://manovich.net/content/04-projects/046-remixability-and-modularity/43\\_article\\_2005.pdf](http://manovich.net/content/04-projects/046-remixability-and-modularity/43_article_2005.pdf). Accessed 26 Mar 2023
21. Riva, G.: *Fake news. Vivere e sopravvivere in un mondo post-verità*. Il Mulino, Bologna (2018)
22. Beattie, V., Jones, M.J.: The impact of graph slope on rate of change judgments in corporate reports. *Abacus* **38**(2), 177–199 (2002)
23. Tufte, E.R.: *The visual display of quantitative information* (2nd ed.). Graphics Press (2002)
24. Meyer, J., Shinar, D., Leiser, D.: Multiple factors that determine performance with tables and graphs. *Hum. Factors* **39**, 268–286 (1997)
25. Carpenter, P.A., Shah, P.: A model of the perceptual and conceptual processes in graph comprehension. *J. Exp. Psychol. Appl.* **4**(2), 75 (1998)
26. Cleveland, W.S., McGill, R.: Graphical perception and graphical methods for analyzing scientific data. *Science* **229**(4716), 828–833 (1985)
27. Freedman, E.G., Shah, P.: Toward a model of knowledge-based graph comprehension. In: Hegarty, M., Meyer, B., Narayanan, N.H. (eds.) *Diagrammatic Representation and Inference*. Diagrams 2002. LNCS, vol. 2317, pp. 59–141. Springer, Berlin (2002). [https://doi.org/10.1007/3-540-46037-3\\_3](https://doi.org/10.1007/3-540-46037-3_3)
28. Glazer, N.: Challenges with graph interpretation: A review of the literature. *Stud. Sci. Educ.* **47**(2), 183–210 (2011)
29. Friel, S.N., Curcio, F.R., Bright, G.W.: Making sense of graphs: Critical factors influencing comprehension and instructional implications. *J. Res. Math. Educ.* **32**(2), 124–158 (2001)
30. Shah, P., Hoeffner, J.: Review of graph comprehension research: Implications for instruction. *Educ. Psychol. Rev.* **14**(1), 47–69 (2002)
31. Lupton, E.: *Visual dictionary*. In: *ABC's of the Bauhaus: The Bauhaus and Design Theory*. Princeton Architectural Press, Princeton (2019)
32. Drucker, J.: *Graphesis: Visual Forms of Knowledge Production*. Harvard University Press, Cambridge (2014)
33. Oliverio, S.: *Pedagogia e Visual Education*. Unicopli, Milan (2006)
34. Moholy-Nagy, L.: *Vision in Motion*. Paul Theobald, Chicago (1946)
35. Kepes, G.: *Il linguaggio della visione*, vol. 2. Edizioni Dedalo, Bari (1990)
36. Bertin, J.: *Semiology of Graphics*. Amsterdam University Press, Amsterdam (2011)
37. Balchin, W., Coleman, A.M.: Graphicacy should be the fourth ace in the pack. *Cartographica. Int. J. Geograph. Inf. Geovisual.* **3**(1), 23–28 (1966)
38. Schön, D.: *The Reflective Practitioner*. Temple-Smith, London (1983)
39. Culbertson, H.M., Powers, R.D.: A study of graph comprehension difficulties. *Audio Vis. Commun. Rev.* **7**, 97–110 (1959). <https://doi.org/10.1007/BF02767016>
40. Bowen, G.M., Roth, W.M.: Graph interpretation practices of science and education majors. *Can. J. Sci. Math. Technol. Educ.* **3**(4), 499–512 (2003)
41. Preece, J.: A survey of graph interpretation and sketching errors. Open University




42. Bowen, G.M., Roth, W.M.: Graph interpretation practices of science and education majors. *Can. J. Sci. Math. Technol. Educ.* **3**(4), 499–512 (2003)
43. Åberg-Bengtsson, L., Ottosson, T.: What lies behind graphicacy? Relating students' results on a test of graphically represented quantitative information to formal academic achievement. *J. Res. Sci. Teach.* **43**(1), 43–62 (2006)
44. Lau, A., Moere, A.V.: Towards a model of information aesthetics in information visualization. In: 2007 11th International Conference Information Visualization (IV2007), pp. 87–92. IEEE (2007)
45. Cox, R., Brna, P.: Supporting the use of external representations in problem solving: the need for flexible learning environments. *J. Artif. Intell. Educ.* **6**, 239–302 (1995)
46. Dahmen, N.S., Mielczarek, N., Perlmutter, D.D.: The influence-network model of the photojournalistic icon. *J. Commun. Monographs* **20**(4), 264–313 (2018)
47. Hemsley, J., Snyder, J.: Dimensions of visual misinformation in the emerging media landscape. In: Southwell, B., Thorson, E.A., Sheble, L. (eds.) *Misinformation and Mass Audiences*. University of Texas Press, Austin (USA) (2018)
48. Falcinelli, R.: *Critica portatile al visual design: da Gutenberg ai social network: [come informano, narrano e seducono i linguaggi che ci circondano]*. Einaudi, Turin (2014)
49. Brennen, J.S., Simon, F.M., Nielsen, R.K.: Beyond (mis) representation: visuals in COVID-19 misinformation. *Int. J. Press/Politics* **26**(1), 277–299 (2021)
50. Fontana, A.: *Fake news: sicuri che sia falso? Gestire disinformazione, false notizie e conoscenza deformata*. Hoepli, Milan (2018)
51. Eco, U.: *Sémiologie des messages visuels*. Communications **15**(1), 11–51 (1970)
52. Curcio, F.R.: Comprehension of mathematical relationships expressed in graphs. *J. Res. Math. Educ.* **18**, 382–393 (1987)
53. Gal, I.: Adults' statistical literacy: meanings, components, responsibilities. *Int. Stat. Rev.* **70**(1), 1–25 (2002)
54. Gillies, D., Giorello, G.: *La filosofia della scienza del XX secolo*. Bari: Laterza (1995)
55. Cristallo, V., Mariani, M.: From data gate to story gate. Territory visualization models and processes for design driven actions. In: 3rd International Conference on Environmental Design, Mediterranean Design Association, Marsala (2019)
56. Ware, C.: Visual queries: the foundation of visual thinking. In: Tergan, S.O., Keller, T. (eds.) *Knowledge and Information Visualization*. LNCS, vol. 3426, pp. 27–35. Springer, Berlin, Heidelberg (2005). [https://doi.org/10.1007/11510154\\_2](https://doi.org/10.1007/11510154_2)
57. Nickerson, R.S.: Confirmation bias: a ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* **2**(2), 175–220 (1998)
58. Kim, A., Moravec, P.L., Dennis, A.R.: Combating fake news on social media with source ratings: the effects of user and expert reputation ratings. *J. Manag. Inf. Syst.* **36**(3), 931–968 (2019)
59. Rajsic, J., Wilson, D.E., Pratt, J.: Confirmation bias in visual search. *J. Exp. Psychol. Hum. Percept. Perform.* **41**(5), 1353 (2015)
60. Zollo, F., Bessi, A., Del Vicario, M., Scala, A., Caldarelli, G., Shekhtman, L.: Debunking in a world of tribes. *PLoS ONE* **12**(7), e0181821 (2017)
61. Quattrociochi, W., Vicini, A.: *Misinformation. Guida alla società dell'informazione e della credulità*. FrancoAngeli, Milan (2018)
62. Wobbrock, J.O., Hattatoglu, L., Hsu, A.K., Burger, M.A., Magee, M.J.: The Goldilocks zone: young adults' credibility perceptions of online news articles based on visual appearance. *New Rev. Hypermedia Multimedia* **27**(1–2), 51–96 (2021)
63. Roth, W.M., McGinn, M.K.: Inscriptions: toward a theory of representing as social practice. *Rev. Educ. Res.* **68**(1), 35–59 (1998)
64. Perondi, L.: *Sinsemie: scritture nello spazio*. Stampa Alternativa/Nuovi Equilibri, Viterbo (2012)

65. Anceschi, G.: *L'oggetto della raffigurazione*. Etaslibri, Milan (1992)
66. Brooke, J.: SUS: a “quick and dirty” usability. *Usab. Eval. Indus.* **189**(3), 6 (1996)
67. Lewis, J.R., Sauro, J.: Item benchmarks for the system usability scale. *J. Usab. Stud.* **13**(3), 158–167 (2018)
68. Kosslyn, S.M.: *Elements of graph design*. W.H. Freeman and Company, New York (1994)
69. Cairo, A.: Uncertainty and graphicacy: how should statisticians, journalists, and designers reveal uncertainty in graphics for public consumption?. In: Errea, J.G. (ed.) *Visual Journalism: Infographics from the World’s Best Newsrooms and Designers* (Translation ed.). Gestalten, Neustadt (2017)
70. Horn, R.E.: *Visual language: global communication for the 21st century*. MacroVU, Bainbridge Island (1998)
71. Wilmot, D.: Investigating children’s graphic skills: a south african case study. *Int. Res. Geograp. Environ. Educ.* **11**(4), 325–340 (2002)
72. Carretero Gomez, S., Vuorikari, R., Punie, Y.: *DigComp 2.1: the digital competence framework for citizens with eight proficiency levels and examples of use*. EUR 28558 EN. Publications Office of the European Union, London (2017)
73. Danos, X.: *Graphicacy and Culture: Refocusing on Visual Learning*. Design Press Ltd., Loughborough (2018)
74. Thompson, D.S.: Teaching students to critically read digital images: a visual literacy approach using the DIG method. *J. Vis. Literacy* **38**(1–2), 110–119 (2019)
75. Mauri, M., Colombo, G., Briones, M.D.L.Á., Ciuccarelli, P.: Teaching the critical role of designers in the data society: the DensityDesign approach. In Börekçi, N., Koçyıldırım, D., Korkut, F. Jones, D. (eds.) *Insider Knowledge, DRS Learn X Design Conference 2019*, pp. 9–12 (2019)



# Evaluating the Quality Characteristics of Space Geeks

Abdelbaset Assaf<sup>(✉)</sup> , Lana Issa , and Mohammed Eshtay 

Luminus Technical University College, Amman, Jordan  
{a.assaf, l.issa, m.eshtay}@ltuc.com

**Abstract.** Rapid developments in technology around the world require continuous efforts in academia to develop innovative solutions to enhance the learning of basic and advanced programming concepts. Educators understand the value of preparing highly skilled programmers to join the industry, and this puts a responsibility to develop and enhance teaching techniques in order to get better results, especially in introductory courses where novice programmers face difficulties in understanding basic programming concepts. Serious games were proven to be effective, motivational and beneficial for novice programmers to support them during their learning journey. In this paper, we investigate the issues that students face when learning arrays and based on that, we design and implement Space Geeks, a serious game targeted at teaching arrays for novice programmers. We discuss the design principles used to develop Space Geeks, and we test and evaluate the game in an educational environment. Our findings show that Space Geeks, our developed serious game, is promising to help novice programmers improve in learning basic programming concepts, due to its ease of use and ease of understanding.

**Keywords:** Serious games · Introductory programming · Teaching arrays · Serious games evaluation · Game-based-learning

## 1 Introduction

Teachers in Introductory Programming courses are under a continuous challenge to help learners form a correct understanding of the basics in programming, including understanding the programming environment, and the logic of building programs. Students in introductory programming courses face many difficulties understanding the syntax, concepts, and strategies of programming, and this leads to various misconceptions in programming and could affect students' attitudes towards programming and their scholar achievement [34].

Various research efforts have been put to analyse the different areas of misconceptions that students fall into, such as misconceptions about the syntax [3] and misconceptions about the concepts [34]. Also, many initiatives are taken towards investigating potential factors that could lead to these different misconceptions, in different learning environments and different educational levels.

Arrays are an important topic in introductory programming, which was previously investigated by many researchers to improve the delivery of this topic to students since

it was proven that many students face difficulties in understanding and dealing with arrays [10].

There is an essential need for the continuous development of innovative solutions that could help in delivering programming concepts to students, especially with the changes in many different learning environments around the world and the absence of the traditional scene of classrooms. Serious Games (SG) are considered one of the trending techniques that received the attention of many researchers and educators in the programming education field, due to their effect on students' attitudes, motivation, and encouraging active learning with its interactivity [4].

This paper is an extension of our recent paper [4] where we have proposed a serious educational game targeted at teaching arrays concept to novice programmers. In this paper, we implement and evaluate the serious game, Space Geeks, which is a sci-fi-based game targeted at supporting novice programmers in achieving more in introductory programming courses. Especially in environments where students do not have much experience in programming concepts in schools. The initial design of the game is targeted at learning arrays. According to our previous work, we conducted an initial survey of the rating difficulty of every programming concept in the introduction to Programming course, and 53% of the participants rated arrays as a hard topic. In this paper, we continue the design and implementation of our previously proposed game, and we present our early findings of performing tests and evaluations on the implemented game.

## 2 Identifying Students Misconceptions About Arrays

There have been previous discussions about students' misconceptions of arrays and other introductory programming topics, and that is because of the new concepts they are learning in programming that have no relation to their previous knowledge about maths, physics or any other basic topics they learn in schools [31].

Arrays being the first data structure that students are introduced to, absorbing the idea of data, and data structures, and operations performed on these data structures to build algorithmic solutions, could be challenging, especially for students with no previous background in computing.

Students' mistakes in misconceiving arrays, could be in distinguishing the value from the index [6], understanding the data structure shape itself of consequent elements in the computer memory, understanding the process of direct access to elements inside the array using the index, iterating among array elements using loops, and being able to read and write array elements.

The misconception in arrays could lead to further misconceptions in other topics that are built on and extended from arrays, such as Stacks, Queues, Circular Arrays, ArrayLists, etc., and could affect the whole understanding of data structures later on. Researchers found that some of the reasons students face difficulties in data structures courses are because of their original misconceptions about introductory topics and essential concepts such as arrays, as the first data structure to be presented to students [12].

In order to further investigate the overall understanding of arrays among students, we conduct an evaluation survey for students from three computing majors, Software

Engineering, Cloud Computing, Cyber Security, and Artificial Intelligence. Total of 77 students, all in their first semester and taking an introduction to programming course using Java programming language. In this course, students are introduced to multiple basic concepts in introduction to programming, such as variables, arithmetic operations, conditions, loops, arrays, flowcharts, and algorithms design.

**2.1 Survey Results**

In order to effectively design the game to target the weaknesses among students and develop the right activities that could support students in learning the programming concepts, and avoid the misconceptions that could be formed. We conduct an initial survey targeted specifically at arrays since that we previously found it to be rated by students as one of the hardest topics in the introduction to programming. The survey was conducted on higher education students studying British diploma programs in Software Engineering, Cyber Security, Cloud Computing, and Artificial Intelligence, and all of these majors share the first year as the basic computing courses are common, so all students are similar in terms of the course level.

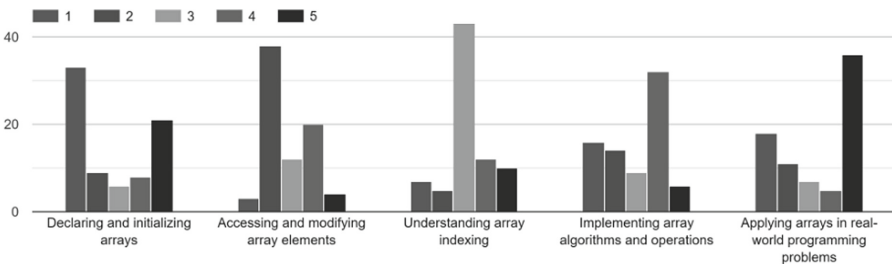
First of all, we asked the 77 students to evaluate the difficulty of the arrays topic, and 65% of students rated arrays as “difficult”, then, we ask further questions to determine the areas of difficulty regarding this topic.

We ask students to select all difficulties that they face during learning arrays, 37% of students selected having difficulty in understanding the syntax of arrays, 35% of students selected having difficulty in manipulating array elements, 28% of students said they have trouble in determining how the array is handled in the computer memory, 28% of students expressed having difficulties in debugging errors when using arrays in code, and 16% of students said they have trouble in distinguishing the index from the element in the array.

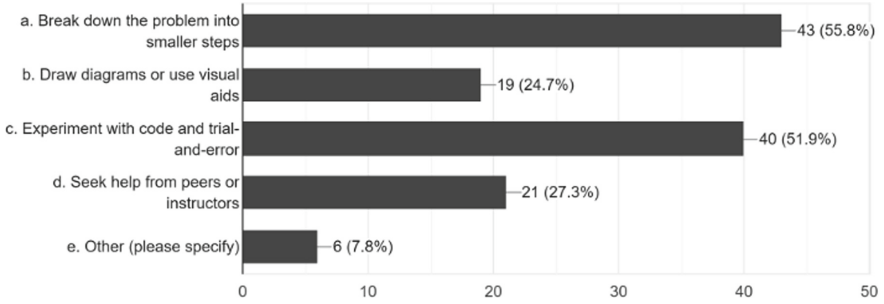
Furthermore, we asked students to rank the difficulty of every activity performed during the process of learning arrays in classes, the results are shown in Fig. 1.

We also asked students if they find difficulty in visualising how array elements are ordered and stored in the computer memory, the results showed that 59% of students answered that they have this difficulty.

To further investigate the students’ behaviour and levels, we asked them to specify how they solve programming challenges that are about arrays, the results are shown



**Fig. 1.** Ranking the difficulties of activities performed when learning arrays.



**Fig. 2.** Students approached to solving programming problems about arrays.

in Fig. 2, as the majority of students use trial-and-error approach and break down the challenge into smaller tasks to be able to understand how every part is constructed in the final answer. Other students chose visualisation by drawing array elements which will help them in visualising how array elements are stored and how they should be manipulated to solve the programming challenge, but we note here that this option was only selected by 27% of the students, and that is expected since that many students expressed in the previous question having difficulty in visualising arrays elements.

In the end, we ask students if they have difficulties learning other concepts in the same course, and the results find that 35% of the students selected had difficulties with understanding and using loops, 20% of students selected had difficulties in understanding and using conditions, and 13% of students selected having difficulties in understanding and using variables in programming, while 46% of students said they have no difficulties, 48% of this category of students are the ones who expressed not having difficulties with arrays in the first place, and the rest are students who only showed having difficulties in arrays topic and not other topics.

Such analysis and investigation have helped us to identify the areas of weaknesses regarding arrays, and which practices students consider hard when studying and practising arrays using a programming language. In the game we present, we provide visualisation using graphics and animation used in the game modelling. We also design the game levels to cover many aspects of learning arrays starting from the basic aspect which is understanding what is an index and what is an element and how they are structured in an array, moving towards manipulating and debugging arrays effectively.

### 3 Designing Space Geeks

Designing Space Geeks was based on different theories and design principles as follows:

#### 3.1 Hero and Narrative

In order to create an engaging video game, it is essential to have a main hero or character and immerse the players in a compelling story. This design principle is a key reason why video games are so successful [17]. With this in mind, we followed this theory

when designing Space Geeks. The game features a main character, and the storyline revolves around leading this character out of a building so that he can escape. Along the way, players must overcome a series of challenges represented by programming tasks. By incorporating this design principle, Space Geeks offers players an immersive and engaging experience.

### **3.2 Curriculum Alignment**

It is vital to align the content of a game with its intended learning objectives. This means designing the game to impart specific knowledge or skills to the player. The mechanics and gameplay must be directly related to the learning objectives to consistently engage and reinforce the player's understanding of the material [29]. Serious game design often neglects this critical principle. However, Space Geeks prioritizes this by focusing on one programming concept (arrays) and providing engaging array programming tasks to strengthen the player's comprehension.

### **3.3 Engagement and Interactivity**

In this programming game, the character encounters a range of tasks that demand both coding and problem-solving skills [33]. Each activity is distinct and poses new challenges. Further, the game ensures the constant engagement of the player through the different tasks that must be completed in order to progress

### **3.4 Mastery and Skill Development**

Serious games are a powerful tool for facilitating skill development and mastery. Through progressive challenges and scaffolded learning experiences, players can gradually build competence and expertise [26]. Space Geeks is a perfect example of this approach in action. Beginning with a straightforward question regarding the declaration of an array of integers, the game gradually increases the complexity of the challenges, providing ample opportunities for players to hone their skills and improve their game. By challenging themselves and pushing their limits, players can reach new levels of competence and mastery.

### **3.5 Familiar Interactions**

To ensure engagement among students, we have incorporated familiar game mechanics from popular video games. The design principle behind this is using controls that are easy to use and are similar to what's commonly seen in the gaming world [27]. For instance, the game allows character movement through keyboard controls and uses the mouse for looking around. This way, players can easily navigate through the game while having a familiar user experience. The keyboard is used as follows:

- W: Pressing the “W” key moves the character forward in the game.
- A: Pressing the “A” key moves the character to the left.
- S: Pressing the “S” key moves the character backwards.

- D: Pressing the “D” key moves the character to the right.
- Spacebar: Pressing the spacebar is used for actions such as jumping or vaulting over obstacles.

The game also allows the use of console controllers for movement. However, a keyboard is needed to complete the tasks which are coding tasks.

### 3.6 Character Animation

In the world of serious games, character animation can be a powerful tool for creating engaging interactions and bringing the game’s characters to life. By using animation to convey emotions, behaviours, and gestures, developers can enhance the storytelling experience and immerse players even further into the game world [20]. Space Geeks is a 3D game developed around a main character with appropriate animation for character movement and engagement in the 3D world.

### 3.7 Purposeful Animation

Serious games should purposefully and intentionally incorporate animation in order to enhance the overall gameplay experience while also supporting the learning objectives. By using animation in a variety of ways such as demonstrating processes, simulating real-world interactions, providing visual feedback, or engaging players through interactive and appealing elements, the educational potential of the game is maximized [14]. For example, in Space Geeks, after successfully completing a task, an animation is played to represent the player’s progress. When a coding task requires the student to print the values of an array, a corresponding animation will demonstrate the process by illustrating how the index is updated with each iteration, displayed through a collection of books and ultimately copied on a board in front of the player.

## 4 Serious Games Evaluation

Evaluating serious games is very important to ensure the validity of serious games, and to measure to what extent the game covers the intended learning goals. Several studies that presented different serious games, discussed many areas of evaluation to assess the different characteristics of serious games. For example, some studies focused on the engagement characteristic [2], others considered usability [16,28], whereas usability is an umbrella that could hold many sub-features to be evaluated in the serious games, such as effectiveness and satisfaction [22]. (Yáñez-Gómez, 2017), talks about different paths followed by previous researchers to conduct usability evaluation over serious games, such as, Questionnaires, focus groups, observations, heuristic evaluations, and many other types of evaluations that could be used to evaluate usability. serious games evaluation also considers features that are related to students feelings towards the serious game such as motivation [32] and enjoyment [35].

On the other hand, we find that many studies evaluated serious games by combining several quality characteristics such as the GameFlow model, introduced by Sweetser



and Wyeth in 2005. It primarily concentrates on the players' subjective experience. It evaluates the quality of serious games by considering factors, such as immersion, challenge, and enjoyment [30]. Additionally, the Serious Game Evaluation Framework, formulated by Connolly in 2012, presents a holistic set of criteria for assessing serious games, encompassing aspects like pedagogy, usability, functionality, and engagement, thereby offering a systematic approach to gauging their quality [11].

In 2013, Bellotti developed another framework called the Game-Based Learning Quality Framework (GBLQF), which emphasizes evaluating the learning effectiveness of serious games. This framework covers aspects like educational alignment, gameplay design, feedback mechanisms, and assessment strategies [5]. Its primary objective is to ensure that serious games achieve educational goals while simultaneously delivering a captivating experience for learners. Furthermore, the Mechanics, Dynamics, and Aesthetics (MDA) framework, proposed by Hunicke in 2004, is a conceptual model designed to assess game design elements. Although not explicitly tailored for serious games, the MDA framework has proven valuable in evaluating their quality characteristics. It sheds light on the relationships between game mechanics, player interactions, and the emotional experiences elicited by the game [15].

While the literature on evaluating serious games is rich with various evaluation frameworks, we have chosen to utilise the five-dimensional evaluation framework by Abdellatif [1] for our evaluation. We found this framework inclusive, direct, and has the potential to enhance the quality of Space Geeks, the serious game we are evaluating. The following section will provide a detailed explanation of each quality characteristic within the five-dimensional evaluation framework.

## 5 Evaluating the Quality Characteristics of Space Geeks

### 5.1 The Five-Dimensional Evaluation Framework

Several quality characteristics can be evaluated in serious games. Calderon and Ruiz (2015) summarised 18 quality characteristics that have been used in evaluating serious games. Further, Abdellatif [1] proposed a five-dimensional evaluation framework by analysing the 18 quality characteristics and then dividing them into primary and secondary characteristics. The five-dimensional evaluation framework evaluates five of the primary quality characteristics which are: Usability, Motivation, Understandability, User Experience and Engagement. The framework assesses the above-mentioned characteristics with three questions each, which represent and measure specific factors for each characteristic.

1. **Usability** When designing serious games, usability is a critical characteristic that directly affects the game's overall effectiveness and user experience. A game with good usability allows players to easily navigate the interface, comprehend the game mechanics, and achieve their learning objectives without frustration or confusion. According to [25], usability is essential in serious games to create an immersive learning environment that enables effective knowledge acquisition and skill development. Additionally, a study by Boyle [7] confirmed that high usability enhances player engagement and enjoyment, leading to increased motivation to invest time

and effort in the game. Prioritizing usability allows serious game designers to create accessible, intuitive, and user-friendly experiences that optimize the game's learning potential and enhance user satisfaction.

2. **Motivation** It is a critical characteristic in serious game design as it significantly impacts players' engagement, persistence, and learning outcomes. Motivated players are more likely to invest effort and time in the game, leading to deeper learning experiences and improved knowledge acquisition. According to Malone [21], intrinsically motivating instruction, such as that provided by well-designed serious games, can enhance learners' motivation and promote active engagement. Additionally, a study by Kiili and Ketamo [19] highlights that motivated players are more likely to perceive serious games as enjoyable and continue using them for learning purposes. By incorporating motivational elements such as challenges, rewards, progression, and social interaction, serious game designers can foster intrinsic motivation, sustaining players' interest and enhancing the effectiveness of the learning experience.
3. **Understandability** It is a vital characteristic in serious game design as it directly affects players' ability to comprehend and make sense of the game's content and mechanics. A game with high understandability ensures that players can easily grasp the instructions, rules, and objectives, enabling them to engage with the learning materials effectively. According to Hainey [14], clear and concise instructions, along with intuitive interface design, contribute to the understandability of serious games, facilitating players' learning process. Additionally, research by Nacke [24] emphasizes that understandable serious games promote player engagement and reduce frustration, leading to enhanced learning outcomes. By prioritizing understandability, serious game designers can create accessible and user-friendly experiences that promote effective learning and ensure players can fully engage with the game's educational content.
4. **User Experience** In serious game design, creating a positive user experience is pivotal to ensure that players are fully engaged and motivated to learn. The overall quality of player interaction with the game is encapsulated in the UX, and therefore it is vital to prioritize its design. Studies have shown that effective user experience design in serious games improves player satisfaction, engagement, and intrinsic motivation, leading to better learning outcomes [13]. Additionally, Nacke and Lindley [23] noted that usability, aesthetics, and emotional engagement must be considered in serious game design to create a memorable and compelling experience for the player. By prioritizing the user experience, serious game designers can guarantee that players have a rewarding and impactful learning journey within the game.
5. **Engagement** It is a decisive component in the creation of effective serious games. It greatly influences players' interest, attention, and active participation in the game environment. Achieving high levels of engagement is vital in order to create an immersive and compelling learning experience that can lead to deep learning. According to Boyle [7], engagement in serious games has a positive impact on motivation, learning outcomes, and knowledge retention. Moreover, Kiili's [18] research highlights that engaged players are more likely to experience a state of flow, characterized by intense focus and enjoyment, resulting in enhanced cognitive and affective involvement. Serious game designers can facilitate player engagement by incorporating elements such as challenges, feedback, interactivity, and meaningful

narratives. This approach ensures that players remain motivated and committed to the learning experience. Therefore, serious game designers play a crucial role in fostering effective learning through engaging game design.

## 5.2 The Evaluation

The five-dimensional evaluation framework was used to evaluate Space Geeks. 49 first-year students played the Space Geeks game for around 20 min. Then they were asked to rate the 15 evaluation factors on a scale from 1 to 10. The students were taking the Introduction to Programming course and they included Bachelor and British Diploma students. The students consisted of 20 female students and 29 male students (Table 1).

**Table 1.** The evaluation.

Quality Characteristic	Question	Rating	Total
Usability	The game is a useful learning tool for computer programming	8.95	8.19
	The game does not contain errors	7.38	
	The game is easy to use	8.24	
Motivation	The game is challenging	6.59	7.59
	Playing the game is enjoyable	8.28	
	Playing the game sparked my curiosity	7.89	
Understandability	The game goals are clear and understood easily	7.26	7.38
	The game offers a set of straightforward steps to be followed	7.18	
	The game can be played individually without the need for assistance	7.71	
Engagement	The purpose of the game is appealing	8.30	8.05
	The idea and the storyline of the game are interesting	7.67	
	The game is self-controlled by the user and allows making decisions	8.18	
User experience	The game is competitive	7.22	6.95
	The game allows social interaction	5.26	
	The game promotes the involvement of the learner	8.36	

Based on the ratings, it is evident that the game is a valuable learning tool for computer programming, receiving an impressive rating of 8.95. However, students reported some errors while using the game. The rating of game error-free average is 7.38 indicating that some errors may still need fixing. Nonetheless, ease of use maintained a positive rating of 8.24. In conclusion, the game's overall usability is rated positively, with room for improvements to address the reported errors.

Based on the ratings received, we can conclude that while playing the game is enjoyable and sparks curiosity, there is room for improvement in terms of the level of challenge. The game received a rating of 6.59 for its level of challenge, indicating that it may not be sufficiently challenging for some users. However, the gameplay received a high rating of 8.28 for being enjoyable and maintained a rating of 7.89 for sparking curiosity among players.

When it comes to the understandability of the game, it seems that there is still some room for improvement. While the game's goals are generally clear and received a rating

of 7.26, there is an opportunity to enhance its clarity even further. Currently, the availability of straightforward steps to follow is moderately rated at 7.38, and the game's ability to be played individually without assistance is higher, with a rating of 7.71. Based on these ratings, it appears that the game could benefit from further improvements to ensure that players can easily understand its objectives and gameplay mechanics.

In terms of engagement, the game's purpose has received a high rating of 8.30, indicating its appeal to players. However, there is room for improvement in the idea and storyline, which gathered a rating of 7.67. To maximize engagement, enhancing these aspects would be key. On another side, the game's self-controlled nature and decision-making capabilities were positively received with a rating of 8.18. Overall, while the game's purpose is already captivating, adding depth to the idea and storyline would further improve player engagement.

As per the users' ratings, the game has a strong competitive component, receiving a solid 7.22 rating. The game's ability to facilitate social interaction, on the other hand, scored a lower 5.26 rating, indicating room for improvement in this aspect. However, the game received a high rating of 8.36 for promoting learners. It can be concluded that the game provides an engaging competitive experience and effectively supports learner engagement. Nevertheless, since social interaction is a crucial aspect of the user experience, addressing its low rating is essential for enhancing the overall user experience.

In summary, the game is a great learning tool that boasts exceptional usability and ease of use. It's not only fun and engaging but also piques the player's curiosity, promoting active involvement. However, there are some areas that require improvement to enhance the overall user experience. Firstly, potential errors need to be addressed, and the level of challenge should be increased. Additionally, the game's concept and storyline could be improved, and its social interaction limitations addressed. Enhancing the understandability of the game would also ensure a more immersive and enjoyable experience. By addressing these aspects, the game can be further improved and provide greater engagement for its players.

The evaluation survey provided users with a section to leave comments about their experience playing the serious game. However, very few students utilized this space to share their thoughts, and those who did mainly expressed their enthusiasm for the game. As our main objective was to assess the game's efficacy, we centred our attention on comments that contained constructive criticism, such as:

- "Change the colors to make it clearer and improve the screen where the code is written"
- "Controlling the player direction is difficult and the arrow keys don't control the player movement"
- "The control should be smooth and collective"

Valuable feedback was received regarding the game's usability, particularly regarding colour choices and character movement. One noteworthy aspect to address is that despite using the well-known gaming keys (WASD) for character movement, many students struggled to figure it out. During gameplay, non-gamers made comments about their unfamiliarity with controlling the character. These insights provide prime opportunities to improve the game's user experience. In order to enhance the gaming experience, it would be advisable to review the colour scheme choices within the game.

Additionally, incorporating arrow keys as a means of controlling the character's movements would be beneficial. It may also be useful to provide a brief introduction to the controls at the start of the game. Additionally, implementing these changes presents an opportunity to study the impact and effectiveness of these enhancements on the game.

Another important comment from a student was:

“Possibility of entry with a group, having a competition with them and having scores after completing any level could get us more excited than just playing with no visible progress”

The student's suggestion of introducing group entry, competition, and visible scoring after completing levels of the game is a valuable one. It emphasizes the importance of social and competitive elements in enhancing player excitement and engagement. By implementing social features like group entry and competitive features such as visible scoring, the game could provide players with a sense of progress and motivate them to strive for higher scores. Such elements make players compare their performance with other players and keep them engaged. Incorporating these features could make the game more appealing and encourage players to continue playing and improving their skills.

## 6 Conclusion

In this paper, we present a new serious game, Space Geeks, that is created to help novice programmers achieve better basic programming concepts such as arrays. First, we investigated the details of misconceptions about arrays among students, and then, based on that, we create a serious game initially targeted at the arrays concept, that is designed following a five-dimensional evaluation framework which considers usability, motivation, understandability, engagement and user experience. Our findings imply that such games are suitable in environments in which novice programmers struggle with basic programming concepts and need support to achieve more. It is also worth noting that combining both game usability characteristics and user-related characteristics (such as engagement and motivation) results in great satisfaction among users and impressive learning achievements.

## References

1. Abdellatif, A.J., McCollum, B., McMullan, P.: Serious games: quality characteristics evaluation framework and case study, In: 2018 IEEE Integrated STEM Education Conference (ISEC), Princeton, NJ, USA, pp. 112-119 (2018)
2. Adamo-Villani, N., Haley-Hermiz, T., Cutler, R.: Using a serious game approach to teach 'operator precedence' to introductory programming students. In: Proceedings of the 17th International Conference on Information Visualisation, pp. 523-526 (2013)
3. Altadmri, A., Brown, C.N.: 37 Million compilations: investigating novice programming mistakes in largescale student data. In: Proceedings of the 46th ACM Technical Symposium on Computer Science Education (SIGCSE'15). ACM, New York, pp. 522-527 (2015)
4. Assaf, A.J., Eshtay, M., Issa, L.: Space geeks: a proposed serious game to teach array concept for novice programming students. *WEBIST* **2022**, 431-438 (2022)
5. Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P., Berta, R.: Assessment in and of serious games: an overview. *Adv. Hum. Comput. Interact.* **2013**, 136864 (2013)

6. Du Boulay, B.: Some difficulties of learning to program. *J. Educ. Comput. Res.* **2**(1), 57–73 (1986)
7. Boyle, E.A., Connolly, T.M., Hainey, T., Boyle, J.M.: Engagement in digital entertainment games: a systematic review. *Comput. Hum. Behav.* **28**(3), 771–780 (2012)
8. Calderon, A., Ruiz, M.: A systematic literature review on serious games evaluation: an application to software project management. *Comput. Educ.* **87**, 396–422 (2015)
9. Cederholm, H., Hilborn, O., Lindley, C., et al.: The aiming game: using a game with biofeedback for training in emotion regulation. In: *Proceedings of the 5th International Conference on Digital Research Association: Think Design Play* (2011)
10. Cheah, C.S.: Factors Contributing to the Difficulties in Teaching and Learning of Computer Programming: a literature review. *Contemp. Educ. Technol.* **12**(2), ep272 (2020). <https://doi.org/10.30935/cedtech/8247>
11. Connolly, T.M., Boyle, E.A., MacArthur, E., Hainey, T., Boyle, J.M.: A systematic literature review of empirical evidence on computer games and serious games. *Comput. Educ.* **59**(2), 661–686 (2012)
12. Danielsiek, H., Paul, W., Vahrenhold, J.: Detecting and understanding students' misconceptions related to algorithms and data structures. In: *Proceedings of the 43rd ACM technical symposium on Computer Science Education*, pp. 21–26 (2012)
13. Deterding, S., Dixon, D., Khaled, R., Nacke, L.: From game design elements to gamefulness: defining gamification. In: *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, pp. 9–15 (2011)
14. Hainey, T., Connolly, T.M., Boyle, E.A., Wilson, A., Razak, A.: A systematic literature review of games-based learning empirical evidence in primary education. *Comput. Educ.* **102**, 202–223 (2016)
15. Hunicke, R., LeBlanc, M., Zubek, R.: MDA: a formal approach to game design and game research. In: *Proceedings of the AAAI Workshop on Challenges in Game AI*, pp. 1–5 (2004)
16. Karavidas, L., Hippokratis, A., Thrasyvoulos, T.: Usability evaluation of an adaptive serious game prototype based on affective feedback. *Information* **13**(9), 425 (2022)
17. Kelleher, C., Pausch, R., Kiesler, S.: Storytelling alice motivates middle school girls to learn computer programming. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1455–1464 (2007)
18. Kiili, K.: Digital game-based learning: towards an experiential gaming model. *Internet High. Educ.* **8**(1), 13–24 (2005)
19. Kiili, K., Ketamo, H.: The effectiveness of exergames: comparing exergame and traditional exercise preferences in young adults. *J. Phys. Act.* **13**(2), 100–106 (2016)
20. Li, X., Atkins, M.S., Stanton, N.A.: Applying the lessons learned from the study of game immersion to virtual environments for learning. *Comput. Educ.* **57**(2), 1685–1693 (2011)
21. Malone, T.W.: Toward a theory of intrinsically motivating instruction. *Cogn. Sci.* **5**(4), 333–369 (1981)
22. Mortara, M., Catalano, C.E., Fiucci, G., Derntl, M.: Evaluating the Effectiveness of Serious Games for Cultural Awareness: The Icura User Study. In: De Gloria, A. (ed.) *GALA 2013. LNCS*, vol. 8605, pp. 276–289. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-12157-4\\_22](https://doi.org/10.1007/978-3-319-12157-4_22)
23. Nacke, L. E., Lindley, C. A.: Flow and immersion in gameful design. In: *Proceedings of the 2009 Annual International Conference on the Foundations of Digital Games*, pp. 32–39 (2009)
24. Nacke, L.E., Bateman, C., Mandryk, R.L.: BrainHex: a neurobiological gamer typology survey. *Entertainment Comput.* **5**(1), 55–62 (2014)
25. Pivec, M.: Usability in serious games for learning: a review. *J. Univ. Comput. Sci.* **20**(1), 6–31 (2014)

26. Plass, J.L., Homer, B.D., Kinzer, C.K.: Foundations of game-based learning. *Educ. Psychol.* **50**(4), 258–283 (2015)
27. Rieber, L.P.: Seriously considering play: designing interactive learning environments based on the blending of microworlds, simulations, and games. *Educ. Technol. Res. Dev.* **44**(2), 43–58 (1996)
28. Rodríguez-Cerezo, D., Sarasa-Cabezuelo, A., Gomez-Albarran, M., Sierra, J.: Serious games in tertiary education: a case study concerning the comprehension of basic concepts in computer language implementation courses. *Comput. Hum. Behav.* **31**, 558–570 (2014)
29. Squire, K.: *Video games and learning: Teaching and participatory culture in the digital age*. Teachers College Press (2011)
30. Sweetser, P., Wyeth, P.: GameFlow: a model for evaluating player enjoyment in games. *ACM Comput. Entertainment* **3**(3), 1–24 (2005)
31. Vrachnos, E., Jimoyiannis, A.: Secondary education students' difficulties in algorithmic problems with arrays: an analysis using the SOLO taxonomy. *Themes Sci. Technol. Educ.* **10**(1), 31–52 (2017)
32. Wangenheim, G.C., Savi, R., Borgatto, F.A.: DELIVER! - an educational game for teaching Earned Value Management in computing courses. *Inf. Softw. Technol.* **54**, 286–298 (2012)
33. Wouters, P., Van Nimwegen, C., Van Oostendorp, H., Van Der Spek, E.D.: A meta-analysis of the cognitive and motivational effects of serious games. *J. Educ. Psychol.* **105**(2), 249–265 (2013)
34. Qian, Y., Lehman, J.: Students' misconceptions and other difficulties in introductory programming: a literature review. *ACM Trans. Comput. Educ.* **18**(1), Article 1 (2017)
35. Zhang, J., Caldwell, R.E., Smith, E.: Learning the concept of Java inheritance in a game. In: *The 18th International Conference on Computer Games (CGAMES)*, pp. 212–216 (2013)

# Author Index

## A

Alencar, Paulo 60  
Assaf, Abdelbaset 248

## B

Buchem, Ilona 24

## C

Caccamo, Alessio 229  
Cajueiro, Daniel Oliveira 98  
Celestino, Victor Rafael Rezende 1, 98  
Cowan, Don 60

## D

de Melo, Maísa Kely 1, 98  
de Oliveira, Flávio Augusto R. 1  
Deutschländer, Jana 209  
Dib, Marcos Vinícius Pinheiro 98

## E

Eshtay, Mohammed 248

## F

Faria, Allan Victor Almeida 1, 98  
Figueira, Alvaro 166  
Fosci, Paolo 142

## I

Issa, Lana 248

## K

Kaili, Michalis 42  
Kanavos, Andreas 84  
Kapitsaki, Georgia M. 42  
Klein, Andreas M. 209

Kollmorgen, Jessica 186  
Kölln, Kristina 209

## M

Melo, Glaucia 60  
Mohasseb, Alaa 84  
Mori, Jun'ichiro 121

## N

Nascimento, Lirielly 166

## O

Ochi, Masanao 121  
Oliveira, Toacy 60

## P

Pillat, Raquel 60  
Psaila, Giuseppe 142

## R

Rauschenberger, Maria 24, 209  
Rocha, Carlos Alberto Alvares 98

## S

Sakata, Ichiro 121  
Schön, Eva-Maria 24  
Schrepp, Martin 186  
Shiro, Masanori 121  
Sostak, Stefano 24

## T

Telemaco Neto, Ulisses 60  
Thomaschewski, Jörg 186

## W

Weigang, Li 1, 98  
Winter, Dominique 209