Scientometrics (2022) 127:7061–7075
https://doi.org/10.1007/s11192-022-04380-z

Constructing a high-quality dataset for automated creation of summaries of fundamental contributions of research articles

Haihua Chen[1] · Huyen Nguyen[1] · Asmaa Alghamdi[2]

Received: 31 October 2021 / Accepted: 9 April 2022 / Published online: 28 April 2022

Abstract

Research contributions, which indicate how a research paper contributes new knowledge or new understanding in contrast to prior research on the topic, are the most valuable type of information for researchers to understand the main content of a paper. However, there is little research using research contributions to identify and recommend valuable knowledge in academic literature for users. Instead, most existing studies mainly focus on the analysis of other elements in academic literature, such as keywords, citations, rhetorical structure, discourse, and others. This paper first introduces a fine-grained annotation scheme with six categories for research contributions in academic literature. To evaluate the reliability of our annotation scheme, we conduct annotation on 5024 sentences collected from Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL Anthology) and an academic journal Information Processing & Management (IP &M). We reach an inter-annotator agreement of Cohen's kappa = 0.91 and Fleiss' kappa = 0.91, demonstrating the high quality of the dataset. We then built two types of classifiers for automated research contribution identification based on the dataset: classic feature-based machine learning (ML) and transformer-based deep learning (DL). Our experimental results show that SCI-BERT, a pretrained language model for scientific text, achieves the best performance with an F1 score of 0.58, improving the best classic ML model (nouns + verbs + tf-idf + random forest) by 2%. This also indicates a comparable power of classic feature-based ML models to DL-based model like SCI-BERT on this dataset. The fine-grained annotation scheme can be applied for large-scale analysis for research contributions in academic literature. The automated research contribution classifiers built in this paper provide the basis for the automatic research contributions extraction and knowledge fragment recommendation. The high-quality research contribution dataset developed in this research is publicly available on Zenodo https://zenodo.org/record/6284137#.YhkZ7-iZO4Q. The code for the data analysis and experiments will be released at: https://github.com/HuyenNguyenHelen/Contribution-Sentence-Classification.

Keywords Academic literature · Research contribution · Dataset · Text classification · Machine learning · BERT

✉ Haihua Chen
haihua.chen@unt.edu

Extended author information available on the last page of the article

Introduction

Academic literature records the research process with a standardized structure and provides clues to track progress in a scientific field (Lindsay 1995; Hofmann 2016). Generally, the main components of academic literature include an abstract, introduction, related work, method, experiment and result, and conclusion (Day 1989; Peat et al. 2002; Sollaci and Pereira 2004). In recent years, academic text mining using content from different components has received increasing attention from researchers. However, most of the existing research focuses on key phrase extraction (Park and Caragea 2020), citation content analysis (Fisas et al. 2016), rhetorical structure analysis (Sateli and Witte 2015), and essential sentence extraction (Mehta et al. 2018), with little attention paid to the research contributions stated in the full text Swales (1990). Research contributions, indicating how a research paper contributes new knowledge or new understanding in contrast to prior research on the topic, are the most valuable type of information for researchers to understand the main content of a paper. A research contribution relates to the research problem addressed by the contribution, the research method, and (at least one) research result (Oelen et al. 2019). For example, "we build a transfer learning framework employing a diverse range of intermediate tasks covering sequence tagging with semantic and syntactic aspects, and natural language inference" and "we achieve competitive performance over both strong baselines and previous works" are two contribution statements (Park and Caragea 2020). Research contributions can help researchers understand the core content of a paper and the growth of innovation in science.

We can easily identify sentences about research contributions from the introduction section by locating such statements as "Our contributions are summarized as follows," "The major contributions of this paper are," or similar phrasings. There are different types of contributions; for example, creating datasets, building new models, performing evaluations, etc. If these contributions can be classified appropriately and automatically, it would be helpful for knowledge recommendation, structured abstract generation, and scientific evolution analysis. However, an annotation scheme, the codebook which defines the annotation categories and the annotation guidelines (Hovy and Lavid 2010), for research contributions is one of the essential requirements for research contribution classification.

We studied the existing annotation schemes for academic literature. We noted that most of them mainly focus on context types (indicate the various roles of a citation context plays in different components of an article) (Angrosh et al. 2012), citation functions (indicate what could the author's intention have been in choosing a certain citation) (Teufel et al. 2006), and future work types (indicate the different categories of future work sentences, such as method, resources, evaluation, application, problem, and others) (Hao et al. 2020). There is no annotation scheme for research contributions. To bridge this gap, we first propose a fine-grained annotation scheme with six categories for research contributions in academic literature. A human annotation experiment (where humans are asked to identify and annotate the data) conducted on 5,024 sentences collected from Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL Anthology) and an academic journal Information Processing & Management (IP &M) demonstrates the reliability of our scheme. Based on the high quality dataset constructed, we built automated research contribution classifiers using classic machine learning (ML) models and transformer-based deep learning (DL) models. The contributions of our paper are as follows:

1. We propose a fine-grained annotation scheme for research contributions in academic literature. The annotation scheme includes six types of research contributions: dataset/resources creation, theory proposal, model construction or optimization, algorithms/methods construction or optimization, performance evaluation, and applications.
2. We conduct a human annotation experiment to evaluate the reliability of our scheme. We reach an inter-annotator agreement of Cohen's kappa = 0.91 and Fleiss' kappa = 0.91.
3. We develop a high-quality research contribution dataset, including 5024 annotated sentences on the six categories. The dataset is publicly available.
4. We apply classic ML and transformer-based DL models for automated research contribution extraction. The experimental results show that the SCI-BERT model achieves the best performance among all the models, with an F1 score of 0.58.

Related work

This paper proposes an annotation scheme for research contributions in academic literature, then builds many classifiers for automated research contribution identification. Therefore, we review the literature from the following sub-topics: annotation schemes for academic literature, research contributions analysis, and ML models for text classification.

Annotation schemes for academic literature

Several annotation schemes have been proposed for academic literature. The annotation schemes are either designed for full texts or citation sentences only. Regarding the full text level, Swales (1990, 2011) produced one of the earliest models (i.e., CARS) for analyzing research papers. CARS model consists of three "moves" (components) with "steps" (sub-components) that most research paper introduction covers. Hao et al. (2020) proposed an annotation scheme with six main categories and 17 sub-categories for future work sentences. D'Souza and Auer (2020) described ten core information units for organizing academic contributions in a knowledge graph (KG): ResearchProblem, Approach, Objective, ExperimentalSetup, Results, Tasks, Experiments, AblationAnalysis, Baselines, and Code. For the citation sentence level, Teufel et al. (2006) proposed an annotation scheme with four categories and 12 fine-grained categories for citation function. The annotation experiment was performed on 320 conference articles and kappa agreement was used to measure the reliability. Alternatively, Angrosh et al. (2012) presented a citation-centric annotation scheme for academic literature. It included six categories for citation sentences (i.e., the sentences include the citation marks) and five categories for non-citation sentences (i.e., sentences surrounding the citation sentences and providing further descriptions of the citation sentences.). A pilot study was carried out using 11 annotators and nine articles. Agreement calculated with Krippendorff's alpha was used to measure the reliability. The above research provides us insights into how to construct a fine-grained annotation scheme for research contribution sentences and how to evaluate the reliability of our scheme.

Research contributions analysis and identification

Research contributions analysis is a new topic, which has recently garnered attention. Auer et al. (2018) constructed the open research knowledge graph (ORKG) where each paper was summarized with its fundamental contribution properties and values. In the ORKG, the contributions were interconnected via the graph, even across papers. It helps users to compare research contributions between different papers while writing an academic literature review (Oelen et al. 2019). Similarly, Vogt et al. (2020) represented research contributions in scholarly knowledge graphs using knowledge graph cells. Compared to the ORKG (Auer et al. 2018), the Research Contribution Model (RCM) can generate a KG whose content is more easily maintained and easier to understand (Vogt et al. 2020). The ontology built by Vogt et al. (2020) provided a reference for defining the annotation categories in our study. However, it identified contributions from abstracts only rather than full texts. D'Souza and Auer (2020) developed an annotation scheme to identify research contributions from natural language processing (NLP) literature with the structure of $< subject, predicate, object >$. In 2021, a scientific information extraction task called NLPContributionGraph was organized on SemEval-2021. The task aimed to build a comprehensive knowledge graph that publishes the research contributions of scholarly publications per paper, and even across papers, where the contributions are connected via the graph (Jaradeh et al. 2019). More than ten teams attended and contributed to this task.

Instead of focusing on research contribution identification, some researchers targeted a similar task: research highlight extraction. Wang et al. (2018) compared the differences between extracting highlights and abstracts from journal articles. However, they relied on classic features such as word frequency, term frequency-inverse document frequency (tf-idf), sentence length, etc. Rehman et al. (2021) conducted a preliminary experimental study using DL models to generate research highlights from scientific abstracts, but the performance still has much room for improvement.

Research contributions have also been applied for evaluating the value and impact of academic literature. Le et al. (2019) applied research contributions identified from citing papers for evaluating the academic value of cited papers. In addition, research contributions also have the potential in assessing the innovation level of an academic literature article. Kok and Schuit (2012) designed a novel approach to map contributions in research articles in the health field to assist stakeholders to better utilize the research. Morton (2015) proposed an empirically grounded framework for assessing the impact of research based on research contributions. If the research contributions can be automatically and accurately extracted from scientific literature, both of the above applications can be easily extended to other fields.

Machine learning and deep learning for text classification

Automated research contribution identification is a text classification task. Models that are fit for short text classification can also be used in this research. Kowsari et al. (2019) conducted a comprehensive review of ML algorithms for text classification from text feature extractions, dimensionality reduction methods, existing algorithms and techniques, and evaluation methods. Li et al. (2020) compared the multiple ML and DL models for text classification. Among which, the transformer-based methods (i.e., ELMo, GPT, BERT), which apply unsupervised methods to mine semantic knowledge automatically and then construct pre-training targets to support semantic understanding, have been widely used and proven effective. Quantitative evaluation showed that BERT-based models get better results on most datasets (Li et al. 2020). Therefore, we opted to try BERT-based models firstly when implementing a text classification task, as suggested by Li et al. (2020). Recently, SCI-BERT, a pre-trained language model for scientific text, was developed to improve performance on downstream scientific NLP tasks (Beltagy et al. 2019).

High-quality research contribution dataset annotation

Data acquisition and preparation

The initial data in this research is from the ACL Anthology (nd 2022) and IP &M (nd 2022b). We select the two sources because the research contributions are clearly claimed and can be easily extracted. We manually identify the sentences which indicate the research contributions. Specifically, we conduct a pre-investigation of the original corpus to summarize the patterns of research contribution sentences, then formulate the labeling specifications. For IP &M, we directly take the research highlights as the contributions. For ACL articles, we use two strategies to identify the contribution sentences: (1) For explicit research contribution sentences, we are able to easily locate them by identifying the contribution block indicators; for example, "Our contributions are summarized as follows," "The major contributions of this paper are," or similar statements. (2) As suggested by Swales (1990), we located implicit contribution sentences as findings in the last paragraph or the second to the last paragraph in the introduction section. We then tease out several verbs or verb phrases that indicate the research contributions, such as "present", "introduce", "compare", "design", "apply", "develop", etc. By following the above strategy, we are able to finally collect 3374 research contribution sentences from ACL and 1650 from IP &M in total.

An annotation scheme for research contributions

We create an annotation scheme with six types of research contributions, which is adapted from the annotation scheme in our International Conference on Scientometrics and Informetrics (ISSI) 2021 paper (Chen and Kanuboddu 2021). The initial annotation scheme included nine categories: dataset creation, theory proposal, model construction, model optimization, new algorithm/ method/ technology, algorithm/ method/ technology/ optimization, performance evaluation, resources, and applications. Our pre-experimental study shows that ML models have difficulty in distinguishing model construction and model optimization, algorithm construction and algorithm optimization, and dataset creation and resources, even though they indeed belong to these different categories. Therefore, we merge these categories as model construction or optimization, algorithm/ method/ construction or optimization, and dataset/ resources creation, respectively. We separate the method and model since they are different concepts, especially in computational linguistics, according to QasemiZadeh and Handschuh (2014). A more detailed explanation of each category with definitions and examples can be found in Table 1.

Human annotation experiment

Annotation procedure

To evaluate the reliability of the research contribution annotation scheme discussed before and create a dataset for automatic research contribution classification, we conduct the

Table 1 Our annotation scheme for research contributions

Category	Description	Example
Dataset/ resources creation	Create or expand datasets and resources	We propose to build multi-task datasets for the News and tweets domains, by unifying the afore-mentioned task-independent datasets.
Theory proposal	Propose a new theory to solve existing problems or for improvement	We suggest viewing learning event embedding as a multi-relational problem, which allows us to capture different aspects of event pairs.
Model construction or optimization	Construct a model or propose strategies to optimize the existing model	We propose generating comments with a graph-to-sequence model that models the input news as a topic interaction graph.
Algorithms/ methods construction or optimization	Develop a new algorithm/ method or optimize an existing one	We present a new method for sentiment lexicon induction that is designed to be applicable to the entire range of typological diversity of the world's languages.
Performance evaluation	Evaluate the performance of a new or existing implementation	We evaluate a broad variety of neural models on the new dataset, establishing strong baselines that surpass previous feature-based models in three tasks.
Applications	Apply the proposed model or theory to other tasks	Our model can also be applied to the cross-domain named entity recognition task, and it achieves better adaptation performance than other existing baselines.

annotation experiment with six annotators who are designers of the scheme and very familiar with the annotation guideline. They also have a background in NLP and ML, which can ensure annotation quality.

The six annotators are divided into two groups for the annotation; in other words, each sentence will be annotated by three annotators. During the annotation, the annotators independently annotated the same number of sentences with the proposed scheme. A majority vote is used to decide the final label for a sentence. If the label of a sentence cannot be confirmed based on the three annotators, another annotator will label the sentence.

Annotation results

We obtain 5,024 annotated sentences in total. We combine Cohen's kappa (Carletta 1996) and Fleiss' kappa (Falotico and Quatto 2015) to measure the agreement. Cohen's kappa is a statistic used to measure inter-rater agreements between two annotators (Carletta 1996). The value of kappa ranges between − 1 and 1. Generally, a kappa of 0.8 is considered stable. Fleiss' kappa is a statistical extension method of Cohen used for determining agreement among more than two annotators (Falotico and Quatto 2015). Overall, we reached an inter-annotator agreement of Cohen's kappa = 0.91 (average of three pairs) and Fleiss' kappa = 0.91. The agreement is quite good, considering the number of categories. To evaluate the annotation quality for each category, we also calculate the Fleiss' kappa of the three annotators for each category. The results are shown in Table 2, further demonstrating the high quality of the annotated research contribution dataset.

Statistical analysis of the annotated dataset

The relative frequency of each category observed in the annotation results is shown in Fig. 1. As can be seen from the figure, the top three categories are theory proposal, model construction or optimization, and algorithms/ methods construction or optimization, with 1340, 1246, and 1041 contribution sentences, respectively, which comes in at 72.5%. The distribution is aligned with the scopes of ACL and IP &M. Since ACL is the top conference while IP &M is the top journal in computer and information science, they both request stronger contributions, especially technical ones, before they can be accepted.

Key terms analysis in each category

We extract the top 20 key terms based on term frequency (as listed in Fig. 2) for the contribution sentences in each category and the key terms are limited to uni-gram.

Table 2 Our annotation scheme for research contributions

Category	Fleiss' kappa
Dataset/ resources creation	0.90
Theory proposal	0.89
Model construction or optimization	0.89
Algorithms/ methods construction or optimization	0.90
Performance evaluation	0.92
Applications	0.77

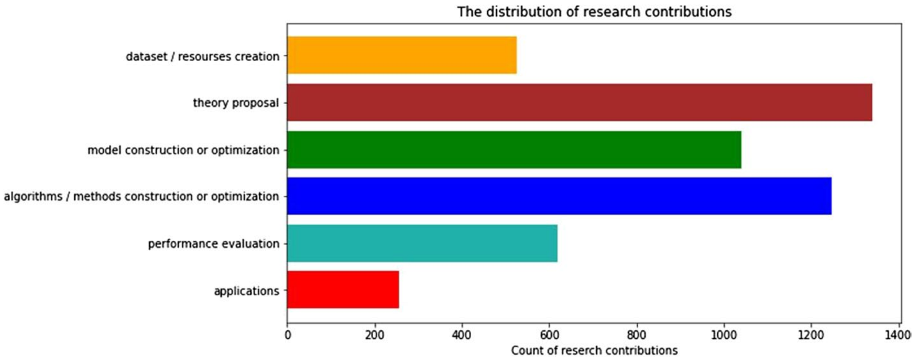

Fig. 1 The distribution of research contributions in each category

As shown in Fig. 2, most of the key terms are nouns and verbs. The noun terms reflect the most important contributions and the verb terms indicate how the research contributions are presented in each category. For example, in the category dataset or resources creation, it is obvious that the key terms “data”, “dataset”, “corpus”, “xxx dataset”, and “annotate” are at the top of the list. Verbs such as “present” and “introduce” are frequently used to introduce a new dataset. Similarly, in the model construction and optimization categories, key terms such as “model”, “neural network model”, and “language model” are

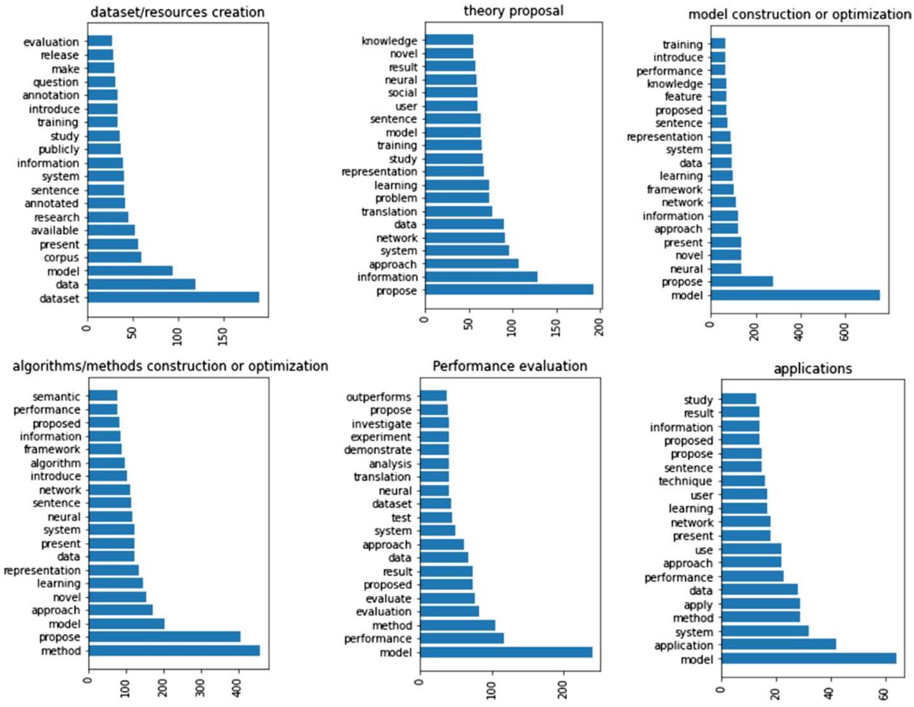

Fig. 2 The top 20 key terms in each category

prevalent. In the performance evaluation category, "performance", "evaluation", "evaluate", "test", and "outperform" are among the top key terms. The analysis of the key terms can provide effective features for automatic research contribution classification in the future.

Automated research contribution classification

Text representation

In this research, we use two methods for text representation: manual features and pre-trained word embeddings. The feature-based method is not only labor-intensive but also sometimes less effective due to highly sparse vectors. In contrast, word embedding-based methods learn feature representation from a large corpus and generate shorter dense vectors that better capture contextual information. Some highly-performed pre-trained models are Word2Vec, GloVe, and BERT. However, these pre-trained models suffer from domain-specific issues, requiring fine-tuning on the domain datasets. Therefore, this study investigates different word presentation and feature extraction methods performed in classic ML and DL models.

For manual features, we incorporate the most frequent nouns (1298 features) and verbs (1344 features) along with tf-idf (1000 features). In addition, Word2Vec (Mikolov et al. 2013), a pre-trained word embedding released by Google, is also applied to encode research contribution sentences to train other classic classifiers. Word2Vec presents each instance with 300 dimensional dense vectors, requiring our classifiers to learn fewer weights than the manual feature-based representation; therefore, it possibly helps with generalization and avoiding overfitting. Moreover, SCI-BERT (Beltagy et al. 2019), a pre-trained language model trained on 1.14M scientific papers from Semantic Scholar, is used to encode text in the DL model. As mentioned, pre-trained embeddings are less effective in some domain-specific datasets, so using SCI-BERT possibly can overcome this limitation.

Classification algorithms

We implement multiple ML and DL classification models. Even though DL has proven its outperformance in most NLP tasks, it requires more training data and computational resources. If effective features can be extracted and selected, ML models can also achieve good performance. Therefore, we compare several manual feature-based ML algorithms and the transformer-based DL model. The classification algorithms used in our study are summarized as follows:

- *Logistic Regression (LR)* is a probabilistic classifier based on a logistic function to learn conditional probability. It assumes the independent relations among features. We use L2 regulation, and lbfgs with maximum iterations of 700 to optimize the model and avoid overfitting.
- *Random Forest (RF)* is an ensemble model that fits several decision tree classifiers on a variety of subsets and eventually takes the average of the performance of classifiers for a more reliable predictive score and avoid overfitting. Parameters used in this model are set by default.

- *K-nearest neighbors (KNN)* is a non-generalizing learning model. The model stores training data points. Classification is based on the majority votes of k nearest neighbors' labels toward the predicting data point. We selected k = 5 as the number of k neighbors.
- *Decision Trees (DT)* makes predictions by learning the decision rules inferred from the training data. The model is highly interpretable but less generalized and unstable since it prioritizes locally optimal decisions.
- *Naive Bayes (NB)* is a probability model based on the Bayes theorem, assuming the conditional independence between features. NB performs very well compared to other complex models in many cases, especially when the training data is small.
- *Support Vector Machines (SVM)* attempts to map training instances into the high-dimension space to maximize the distance between categories (also called margins). Therefore, the model can perform well in high-dimension datasets, even if the number of dimensions is greater than the number of instances. We set 'ovo' (one-versus-one) as a decision function of the model.
- *BERT* is a Bidirectional Encoder Representations from Transformers. We implement two BERT-based models for the purpose of comparison: BERT, and SCI-BERT. We encode sentences with BERT, and SCI-BERT embeddings trained on the BERT architecture and further fine-tune it in our dataset. The model is trained on eight epochs, batch sizes of 32, and an Adam optimizer with a learning rate of 2e-5.

Notice from the data exploratory analysis 1, the "application" class is strongly imbalanced in comparison to other classes. Class imbalance significantly declines the model performance (Weng et al. 2020). Therefore, we implement the oversampling method with the SMOTE algorithm (Chawla et al. 2002) for augmenting enough data of this class to train the model. All classic ML models are trained and validated with the ten-fold cross-validation.

Evaluation metrics

We use recall, precision, and F1 score as metrics to evaluate the performance on each category since they are the most frequently used evaluation metrics for text classification (Li et al. 2020). For the overall performance, we use weighted-average precision, recall, and F1 score. Each class's contribution to the average is weighted by its size.

Results

The overall results are presented in Table 3. Overall, BERT-based models show the best performance compared to all the others, demonstrating the effectiveness of BERT in research contribution identification. In addition, SCI-BERT performs better than the general BERT model, 0.58 in comparison to 0.56 on the F1 score. The results also indicate that the RF model in which tf-idf, most frequent nouns, and verbs were used as features is comparable with the general BERT and performs better than Word2Vec-based ML models. It indicates that the research contribution identification requires a more domain-specific text representation, and the manually-engineered features can capture adequate information for it.

Figure 3 compares accuracy scores of classic ML models with the Word2Vec embedding (left) and handcrafted feature (right) representations over the ten-fold cross-validation.

Table 3 The overall results of research contribution classification

Model	Accuracy	Recall	Precision	F1 score
Manual features + LR	0.49	0.49	0.49	0.49
Manual features + RF	0.56	0.56	0.56	0.56
Manual features + KNN	0.36	0.36	0.39	0.34
Manual features + DT	0.45	0.45	0.45	0.44
Manual features + NB	0.28	0.28	0.33	0.29
Manual features + SVM	0.52	0.52	0.53	0.51
Word2Vec + LR	0.49	0.52	0.49	0.50
Word2Vec + RF	0.49	0.49	0.48	0.48
Word2Vec + KNN	0.28	0.28	0.43	0.28
Word2Vec + DT	0.30	0.30	0.32	0.31
Word2Vec + NB	0.46	0.46	0.46	0.46
Word2Vec + SVM	0.52	0.52	0.52	0.52
BERT	0.57	0.57	0.58	0.56
SCI-BERT	**0.59**	**0.59**	**0.59**	**0.58**

The bold value indicates the best performanceon each measurement

With the Word2Vec embedding, SVM achieves the highest performance, followed by LR. However, given the handcrafted feature extraction, RF outperforms other classic ML models, followed by SVM. This strengthens our assumption about the importance of feature engineering on classic ML models' performance. It is also coincident with the conclusion in Fernández-Delgado et al. (2014) about the highest performance of RF and SVM among 179 classifiers on 121 datasets.

The confusion matrix in Fig. 4 describes the performance of SCI-BERT, our best model, on the test set. It gives us a better idea about how the model performs in the six categories. It indicates a correlation between the data size and the number of true positives. The model performs better in the classes with more data but the worst in the "applications" class, 0.36 on the F1 score. Its data accounts for 4.2% of the dataset, which is insufficient for the model to learn patterns in this class. Even though the SMOTE oversampling is applied to solve the class imbalance in classic ML models, the performance of most classic ML models is even worse in this class, under 0.19 on the F1 score, except for RF with handcrafted

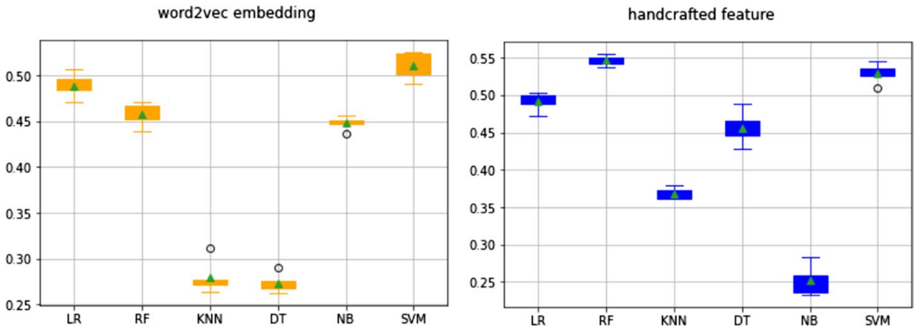

Fig. 3 Cross-validation performance (accuracy) of classic ML models with Word2Vec embeddings (left) and manual features (right)

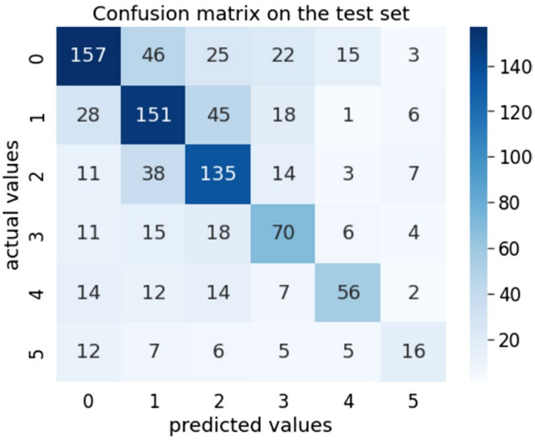

Fig. 4 Confusion matrix of SCI-BERT. From 0 to 5, six categories are theory proposal, algorithms/ methods construction or optimization, model construction or optimization, performance evaluation, dataset/ resources creation, and applications, respectively

features. Meanwhile, we do not use any methods to handle the class-imbalance issues in DL models, indicating that BERT-based models can overcome this issue by themselves.

Discussion

According to Li et al. (2020), word embedding-based models such as BERT can get better performance on most text classification datasets, which means that we can always implement DL models first to get SOTA results. However, this conclusion does not fit the domain-specific text classification tasks, as demonstrated by Chen at al. (Chen et al. 2022) in a legal text classification task. This research further confirms the conclusion, as can be seen from the results of Word2Vec and manual features with machine learning models such as RF and SVM.

Many factors can affect the model selection of domain-specific text classification classification, such as data, performance, computation, and interpretation Chen et al. (2022). In this research, we aim to build a strong baseline for research contribution classification. Therefore, we pay more attention to the performance aspect. For the word embedding-based classification models, in addition to the data quality, the quality of the word embedding will also affect the model performance (Chen et al. 2021). The research contribution datasets used for the classification is of high-quality according to the annotation results in section 3.3. As for the quality of the word embedding, the quality of the pre-training data for BERT-based models is higher since their training data cover much more scientific concepts than the training dataset used for training the Word2Vec embedding. Therefore, the BERT-based models achieve better results than Word2Vec-based models (Shen and Liu 2021).

SCI-BERT, trained on large-scale academic publications, achieves the best performance on scientific NLP tasks, indicating the effectiveness of fine-tuning the general language models with domain unlabeled texts on domain text classification (Beltagy et al. 2019; Chakravarthi 2021). Although manual feature-based machine learning models do not perform as well as SCI-BERT, it generates a similar performance to the BERT model, indicating the effectiveness of our manually-selected features. Notice that the feature-based models are more efficient and take less computational resources than BERT Chen et al. (2022). Moreover, the results from the feature-based models can more easily be interpreted.

Even for the extraction of contribution sentences or non-contribution sentences task, the performance is quite low (< 50%) (Wang et al. 2018). Classifying research contributions into different types is a more challenging task. We believe our research has built a strong baseline for further research. The high-quality dataset and the baseline constructed in this study are intended to be the foundation of research contribution classification and automated creation of summaries of fundamental contributions.

Conclusion and future work

In this paper, we propose a fine-grained annotation scheme with six categories of research contributions. Based on the proposed annotation scheme, we create a high-quality dataset for research contribution identification with 5024 contribution sentences taken from the ACL Anthology and IP &M. The dataset quality and bias elimination are validated with very high kappa values, 0.91 on both Cohen's kappa and Fleiss' kappa. Furthermore, we provide several benchmarking models on the created dataset: classic ML, and DL, using handcrafted features and contextual word embeddings. Our experiments prove an outperformance of the SCI-BERT model, followed by random forest with the manual feature extraction method. In the future, we plan to expand the dataset, especially in some heavily imbalanced classes such as "applications". We will also increase the comprehensiveness of the dataset by including contribution sentences from some other journals. To improve the model performance, we will explore both transfer learning by fine-tuning with more related data and generating more effective features for research contribution representation.

Acknowledgements The paper is a substantially extended version of the ISSI2021 conference paper "A Fine-Grained Annotation Scheme for Research Contribution in Academic Literature". The authors would like to thank Roohia Shahzad, Rubab Shahzad, Aakansha Tallapally, Durga Bhavana Yerrabelli, Nikitha Malladi, and Riyaz Ahmad Shaik at the University of North Texas for participating in the annotation experiment. The authors would like to thank Bhavya Nandana Kanuboddu for contributing to the data analysis and the visualizations of the ISSI conference paper. The authors would like to thank Marie Bloechle at the University of North Texas for editing the language and writing of the paper. The authors are grateful to all the anonymous reviewers for their precious comments and suggestions.

Author contributions HC: Research design, project management, investigation, methodology, writing. HN: Methodology, experiments, data analysis, writing. AA: Data curation, data analysis, review, and editing.

References

Angrosh, M., Cranefield, S., & Stanger, N. (2012). A citation centric annotation scheme for scientific articles. *Proceedings of the Australasian Language Technology Association Workshop, 2012*, 5–14.

Auer, S., Kovtun, V., Prinz, M., Kasprzik, A., Stocker, M., & Vidal, M. E. (2018). Towards a knowledge graph for science. In Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, (pp 1–6).

Beltagy, I., Lo, K., & Cohan, A. (2019). Scibert: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics, 22*(2), 249–254.

Chakravarthi, B.R. (2021). Domain identification of scientific articles using transfer learning and ensembles. In Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2021 Workshops, WSPA, MLMEIN, SDPRA, DARAI, and AI4EPT, Delhi, India, May 11, 2021 Proceedings, (vol 12705, p. 88). Springer Nature.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority oversampling technique. *Journal of artificial intelligence research, 16*, 321–357.

Chen, H., & Kanuboddu, B. N. (2021). A fine-grained annotation scheme for research contribution in academic literature. In Proceedings of the 18th International Conference on Scientometrics and Informetrics, (pp 241–248).

Chen, H., Chen, J., & Ding, J. (2021). Data evaluation and enhancement for quality improvement of machine learning. *IEEE Transactions on Reliability, 70*(2), 831–847.

Chen, H., Wu, L., Chen, J., Lu, W., & Ding, J. (2022). A comparative study of automated legal text classification using random forests and deep learning. *Information Processing & Management, 59*(2), 102798.

Day, R. A., et al. (1989). The origins of the scientific paper: the imrad format. *Journal of the American Medical Directors Association, 4*(2), 16–18.

D'Souza, J., & Auer, S. (2020). Nlpcontributions: An annotation scheme for machine reading of scholarly contributions in natural language processing literature. In EEKE@JCDL'20 - Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents.

Falotico, R., & Quatto, P. (2015). Fleiss' kappa statistic without paradoxes. *Quality & Quantity, 49*(2), 463–470.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research, 15*(1), 3133–3181.

Fisas B, Ronzano F, Saggion H (2016) A multi-layered annotated corpus of scientific papers. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), (pp. 3081–3088).

Hao, W., Li, Z., Qian, Y., Wang, Y., & Zhang, C. (2020). The acl fws-rc: A dataset for recognition and classification of sentence about future works. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in, 2020*, 261–269.

Hofmann, A. H. (2016). *Scientific writing and communication: papers, proposals, and presentations* (3rd ed.). Oxford, United Kingdom: Oxford University Press.

Hovy, E., & Lavid, J. (2010). Towards a 'science'of corpus annotation: a new methodological challenge for corpus linguistics. *International Journal of Translation, 22*(1), 13–36.

Jaradeh, M.Y., Oelen, A., Farfar, K.E., Prinz, M., D'Souza, J., Kismihók, G., Stocker, M., & Auer, S. (2019). Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. In Proceedings of the 10th International Conference on Knowledge Capture, ACM, (pp. 243-246), https://dl.acm.org/doi/10.1145/3360901.3364435.

Kok, M. O., & Schuit, A. J. (2012). Contribution mapping: A method for mapping the contribution of research to enhance its impact. *Health Research Policy and Systems, 10*(1), 1–16.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information, 10*(4), 150.

Le, X., Chu, J., Deng, S., Jiao, Q., Pei, J., Zhu, L., & Yao, J. (2019). Citeopinion: Evidence-based evaluation tool for academic contributions of research papers based on citing sentences. *Journal of Data and Information Science, 4*(4), 26–41.

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., & He, L. (2020). A survey on text classification: From shallow to deep learning. arXiv preprint arXiv:2008.00364.

Lindsay, D. (1995). Scientific Writing. Longman Cheshire.

Mehta, P., Arora, G., & Majumder, P. (2018). Attention based sentence extraction from scientific articles using pseudo-labeled data. CoRR arXiv:1802.04675

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Morton, S. (2015). Progressing research impact assessment: A 'contributions' approach. *Research Evaluation, 24*(4), 405–419.

nd (2022a) Annual meeting of the association for computational linguistics (acl). Retrieved February 18, 2022, from https://aclanthology.org/venues/acl/

nd (2022b) Information processing & management. Retrieved February 18, 2022, from https://www.journals.elsevier.com/information-processing-and-management

Oelen, A., Jaradeh, M. Y., Farfar, K. E., Stocker, M., & Auer, S. (2019). Comparing research contributions in a scholarly knowledge graph. In CEUR Workshop Proceedings 2526 (2019), (vol 2526, pp. 21–26). Aachen: RWTH Aachen.

Park, S., & Caragea, C. (2020). Scientific keyphrase identification and classification by pre-trained language models intermediate task transfer learning. In Proceedings of the 28th International Conference on Computational Linguistics, (pp. 5409–5419).

Peat, J., Elliott, E., Baur, L., & Keena, V. (2002). *Scientific writing: Easy when you know how* (1st ed.). London, United Kingdom: BMJ Books.

QasemiZadeh, B., & Handschuh, S. (2014). The acl rd-tec: a dataset for benchmarking terminology extraction and classification in computational linguistics. In Proceedings of the 4th International Workshop on Computational Terminology (Computerm), (pp. 52–63).

Rehman, T., Sanyal, D. K., Chattopadhyay, S., Bhowmick, P. K., & Das, P. P. (2021). Automatic generation of research highlights from scientific abstracts. In EEKE@JCDL'21 - Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents.

Sateli, B., & Witte, R. (2015). What's in this paper? combining rhetorical entities with linked open data for semantic literature querying. In Proceedings of the 24th International Conference on World Wide Web, (pp. 1023–1028).

Shen, Y., & Liu, J. (2021). Comparison of text sentiment analysis based on bert and word2vec. In 2021 IEEE 3rd International Conference on Frontiers Technology of Information and Computer (ICFTIC), IEEE, (pp. 144–147).

Sollaci, L. B., & Pereira, M. G. (2004). The introduction, methods, results, and discussion (imrad) structure: A fifty-year survey. *Journal of the Medical Library Association, 92*(3), 364.

Swales, J. (1990). Genre analysis: English in academic and research settings. Cambridge University Press.

Swales, J. M. (2011). *Aspects of article introductions, michigan* (classics). University of Michigan Press.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). An annotation scheme for citation function. In Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, (pp. 80–87).

Vogt, L., D'Souza, J., Stocker, M., & Auer, S. (2020). Toward representing research contributions in scholarly knowledge graphs using knowledge graph cells. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in, 2020*, 107–116.

Wang, W. M., See-To, E. W. K., Lin, H. T., & Li, Z. (2018). Comparison of automatic extraction of research highlights and abstracts of journal articles. In Proceedings of the 2nd International Conference on Computer Science and Application Engineering, (pp. 1–5).

Weng, W.H., Deaton, J., Natarajan, V., Elsayed, G. F., & Liu, Y. (2020). Addressing the real-world class imbalance problem in dermatology. In Machine Learning for Health, PMLR, (pp. 415–429).

Authors and Affiliations

Haihua Chen[1] · Huyen Nguyen[1] · Asmaa Alghamdi[2]

[1] Department of Information Science, University of North Texas, Denton, Texas 76203, USA

[2] Department of Computer Science and Engineering, University of North Texas, Denton, Texas 76203, USA