








Utilizing Out-Domain Datasets to Enhance Multi-task Citation Analysis

Dominique Mercier^{1,2}(✉) , Syed Tahseen Raza Rizvi^{1,2} , Vikas Rajashekar¹ , Sheraz Ahmed¹ , and Andreas Dengel^{1,2} 

¹ German Research Center for Artificial Intelligence (DFKI) GmbH,
Trippstadter Straße 122, 67663 Kaiserslautern, Germany

{dominique.mercier, syed.rizvi, vikas.rajashekar, sheraz.ahmed,
andreas.dengel}@dfki.de

² TU Kaiserslautern, Erwin-Schrödinger-Straße 52, 67663 Kaiserslautern, Germany

Abstract. Citations are generally analyzed using only quantitative measures while excluding qualitative aspects such as sentiment and intent. However, qualitative aspects provide deeper insights into the impact of a scientific research artifact and make it possible to focus on relevant literature free from bias associated with quantitative aspects. Therefore, it is possible to rank and categorize papers based on their sentiment and intent. For this purpose, larger citation sentiment datasets are required. However, from a time and cost perspective, curating a large citation sentiment dataset is a challenging task. Particularly, citation sentiment analysis suffers from both data scarcity and tremendous costs for dataset annotation. To overcome the bottleneck of data scarcity in the citation analysis domain we explore the impact of out-domain data during training to enhance the model performance. Our results emphasize the use of different scheduling methods based on the use case. We empirically found that a model trained using sequential data scheduling is more suitable for domain-specific usecases. Conversely, shuffled data feeding achieves better performance on a cross-domain task. Based on our findings, we propose an end-to-end trainable multi-task model that covers the sentiment and intent analysis that utilizes out-domain datasets to overcome the data scarcity.

Keywords: Artificial intelligence · Natural language processing · Scientific citation analysis · Multi-task · Transformers · Sentiment analysis · Intent analysis · Multi-domain

1 Introduction

Neural Networks have recently been applied to tasks from a wide range of domains. They are also notorious for their desire for very large amounts of annotated data, one of the key requirements to use neural networks is the availability of annotated data. While the process of data annotation can be automated in some domains to ensure the

D. Mercier, S. T. R. Rizvi and V. Rajashekar—Equal Contribution.

© Springer Nature Switzerland AG 2022

A. P. Rocha et al. (Eds.): ICAART 2021, LNAI 13251, pp. 113–134, 2022.

https://doi.org/10.1007/978-3-031-10161-8_6

availability of the necessary data. However, it is not always possible and the quality of automatically annotated data can not be ensured as mentioned in [27].

Citations data for sentiment analysis is a particular example of such a scenario where the data is already very scarce and challenging to collect and annotate using automated approaches. While the annotation of product reviews can be automated, the automated annotation of texts without additional features like stars ratings and emojis, etc. is a significantly more complex task [25].

Scientific publications play an important role in the progress of a community. The “Publish or Perish” principle continuously pushes the researchers to periodically publish their scientific contributions which resulted in a boom of publications. This exponential increase in the amount of existing scientific publications has posed a challenge of evaluating the impact of each contribution in this publication outburst. Despite the existence of various metrics, including the h-index, aspects such as sentiment and intent are rarely evaluated. It is a well-established fact that most of the existing metrics heavily rely on citation counts and therefore only take quantitative aspects of a citation into consideration [5]. However, the quality of a scientific contribution should not entirely depend on quantitative aspects rather on the content and the results [12,36].

Such qualitative facet greatly assists in the citation impact measurements by enriching them and therefore resulting in more sophisticated significance rankings [36]. The task of sentiment classification offers contextual insights into a given text corpus and is applied on various domains such as movie review, product reviews, and Twitter data [3,11,16,19,32]. Performing sentiment analysis on objective citation data is still challenging due to the objectivity of the text and the limited amount of annotated data.

The intent of a citation found in scientific literature refers to the purpose of citing the existing scientific artifacts. Citation intent analysis serves a dual purpose. Besides the intention of a citation i.e. approach, dataset, survey, or related work, it also plays a crucial role in identifying the sentiment [20] of that citation based on its occurrence position in the paper. For instance, citations found in the evaluation and discussion section are more likely to be negative, as the citing authors usually compare the results of their approach in evaluation to prove the superiority of their proposed approach.

Despite the recently published approaches [4] there is still a scarcity of methods and datasets for the task of scientific citation analysis. There are a couple of factors that caused this data scarcity. Firstly, the high costs of manual annotation and the highly objective text make it impossible to automatically annotate it with a high quality. Secondly, there is no formal definition of intention used to classify citations properly.

In our previous paper ImpactCite [21], we contributed by releasing a cleaned citation sentiment dataset for the task of citation sentiment analysis. In addition, we proposed a transformer-based approach for classifying the sentiment or intent of a given citation string. Even with our dataset contribution, the scarcity of citation sentiment data was not eliminated. Therefore, in this paper, we further investigated the usage of out-domain sentiment datasets to learn and transfer their knowledge to the citation sentiment task. For this purpose, we utilize the cleaned dataset proposed in ImpactCite [21] and extend its analysis using out-domain data to further improve the performance of the citation analysis model. We also evaluate different scheduling methods to train models on the data and investigate the impact of those strategies concerning the model per-

formance. Furthermore, we investigate the impact of training a single model for two different tasks of the same domain to enhance the accuracy of the task with limited annotated data. It will significantly save both time and computation resources. To our knowledge, it is the first endeavor of investigating sentiment and intent classification on scientific data including out-domain data integration. Citation sentiment analysis would benefit greatly if sentiment datasets from other domains had a positive effect on citation analysis models. The contributions of this publication are as follows:

1. Evaluation of out-domain data usage during training
2. Evaluation of different scheduling methods
3. An end-to-end sentiment and intent citation classification multi-task model.

2 Related Work

This section discusses the literature related to three relevant aspects. Firstly, we will discuss the works related to the sentiment classification followed by the intent classification. Later, we cover literature related to the use of out-domain data, transfer learning, and the possible impact of these approaches concerning data scarcity.

2.1 Sentiment Classification

Sentiment analysis has been a notable task in natural language processing. Several approaches have been proposed in the existing literature which focused on tackling the task of Sentiment classification. The most common use cases for the task of sentiment classification include the sentiment analysis of tweets, movies, and products reviews. Tang et al. [30] proposed a word embeddings-based approach contingent on sentiment present in a tweet. These sentiment-oriented word embeddings make this approach very suitable for the task of sentiment classification. Thongtan et al. [31] took it one step further and applied document embeddings instead of word embeddings. These document embeddings were trained using cosine similarity as the similarity measure. The effectiveness of this approach was demonstrated by applying it to a dataset consisting of movie reviews. Cliche [6] adopted a slightly different path and employed an ensemble of Convolutional Neural Network (CNN) and Long-Short Term Memory (LSTM) models. This approach was trained and fine-tuned on a large corpus of unlabeled tweets for the task of sentiment classification.

BERT [9] is considered as the most popular choice for different natural language processing (NLP) tasks. It was trained on a large corpus of unlabeled data. Owing to its success in resolving other NLP problems, BERT has also been applied to the task of sentiment analysis. Several approaches [23,33,37] took advantage of the baseline BERT model and further tapped the potential of the model by incorporating different modules like pre-processing, attention, and structural features, etc. These modules provided some additional information to the model which in turn helped the model to better predict the resultant label.

Most of the research related to the task of sentiment analysis is performed for the domains of movie/product review or Twitter data sentiment analysis. However, a minor

fraction of the literature also targets a different domain for the task of sentiment analysis. Citation sentiment analysis is also of extreme importance as it helps us in understanding the impact of research artifacts in a scientific community. Citation sentiment analysis is vastly different from movie/product review or Twitter sentiment analysis, unlike reviews and tweets, citations appear in the scientific literature which is a quite formal form of text. Esuli and Sebastiani [10] proposed the idea that sentiment classification has striking similarities with opinion and subjectivity mining. They further discussed that an individual can premeditate a seemingly positive or negative citation by only using their inclinations and writing style.

Athar et al. [2] explored the idea of using sets of several features like science lexicon, contextual polarity, dependencies, negation, sentence splitting, and word-level features for citation sentiment classification. They performed several experiments to establish a set of most suitable features which has optimal performance in classifying citation sentiment found in scientific literature. On a similar line, Xu et al. [34] carried out a citation sentiment analysis on the clinical trial literature. For this task, they employed a different set of features like n-grams, sentiment lexicon, and structure information. The task of sentiment classification is particularly hard for citation data due to the inadequate number of datasets available which have a very limited number of samples for sufficiently training a model. Finding a sentiment in a text that is written to be analytical and objective is substantially different from doing so in highly subjective text pieces like Twitter data.

2.2 Intent Classification

Intent classification and sentiment classification seem to be nearly identical tasks. However, both tasks are inherently different as intent classification is more inclined towards motive behind citation which is generally closely related to the section in which the citation string appears. Intent classification has become a more challenging task due to the increasing usage of compound section titles. Cohan et al. [7] employed bi-directional LSTM equipped with an attention mechanism. Additionally, they proposed to use ELMo vectors and structural scaffolds i.e. citation worthiness and section title.

Another interesting work is SciBERT which is a variation of BERT specifically optimized for scientific publications and was proposed by Beltagy et al. [4]. The model was trained on 1.14 million scientific publications containing 3.17 billion tokens. The training data originates from two different domains, namely the computer science and biomedical domain. SciBERT was successfully applied on several NLP tasks including the classification of sections.

Furthermore, Mercier et al. [20] tackled the sentiment and intent classification using a fusion approach of different baseline classifiers such as a Support Vector Machine (SVM) and a perceptron. They used a set of textual features consisting of adjectives, hypernyms, type, length of tokens, capitalization, and synonyms. Closely related to that, Abu-Jabra et al. [1] proposed an SVM-based approach to perform the intent classification of citations. They stated that structural and lexical features in their experiments have shown to be of very high significance when it comes to the intent of a citation.

2.3 Out-Domain Data Utilization

Su et al. [29] presented in their work to study the impact of out-domain data for question answering. They investigated different training schedules and their impact on accuracy. The main focus of their work was a better generalization. Another work that conducted experiments related to the robust training using in-domain and out-domain data was proposed by Li et al. [15]. Their proposed method provides the capabilities to learn domain-specific and general data in conjunction to overcome the convergence towards domain-specific properties. Sajjad et al. [26] proposed an approach that first learns of different out-domain data and finally fine-tunes on in-domain data to achieve the optimal results. This approach intuitively utilizes the data of the different domains and therefore has a much larger training corpus for a better generalization.

Khayrallah et al. [13] addressed the amount of out-domain vocabulary. Their findings showed that with the use of out-domain data and a continuous adaption towards the domain, the number of words not included in the vocabulary can be reduced efficiently. For this purpose, they used an out-domain model and trained it with a modified training objective continuously on the in-domain data. Furthermore, Mrkšić et al. [22] showed that using the out-domain data can yield significant improvements for very small datasets. And therefore makes it possible to train models using these sets when it is not possible to do that without the use of out-domain data.

3 Datasets

This paper mainly focuses on the task of sentiment and intent analysis. Therefore we selected a range of datasets suitable for sentiment classification and also for intent classification.

3.1 Sentiment Datasets

For the task of Sentiment classification, we employed various datasets for our experiments. Our target domain is the scientific literature. However, we selected some out-domain datasets to overcome the data scarcity. Following are the datasets selected for the sentiment classification task:

1. Movie reviews
2. Product reviews
3. Twitter data
4. Scientific data.

To standardize the labels of selected datasets, a preprocessing step was essential. For experiments evaluating out-domain knowledge transfer and sequential training, we preprocessed the selected datasets for binary sentiment classification tasks i.e. positive and negative. It enabled us to train and test models across different datasets. To do so, we excluded the neutral class and grouped different labels if the datasets had multiple classes that correspond to the positive or negative label e.g. ‘good’ and ‘very good’ or 4 out of 5 and 5 out of 5 stars. However, we used all three classes i.e. positive, negative, and neutral for the multi-task experiments. The details of the selected sentiment datasets are as follows:

Movie Reviews. From the domain of movie reviews, we decided to use three popular datasets that quantified both positive and negative reviews in the form of a numerical score. The IMDB [17] dataset contains about 25,000 training and 25,000 test instances of highly polar reviews. It is the largest dataset by volume in the selected datasets. The second dataset we used in our experiments is the Cornell movie review data [24]. It is a considerably small dataset as compared to IMDB. However, it has an even distribution of 1,000 samples for each of the positive and negative classes. The last dataset that we selected from movie reviews is the Stanford Sentiment Treebank dataset [28]. For this dataset, we had to discard the samples not related to either negative or positive classes. All three above-mentioned datasets are related to the same task from the same domain and therefore their underlying structure should be rather similar.

Product Reviews. To include a dataset from a different domain than the Movie reviews, we selected the amazon product review dataset [18]. This dataset consists of various product categories. Some of the categories in the amazon data are closely related to the movie reviews such as Books, TV, and Movies. On the other hand, some categories are completely different from movie reviews such as Beauty, Electronic, and Video Games. For our experiments, we selected one category from amazon data that was unrelated to the movie reviews. The chosen category was related to the instrument reviews. The product reviews were quantified in the form of 1 – 5 stars. For our experiments, we converted the star ratings into positive and negative classes while skipping the neutral class. Product reviews with ratings with 4 and 5 stars were labeled as positive. On the other hand, product reviews with 1 and 2 stars were labeled as negative. However, product reviews with a star rating of 3 were skipped as they belonged to the neutral class and were not relevant for our experiments.

Twitter Data. Sentiment analysis on Twitter data is a quite popular task. For this purpose, we selected a couple of Twitter datasets. Intuitively, we assume that the Twitter datasets are the most subjective ones in our selection as their language style differs significantly from the scientific and other domain datasets. The first dataset is related to airline reviews in form of tweets. The data was taken from Kaggle¹ and contains three classes i.e. positive, negative, and neutral. Similar to other datasets, we removed the neutral class. The same class elimination was performed for the second dataset Sentiment140 dataset². This dataset was composed using 1.6 Million general tweets collected from Twitter along with their sentiment.

Scientific Data. From the scientific domain, we selected a dataset called Citation Sentiment Corpus (CSC-Clean). It was proposed in our previous paper [21]. There have been very limited contributions for the citation sentiment analysis task as the number of available datasets is almost not existent. Although there exist some datasets proposed by Xu et al. [34] and Athar [2] those are either not publicly available or suffer from bad

¹ Twitter US Airline Sentiment: <https://www.kaggle.com/crowdflower/twitter-airline-sentiment>.

² Sentiment140: <https://www.kaggle.com/kazanova/sentiment140>.

Table 1. Comparison of citation sentiment corpus (CSC) and citation sentiment clean (CSC-C) dataset. Taken from [21].

Classes	CSC	CSC-Clean	CSC-Clean Dist.	Removed [%]
Positive	829	728	9.12%	101 (12.18)
Neutral	280	253	87.71%	27 (9.64)
Negative	7,627	6,999	3.17%	629 (8.25)

quality. The reason for this data scarcity is the expensive and complicated labeling process. We decided to use the CSC dataset [2] as a baseline. Upon careful dataset analysis, we found out that there exist duplicate instances with an occasionally same or different label in the CSC dataset. These instances also often exist in different data splits such as training and test set. We identified these quality issues and cleaned the dataset to achieve a better quality throughout the same corpus. Table 1 shows the original sample count, number of removed instances concerning duplicates, and the remaining number of samples. In addition, we show the updated dataset distribution and the percentage of removed instances concerning each class. In total, we removed 757 instances which are 8.67% of the data. For duplicates with two different labels, we removed both the original and the duplicated instances as this is the only appropriate solution to avoid a subjective bias from our side. Including one of the instances would bias the data and results. The resultant dataset is referred to as CSC-Clean and publicly available³.

Sentiment Dataset Statistics. In Table 2 we show the statistics of each sentiment dataset after pre-processing them to exclude the neutral class and existing duplicates. These statistics include the number of samples used to train, validate, and test our models. In addition, the table also shows the dataset distribution highlighting that datasets such as the Instruments, US Airline, and CSC-Clean are heavily biased towards one of the two classes. Another characteristic is that the collected datasets differ largely in their size. This resulted in the need to upsample or downsample the data for some experiments to make the results comparable.

3.2 Intent Dataset

From the scientific domain, we selected a dataset related to citation intent analysis called SciCite. The SciCite dataset proposed in [7] is a famous benchmark for citation intent classification. It was curated using medical and computer science publications and is publicly available. The size of this dataset is sufficient to train any deep learning model and the existing benchmarks emphasize the high quality of the dataset. However, the dataset has an imbalanced sample distribution in which the vast majority of the samples are assigned to the ‘Background’ class. Another, important aspect of the dataset is the coarse-grained label process which was applied to create that dataset. According to

³ <https://github.com/DominiqueMercier/ImpactCite>.

Table 2. Comparison all used datasets. Only including the positive and negative class. Neutral class for CSC-Clean was excluded in this table.

Domain	Dataset	Train	Val	Test	Positive [%]	Negative [%]
Movie Reviews	IMDB	19,923	4,981	24,678	50.19	49.81
	Cornell	6,823	1,706	2,133	50.0	50.0
	Stanford Sent.	6,911	872	1,819	51.64	48.36
Product Reviews	Instruments	6,068	1,507	1,897	95.07	4.93
Twitter Data	US Airline	7,243	1,811	2,264	19.81	80.19
	Sentiment140	10,161	2,541	3,176	49.94	50.06
Scientific Data	CSC-Clean	797	89	95	74.21	25.79

Table 3. SciCite [7]. Number of instances and class distribution. Taken from [21].

Classes	Training	Validation	Test	Total	Percentage
Result	1,109	123	259	1,491	13.53
Method	2,294	255	605	3,154	28.62
Background	4,840	538	997	6,375	57.85

the authors, the distribution follows the real-world distribution and the number of samples is large enough to sufficiently learn the concepts of each class. Detailed information about the dataset can be found in Table 3. We mainly employed SciCite along with the CSC-Clean dataset to demonstrate the capability of training a multi-task model, where tasks are different and yet from the same domain.

4 Contributions

We divided this section into three main parts. The first part discusses the baseline work from our previous paper ImpactCite [21]. Secondly, we will discuss the impact of training a model on out-domain data. And the third part covers a fusion approach to combine sentiment and intent. We further show that both methods rely on different aspects of the task and highlight their advantages.

4.1 ImpactCite

Our previously proposed approach, ImpactCite [21] served as a baseline for this paper. It is an XLNet [35] based approach for analysis of sentiment and intent of citations found in scientific literature. ImpactCite utilizes two separate XLNets to provide a citation sentiment and intent analysis. To the best of our knowledge, there exists only limited work concerning scientific citation analysis.

The task of citation analysis involves two challenging dataset characteristics. First, the dependency on sentences next to the actual citation. Taking into account that most of the citation sentiment origins from neighborhood sentences lead to longer sequences. Secondly, the model needs to cover dependencies in both directions as in the scientific

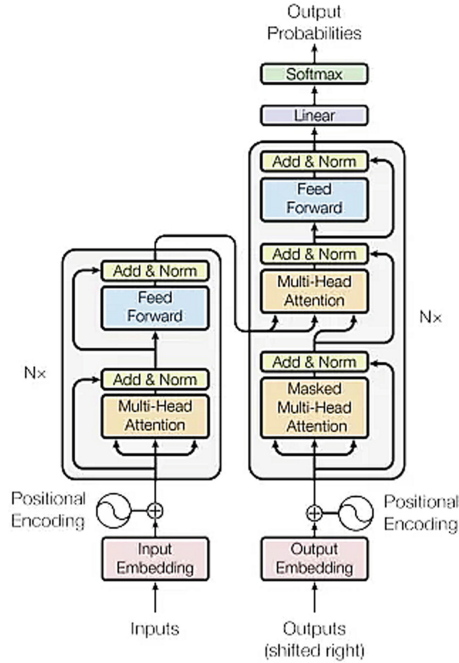


Fig. 1. Transformer-XL architecture [8]. Each of the Multi-Head Attention layers is composed of multiple attention heads that apply a linear transformation and compute the attention.

world the sentiment might be given before or even after the actual citation sentence. Taking into account these essential properties we decided to use XLNet [35] as a model for our experiments. XLNet is a well-known transformer-based network structure that can cover long sequences and bi-directional dependencies. The auto-regressive model is based on a Transformer-XL [8] as the backbone. The Transformer XL architecture is shown in Fig. 1. In addition, there exist many pre-trained XLNet models which is essential for the sentiment classification as the number of datasets for scientific sentiment citation is not sufficient to train such a model from scratch. Precisely, we decided to use the XLNet-Large model to make sure that the model is large enough to cover the whole context. XLNet-Large consists of 24-layers, 1,024 hidden units, and 16 heads. During our experiments, we only fine-tune the pre-trained model according to the different tasks involving cross-domain sentiment analysis, scientific sentiment classification, and scientific intent classification. As the language of the pre-trained model and the data used to fine-tune it we benefit from the pre-trained weights as the general language structure is similar and only needs small adjustments concerning the domain and task.

Separating these two tasks enables us to fine-tune the corresponding model to each task and achieve the best possible results for that task. This is especially beneficial for the intent as the amount of sentiment citation data is limited. However, the major drawback is that two separate models are required for this purpose and the sentiment does not benefit from the intent model although both tasks are from the same domain.

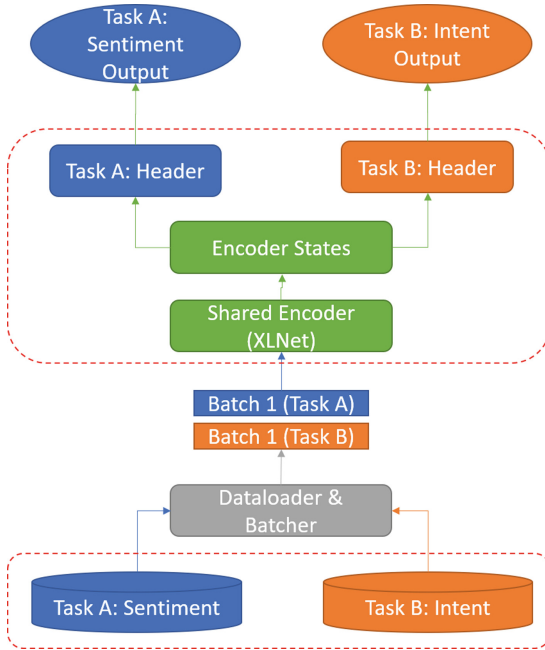


Fig. 2. Multi-Task setup combining sentiment and intent task. The same encoder is used for both tasks and a task specific head is trained.

4.2 Overcoming Data Scarcity and Data Feeding Techniques

In this paper, we investigated the techniques to overcome the scarcity of data for certain domains. Particularly for sentiment analysis of scientific citations, there are not many datasets available. In this paper, we propose that training on out-domain data and later finetuning on target domain results in better model performance, therefore, bridging the data scarcity gap. Additionally, we experimented with different data feeding methods to analyze their impact on the performance of the final model.

4.3 Fusion Approach

Lastly, in this paper, we propose that although the citation sentiment and intent analysis are different tasks. However, we believe that the underlying text structure concerning the sentiment and intent task on scientific data is similar. Based on the cross-domain sentiment classification we show that the addition of data addressing the same task or the same domain can enhance the scientific sentiment classification. Ultimately, we train a single XLNet model on both the sentiment and intent datasets that performs the complete citation analysis and resolves the dataset size issues. The pipeline is visualized in Fig. 2.

5 Experiments and Analysis

In this section, we will discuss our experiments and their results. All experiments are classified into four sets. The first set discusses the performance benchmark of XLNet [35] for the task of intent classification. The second set discusses the experiments related to the performance benchmark of XLNet for sentiment classification. We performed these benchmarking experiments using several other models ranging from the baseline models i.e. CNN to highly sophisticated language models i.e. BERT [9], ALBERT [14] and XLNet [35]. In the third set of experiments, we will discuss the experiments related to training on out-domain data and testing on several different domains dataset, which also includes finetuning on the target domain dataset. Additionally, a collection of experiments discussing the effects of different data feeding techniques are also discussed in this set of experiments.

Finally, we discuss experiments combining the sentiment and intent modality and serve a single model that processes both tasks. Doing so requires a deep understanding of multiple aspects such as domain dependency, model selection, and task relation. In our case, the first two aspects are covered by benchmarking and the out-domain evaluation. In addition, it has been shown that the tasks are related to each other [20].

5.1 Intent Classification

Experiment1: Performance Benchmarking. To evaluate the performance of different model architectures on the intent classification task we decided to use the SciCite dataset [7]. We used the original train and test splits provided by the dataset and divided our models into two categories. The first category includes all baseline models. We explored different setups of CNNs, LSTMs, and RNNs. These models were trained from scratch using the SciCite dataset. In addition, we also trained BERT [9], ALBERT [14] and ImpactCite (XLNet). The second category of models was pre-trained and is a member of the transformer-based solutions. These models were only fine-tuned on the SciCite dataset. Due to high imbalance data, we employed the micro-f1 and macro-f1 scores for performance comparison. Furthermore, initial experiments using the CNN, LSTM, and RNN approaches have shown that their performance using pre-trained embeddings e.g. GloVe⁴ did not improve compared to newly initialized embeddings. We emphasize that one of the reasons for this might be the domain discrepancy between the pre-trained embeddings and the scientific domain.

Results and Discussion. Table 4 shows the performance benchmark results of different selected architectures for the intent classification task. It is evident from the results that both the LSTM and RNN are not able to compete with the CNN. A reason for the inferior performance of the RNN is the length of the sequences resulting in vanishing gradients for the RNN. The LSTM on the other hand suffers from the bi-directional influences between the sentences that are not completely covered by the architecture. We further explored different layer and filter sizes for baseline models. However, there is only an insignificant difference when tuning the parameters. Concerning the time

⁴ <https://nlp.stanford.edu/projects/glove/>.

Table 4. Performance evaluation on SciCite [7] (intent) dataset. L = Layer, F = Filter, C = convolution size. Taken from [21].

Topography	Architecture	Class-based accuracy			micro-f1	macro-f1
		Result [%]	Method [%]	Background [%]		
CNN	L 3 F 100 C 3,4,5	79.92	76.53	79.24	78.50	78.56
CNN	L 3 F 100 C 2,4,6	81.85	77.69	81.14	80.12	80.22
CNN	L 3 F 100 C 3,3,3	64.09	71.74	85.46	78.05	73.76
CNN	L 3 F 100 C 3,5,7	76.45	74.05	85.46	80.49	78.65
CNN	L 3 F 100 C 3,7,9	68.34	70.58	87.26	79.20	75.39
LSTM	L 2 F 512	73.75	73.55	79.54	76.80	75.61
LSTM	L 4 F 512	75.29	69.59	82.95	77.54	75.94
LSTM	L 4 F 1024	68.73	70.91	84.25	77.75	74.63
RNN	L 2 F 512	25.10	56.86	62.19	55.30	48.05
BERT [9]	Base	84.56	75.37	89.47	84.20	83.13
ALBERT [14]	Base	83.78	77.03	87.06	83.34	82.62
ImpactCite [21]	Base	92.67	85.79	88.34	88.13	88.93
BiLSTM-Att [7]	*	*	*	*	*	82.60
Scaffolds [7]	*	*	*	*	*	84.00
BERT [4,9]	Base	*	*	*	*	84.85
SciBert [4]	*	*	*	*	*	85.49

consumption, the CNN shows superior performance over the other baseline approaches as it can compute things in parallel as compared to LSTMs and RNNs.

The second category presented in Table 4 shows the complex language models. We were able to achieve a new state-of-the-art performance using ImpactCite [21]. It significantly outperformed the other fine-tuned language models by up to 3.9% micro-f1 and 5.8% macro-f1 score. Especially, the increase in the minority classes has shown a significant difference of 10%. Summarizing the findings, we have demonstrated that ImpactCite (XLNet) was able to outperform the CNN by 8.71% and the language models by 3.9% macro-f1 score and significantly increased the performance for the minority class. This highlights the significantly better capabilities of the larger transformer-based model pre-trained on a different domain and later fine-tuned.

5.2 Sentiment Classification

In this section, we will discuss the experiments conducted for the task of scientific sentiment classification. There were two datasets used in these experiments namely Citation Sentiment Corpus (CSC) and our proposed clean version of the dataset called CSC-C.

Experiment 1: Fixed Dataset Split on CSC Sentiment Dataset. For this experiment, we employed a fixed 70/30 data split for the CSC dataset excluding any additional dataset cleansing. We evaluated the performance of each previously used model. Additionally, we employed several sample strategies i.e. focal loss, SMOTE & upsampling, and analyzed their impact concerning the imbalanced data.

Table 5. Performance: Citation Sentiment Corpus (CSC). Taken from [21].

Topography	Modification	Class-based accuracy		
		Positive [%]	Negative [%]	Neutral [%]
CNN	*	28.2	21.3	94.8
CNN	Focal	36.9	16.9	94.3
CNN	SMOTE	39.4	20.2	84.2
CNN	Upsampling	36.1	6.7	92.8
LSTM	*	32.8	12.4	93.9
LSTM	Focal	42.7	19.1	82.8
LSTM	SMOTE	42.3	20.2	83.7
LSTM	Upsampling	26.1	11.2	97.0
RNN	*	24.5	21.3	72.7
BERT [9]	*	38.6	20.4	96.4
ALBERT [14]	*	44.3	28.8	95.8
ImpactCite [21]	*	78.9	85.7	75.4

Results and Discussion. The results of this experiment are shown in Table 5. We observed that all models mainly captured the concept of neutral citations. Additionally, we also observed that the methods like focal loss and SMOTE sampling increased the performance of the CNNs and LSTMs. Furthermore, upsampling does not help to improve the performance of the model. However, ImpactCite [21] effectively learned representations of each class. Especially, the negative class was captured in a much better way by ImpactCite. Although ImpactCite showed slightly worse performance on the neutral class, it performed significantly better for positive and negative classes. We conclude that ImpactCite is able to deal with the large class imbalance and show that the complex language models are superior to the baseline approaches enhanced with sampling and focus strategies for the CSC dataset.

Experiment 2: Cross-Validation on CSC-Clean Sentiment Dataset. In order to compare our proposed ImpactCite with the results of Athar [2] we used a 10-fold-cross validation. However, due to the missing split information and the duplicates that exist in the original CSC dataset, we decided to perform the experiment on the CSC-C dataset. Although the results are not directly comparable, the approach [2] is favored due to the duplicates that appear in the training and test data. For the sake of completion, we included [2] as a reference. During the 10-fold cross-validation, we used nine splits as training and one split as a test dataset for each run and averaged the results at the end. A collection of experiments were performed employing a variety of models ranging from baseline CNN models to complex BERT language models. In order to successfully apply the baseline methods, we used the class weights as they have shown superior performance in previous experiments.

Table 6. Cross validation performance: Sentiment citation corpus (CSC-C). Taken from [21].

Topography	Class-based accuracy			micro-f1	macro-f1
	Positive [%]	Negative [%]	Neutral [%]		
CNN	40.2	24.9	95.0	88.6	43.4
LSTM	34.8	19.0	92.1	84.6	46.1
RNN	20.7	17.9	86.0	77.9	41.5
BERT [9]	72.8	80.2	70.3	74.4	74.4
ALBERT [14]	71.1	72.5	67.6	70.4	70.4
ImpactCite [21]	64.6	86.6	82.0	77.7	77.7
SVM [2] ^a	*	*	*	89.9	76.4

^a Trained and tested on CSC

Results and Discussion. The results of this experiment are shown in Table 6. Interestingly, the baseline models were not able to achieve comparable performance even though the class weights were employed. In order to resolve the class imbalance issue, we pre-processed the folds for the baseline approaches such that the number of positive and neutral training samples was decreased to the number of negative samples. Doing so resulted in the performances shown in the table. Additionally, we observed that the complex language models performed much better on the small dataset. They significantly outperformed the baseline methods and achieved good results across all three classes. In addition, ImpactCite~[21] outperformed all other selected models and sets a new state-of-the-art for citation sentiment classification on the CSC-Clean. For the sake of completeness, we included the SVM used by Athar evaluated on the CSC dataset.

5.3 Out-Domain: Evaluating Impact of Additional Data

In this section, we present our results using out-domain data to evaluate its impact on the model performance. We investigate multiple scenarios of cross dataset training and testing on datasets from different domains. Furthermore, we conducted experiments concerning the use of multiple datasets and an optimal schedule strategy to enlarge the corpus size. We also discuss details of some experiments related to different data feeding methods.

Experiment 1: Out-domain Testing. In this experiment, we employed a pre-trained XLNet for each dataset and fine-tune it on one dataset. Once the model is trained, we evaluated its performance across all datasets to find out which datasets are semantically closer to each other. The goal is to better understand the correlation of the dataset and to what extent it is possible to use the model trained on an out-domain dataset for the prediction of sentiment across other domains. In this experiment, we trained each model for 40 epochs with a batch size of 24. In addition, we also used an early stopping mechanism such that if the model converges before 40 epochs then it will stop further training to prevent over-fitting. It has to be mentioned, that in this experiment the datasets had different sizes, as shown in Table 2.

Table 7. Results for testing on out-domain data using XLNets trained on a single dataset. Results are macro f1-scores in percent.

Train \ Test	Movie			Product	Twitter		Scientific
	IMDB	Cornell	Stanford	Instruments	Us Airline	Sentiment140	CSC-Clean
IMDB	94.38	81.58	83.66	70.20	64.53	62.72	54.16
Cornell	92.05	89.69	94.39	57.46	87.69	69.15	60.28
Stanford	91.71	89.49	92.85	63.68	86.46	68.89	63.76
Instruments	86.51	55.53	57.14	82.73	52.63	57.52	49.71
US Airline	56.80	71.45	79.47	43.80	92.21	68.39	43.45
Sentiment140	79.28	72.63	76.95	65.13	77.14	80.57	62.42
CSC-Clean	85.04	62.91	62.79	64.60	63.88	62.13	76.67

Results and Discussion. In Table 7 we show the results when using a single training set and testing across all datasets. Overall the best performance was achieved using the same dataset for training and testing, the only exception is the Stanford dataset. Interestingly, the performance for the Stanford dataset is surprisingly good when the model is trained on the Cornell data. It has to be mentioned, that both datasets are from the same domain. This shows that training on more domain data without fine-tuning on a specific dataset can result in a pretty good model for that dataset which is taken from the same domain. Overall training on the Stanford dataset was not successful. In general training on a dataset of the same domain without fine-tuning the model resulted in a good performance on their own domain however it is not the case when trained on out-domain data. One reason for this is the correlation between the data within the same domain. The results further show that the correlation across domains is in general lower but in the case of the Instruments dataset, the correlation is high enough to achieve superior performance using a dataset that is more balanced from the movie review domain. This suggests that a correlation between movie reviews and instrument reviews (product reviews) exists. Intuitively, this is the case because the understanding of positive and negative in the scientific domain is fundamentally different compared to review data or tweets.

Experiment 2: Sequential Training. In this experiment, we evaluated the impact of a sequential training scheme. The idea is that if a dataset is very small and therefore it is not possible to train only on that dataset, we enhance the data size by using additional datasets. There are two interesting aspects to using additional datasets. One it will increase the amount of data available for training and secondly, we also want to evaluate the impact of the dataset sequence in which data is fed to the network. Intuitively, the last dataset category in the training sequence should be favored with respect to the performance as the gradients are optimized on it. We performed this for a fixed sequence of datasets and categories and used several permutations of the sequence of categories to have comparable results. In addition, we performed these experiments twice, once for the upsampled datasets and once for the downsampled. The reason for this procedure is that it is important to make all datasets the same size such that they can contribute

Table 8. Macro f1-scores for sequential training. Sequence within the categories: [P]roduct (Instruments), [M]ovie (Cornell, IMDB Stanford Sent.), [S]cientific (CSC-Clean), [T]witter (Sentiment140, US Airline). ‘Up’ corresponds to the upsampled training data and ‘Down’ to the down-sampled training data.

Test Train	Movie			Product	Twitter		Scientific
	IMDB	Cornell	Stanford	Instruments	Us Airline	Sentiment140	CSC-Clean
Up							
STPM	93.05	88.51	90.87	80.22	89.69	75.45	78.18
MSTP	92.94	86.98	89.35	80.25	86.97	77.16	69.16
PMST	91.62	87.81	90.05	74.32	89.84	76.25	70.39
TPMS	92.19	88.19	91.26	77.72	90.04	76.08	76.97
Down							
STPM	92.38	87.29	89.98	80.45	85.93	76.96	75.55
MSTP	92.26	85.65	88.27	78.69	88.38	76.13	75.55
PMST	90.55	85.94	89.27	65.93	88.79	75.33	66.73
TPMS	88.95	83.00	86.98	72.07	87.11	75.18	67.34

the same amount to the training. With the initial dataset sizes, this would not be the case and a few datasets would dominate the training due to their size. In the upsampled version we used 3,000 samples whereas for the downsampling experiment we used the number of instances of the smallest dataset as a reference number. For some datasets, this means we had to select a subset of the training instances. This means we do not preserve the individual class distribution. The sequence of the datasets is shown in the corresponding results tables.

Results and Discussion. In Table 8 we present the results for sequential training. The upper part of the table covers the training results using the upsampling whereas the lower part covers the downsampling results. Our results for the upsampling showed that putting the movie review data at the end achieved the best scores for three out of the seven datasets. The performances overall were superior to the scores of the downsampling. Using the movie data as the last dataset in the training resulted in a 78.18% macro f1-score for the scientific data which is 1.21% better compared to setting the scientific data at the end of the sequence. The downsampled part shows that the training with the product data, in the end, has shown the best performance for the three datasets. Interestingly, the performance on the scientific data was 8.21% better using the downsampled either the product or movie datasets in the end compared to using its own dataset as last in the downsampled scenario. Except for the testing on Instruments and the CSC-Clean dataset, the performances of the other datasets did not change dramatically based on the feeding sequence. Another interesting finding was that putting the movie reviews in the end for the downsampled experiments did not result in a bad performance for all other dataset categories and led to a maximum drop of 2.86% for the US Airline dataset compared to the best performance for that dataset. In general, it was not the case that the models shows a bias towards the dataset that was used last in the training epoch. It is to be noted that due to the computational effort we did not try every combination but selected a subset that puts every category once at each position.

Table 9. Macro f1-scores for shuffled training. ‘Up’ corresponds to the upsampled training data and ‘Down’ to the downsampled training data.

Test \ Train	Movie			Product	Twitter		Scientific
	IMDB	Cornell	Stanford	Instruments	Us Airline	Sentiment140	CSC-Clean
Up	93.65	88.04	91.81	88.90	89.99	77.13	74.45
Down	97.80	87.07	88.42	83.40	86.96	76.65	73.73

Furthermore, the general finding of this experiment series is that unexpectedly the network does not work better when trained last on the evaluating dataset. Although most of the achieved accuracies are comparable it is not easy to predict which sequence works best for which testing set. Generally, upsampling was superior for most of the datasets. However, it requires much more training time. In our case, the dataset size is 3,000 compared to 797 samples for the downsampled version.

Experiment 3: Shuffled Training. In addition to the sequential data feeding experiment, we performed similar experiments by shuffling the data. The major difference compared to the previous experiment was that there is no sequence preserved, neither within the categories nor between the categories. Therefore, the gradients can align to each of the data samples and are not biased towards the last category in the setup.

Results and Discussion. In Table 9 we show the results of the shuffled upsampling and downsampling experiments. Surprisingly, the macro-f1 scores are close to each other. In these experiments, the downsampled data used about 800 instances of each dataset whereas the upsampled 3,000. Even more interesting is that the shuffled model performed well across all datasets. The largest accuracy drop compared to the single dataset training models was about 3.44% for the Sentiment140 dataset. Comparing the performances of the downsampled model to the models trained exclusively on those datasets, the accuracy of the shuffled model is impressively good. The same holds for the upsampled model. In general, the shuffled model holds a better generalization as it can be applied on all the datasets even without fine-tuning and sticks to good performance.

5.4 Multi-task Model: Fusing Scientific Sentiment and Intent

Experiment 1: Multi-domain Usage. We further experimented with the unified model for the sentiment and intent classification. This experiment combines both tasks into a single model. The motivation behind this experiment is to handle the increased amount of computation resource and inference time when using two separated models as proposed in ImpactCite. However, due to the small size of the CSC-Clean dataset it is not possible to train it directly in conjunction with the intent task. Therefore, we utilized the previous findings and combined the citation sentiment data with the sentiment datasets from other domains to enlarge the training set. Therefore, the sentiment task covers the sentiment classification for all used datasets that included a neutral class.

Table 10. Macro f1-scores sentiment and intent classification. Shows that the single task model is superior for the individual tasks.

Setup \ Task	Mutli-Task		Single-Task (ImpactCite)	
	All sent. datasets	CSC-Clean + Stanford	CSC-Clean	SciCite
Sentiment	64.00	56.00	80.41	*
Intent	78.00	78.00	*	88.93

Results and Discussion. Results in Table 10 show that the unified multi-task model has advantages however it is achieved with certain limitations. Firstly, the advantage of the multi-task model is that only a single model is used and two different heads are trained. This makes inference twice as fast as only one forward pass is needed and reduces the required hardware. However, the only impediment is that the model is trained on the conjunction of sentiment data and therefore the bias of the out-domain context can hinder the intent performance. It is to be noted that the model is robust against out-domain data for the sentiment task.

6 Discussion

In our previous paper [21] we have shown that our approach is capable to perform well on both the sentiment and intent classification. The results clearly highlighted the problems with the scientific sentiment domain and the lack of data. Additionally, the unbalanced datasets resulted in difficulties to converge for all evaluated methods except ImpactCite [21]. Neither ALBERT [14] nor BERT [9] were able to converge up to a state that provides a sufficient performance across all tested classes. While an intent classification using those models works well this is not the case for sentiment classification as some classes were not captured by the models. Especially, the negative class was identified as one of the major shortcomings. However, we were able to overcome this data shortcoming up to a certain extent using ImpactCite [20]. We achieved a new state-of-the-art performance for both tasks emphasizing the gains using XLNet [35] when the existing data is limited and unbalanced. In addition, these findings served as a baseline for qualitative citation analysis which is most times not considered due to the lack of available datasets.

In this paper we mainly focused on the utilization of out-domain data to enhance the sentiment classification in the scientific domain which suffers from the lack of existing annotated datasets. Our experiments have shown that without a specific fine-tuning the correlation between in-domain datasets is stronger compared to out-domain datasets and it is possible to achieve surprisingly good results training a classifier on a dataset of the same domain even without fine-tuning. Interestingly, in some cases, the larger quality datasets have shown better performance on some test sets than using the original training set. Going one step ahead, we evaluated different scheduling techniques to better understand the impact of data fusion. First, we tried different sequential concatenations resulting in better-generalized models that we are able to perform well across all datasets. Although the sequence has been shown to bias the performance slightly

towards the last category the results showed that the movie data as the last set in the sequence performed best. In addition, the difference between the upsampled and down-sampled training dataset versions highlighted that if the number of datasets concatenated is sufficient then this approach works for very small datasets below 800 samples. Next, we mixed all sentiment training data to avoid preserving sequence to favor any of the domains which resulted in a superior model with respect to the generalization. Shuffling all the data removed the convergence towards a single domain. Although it would be possible to fine-tune the model on a single dataset. We demonstrate that our solution is more robust as it is confronted with out-domain data during the training and further utilizes this data to establish a more general understanding of the underlying language concepts that are not bound towards one domain.

Ultimately, the combination of tasks within a single model can be very complex. During our experiments, we faced several challenges while combining the sentiment and intent tasks. It was not possible to train a model that is capable to converge using only the scientific sentiment and intent data. This is the case as the sentiment data is very small and when combined with the intent task, the network is not able to learn the concept of sentiment, especially negative sentiment, due to a large amount of unrelated data. Although we have shown in our previous work [21] that the use of two separate models is possible this might not be desired as the hardware required to run two models parallel is expensive. Furthermore, a sequential inference suffers from time delay. As a feasibility study, we combined the sentiment data with the out-domain sentiment data and trained the multi-task model. Ultimately, the proposed model is capturing multiple tasks and domains.

7 Conclusion

Utilizing our previous conducted experiments and findings presented in [21] we evaluated the impact of out-domain data usage during the training to enhance datasets and overcome data scarcity in less popular domains. Specifically, the issues faced in the sentiment analysis motivated us to evaluate the combination of different domain sentiment tasks and our results show impressive performances when the training procedure is aligned to work with the multiple concatenated datasets. Our first finding highlights that training using an in-domain dataset can already result in a suitable classifier for the target dataset even without fine-tuning due to the correlation of the data within the same domain. Going one step further we evaluated the impact of mixed datasets across the domains to enlarge the available amount of data. Doing so we found that the results for some datasets could be improved using a sequential approach in which the datasets with higher quality at the end boost the classifier. Furthermore, shuffling the datasets resulted in a powerful cross-domain model showing a good performance across all datasets. In contrast to the sequential scheduling, the performance of the shuffled approach was more balanced and not biased towards a single domain. Ultimately, we have shown in a feasibility study that multi-task models can be enhanced using out-domain data to enlarge the dataset. It was impossible to combine the scientific sentiment and citation data directly using the sentiment data due to the scarcity of the data. However, with out-domain data mixing, and a shuffled schedule we were able to come up with

a fully converge sentiment and intent model. One benefit of this model is the shared encoder resulting in much lower hardware requirements, faster training, and inference. In contrast to that, the separately trained models better converge for their specific task resulting in higher accuracies. We aim for the optimization of the dataset combinations and task combinations to achieve a better multi-task model open for future research.

References

1. Abu-Jbara, A., Ezra, J., Radev, D.: Purpose and polarity of citation: towards NLP-based bibliometrics. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 596–606. Association for Computational Linguistics, Atlanta, June 2013. <https://www.aclweb.org/anthology/N13-1067>
2. Athar, A.: Sentiment analysis of citations using sentence structure-based features. In: Proceedings of the ACL 2011 Student Session, pp. 81–87. Association for Computational Linguistics, Portland, June 2011. <https://www.aclweb.org/anthology/P11-3015>
3. Bahrainian, S.A., Dengel, A.: Sentiment analysis and summarization of Twitter data. In: 2013 IEEE 16th International Conference on Computational Science and Engineering, pp. 227–234. IEEE (2013)
4. Beltagy, I., Lo, K., Cohan, A.: SciBERT: a pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3606–3611 (2019)
5. Bornmann, L., Daniel, H.D.: What do we know about the h index? *J. Am. Soc. Inform. Sci. Technol.* **58**(9), 1381–1385 (2007)
6. Cliche, M.: BB-twtr at SemEval-2017 task 4: Twitter sentiment analysis with CNNs and LSTMs. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 573–580. Association for Computational Linguistics, Vancouver, August 2017. <https://doi.org/10.18653/v1/S17-2094>, <https://www.aclweb.org/anthology/S17-2094>
7. Cohan, A., Ammar, W., van Zuylen, M., Cady, F.: Structural scaffolds for citation intent classification in scientific publications. arXiv preprint [arXiv:1904.01608](https://arxiv.org/abs/1904.01608) (2019)
8. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q.V., Salakhutdinov, R.: Transformer-XL: attentive language models beyond a fixed-length context. arXiv preprint [arXiv:1901.02860](https://arxiv.org/abs/1901.02860) (2019)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018)
10. Esuli, A., Sebastiani, F.: Determining term subjectivity and term orientation for opinion mining. In: 11th Conference of the European Chapter of the Association for Computational Linguistics (2006)
11. Feldman, R.: Techniques and applications for sentiment analysis. *Commun. ACM* **56**(4), 82–89 (2013)
12. Garfield, E.: Is citation analysis a legitimate evaluation tool? *Scientometrics* **1**(4), 359–375 (1979)
13. Khayrallah, H., Thompson, B., Duh, K., Koehn, P.: Regularized training objective for continued training for domain adaptation in neural machine translation. In: Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pp. 36–44 (2018)
14. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: a lite bert for self-supervised learning of language representations. arXiv preprint [arXiv:1909.11942](https://arxiv.org/abs/1909.11942) (2019)

15. Li, Y., Baldwin, T., Cohn, T.: What's in a domain? Learning domain-robust text representations using adversarial training. arXiv preprint [arXiv:1805.06088](https://arxiv.org/abs/1805.06088) (2018)
16. Lin, C., He, Y.: Joint sentiment/topic model for sentiment analysis. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 375–384 (2009)
17. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 142–150. Association for Computational Linguistics, Portland, June 2011. <http://www.aclweb.org/anthology/P11-1015>
18. McAuley, J., Targett, C., Shi, Q., Van Den Hengel, A.: Image-based recommendations on styles and substitutes. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 43–52 (2015)
19. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: a survey. *Ain Shams Eng. J.* **5**(4), 1093–1113 (2014)
20. Mercier, D., Bhardwaj, A., Dengel, A., Ahmed, S.: SentiCite: an approach for publication sentiment analysis. arXiv preprint [arXiv:1910.03498](https://arxiv.org/abs/1910.03498) (2019)
21. Mercier, D., Rizvi, S.T.R., Rajashekar, V., Dengel, A., Ahmed, S.: ImpactCite: an XLNet-based solution enabling qualitative citation impact analysis utilizing sentiment and intent. In: Proceedings of the 13th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART, pp. 159–168. INSTICC, SciTePress (2021). <https://doi.org/10.5220/0010235201590168>
22. Mrkšić, N., et al.: Multi-domain dialog state tracking using recurrent neural networks. arXiv preprint [arXiv:1506.07190](https://arxiv.org/abs/1506.07190) (2015)
23. Munikar, M., Shakyia, S., Shrestha, A.: Fine-grained sentiment classification using BERT. In: 2019 Artificial Intelligence for Transforming Business and Society (AITB), vol. 1, pp. 1–5 (2019)
24. Pang, B., Lee, L.: A sentimental education: sentiment analysis using subjectivity. In: Proceedings of ACL, pp. 271–278 (2004)
25. Ranjan, H., Agarwal, S., Prakash, A., Saha, S.K.: Automatic labelling of important terms and phrases from medical discussions. In: 2017 Conference on Information and Communication Technology (CICT), pp. 1–5. IEEE (2017)
26. Sajjad, H., Durrani, N., Dalvi, F., Belinkov, Y., Vogel, S.: Neural machine translation training in a multi-domain scenario. arXiv preprint [arXiv:1708.08712](https://arxiv.org/abs/1708.08712) (2017)
27. Snow, R., O'connor, B., Jurafsky, D., Ng, A.Y.: Cheap and fast-but is it good? Evaluating non-expert annotations for natural language tasks. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp. 254–263 (2008)
28. Socher, R., et al.: Recursive deep models for semantic compositionality over a sentiment tree-bank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1631–1642 (2013)
29. Su, D., et al.: Generalizing question answering system with pre-trained language model fine-tuning. In: Proceedings of the 2nd Workshop on Machine Reading for Question Answering, pp. 203–211 (2019)
30. Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., Qin, B.: Learning sentiment-specific word embedding for twitter sentiment classification. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1555–1565 (2014)
31. Thongtan, T., Pienthrakul, T.: Sentiment classification using document embeddings trained with cosine similarity. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pp. 407–414. Association for Computational Linguistics, Florence, July 2019. <https://doi.org/10.18653/v1/P19-2057>, <https://www.aclweb.org/anthology/P19-2057>

32. Wu, Z., Rao, Y., Li, X., Li, J., Xie, H., Wang, F.L.: Sentiment detection of short text via probabilistic topic modeling. In: Liu, A., Ishikawa, Y., Qian, T., Nutanong, S., Cheema, M.A. (eds.) DASFAA 2015. LNCS, vol. 9052, pp. 76–85. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22324-7_7
33. Xie, Q., Dai, Z., Hovy, E.H., Luong, M., Le, Q.V.: Unsupervised data augmentation. CoRR abs/1904.12848 (2019). <http://arxiv.org/abs/1904.12848>
34. Xu, J., Zhang, Y., Wu, Y., Wang, J., Dong, X., Xu, H.: Citation sentiment analysis in clinical trial papers. In: AMIA Annual Symposium Proceedings, vol. 2015, p. 1334. American Medical Informatics Association (2015)
35. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: XLNet: generalized autoregressive pretraining for language understanding. In: Advances in Neural Information Processing Systems, pp. 5754–5764 (2019)
36. Yousif, A., Niu, Z., Tarus, J.K., Ahmad, A.: A survey on sentiment analysis of scientific citations. *Artif. Intell. Rev.* **52**(3), 1805–1838 (2017). <https://doi.org/10.1007/s10462-017-9597-8>
37. Zhou, P., et al.: Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (volume 2: Short Papers), pp. 207–212 (2016)