ImpactCite: An XLNet-based Solution Enabling Qualitative Citation Impact Analysis Utilizing Sentiment and Intent

Dominique Mercier[1,2,*][a], Syed Tahseen Raza Rizvi[1,2,*][b], Vikas Rajashekar[1],
Andreas Dengel[1,2][c] and Sheraz Ahmed[2][d]

[1] *Technische Universität Kaiserslautern, 67663 Kaiserslautern, Germany*

[2] *German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany*

Keywords: Deep Learning, Natural Language Processing, Intent Classification, Sentiment Classification, Document Processing.

Abstract: Citations play a vital role in understanding the impact of scientific literature. Generally, citations are analyzed quantitatively whereas qualitative analysis of citations can reveal deeper insights into the impact of a scientific artifact in the community. Therefore, citation impact analysis including sentiment and intent classification enables us to quantify the quality of the citations which can eventually assist us in the estimation of ranking and impact. The contribution of this paper is three-fold. First, we provide ImpactCite, which is an XLNet-based method for citation impact analysis. Second, we propose a clean and reliable dataset for citation sentiment analysis. Third, we benchmark the well-known language models like BERT and ALBERT along with our proposed approach for both tasks of sentiment and intent classification. All evaluations are performed on a set of publicly available citation analysis datasets. Evaluation results reveal that ImpactCite achieves a new state-of-the-art performance for both citation intent and sentiment classification by outperforming the existing approaches by 3.44% and 1.33% in F1-score. Therefore, the evaluation results suggest that ImpactCite is a single solution for both sentiment and intent analysis to better understand the impact of a citation.

1 INTRODUCTION

Scientific publications play an important role in the development of a community. An exponential increase in scientific literature has posed a challenge of evaluating the impact of a publication in a given scientific community. Citations majorly contribute towards the eminence of an author as well as the impact of their publications in a society. However, counting citations serves as a quantitative metric and therefore does not provide qualitative insights into the citations. In order to get a qualitative insight, the sentiment of a given citation is identified which refers to the opinion of the citing author about the cited literature.

We emphasize that using a qualitative metric taking into account different aspects of the citation leads to a much more sophisticated representation of the importance of an citation. Therefore, the sentiment and intent are used exemplary as two meaningful features that can enhance the existing approach. Precisely, the existing approach does not take into account the publication date, community background or availability of the work. However, the quality of a work should not depend on these aspects but rather on the content and the results. In short, to create a good metric it is important to cover additional aspects independent of the number of citations.

Sentiment classification provides us contextual insight into each of the literature citations. Sentiment classification is commonly applied to different domains (Bahrainian and Dengel, 2013; Wu et al., 2015; Feldman, 2013; Lin and He, 2009; Medhat et al., 2014) i.e. movie reviews, product reviews, citations, etc. where a given text string is classified based on its hidden sentiment. Therefore, it is possible to classify sentiments as either subjective & objective or a more fine-grained classification into positive, neutral, and negative depending on the domain and instances. However, sentiment classification can also induce subjectivity to the opinion.

[a] https://orcid.org/0000-0001-8817-2744
[b] https://orcid.org/0000-0002-4359-4772
[c] https://orcid.org/0000-0002-6100-8255
[d] https://orcid.org/0000-0002-4239-6520
[*] Authors contributed equally.

Mercier, D., Rizvi, S., Rajashekar, V., Dengel, A. and Ahmed, S.
ImpactCite: An XLNet-based Solution Enabling Qualitative Citation Impact Analysis Utilizing Sentiment and Intent.
DOI: 10.5220/0010235201590168
In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART 2021) - Volume 2*, pages 159-168
ISBN: 978-989-758-484-8

Sentiment classification provides us a deeper qualitative insight into a given literature citation. However, to get even deeper insights and to evade the likelihood of subjectivity, intent could be identified. The intent of a literature citation refers to the purpose of citing the existing literature. An author can cite a published manuscript for a number of reasons i.e. describing related works, using, extending, or comparing existing approaches and to contradict the claims from previous literature. Intent classification plays a crucial role in validating predicted sentiment of a given citation. The positioning of the citation plays an important role in identifying the sentiment. For instance, citations usually 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 approach.

Despite the recently published approaches e.g. Beltagy et al. (Beltagy et al., 2019) there is still a lack of methods and dataset used for scientific citation analysis. This lack of data originates from the effort mandatory to annotate scientific citations. Furthermore, most sentiment analysis cover domains in which the data is highly subjective and the annotation can be automated. Besides, there is no common definition of intention used to classify publications properly. In this paper, we cleaned a publicly available dataset for citation sentiment analysis and benchmarked the performance of several models ranging from simple CNN to more sophisticated transformer networks for sentiment and intent classification. By doing so, we achieved a new state-of-the-art for both sentiment and intent classification. We also present the new state-of-the-art as a single solution to be separately trained for sentiment and intent classification. The contributions of this paper are as follows:

- We propose one solution for both tasks in hand i.e. sentiment and intent classification. The proposed model can be separately trained for both tasks.
- We removed the discrepancies and the redundancies present in the previous version of the dataset and made a cleaned and reliable dataset for citation sentiment analysis publicly available[1] for the community.
- We conducted performance benchmarking of a set of models ranging from simple CNN based models to sophisticated transformer networks and achieving state-of-the-art performance for both sentiment and intent classification.

2 RELATED WORK

In this section, we discuss the existing literature for sentiment and intent classification. We also highlight the key aspects of each existing approaches.

2.1 Sentiment Classification

Sentiment classification is a popular task and due to its wide range of applications, there exist numerous publications to address this problem. Tang et al. (Tang et al., 2014) proposed sentiment-specific word embeddings for performing sentiment classification of tweets. Therefore highlighting that the use of highly specialized word embeddings can improve performance for sentiment classification. Thongtan et al. (Thongtan and Phienthrakul, 2019) employed document embeddings trained with cosine similarity to perform sentiment classification on a movie review dataset. Cliche (Cliche, 2017) proposed a sentiment classifier for tweets consisting of an ensemble of CNN and LSTM models trained and finetuned on a large corpus of unlabeled data.

With the popularity of transformer networks, BERT(Devlin et al., 2018) became a famous choice among the community for a range of Natural Language Processing (NLP) tasks. The BERT model was trained on a large volume of unlabeled data. Therefore, recent literature in the sentiment analysis domain makes use of the BERT model to improve the performance for the task in hand. In (Munikar et al., 2019; Zhou et al., 2016; Xie et al., 2019), the authors take advantage of transfer learning to adapt pre-trained BERT model for sentiment classification and further boost the performance by complementing it with pre-processing, attention modules, structural features, etc.

The literature discussed so far dealt with sentiment classification in tweets or movie reviews. On the other hand, citation sentiment classification is quite different from review sentiment classification, as the text in scientific publications is formal. Esuli and Sebastiani (Esuli and Sebastiani, 2006) defined that the sentiment classification is analogous to opinion mining and subjectivity mining. They further discussed that personal preferences and writing style of an author can induce subjectivity in the citations as an author can deliberately make a citation sounding positive or negative. Athar (Athar, 2011) performed different experiments using sets of various features like science lexicon, contextual polarity, dependencies, negation, sentence splitting and word-level features to identify an optimal set of features for sentiment classification in scientific publications. Xu et

[1] https://github.com/DominiqueMercier/ImpactCite

al. (Xu et al., 2015) performed sentiment analysis of citations in clinical trial papers by using textual features like n-grams, sentiment lexicon, and structure information. Sentiment classification is significantly important in the domain of scientific citation analysis due to the scarcity of scientific datasets suitable for scientific sentiment classification and the shallow definition of sentiment for this domain. 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

The basic concepts of intent classification are the same as sentiment classification. However, contrary to the sentiment classification, the definition of the citation intent classification is much sharper and the label acquisition is strongly related to the sections of a paper where it appears. Usually, section title provides a good understanding of the intent of the citation. However, compound section titles in scientific work can prove to be challenging for identifying the intent. Cohan et al. (Cohan et al., 2019) performed citation intent analysis by employing bi-directional LSTM with attention mechanism and consolidating it with ELMo vectors and structural scaffolds like citation worthiness and section title.

Beltagy et al. (Beltagy et al., 2019) proposed *SciBERT*, which is a variation of BERT optimized for scientific publications and trained on 1.14 Million scientific publications containing 3.17 Billion tokens from biomedical and computer science domains. SciBERT was applied to a group of NLP tasks including text classification to sections. Mercier et al. (Mercier et al., 2019) employed a fusion of Support Vector Machine (SVM) and perceptron based classifier to classify the intent of the citations. They used a set of textual features consisting of type & length of tokens, capitalization, adjectives, hypernyms, and synonyms. Similarly, Abu-Jabra et al. (Abu-Jbara et al., 2013) also employed SVM to perform the intent classification of citations. They suggested that lexical and structural features play a crucial role in identifying the intent of a given citation.

3 DATASETS

This paper deals with two important aspects concerning citation analysis namely the citation sentiment and intent. For this purpose, we used the following datasets to carry out the evaluations. We identified some inconsistencies in the sentiment dataset, which was later thoroughly cleaned and is being released along with this paper. However, despite the dataset limitation we decided to stick with the sentiment dataset and improve its quality to propose a cleaned version usable to perform citation sentiment analysis using deep neural models.

Table 1: SciCite (Cohan et al., 2019). Number of instances and class distribution.

	Classes		
	Result	Method	Background
Num. Train	1109	2294	4840
Num. Val	123	255	538
Num. Test	259	605	997
Num. Total	1491	3154	6375
Class dist.	13.53%	28.62%	57.85%

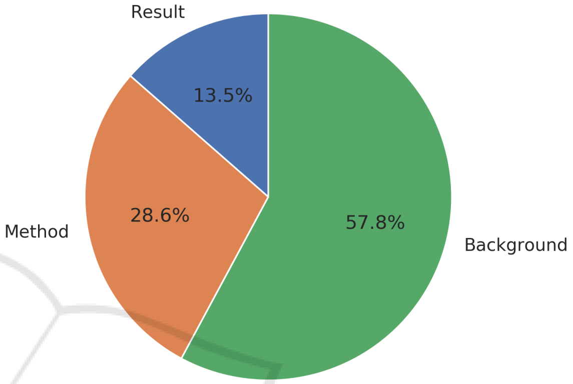

Figure 1: SciCite class distribution.

3.1 SciCite: An Intent Classification Dataset

In intent classification, we performed our experiments on the SciCite dataset which was proposed by Choan et al. (Cohan et al., 2019) and covers medical and computer science publications. We chose this dataset for the following reasons: SciCite is a well known publicly available dataset and covers computer science citations. Additionally, it has strong results emphasizing its quality and it is large enough to be used with state-of-the-art deep learning approaches.

The SciCite dataset has an unbalanced class distribution and consists of coarse-grained labels obtained by clustering citations based on their parent section. According to the authors (Cohan et al., 2019), three classes provide a scheme that covers the different intents. Table 1 shows detailed information about the number of samples used for training, validation and test. This dataset consists of three classes i.e. Result,

Table 2: Citation sentiment corpus (Athar, 2011). Number of instances and class distribution.

	Classes		
	Positive	Negative	Neutral
Avg. Length	229.4	221.8	219.6
Num. of samples	829	280	7627
Class dist.	9.49%	3.21%	87.30%

Method and Background. Where each class label represents the section where the citation was present and depict its respective intent to whether compare, extend or simply refer to the existing literature. Fig 1 shows the distrubution of the SciCite dataset where most of the samples belong to the Background class with 57.8%, while the Method and result class have relatively less number of samples. The background section provides the majority of citations whereas only a small amount of citations are classified as result or method. According to the authors, the distribution follow the real-world distribution and the number of samples is large enough to correctly learn the concepts of each class.

3.2 CSC: A Citation Sentiment Corpus

When it comes to the task of citation sentiment classification using publicly available high-quality datasets there is a lack of data. Although, there exist datasets for scientific papers e.g. the dataset proposed by Xu et al. (Xu et al., 2015) or the sentiment citation corpus proposed by Athar (Athar, 2011) these are either not publicly available or have quality issues. Precisely, this problem origins because of the data acquisition and labeling of scientific text as is can not be automated. Conversely, it is straight forward to acquire twitter or movie review data and label it. Due to the lack of alternate solutions, we had to stick to the dataset proposed by Athar (Athar, 2011) although this dataset has a very unbalanced class distribution as shown in Table 2. Fig 2 shows the distribution of samples among different classes. In the following sections, we refer to this dataset as CSC.

The CSC dataset consists of three classes Positive, Negative and Neutral. Where each class label represents the opinion of the citing author about the cited literature. Figure 3 shows the variation in token length and their distribution among different classes. It can be observed that the token length of the samples shows that the sample length is not an indicator for the label. In addition, these numbers demonstrate that a citation contains multiple sentences resulting in an additional context that can be utilized. Extracting only the sentence containing the citation would result in a potential information loss as the sentiment can be included in a follow-up or previous sentence. Therefore, we decided to keep the instances as they are providing us instances of multiple sentences to assure that the content relation can be learned correctly.

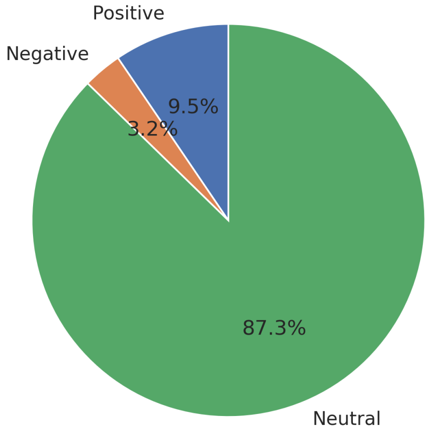

Figure 2: Citation sentiment corpus class distribution.

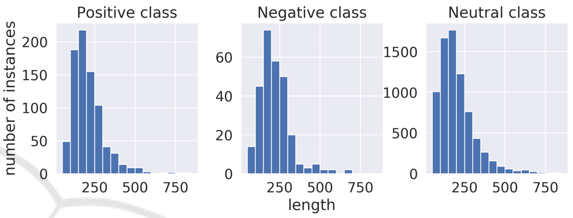

Figure 3: Citation sentiment corpus. Sample length class-wise.

3.3 CSC-Clean: A Cleaned Citation Sentiment Corpus

During the experimentation phase for this paper we identified several discrepancies concerning duplicated instances, wrong data splits, and samples with impressively bad quality concerning their label consistency. Therefore, it was not possible to compare our approach with the existing results published for the citation sentiment corpus and we decided to clean the dataset to create a improved dataset with better quality covering the same corpus. To do so, we applied the following two steps for dataset cleansing:

1. Removing duplicate samples with different labels
2. Removing duplicate samples with same labels

During dataset cleansing, we removed 756 instances as shown in Table 3. The removed instances were either identical duplicates of existing instances or provided different labels for the same text. In the case of samples with inconsistent labels, we removed all

Table 3: Comparison of citation sentiment corpus and clean citation sentiment dataset.

	Classes		
	Positive	Negative	Neutral
Citation sentiment corpus	829	280	7627
Clean citation sentiment dataset	728	253	6999
Removed instances	101	27	628

appearances as a manual selection of a specific instance would induce a bias. Although, this reduces the number of available instances, however it is the most appropriated solution to exclude possible subjectivity when it comes to the decision which instances label is correct as for the evaluation it is not suitable to keep both instances. We propose the dataset without any duplicates or inconsistent labels enabling to produce fair and meaningful results using cross-validation to overcome the limited amount of instances for the minority classes. In this paper, we will refer to this dataset as CSC-Clean. The cleaned dataset will be publicly available on the following link: https://github.com/DominiqueMercier/ImpactCite.

4 PROPOSED APPROACH: ImpactCite

In this section, we will briefly describe the neural network architecture adopted for our experiments. Furthermore, we substantiate the architecture choice and provide information about the training procedure.

4.1 Citation Analysis based on XLNet

To tackle the problem of sentiment and intent analysis we propose ImpactCite, an XLNet-based approach. XLNet is a famous choice for several NLP related tasks (Yang et al., 2019). XLNet is an auto-regressive language model contains bi-directional attention and is pre-trained on a large amount of data. The bi-directional attention makes it possible to understand relations within the sentences that can be drawn from left-to-right and vice versa. Due to the permutation generalization approach and the use of Transformer-XL (Dai et al., 2019) as the backbone model, XLNet can achieve excellent performance for language tasks involving long context. The Transformer XL architecture is shown in Figure 4. Especially, the capability to handle long context is important for the sentiment classification task as the sentiment of a citation can depend on the content of preceding or the proceeding sentences.

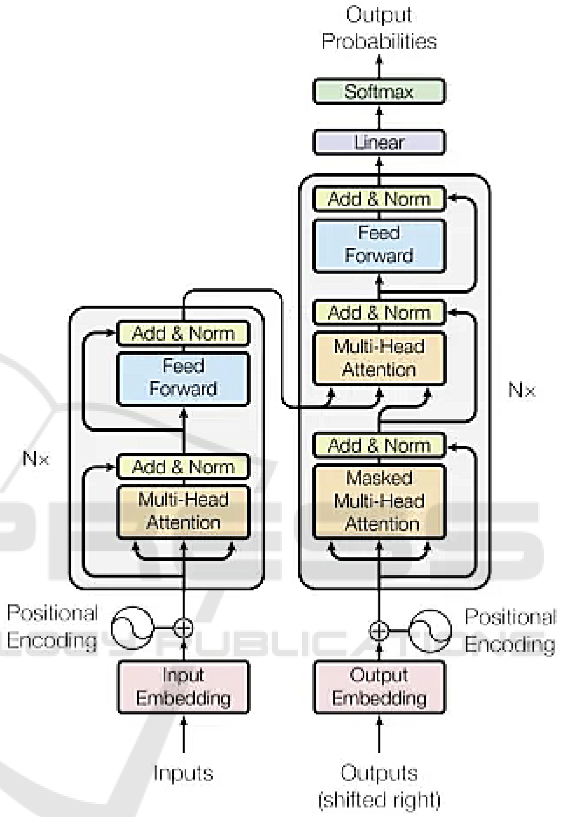

Figure 4: Transformer-XL architecture (Dai et al., 2019). Each of the Multi-Head Attention layers is composed of multiple attention heads that apply a linear transformation and compute the attention.

4.2 Model Architecture and Training Process

There are several variations of XLNet that differ slightly in the number of layers and units. For our experiments, we decided to use two XLNet-Large models. As our tasks cover a long context we decided to use the large version of XLNet. XLNet-Large consists of 24-layers, 1024 hidden units, and 16 heads. During our experiments, we rely on a pre-trained version of the model and fine-tune it according to the citation classification task. We start with a warm-up phase using a fixed learning rate followed by a slow learning rate decay to adjust the weights. This makes it possible to fine-tune the large model on a small dataset as the general language structure is already learned by the pre-trained model and we only adjust the weights to the new domain and task.

In this paper, we used one model for the sentiment classification and the second model for the intent classification. Doing so, we can infer the sentiment and intent for a given instance as an extension to the existing number of citations. To the best of our knowledge, there exists only a limited amount of work that covers the sentiment citation.

5 EXPERIMENTS AND ANALYSIS

In this section, we will discuss all the benchmarking experiments performed for two different text classification tasks namely citation intent and sentiment classification. We employed models ranging from the baseline models i.e. CNN to highly sophisticated language models i.e. BERT (Devlin et al., 2018), ALBERT (Lan et al., 2019) and XLNet (Yang et al., 2019) based ImpactCite.

5.1 Intent Classification

5.1.1 Experiments

For citation intent classification, we performed a bunch of experiments using different models. All the models were trained and evaluated on the SciCite dataset (Cohan et al., 2019). We used the train/test split provided with the SciCite dataset and trained three baseline models i.e. CNN, LSTM, and RNN from scratch using a different number of layers, filters, and convolution sizes. In addition, BERT (Devlin et al., 2018), ALBERT (Lan et al., 2019) and ImpactCite used pre-trained model initialization weights and were later finetuned on SciCite dataset. We computed the micro-f1 and macro-f1 as well as the accuracy for each label and each network. Initial experiments using the CNN, LSTM, and RNN approaches have shown that their performance using pre-trained embeddings e.g. GloVe[2] did not improve compared to new initialized embeddings. We emphazise, that the domain discrepancy could be the reason for the insignificant performance differences.

5.1.2 Results and Discussion

All the evaluation results of citation intent classification are shown in Table 4. it can be observed that the CNN clearly outperformed both the LSTM and RNN. A reason for the worse performance of the RNN is the length of the instances resulting in the vanishing gradient problem, whereas the LSTM processed the citation only in one direction and could not cover the influence on proceeding tokens. We explored different layer and filter sizes for baseline models, however, there is only an insignificant difference when tuning the parameters. Furthermore, CNN was not only superior in performance but also efficient in time complexity than the LSTM and RNN due to the good parallelization.

The second block of Table 4 shows the complex language models whereas the third block shows results from existing literature for citation intent classification. With the fine-tuning of complex language models, we achieved a new state-of-the-art performance by using ImpactCite. ImpactCite significantly outperformed fine-tuned BERT and ALBERT by 3.93% and 4.79% micro-f1 and 5.8% and 6.31% macro-f1 on SciCite dataset. It has to be noted that the accuracy for the classes with less representation in the dataset showed an improvement of about 10% as shown in Figure 5, stating that the generalization worked quite well. The performances of our fine-tuned BERT and ALBERT were close to each other showing only an insignificant difference. To conclude, ImpactCite outperformed CNN by 8.71% which 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 experiment designs for citation sentiment classification and their evaluations in detail. We adopted a couple of splitting strategies to partition the dataset into training and test set. We performed experiments on the original (CSC) and cleaned the citation sentiment corpus (CSC-Clean).

5.2.1 Experiment 1: Fixed Dataset Split on CSC Sentiment Dataset

In this experiment we used a fixed $70/30$ training/test split for the existing citation sentiment corpus proposed by Athar (Athar, 2011) without any additional data cleaning. This version of the dataset contained the duplicates and inconsistent labels. Similar to citation intent classification, we used three baseline models and three complex language models to perform the experiments for citation sentiment classification. In addition, for the baseline networks, we employed several sample strategies i.e. focal loss, SMOTE & up-

[2] https://nlp.stanford.edu/projects/glove/

Table 4: Evaluation results of intent classification on SciCite (Cohan et al., 2019) dataset. L = Layer, F = Filter, C = convolution size.

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	55.3	48.05
BERT (Devlin et al., 2018)	Base	84.56	75.37	89.47	84.20	83.13
ALBERT (Lan et al., 2019)	Base	83.78	77.03	87.06	83.34	82.62
ImpactCite	Base	92.67	85.79	88.34	**88.13**	**88.93**
BiLSTM-Att (Cohan et al., 2019)	*	*	*	*	*	82.60
Scaffolds (Cohan et al., 2019)	*	*	*	*	*	84.00
BERT (Beltagy et al., 2019; Devlin et al., 2018)	Base	*	*	*	*	84.85
SciBert (Beltagy et al., 2019)	*	*	*	*	*	85.49

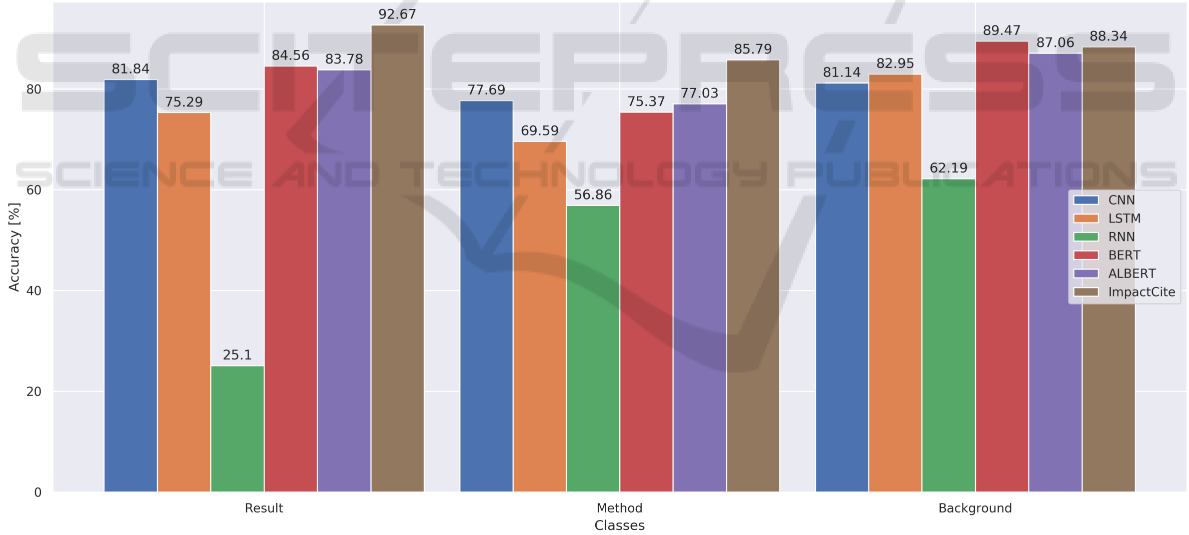

Figure 5: SciCite classwise accuracy.

sampling, and analyzed their impact concerning the imbalanced data.

5.2.2 Results and Discussion

In Table 5 we present the results by using the enhanced baseline approaches as well as three complex language networks. It can be observed that all models were able to capture the concept of neutral citations. Although, the focal loss or SMOTE sampling improved the performance for both the LSTM and the CNN. However, all the models except ImpactCite were unable to classify positive and negative samples with high accuracy. Additionally, we observed that the upsampling method did not improve the performance and rather had a negative impact. ImpactCite showed slightly worse performance on the neutral class, however, it performed significantly better for positive and negative classes. This highlights that only ImpactCite was able to overcome the class-

Table 5: Performance: Sentiment citation corpus (CSC).

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 (Devlin et al., 2018)	*	38.6	20.4	96.4
ALBERT (Lan et al., 2019)	*	44.25	28.81	95.84
ImpactCite	*	78.94	85.71	75.43

imbalance problem. Furthermore, these experiments show that even when used with additional sampling methods the complex language models are superior as they are pre-trained using a large amount of data.

5.2.3 Experiment 2: Cross Validation on CSC-Clean Sentiment Dataset

In this experiment, we will discuss cross-validation performed on the CSC-Clean. Due to a lack of train/test split of the dataset, Athar (Athar, 2011) performed 10-fold cross-validation on the original dataset. However, in our case, we performed 10-fold cross-validation on our CSC-Clean. Nevertheless, we include Athar's approach (Athar, 2011) as a reference and compare it with our clean dataset results. Therefore, we performed ten experiments each using nine out of the ten folds as training and one as test set and averaged their results to compute the overall accuracy. A bunch of experiments was performed employing a variety of models range from baseline CNN models to complex BERT language models.

5.2.4 Results and Discussion

In Table 6 we show the results for the cross-validation of selected models on CSC-Clean. For all baseline models i.e. CNN, RNN, and LSTM, we implemented the class weights to handle the class imbalance problem. Conversely, the results suggest that even after complementing baseline models with elaborated class weights, they are unable to tackle the class weights problem. Therefore, we pre-processed the training set for each fold in which the samples from positive and neutral classes were decreased to the number of negative samples present in the training set of that fold. This pre-processing helped in assuring that each class has equal representation in the training set. It has to be noted that pre-processing was performed on the training set only, whereas keeping the test set intact.

Additionally, complex language models i.e. BERT, ALBERT & ImpactCite can effectively finetune on small training data as they use their respective pre-trained models. Our results highlight that the baseline-approaches were not able to learn the concept of each class whereas the pre-trained models were able to achieve good results for all classes. Fig 6 shows the performance comparison of different approaches for each class. As a result, ImpactCite 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 Overall Discussion

Our evaluation results show that ImpactCite achieved solid results for both the sentiment and intent classification task. ImpactCite was able to handle the long instances and and cover the relation between the sentences within a citation to understand the global context. Conversely, BERT (Devlin et al., 2018) and ALBERT (Lan et al., 2019) were not able to do so. However, for the sentiment classification, it is especially

Table 6: Cross validation performance: Sentiment citation corpus (CSC-C).

Topography	Class-based accuracy			micro-f1	macro-f1
	Positive (%)	Negative (%)	Neutral (%)		
CNN	40.2	24.9	95.0	88.6	43,37
LSTM	34.8	19.0	92.1	84.6	46.13
RNN	20.7	17.9	86.0	77.9	41.53
BERT (Devlin et al., 2018)	72.8	80.2	70.3	74.4	74.4
ALBERT (Lan et al., 2019)	71.1	72.5	67.6	70.4	70.4
ImpactCite	64.6	86.6	82.0	77.7	**77.73**
SVM (Athar, 2011)[3]	*	*	*	89.9	76.4

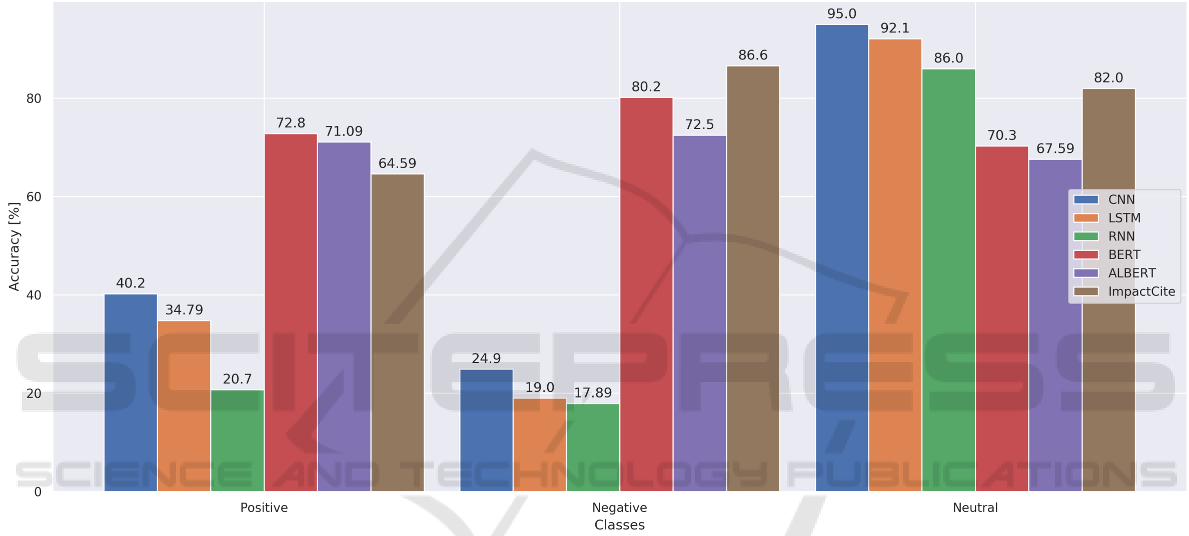

Figure 6: Sentiment citation corpus (CSC-C) classwise accuracy.

important to process the text from both sides to generalize well and deal with the influence of the preceding and following sentences. Additionally, it can utilize the permutations to create synthetic samples to overcome the small amount of data provided for the sentiment task. Therefore, ImpactCite achieved a state-of-the-art performance for both tasks. We propose ImpactCite, an ImpactCite-based solution covering both the sentiment and intent classification which leads to a qualitative citation analysis.

6 CONCLUSION

Our comprehensive experiments show the improvements in both the sentiment and the intent classification task for citations in scientific publications encouraging the use of those two properties to provide better information about the influence of papers. Also, we achieved state-of-the-art performance for the intent classification on SciCite (Cohan et al., 2019) dataset and sentiment classification on our clean citation sentiment dataset. Our results increased the SOTA for SciCite to 88.93% using ImpactCite which is an improvement of 3.44% compared to previous state-of-the-art. Furthermore, for the sentiment citation corpus, we pushed the old state-of-the-art result of 76.4% to 77.73%. Also, we compared the results for the different classes to highlight that the performance for two out of the three classes improved significantly. Our study emphasizes that recent transformer-based and auto-regressive models are far superior compared to simpler approaches like LSTM or CNN. Concerning the sentiment classification, we emphasize that the ImpactCite is much more robust for small or large datasets with long sequences

[3] Trained and tested on CSC.

and significantly outperforms other existing methods.

REFERENCES

Abu-Jbara, A., Ezra, J., and Radev, D. (2013). 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*, pages 596–606, Atlanta, Georgia. Association for Computational Linguistics.

Athar, A. (2011). Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 Student Session*, pages 81–87, Portland, OR, USA. Association for Computational Linguistics.

Bahrainian, S.-A. and Dengel, A. (2013). Sentiment analysis and summarization of twitter data. In *2013 IEEE 16th International Conference on Computational Science and Engineering*, pages 227–234. IEEE.

Beltagy, I., Lo, K., and Cohan, A. (2019). 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)*, pages 3606–3611.

Cliche, M. (2017). 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)*, pages 573–580, Vancouver, Canada. Association for Computational Linguistics.

Cohan, A., Ammar, W., van Zuylen, M., and Cady, F. (2019). Structural scaffolds for citation intent classification in scientific publications. *arXiv preprint arXiv:1904.01608*.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., and Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Esuli, A. and Sebastiani, F. (2006). Determining term subjectivity and term orientation for opinion mining. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Lin, C. and He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 375–384.

Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Mercier, D., Bhardwaj, A., Dengel, A., and Ahmed, S. (2019). Senticite: An approach for publication sentiment analysis. *arXiv preprint arXiv:1910.03498*.

Munikar, M., Shakya, S., and Shrestha, A. (2019). Fine-grained sentiment classification using bert. In *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, volume 1, pages 1–5.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., and Qin, B. (2014). 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)*, pages 1555–1565.

Thongtan, T. and Phienthrakul, T. (2019). 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*, pages 407–414, Florence, Italy. Association for Computational Linguistics.

Wu, Z., Rao, Y., Li, X., Li, J., Xie, H., and Wang, F. L. (2015). Sentiment detection of short text via probabilistic topic modeling. In *International Conference on Database Systems for Advanced Applications*, pages 76–85. Springer.

Xie, Q., Dai, Z., Hovy, E. H., Luong, M., and Le, Q. V. (2019). Unsupervised data augmentation. *CoRR*, abs/1904.12848.

Xu, J., Zhang, Y., Wu, Y., Wang, J., Dong, X., and Xu, H. (2015). Citation sentiment analysis in clinical trial papers. In *AMIA annual symposium proceedings*, volume 2015, page 1334. American Medical Informatics Association.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., and Xu, B. (2016). 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)*, pages 207–212.