Sentiment Classification of Scientific Citation Based on Modified BERT Attention by Sentiment Dictionary

Dahai Yu[1], Bolin Hua[1,*]

[1]Department of Information Management, Peking University, Beijing, 100871, China

Abstract
Citation analysis methods mainly focus on the quantitative indicators, such as the cited number and the H-index, while ignoring the deeper information such as citation function and citation sentiment. Therefore, studying and analyzing the functions and sentiments of citations can more effectively evaluate an article and uncover its underlying information. As for data, this study investigated the existing dataset of citation sentiment classification (CSC), collected and organized a high-quality and available dataset. As for model, based on the pre-trained language model BERT and its variants, a model called DictSentiBERT is proposed to modify attention mechanism using sentiment dictionary, and a series of baseline models are designed for comparative experiments. The experimental results show that compared to the original BERT and baseline models such as RNN and TextCNN, the DictSentiBERT improves the accuracy of CSC and maintains the highest Macro-F1 score.

Keywords
sentiment classification, informetrics, NLP, BERT, pre-trained language model

1. Introduction

Authors of academic articles establish various relations between different papers by citing concepts, methods, conclusions, and experimental processes to support their work, or introducing their own work by pointing out shortcomings in previous works. Therefore, studying these relations is of great significance for exploring implicit information or evaluating the quality and influence of papers. The analysis and mining of citation behavior can help reveal knowledge structures, research hotspots, research trends, and academic exchange networks within the research field.

The demand and function of citation analysis in academic community are gradually increasing, and citation analysis is no longer just to evaluate the academic value of research results. However, the traditional citation analysis methods mainly focus on the quantitative indicators, such as the cited number and the H-index, ignoring the deeper information of citation function and citation emotion. The work of Radicchi Filippo[1] and Baird L M[2] further demonstrates the limitations of the cited number, such as the fact that flawed or controversial paper tends to receive higher citations, while the cited number cannot reflect this information. Therefore, studying and analyzing the functions and sentiments of citations can more effectively evaluate an article and uncover its underlying information. Plus, researchers need to review and analyze existing papers to understand the current research status and development trends. Scientific citation sentiment classification can help researchers better understand others' attitudes and perspectives towards specific research fields, which helps to determine the quality and reliability of existing research, as well as evaluate research trends in the field. Citation sentiment refers to the author's emotional attitude towards the cited paper, such as approval, opposition, or neutrality. Citation sentiment analysis reveals this emotional attitude through various methods, such as SVM, Naive Bayes, TextCNN, BERT, etc. The dataset, code and logs have been uploaded to GitHub[1].

2. Recent Work

Research on the sentiments classification of text gradually emerged and increased significantly after 2009. Product review, social media conversations, news, and blogs are the most concerned fields[3]. According to Yousif[4] et al.'s research, sentiments classification of scientific citation first appeared around 2011.

Sentiment dictionary, machine learning, and deep learning are the three most common methods. Small[5] et al. used one to three sentences as citation contexts to assist in analyzing citation emotions, in order to understand the structure and potential cognitive processes of the citation. He used a dataset composed of a large number of prompt words or phrases to analyze in detail the functions and sentiments of 20 papers. Athar[6] classified citations into three categories: Positive, Negative and Neutral using SVM classifiers with different citation

Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 3rd AI + Informetrics (EEKE-AII2023), June 26, 2023, Santa Fe, New Mexico, USA and Online
*Corresponding author.
yudahai@pku.edu.cn (D. Yu); huabolin@pku.edu.cn (B. Hua)

[1]https://github.com/UFOdestiny/DictSentiBERT

sentiment detection features, and constructed a corpus containing 8736 instances. Poria[7] et al. proposed using CNN to extract features from multi-modal content and providing these features to a multi-core learning classifier for sentiment detection, which also achieved good results on different datasets.

The method of pre-trained models is gradually becoming popular. Beltagy[8] et al. used a large scientific corpus including a total of 1.14 million scientific papers of the biomedical (82%) and computer science (12%), rather than a general corpus to pre-train BERT. To some extent, the SCIBERT is more suitable for NLP tasks of scientific papers, significantly improving the effect of classifying scientific citations.

This study focuses on integrating prior knowledge into pre-trained models. Tingyu Xia[9] et al. found through analysis that the first layer of BERT has the worst ability to capture semantic similarity and lacks synonym information. Therefore, the author directly guided the attention of the first layer of BERT through prior knowledge. This method improves the performance of semantic matching, especially in small data. Weijie Liu[10] et al. proposed an article that applied knowledge graph to BERT and created K-BERT to solve the problem of poor performance of BERT in professional fields, and solved the two major problems of heterogeneous embedding space and knowledge noise in one fell swoop.

In summary, these studies have made outstanding contributions in the field of CSC. The method of sentiment dictionary is relatively simple, but it is limited by the quality and coverage of the dictionary, making it difficult to adapt to constantly changing themes. Methods of machine learning can achieve high accuracy, which, however, rely heavily on feature engineering, requiring manual selection and they may face challenges in efficiency and generalization when processing large-scale data. Deep learning methods perform well, but their application may be limited for tasks that lack large-scale annotated data. The pre-trained model does not require large-scale data, but if there is a significant difference between the trained corpus and the task corpus, its effectiveness will also be greatly reduced. Integrating the prior knowledge into the task of CSC can further improve the effectiveness. Among them, integrating knowledge graphs or constructing domain ontologies with BERT can achieve better results. However, there are also some problems: building and maintaining knowledge graphs requires a large amount of domain expert knowledge and data, resulting in higher maintenance costs. In addition, the updating process of the knowledge graph is relatively complex and time-consuming, so it may not be able to adapt to new fields or topics in a timely manner, limiting the model's adaptability to constantly changing text data.

Based on the aforementioned research's shortcomings and gaps, this study aims to explore the application of pre-trained model methods guided by prior knowledge in CSC. Using a sentiment dictionary to annotate the emotional intensity of each word in a sentence and adjust the attention matrix accordingly, the DictSentiBERT is introduced to combine the advantages of emotional knowledge and pre-trained models to improve classification performance without requiring a large amount of additional annotated data.

3. Data

The processing of data includes three stages: Source and Supplement, Preprocessing and Manual Screening, as shown in Figure 1. This study firstly investigated the existing datasets through checking academic papers, search engines, etc. It was found that the existing publicly available datasets have neither good quality nor good quantity. Therefore, we selected two datasets with relatively higher quality and better usability. Then, we used SCICite to supplement the citation sentiments corpus proposed by Athar. Afterwards, the dataset is subjected to a series of processing steps, including data deduplication and so on. Finally, we manually filtered these data.

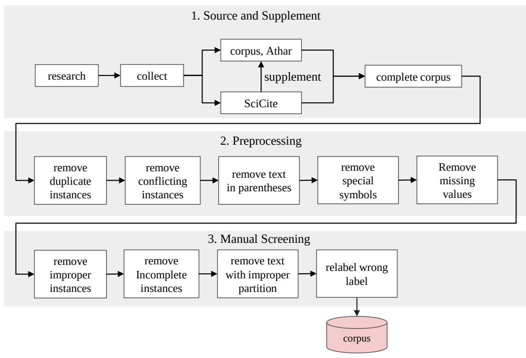

Figure 1: Flow Chart of Data Processing

3.1. Source and Supplement

After conducting detailed research, it was found that although there are many studies on CSC, such as the dataset collected by Xu[11] et al., the dataset annotated by Budi[12] et al., the dataset studied by Yaniasih[13] et al., or the emotional citation corpus proposed by Athar, these datasets are either not publicly available or have terrible quality. This may be due to the lack of unified and standardized annotation for data collection and labeling of scientific citation texts, making it difficult to achieve automation, or it may be due to lack of research in this field.

It's natural to think about transfer learning because obtaining data of movie reviews, social media reviews, or e-commerce reviews is simple, direct and easy. But this is problematic beacuse there are significant differences in language style, purpose, structure, etc. between the texts of scientific papers and those of film reviews.

1) Scientific papers usually adopt formal, professional and objective language style, and try to avoid subjective and emotional expression. Film reviews, on the other hand, place more emphasis on emotional expression and personal subjective opinions.
2) Scientific papers usually adopt standard structured forms,while film reviews are more liberal and typically include content such as movie introductions, personal impressions, and ratings, without a fixed structure and format.
3) The subject range of scientific papers and film reviews is also different. Scientific papers usually involve various professional fields in the academic field, including biology, chemistry, physics, etc., while film reviews mainly involve film, television industry and related topics.

To sum up, the differences between scientific papers and film reviews are multifaceted, involving the purpose, mode, intensity, object and audience of emotional expression. So the use of Transfer learning is not effective. Due to the lack of other solutions, this study still insists on using the dataset[2] proposed by Athar[6]. This corpus contains 8736 pieces of data, with each citation manually annotated as positive, negative, or neutral based on emotions. These citation sentences have been extracted from the ACL Anthology Network corpus.

In order to further improve the accuracy of training at the content level, after conducting comprehensive research on multiple publicly available datasets, we consider using the SCICite dataset proposed by Arman[14] et al. for data supplementation. SCICite contains a training set of approximately 10000 citation sentences and a testing set of approximately 1000 sentences, which are divided into three categories in terms of intention: method, background, and result. This dataset also provides another classification scheme: supportive and not supportive and this scheme fits this task very well. As a result, we extracted approximately 1000 sentences from SCICite to supplement the corpus proposed by Athar (not every sentence has that classification scheme).

3.2. Preprocessing

The study by Mercier[15] et al. indicates that the dataset contains many duplicate instances, incorrect data segmentation, and poor quality of label consistency, which may be caused by unclear division of labor for manual annotation. So, we clean the dataset and do some preprocessing according to their work.

1) Remove missing values.
2) Remove instances with same text but different labels.
3) Remove instances with duplicate text and labels.
4) Remove text within parentheses by regular expressions because the content are unrelated to sentiment analysis, such as "The two systems we use are ENGCG (Karlsson et al., 1994)"
5) Remove various types of special symbols, only retaining English text and numbers. Actually symbols can also provide some emotional information, such as question marks and exclamation marks. In addition, some network symbols may also reflect emotions. However, BERT seems not sensitive to punctuation and other symbols according to Adam's[16] work and the information carried by sybmols is also not very evident in this dataset. Therefore, we decided to exclude special symbols from the whole process.

3.3. Manual Screening

Due to low quality of the dataset, some obvious problematic data were still discovered after preprocessing, which is as lised in Tabel 1.

Table 1
Question and Example

Question	Example
1) Few text	on test BLEU BP BLEU BP pair-CI 95% BLEU BP 3 01 03 32.98 0.92 33.03 0.93 [-0.23, +0.34] 33.6 . . .
2) Incomplete text	Ruge, 1992; Rapp, 2002)).
3) Improper partition	3.1 Part-of-Speech (POS) of NeighboringWordsWeuse 7 featurestoencode this knowledge source: a0a23, . . .
4) Wrong label	When tested on f-structures for all sentences from...,the techniques described in this paper improve BLEU score from 66.52 to 68.82.

For classification tasks, the accuracy of machine learning models depends on the quality of training data. Therefore, we attempts to maximize data quality through manual review and screening. For the above questions (1), (2) and (3), the original data will be directly deleted, and for the question (4), those sentences will be re-labeled. To be precise, there are around 134 sentences with wrong label and I re-labelled them all by by self. Finally, the compiled dataset consists of 7912 sentences, including 1237 positive, 347 negative, and 6328 neutral.

[2] https://cl.awaisathar.com/citation-sentiment-corpus/

4. Model Design

The idea of DictSentiBERT is to integrate the prior information of sentiment dictionary into the BERT, adjust attention mechanisms to better capture and understand emotional information of scientific citations, and achieve higher accuracy in the classification. As shown in Figure 2, the model adopts the following architecture, including input layer, BERT layer, modified attention layer, and output layer.

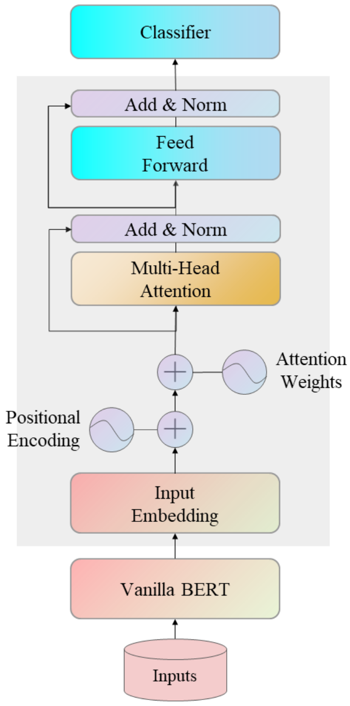

Figure 2: Structure of DictSentiBERT

4.1. Input Layer

In the input layer, coefficient for adjusting attention weights is calculated in advance by SentiWordNet and pos_tag. SentiWordNet is a dictionary used for sentiment analysis, which assigns an emotional intensity score to the three dimensions of positivity, negativity, and objectivity for each word in WordNet. However, the dictionary itself does not have the ability to handle polysemy, so we introduce NLTK's tagging tool: pos_tag. Firstly, we use BERT's tokenizer for word segmentation, converting the sentences into the standard form of BERT input. Next, we annotate each word and assign weight to it with SentiWordNet according to its part of speech. If there is only neutral intensity score, the weight is assigned to *1*. If there are negative or positive intensity, they are added together, and *1* is added to obtain the final score. For example, if the word "book" does not have polarity, then the weight is assigned to *1*. While the word "good" has a positive intensity of *0.5* and a negative intensity of 0, resulting in a final weight of *1.5*.

4.2. BERT Layer

The BERT layer consists of two main structures: embedding and encoder. The input vector is composed of three different embedding, namely wordpiece embedding, position embedding, and segment embedding. The encoder of a transformer consists of a multi-head attention layer, a regularization layer and a forward propagation layer. In the standard BERT model, there are 12 layers of encoders and the word vector dimension is 768. In this study we use the vanilla base-bert.

4.3. Modified Attention Layer

Due to the varying importance of vocabulary and feature weights in the text, attention mechanism is introduced to learn the dependency relationships between vocabulary and pay special attention to the important vocabulary. Therefore, the accuracy of classification can be further improved by assigning different weights to focus on important parts of the context and the specific calculation formula is as follows:

$$scores = QK^T + MASK \tag{1}$$

$$Attention(Q, K, V) = softmax(\frac{scores}{\sqrt{d_k}})V \tag{2}$$

In this step the obtained weights matrix is applied to the original attention matrix. DictSentiBERT processes the input sentences and calculates attention scores as follows:

$$scores = QK^T \odot W + MASK \tag{3}$$

$$Attention(Q, K, V) = softmax(\frac{scores}{\sqrt{d_k}})V \tag{4}$$

Where W is the obtained attention weight matrix and the process of $QK^T \odot S$ is as follows:

$$QK^T \odot W = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ & \ddots & \vdots \\ a_{n1} & & a_{nn} \end{bmatrix}_{n \times n} \odot \begin{bmatrix} w_1 & \dots & w_n \\ & \ddots & \vdots \\ w_1 & & w_n \end{bmatrix}_{n \times n} \tag{5}$$

As for a sequence $s_1 \dots s_n$, the wights of them is $w_1 \dots w_n$ calculated by SentiWordNet in the input layer. The other lines of wights matrix equal the first line.

4.4. Output Layer

The output layer is a fully connected layer that connects and transforms the outputs of the model, and uses the softmax function to calculate the probability score for each category. The final output is the label of input, including neutral, positive, and negative. Some examples are listed in the appendix.

5. Experiment

5.1. Baseline

Two basic pre-trained models: BERT and SCIBERT are used. On this basis, FeedForward NN (FNN), LSTM, TextCNN, Self-Attention and DictSensiBERT proposed in this paper are designed for experiments.

5.2. Arguments

The code was written with PyTorch v1.10 in Python v3.7. And the model was trained on a 16GB RTX A4000 for 50 epochs each with 80% training set and 20% test set. The batch size was set to 32, the learning rate was 5e-6. The AdamW optimizer with a warm-up rate of 0.1 and the cross-entropy loss function were used for optimization.

5.3. Results

Table 2
Experimental Result

Model	BERT		SCIBERT	
	Acc	F1	Acc	F1
FNN	93.05	80	95.14	**86**
LSTM	93.11	80	94.63	84
TextCNN	83.20	52	94.57	**86**
Attention	93.30	80	94.44	84
DictSentiBERT	**93.49**	**81**	**95.20**	**86**

As shown in Table 2, The average accuracy of native BERT is 91.23%, with an average Macro-F1 score of 75%. SCIBERT performs better, with an average accuracy of 94.80% and an average Macro-F1 score of 85%. This indicates that SCIBERT trained in scientific texts is more suitable for CSC. On the other hand, it can also be observed that under the same basic pre-trained model, the performance of DictSentiBERT has also been improved to a certain extent, which proves that the pre-trained model incorporating sentiment dictionary is more conducive to extracting emotional information.

6. Conclusion

This study proposes DictSentiBERT, which adjusts attention mechanism based on sentiment dictionary, and applies it to sentiments classification of scientific citation. We conducted research and organized a high-quality CSC dataset, designed the DictSentiBERT model and a series of baseline models for comparative experiments. Results indicate that pre-trained models can effectively classify sentiments of scientific citations, and SCIBERT performs better than native BERT on this task. Furthermore, DictSentiBERT can improve classification accuracy while maintaining high Macor-F1 score. In summary, this study provides a high-quality CSC dataset and a new model for the sentiments classification of scientific citations. However, this study still suffers from quantity and quality of dataset and a larger corpus is needed to make further improvement and experiment. In the future, we can try to imitate the training process of SCIBERT, collect large-scale scientific citation texts, and adjust the MASK mechanism to focus on emotional words of MLM tasks. Then, we can use the official tool set provided by Google to train BERT from scratch. Alternatively, we can try relying on syntax trees and other methods to focus on the characteristics of sentiment analysis from the perspectives of syntax, grammar, and morphology. Finally, the latest GPT large model can also be combined to use AIGC to modify and guide pre-trained models.

Acknowledgments

This research is supported by High-performance Computing Platform of Peking University. The work is also supported by "National Social Science Foundation of China" Big Data-Driven Research on Semantic Evaluation System of Scientific and Technological Literature (Grant No.21&ZD329).

References

[1] L. M. Baird, C. Oppenheim, Do citations matter?, Journal of information Science 20 (1994) 2–15.

[2] F. Radicchi, In science “there is no bad publicity”: Papers criticized in comments have high scientific impact, Scientific reports 2 (2012) 815.

[3] R. Piryani, D. Madhavi, V. K. Singh, Analytical mapping of opinion mining and sentiment analysis research during 2000–2015, Information Processing & Management 53 (2017) 122–150.

[4] A. Yousif, Z. Niu, J. K. Tarus, A. Ahmad, A survey on sentiment analysis of scientific citations, Artificial Intelligence Review 52 (2019) 1805–1838.

[5] H. Small, Interpreting maps of science using cita-

tion context sentiments: A preliminary investigation, Scientometrics 87 (2011) 373–388.
[6] A. Athar, Sentiment analysis of citations using sentence structure-based features, in: Proceedings of the ACL 2011 student session, 2011, pp. 81–87.
[7] S. Poria, I. Chaturvedi, E. Cambria, A. Hussain, Convolutional mkl based multimodal emotion recognition and sentiment analysis, in: 2016 IEEE 16th international conference on data mining (ICDM), IEEE, 2016, pp. 439–448.
[8] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, arXiv preprint arXiv:1903.10676 (2019).
[9] T. Xia, Y. Wang, Y. Tian, Y. Chang, Using prior knowledge to guide bert's attention in semantic textual matching tasks, in: Proceedings of the Web Conference 2021, 2021, pp. 2466–2475.
[10] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, P. Wang, K-bert: Enabling language representation with knowledge graph, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 2901–2908.
[11] J. Xu, Y. Zhang, Y. Wu, J. Wang, X. Dong, H. Xu, Citation sentiment analysis in clinical trial papers, in: AMIA annual symposium proceedings, volume 2015, American Medical Informatics Association, 2015, p. 1334.
[12] I. Budi, Y. Yaniasih, Understanding the meanings of citations using sentiment, role, and citation function classifications, Scientometrics 128 (2023) 735–759.
[13] Y. Yaniasih, I. Budi, Systematic design and evaluation of a citation function classification scheme in indonesian journals, Publications 9 (2021) 27.
[14] A. Cohan, W. Ammar, M. Van Zuylen, F. Cady, Structural scaffolds for citation intent classification in scientific publications, arXiv preprint arXiv:1904.01608 (2019).
[15] D. Mercier, S. T. R. Rizvi, V. Rajashekar, A. Dengel, S. Ahmed, Impactcite: an xlnet-based method for citation impact analysis, arXiv preprint arXiv:2005.06611 (2020).
[16] A. Ek, J.-P. Bernardy, S. Chatzikyriakidis, How does punctuation affect neural models in natural language inference, in: Proceedings of the Probability and Meaning Conference (PaM 2020), 2020, pp. 109–116.

A. Output of DictSentiBERT

1) Input: The resulting net increase in ATF4 and CHOP is significantly less than that observed with a bona fide ER stress inducer, such as TG.
   Output: 0 (Negative)
2) Input: While this method is known to be generally reliable, there are some questions about the representativeness of the data used.
   Output: 0 (Negative)
3) Input: Translation performance was measured using the BLEU score, which measures n-gram overlap with a reference translation.
   Output: 1 (Neutral)
4) Input: A totally different approach uses the idea of self-training described in the paper.
   Output: 1 (Neutral)
5) Input: This is an important feature from the MT viewpoint, since the decomposition into translation model and language model proved to be extremely useful in statistical MT.
   Output: 2 (Positive)
6) Input: From a strategic viewpoint, layered modular architectures have the competitive advantage, as well as the challenge, in being doubly distributed.
   Output: 2 (Positive)