DSI DATA SCIENCE AND INFORMETRICS

RESEARCH ARTICLE

# A comparative experimental study of citation sentiment identification based on the Athar-Corpus

Xinyue Wang, Danqun Zhao*

Department of Information Management, Peking University, Beijing, China

## ABSTRACT

A series of comparative experiments on citation sentiment identification/analysis (CSI/CSA) are carried out based on comparing and integrating sentiment lexicon and machine learning methods in our paper, and a fusion of the two methods is also explored. We design four groups of comparative experiments for the key steps in current CSI/CSA research, involving sentiment lexicon expansion, text feature extraction, data resampling and method fusion in order to find out combinations of methods with better identification effects. Our experiments' details are as follows: using open citation corpus founded by Athar; selecting SentiWordNet and SO-PMI as original sentiment lexicon and its expansion method; choosing TF-IDF, Word2Vec, BERT for text feature extraction and "SMOTE+Undersampling" as main method for data resampling. Nine frequently-used machine learning algorithms(models), including support vector machine, random forest, decision tree, linear classification, AdaBoost, extremely randomized trees, stochastic gradient descent, long short-term memory network and convolutional neural network, are finally used in our comparative experiments. The experimental results and main findings include: ① The extended sentiment lexicon by SO-PMI is better than the original one for CSI/CSA; ② As a simple method for text feature extraction, TF-IDF is generally better than Word2Vec and BERT; ③ The use of "SMOTE+Undersampling" can better solve data imbalance problem in Athar-corpus; ④ The integration of sentiment lexicon and machine learning can improve the effect of CSI/CSA, specifically shown in their higher index values both of accuracy and Macro-F1.

## KEYWORDS

Citation Sentiment Identification/Analysis (CSI/CSA); Sentiment Lexicon; Machine Learning; SMOTE

* Corresponding Author: zdq@pku.edu.cn

# 1 Introduction

Citation Sentiment Identification/Analysis (CSI/CSA) aims to identify from citation corpus of academic papers their attitudes, personal opinions, or sentimental tendencies expressed by the authors when citing or referring to other documents (Yousif et al., 2019). In the era of full-text metrics, citation sentiment identification has become an important research topic in Citation Content/Context Analysis (CCA). On one hand, it relies on the technological advancement of upstream tasks (e.g., automatic extraction of citation sentences and their contexts, construction of sentiment lexicons, etc.); On the other hand, it also strongly supports the solution of downstream tasks (e.g., academic evaluation, mapping knowledge domain, etc.). Additionally, it intertwines closely with some midstream tasks, such as citation motivation/function identification, citation topic analysis and automatic citation abstracting, etc. Given its important position among CCA and generally euphemistic and implicit sentiment expressions in academic discourse, citation sentiment identification is often considered to be a challenging task with difficulty.

Currently, main approaches employed in CSI/CSA can be broadly categorized into two types: sentiment lexicon and machine learning (including deep learning). The former related studies include Ikram et al. (2018) and Dehdarirad & Yaghtin (2022), who leveraged the SentiWordNet lexicon to discern citation sentiment among computational linguistics literature. Goodarzi et al. (2014) further extended this approach by integrating SentiWordNet with AFINN and Bingliu in biomedical domain texts. Since existing sentiment lexicons are all general-purposed, researchers have advocated for their extension using domain-specific academic literature. For example, Hassan et al. (2020) expanded SentiStrength with about 80 positive and negative words; Zuo et al. (2022) expanded Opinion Finder by SO-PMI algorithm and used it in the field of computational linguistics.

Subsequently, the latter (i.e. machine learning models) have gradually become the mainstream of CSI/CSA research and have achieved higher accuracy. Commonly used traditional models include support vector machine (SVM) (Athar, 2011), linear regression (LR) (Abu-Jbara et al., 2013), native bayes (NB) (Sula & Miller, 2014), random forest (RF) (Raza et al., 2019), decision tree (DT) (Ghosh et al., 2017), and k-nearest neighbor (KNN) (Mehmood et al., 2019). To obtain better results, researchers have conducted comparative experiments for these traditional models with the same feature inputs. Many experiment results from Abu-jbara et al. (2013), Muppidi et al. (2021), and Amjad & Ihsan (2020) found that SVM outperforms LR, NB, RF, and DT.

In recent years, some deep learning models, such as convolutional neural networks (CNN) and recurrent neural networks (RNN), have begun to be used for CSI/CSA task, especially CNN and RNN-based long short-term memory (LSTM). Some experiment results showed that LSTM have higher classification accuracy than SVM (Munkhdalai et al., 2016). Subsequently, CNN have also been shown to possess better results than SVM (Lauscher et al., 2017). Up to now, most studies concluded that deep learning models outperform traditional models on CSI/CSA tasks.

Currently, researchers are gaining a more nuanced understanding of the strengths, weaknesses, and complementary nature of various approaches: sentiment lexicon method relies too much on the quality and scale of lexicons, while traditional machine learning and deep learning often struggle with the need for large-scale annotated corpus. Given these insights, integrating different methods to mitigate their own limitations has emerged as a much-talked-about research. This paper aims to delve into this hot topic through a series of comparative experiments. Our objectives focus on the following key questions: ① Can the expansion of sentiment lexicon improve the accuracy of

CSI/CSA? ② Which text feature extraction method is more applicable in CSI/CSA? ③ Which resampling method can better solve the imbalance of Athar-Corpus? ④ How to fuse sentiment lexicon and machine learning method? How does the fusion method improve or enhance our experiment effect of CSI/CSA?

## 2 Experimental Design

This paper intends to use the corpus created by Athar containing 8736 citation sentiment data (hereinafter referred to as Athar-Corpus) (Athar, 2011) to carry out four groups of comparative experiments, which correspond to the above four key aspects, including sentiment lexicon expansion, text feature extraction, data resampling, and method fusion. The overall experimental design is shown in Figure 1, and four groups of comparative experiments are briefly described as follows.

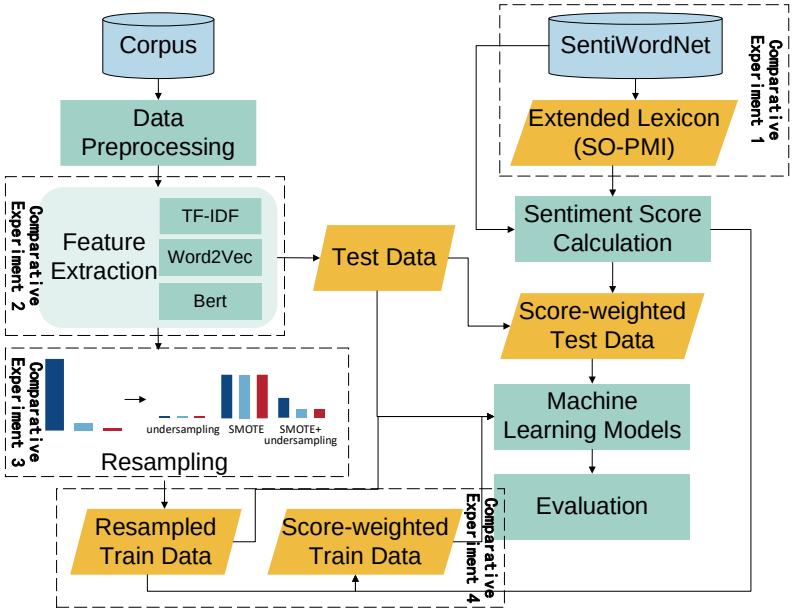



**Figure 1** Diagram of overall experimental design in our paper

### 2.1 Comparative Experiment 1: CSI/CSA Based on the Original Sentiment Lexicon and its Extended Version

Firstly, SentiWordNet, a general-purpose sentiment lexicon, was initially selected. However, given that Athar-Corpus comprises scientific paper data from the field of computational linguistics (CL), there exist inherent disparities in word usage between the two. Therefore, experiment 1 aims to extend SentiWordNet and create an expanded sentiment lexicon tailored specifically for Athar-Corpus. This expansion is achieved utilizing the Semantic Orientation Pointwise Mutual Information (SO-PMI) method. The steps involved in constructing this expanded lexicon are as follows: first, the sentiment scores provided in the original SentiWordNet lexicon are utilized to sort all words in ascending order based on these scores. From this sorted list, 500 positive and 500 negative sentiment seed words with the highest sentiment intensity are extracted. Next, Athar-Corpus undergoes data preprocessing, during which the frequency of occurrence of each word is calculated.

Words with high frequencies are then identified as candidate words. Using Equation 1, the co-occurrence intensity between these candidate words and the positive/negative sentiment seed words is calculated. Subsequently, Equation 2 is applied to determine the difference in co-occurrence intensity between the candidate word and positive/negative sentiment seed words. If the difference favors a higher co-occurrence intensity with positive sentiment seed words, the candidate word is classified as a positive sentiment word, and vice versa. In equation 1 and 2, *pos* and *neg* denote positive and negative sentiment seed words, and *word* denotes a candidate word. Finally, 2702 candidate words with high sentiment scores are filtered and added to the lexicon.

$$PMI(word1, word2) = \log_2\left(\frac{P(word1, word2)}{P(word1)P(word2)}\right) \quad \text{(Equation 1)}$$

$$SO - PMI(word) = \sum_{i=1}^{num(pos)} PMI(word, pos_i) - \sum_{i=1}^{num(neg)} PMI(word, neg_i) \quad \text{(Equation 2)}$$

Comparative experiment 1 focuses on CSI/CSA based on the original sentiment lexicon and an extended lexicon and is used to determine whether the accuracy can be improved based on the latter. This group of experiments is a separate application of the sentiment lexicon method.

## 2.2 Comparative Experiment 2: CSI/CSA Based on Different Text Feature Extraction Methods

Text feature extraction (or feature selection) is a fundamental operation to carry out various natural language processing (NLP) tasks based on machine learning methods. The extraction methods are mainly divided into two categories: based on statistics and based on language models (see Table 1 for details), of which TF-IDF and Word2Vec are the most used methods, and have achieved great research results. In 2018, proposed by Google's BERT has an excellent performance in various NLP tasks. Therefore, comparative experiment 2 will mainly select TF-IDF, Word2Vec, and BERT to carry out CSI/CSA. This group of experiments is a separate application of machine learning methods, and by comparing their sentiment identification effects on nine different machine learning models (see Table 2 for details), the better performer will be selected as the feature extraction method for the subsequent experiments.

**Table 1** Main text feature extraction methods

| Category | Name | Description |
|---|---|---|
| Based on statistics | TF-IDF | The product of TF (term frequency) and IDF (inverse document frequency). The more frequently a word appears in a document and the fewer documents that contain it, the more important the word is. |
| | Glove | An unsupervised learning algorithm, the main idea of which is to realize the vectorized representation of words through the co-occurrence statistical information of words. |
| | BOW | For a collection of text, it is viewed only as a collection of words, with the frequency of word occurrences as the representation of the words. |
| Based on language models | Word2Vec | A neural network is trained and the parameters in the neural network are extracted as word vectors of the words. |

| | |
|---|---|
| | Training methods for neural networks include CBOW and skip-gram. |
| BERT | Taking the original text as input, the system extracts features and outputs a sequence of vectors to realize the vectorization of the text and consider the contextual content of the text. |
| FastText | Similar to Word2Vec, it is also a vectorization method by training a neural network and using the parameters of the neural network as word vectors. |

## 2.3 Comparative Experiment 3: CSI/CSA Based on Different Data Resampling Methods

The crucial difference between citation corpus and other text corpus lies in the fact that citation sentiment is predominantly neutral, resulting in a dearth of explicit sentiment tendencies. This leads to a significant imbalance in the number of data points across three sentiment classes: positive, negative, and neutral. This imbalance can adversely impact the final results of CSI/CSA. The data imbalance problem is generally solved by resampling, which includes undersampling and oversampling. Undersampling refers to deleting some samples from the majority class (citation neutral class); and oversampling refers to adding more samples to the minority class (citation positive and negative classes). Comparative experiment 3 proposes to use the synthetic minority oversampling technique (SMOTE) proposed by Chawla et al. (2002) to mitigate the effect of citation data imbalance. As an optimization of the random oversampling method, SMOTE selects, for each minority class sample x, a random sample y from its k-nearest neighbors and synthesizes a new sample at a randomly selected point on the (x, y) concatenation. This approach reduces the risk of overfitting. In addition, this paper also considers combining the SMOTE method with the undersampling method: using SMOTE for the minority class samples to increase the sample size and at the same time using undersampling for the majority class samples to reduce the sample size, with the expectation of achieving better experimental results.

Eventually, comparative experiment 3 will conduct CSI/CSA experiments based on three different resampling methods including undersampling, SMOTE, and "SMOTE+Undersampling", and compare them with the (unbalanced) original dataset to find out a better way to solve the data imbalance problem.

It should be emphasized that the above resampling treatment of the corpus is limited to the training set and does not involve the test set in order to keep its composition and sample distribution in line with the real situation. Since the data imbalance in the citation corpus is very serious, if SMOTE is performed on the whole dataset first, it will lead to the fact that most of the data in the test set is generated at a later stage, which obviously deviates from the real situation. If not avoided in the experiment, the final sentiment identification accuracy will be greatly improved, but the research's reference value and significance would be considerably diminished.

## 2.4 Comparative Experiment 4: CSI/CSA Based on Multiple Machine Learning Methods and Fusion Methods

As mentioned earlier, there are advantages and disadvantages of conducting CSI/CSA using sentiment lexicon methods and machine learning methods alone, and there are certain complementarities between the two types of methods. For this reason, in comparative experiment

4, this paper integrates the two types of methods and compares its experimental results with those of using machine learning methods alone. Referring to the statistical results for sentiment identification algorithms in the CSI/CSA (Wang & Zhao, 2024), four traditional machine learning models (SVM, LR, DT and SGD), two widely-used deep learning models (LSTM and CNN) and three ensemble models (AB, RF and ET) with better results are selected in our experiment. Table 2 provides a summary description of these nine models.

**Table 2** Nine frequently-used machine learning models

| Name | Description |
|---|---|
| SVM(Support Vector Machine) | A generalized linear classifier whose decision boundary is a maximal margin hyperplane solved for learning samples. |
| DT(Decision Tree) | A commonly used categorical regression method that forms a binary tree by calculating entropy with a high level of model comprehensibility. |
| LR(Logistic Regression) | A method of statistical analysis in which a linear regression model is mapped to a discrete domain to determine the relationship between a sample's attributes and its categories. |
| RF(Random Forest) | A classifier that utilizes multiple decision tree models to train and predict samples, where the output categories are determined by the plurality of the individual tree output categories, is an integrative algorithm. |
| AB(AdaBoost) | The core idea is to train different weak classifiers for the same training set and aggregate them to form a stronger final classifier. |
| ET(Extra Trees) | A variant of Random Forest, which uses the original training set for each decision tree and randomly selects the eigenvalues to divide the decision tree, has an improved generalization capability. |
| SGD(Stochastic Gradient Descent) | Each iteration randomly draws a sample from the training set so that each iteration moves toward the overall optimum. |
| LSTM(Long Short-Term Memory) | An optimization of RNN aimed at solving the long-term dependency problem it suffers from. |
| CNN(Convolutional Neural Network) | Feedforward Neural Networks containing convolutional computation with deep structure |

The fusion method for citation sentiment identification used in this paper is described as follows. First, sentiment scores corresponding to sentiment words are obtained from the extended sentiment lexicon. Since there may be multiple lexemes for the same word in the lexicon, and multiple lexical meanings under the same lexeme, all scores of the word within a particular lexeme are weighted (with higher weights assigned to lexical meanings of higher order), and sentiment scores of sentiment words in the test text are calculated. See Equation 3 for specific calculations, where $pos(j)$、$neg(j)$ refer to the positive and negative sentiment scores of the $j_{th}$ lexical sense of the $i_{th}$ word. Aggregate the sentiment score $senti_i$ of each word to get the sentiment score

matrix $senti$.

$$senti_i = \frac{\sum_{j=1}^{n} \frac{pos(j)-neg(j)}{j+1}}{\sum_{j=1}^{n} \frac{1}{j+1}} \quad \text{(Equation 3)}$$

Subsequently, the feature extraction method is used to obtain the numerical type vector $w$ of the test text, and $senti$ is used as the weight to weight $w$ to obtain the final feature vector $feature = senti \times w$.

Finally, $feature$ will be used as input into the machine learning model for training and testing to realize the integration of sentiment lexicon and machine learning methods.

Comparative experiment 4 will use $feature$ and $w$ as model inputs respectively, to see if the fusion method improves over machine learning alone.

# 3 Experimental Data and Related Tools

## 3.1 Dataset Selection

Experimental datasets for CSI/CSA include two categories: self-constructed sets and open or public corpus. Self-constructed sets generally have a smaller amount of data, but have a higher fit with the research domain, and are more suitable for sentiment lexicon methods that do not need to annotate the data. Public corpus containing citation sentiment annotation information are larger in size and higher in quality and are more suitable for a variety of supervised machine learning methods. Table 3 summarizes the basic situation of five commonly used public corpus.

In this paper, we choose Athar-Corpus, which is the largest and most widely used public corpus with 8736 instances, and its data comes from the CL domain, including source_paper, target_paper, sentiment, and citation_text, with sentiment labeling data, which is convenient for various supervised machine learning methods. The corpus consists of three sentiment categories, positive, negative, and neutral, and has a high degree of imbalance in the distribution of data for each category, with numbers of 829, 280, and 7627, respectively, as shown in Table 4. Table 5 details examples of each categorization in the dataset.

**Table 3** Five Public Corpus often-used in CSI/CSA

| Author | Year | Number of instances | Citation Context | Subject | Percentage of subjective emotions |
|---|---|---|---|---|---|
| Athar (2011) | 2011 | 8736 | No | CL | 13% |
| Abu-jbara et al. (2013) | 2013 | 3568 | Yes | CL | 42% |
| Dong & Schäfer (2011) | 2011 | 1768 | No | CL | 14% |
| Xu et al. (2015) | 2015 | 4182 | Yes | Biomedical | 24% |
| Jochim & Schütze (2012) | 2012 | 2008 | No | CL | 100% |

**Table 4** Distribution of Sentiment in Athar-Corpus dataset

| Citation Sentiment | Number of citation sentences |
|---|---|
| Positive(p) | 829 |
| Negative(n) | 280 |
| Neutral(o) | 7627 |

Table 5 Examples of data from Athar-Corpus dataset

| source_paper | target_paper | sentiment | citation_text |
|---|---|---|---|
| I05-2009 | A00-2024 | O | 5.3 Related works and discussion Our two-step model essentially belongs to the same category as the works of (Mani et al., 1999) and (Jing and McKeown, 2000). |
| J02-4005 | A00-2024 | P | But in fact, the issue of editing in text summarization has usually been neglected, notable exceptions being the works by Jing and McKeown (2000) and Mani, Gates, and Bloedorn (1999). |
| J02-4005 | A00-2024 | N | Jing and McKeown (2000) have proposed a rule-based algorithm for sentence combination, but no results have been reported. |

## 3.2 Data Preprocessing

To enhance the accuracy of sentiment identification, certain preprocessing steps are required after acquiring citation data, specifically including word segmentation, de-duplication, de-specialization of characters and numbers, stemming, lemmatization and Part-Of-Speech tagging. Since the corpus is in English, word segmentation is straightforward and can be achieved using spaces as separators. For frequently occurring but meaningless words in the citation corpus, the NLTK corpus's stop word list is primarily employed for removal, and regular expressions are used to eliminate special characters and numbers from the corpus. NLTK is then used for stemming (removing affixes to get roots) and lemmatization to restore words to their most basic form and eliminate the effects of word morphology. In addition, considering that the affective lexicon method requires word lexicality labeling, NLTK's pos_tag function is utilized to categorize words into nouns, verbs, adjectives, and adverbs.

## 3.3 Algorithm Parameters

According to the algorithm characteristics and tuning experience of the nine classification models, the specific parameter settings are shown in Table 6.

Table 6 Main parameters for nine classification models

| Name | Parameters |
|---|---|
| SVM | 'C': 10, 'gamma': 10, 'kernel': 'rbf', 'max_iter': 1000000 |
| DT | criterion:'entropy',max_depth=9,max_features=7,min_samples_split=4 |
| LR | penalty='l2', solver='liblinear', tol=0.0001, C=1.0 |
| RF | n_estimators=10,criterion='gini', min_samples_split=2 |
| AB | base_estimator=DecisionTreeClassifier, n_estimators=50, learning_rate=1.0, algorithm='SAMME.R' |
| ET | criterion='gini', min_samples_split=2 |
| SGD | loss='hinge', penalty='l2', alpha=0.0001, tol=0.001 |
| LSTM | loss="categorical_crossentropy",optimizer="adam",batch_size=256, epoch=50 |
| CNN | loss="categorical_crossentropy",optimizer="adam",batch_size=32, filters=64, |

| kernel_size1=5, kernel_size2=3, pool_size=3, epoch=50 |
|---|

Note: SVM = Support Vector Machine, DT = Decision Tree, LR = Logistic Regression, RF = Random Forest, AB = AdaBoost, ET = Extra Trees, SGD = Stochastic Gradient Descent, LSTM = Long Short-Term Memory, CNN = Convolutional Neural Network

## 3.4 Evaluation Metrics

The experimental results are often evaluated by using Macro-F1, which is the reconciled mean value of Precision and Recall, and all sentiment categories are treated equally. Due to the high data imbalance of Athar-corpus, this index is used to better evaluate the actual effectiveness of classification models. The calculation of Macro-F1 is shown in Equation 4-Equation 7, where $senti_{right_i}$ is the number of quotations correctly assigned to category $c_i$, $senti_{wrong_i}$ is the number of quotations incorrectly assigned to category $c_i$ from other categories, and $senti_{all_i}$ is the number of quotations actually contained in category $c_i$.

$$precision_i = \frac{senti_{right_i}}{senti_{right_i}+senti_{wrong_i}} \times 100\% \quad \text{(Equation 4)}$$

$$recall_i = \frac{senti_{right_i}}{senti_{all_i}} \times 100\% \quad \text{(Equation 5)}$$

$$F1_i = \frac{2\times precision_i \times recall_i}{precision_i + recall_i} \quad \text{(Equation 6)}$$

$$Macro - F1 = \frac{\sum_{i=1}^{senti} F1_i}{senti} \quad \text{(Equation 7)}$$

## 3.5 Sentiment Lexicon Selection

The sentiment lexicons that are widely used in CSI/CSA studies are SentiWordNet (Esuli & Sebastiani, 2006), Opinion Finder (Wilson et al., 2005), and SentiStrength (Thelwall et al., 2010). Among them, SentiWordNet is the sentiment annotation of WordNet 3.0, which contains more than 110,000 words, such as adjectives (a), adverbs (r), verbs (v), and nouns (n). The lexicon data include word lexemes, IDs, positive sentiment scores, negative sentiment scores, the word text, and descriptive information, which is the most widely used in citation sentiment identification. So we choose to use SentiWordNet as the original sentiment lexicon.

Despite its extensive vocabulary and noteworthy success in citation sentiment identification, SentiWordNet still exhibits certain limitations: (1) as a general lexicon, it may differ from the specific domains encompassed in Athar-Corpus, resulting in the exclusion of domain-specific common words or the misrepresentation of their particular meanings; (2) words in the lexicon often possess multiple meanings, and these different meanings may yield diverse sentimental connotations; (3) while SentiWordNet assigns a sentiment score to each word, it lacks the capacity to consider contextual information, potentially impacting the expression of the sentiment of the word.

For limitation 1, the SO-PMI method detailed in Section 2.1 is used to account for the frequently occurring words in Athar-Corpus and form an extended lexicon, which is more suitable for the classification task of this corpus. For limitation 2, the weighting method described in Section 2.4 is used to consider different lexemes of a word and different lexical meanings of the same lexeme. For limitation 3, the contextual context information in the text is considered in the experiments by adopting the LSTM model. By implementing these solutions, the potential biases of SentiWordNet can be effectively mitigated, leading to improved accuracy in citation sentiment identification.

When using sentiment lexicons for citation sentiment identification, the two main factors of degree adverbs and negative words also need to be considered. In this paper, HowNet (Dong & Dong, 2001) English adverbial lexicon is chosen, which contains 170 adverbs of degree, and is divided into "extremely/most", "very", "more", "super", " slightly ", and "lack". The corresponding degree coefficients are 2, 1.75, 1.5, 1.25, 0.8, and 0.5. The negative word list is constructed by this paper, which mainly contains 17 common negative words such as "no", "not", "never", etc.

# 4 Experimental results and discussion

Based on the aforementioned experimental design, this paper completed 4 sets of comparative experiments in turn.

## 4.1 Comparative Experiment 1: its Results and Discussion

Utilizing the SO-PMI method for sentiment lexicon expansion resulted in an augmentation of 2702 additional CL domain sentiment words (1344 positive and 1358 negative). Combining it with the original lexicon SentiWordNet yields an extended lexicon that fits better with Athar-Corpus, containing about 120,000 words in total. Comparative experiment 1 was conducted on Athar-Corpus using the original lexicon SentiWordNet and its extended version. The results of the experiment revealed a 17% improvement in the Macro-F1 value of 0.35 for the former and 0.41 for the latter (see Figure 2). The segmentation recognition results for three different sentiment polarities (positive p, negative n, and neutral o) are shown in Table 7.

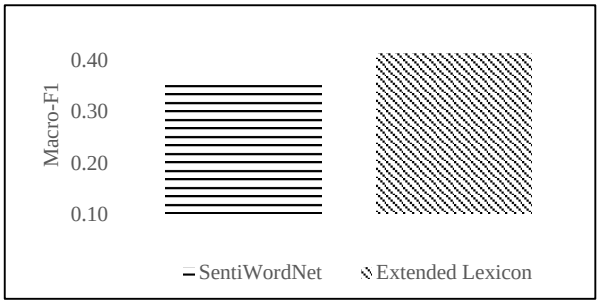



**Figure 2** Results from our Comparative Experiment 1 (Macro-F1 value)

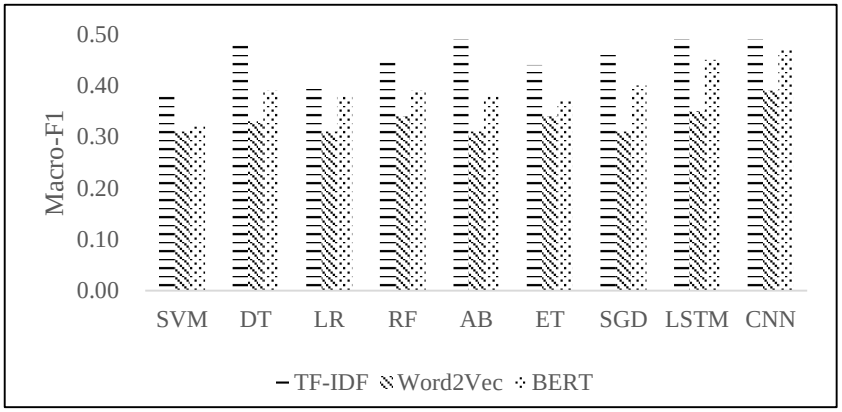
**Table 7** F1 values for different sentiment polarity from Comparative Experiment 1

| Sentiment Lexicon | F1-p | F1-o | F1-n |
|---|---|---|---|
| SentiWordNet | 0.0813 | 0.9223 | 0.0466 |
| Extended Lexicon | 0.2331 | 0.8638 | 0.1401 |

As can be seen in Table 7, the extended lexicon has a greater improvement in both positive and negative sentiment recognition effects than the original sentiment lexicon, and the overall Macro-F1 value is also improved. So it can be considered that the extended lexicon is more compatible with the citation corpus, and it is more suitable for citation sentiment identification task.

## 4.2 Comparative Experiment 2: its Results and Discussion

Comparative experiment 2 uses TF-IDF, Word2Vec, and BERT to realize text feature extraction, completes CSI/CSA on nine machine learning models, and the specific experimental results are shown in Figure 3.



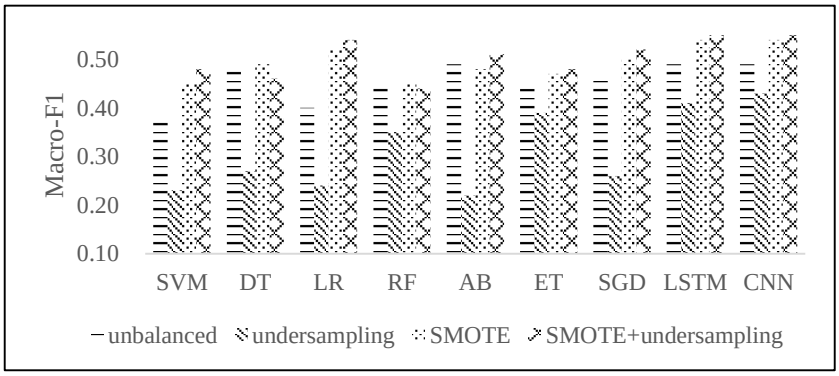

Note: SVM = Support Vector Machine, DT = Decision Tree, LR = Logistic Regression, RF = Random Forest, AB = AdaBoost, ET = Extra Trees, SGD = Stochastic Gradient Descent, LSTM = Long Short-Term Memory, CNN = Convolutional Neural Network

**Figure 3** Results from our Comparative Experiment 2 (Macro-F1 value)

As seen in Figure 3, both Word2Vec and BERT methods work best in the CNN model with Macro-F1 values of 0.39 and 0.47. However, the TF-IDF method outperforms Word2Vec and BERT on all nine models and achieves the highest Macro-F1 value of 0.49 on the AB, LSTM and CNN models, which is significantly higher than the remaining two methods. This experimental result shows that TF-IDF achieves better classification results in citation sentiment identification task. Cunha et al. (2021) also found from comparative experiments of various types of feature extraction methods that the traditional TF-IDF results are better than BERT on smaller datasets (less than 100,000 data), and the results of the present experiments are in line with them. Therefore, the latter 2 sets of comparative experiments choose the TF-IDF method for text feature extraction.

## 4.3 Comparative Experiment 3: its Results and Discussion

Comparative experiment 3 was designed to solve the data imbalance problem of the corpus, and three main strategies were adopted, namely, undersampling, SMOTE, and "SMOTE+undersampling". Usually, undersampling results in a significant loss of data content, while SMOTE may introduce noisy elements when dealing with highly imbalanced datasets. The combined approach of "SMOTE+undersampling" aims to increase the representation of the minority class while reducing the excess of the majority class. In the specific experimental process, three resampled datasets and the original (unbalanced) dataset are used as inputs for nine machine learning models, text feature extraction is unified using TF-IDF, and the final experimental results are shown in Figure 4.



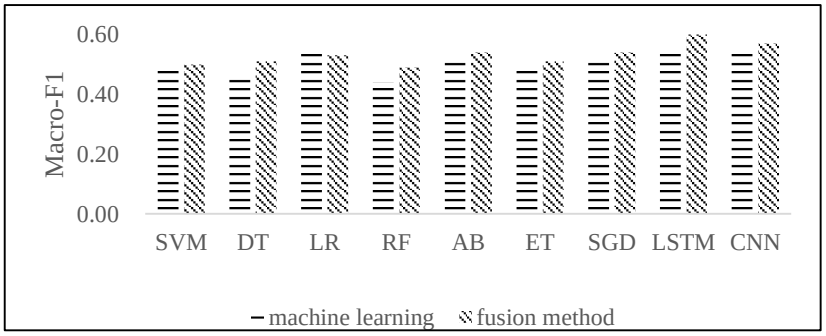
Note: SVM = Support Vector Machine, DT = Decision Tree, LR = Logistic Regression, RF = Random Forest, AB = AdaBoost, ET = Extra Trees, SGD = Stochastic Gradient Descent, LSTM = Long Short-Term Memory, CNN = Convolutional Neural Network

**Figure 4** Results from our Comparative Experiment 3 (Macro-F1 value)

As can be seen in Figure 4, the undersampling method is the least effective on all machine learning models, followed by the original unbalanced dataset, the SMOTE method improves the effect compared to the former two, while the fusion method of "SMOTE+undersampling" is the most effective, obtaining the highest Macro-F1 score of 0.55 on the LSTM and CNN model. This is in line with the experimental findings of Chawla et al. (2002). In contrast, in comparative experiment 4, "SMOTE+undersampling" will be used as the solution strategy for the data imbalance problem.

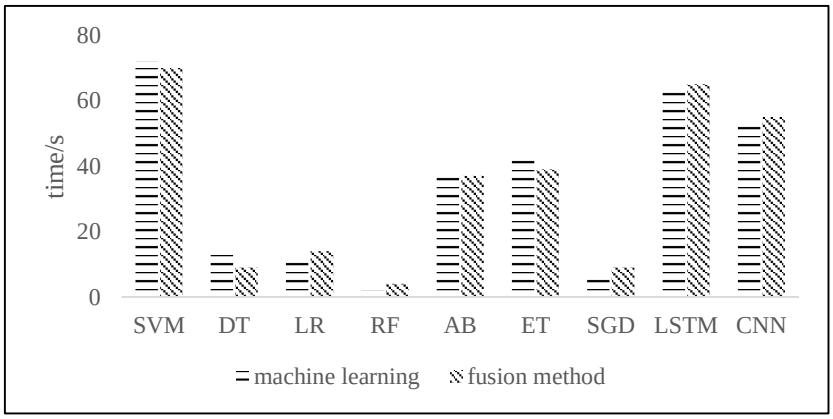
## 4.4 Comparative Experiment 4: its Results and Discussion

The results of three comparative experiments show that better CSI/CSA results will be obtained based on the extended sentiment lexicon, selecting TF-IDF for text feature extraction and using "SMOTE+undersampling" to solve the data imbalance problem. Therefore, comparative experiment 4 chooses to use sentiment scores obtained by the extended sentiment lexicon method to weight the text vectors obtained by TF-IDF, realizing the fusion of the sentiment lexicon method and the machine learning method, and the rest of the experimental strategies remain unchanged. Finally, the results of the fusion method are compared with those of the machine learning method, as detailed in Figure 5. A comparison of the time complexity of the algorithms is shown in Figure 6.




Note: SVM = Support Vector Machine, DT = Decision Tree, LR = Logistic Regression, RF = Random Forest, AB = AdaBoost, ET = Extra Trees, SGD = Stochastic Gradient Descent, LSTM = Long Short-Term Memory, CNN = Convolutional Neural Network

**Figure 5** Results from our Comparative Experiment 4 (Macro-F1 value)




Note: SVM = Support Vector Machine, DT = Decision Tree, LR = Logistic Regression, RF = Random Forest, AB = AdaBoost, ET = Extra Trees, SGD = Stochastic Gradient Descent, LSTM = Long Short-Term Memory, CNN = Convolutional Neural Network

**Figure 6** Comparison of average time complexity of algorithms (running 50 times)

As seen in Figure 5, the experimental results of our fusion method outperform most pure machine learning models (except LR), and the highest Macro-F1 score reaches 0.60 (LSTM model). The difference between the fusion and machine learning methods is small, this is due to the fact that our fusion strategy does not affect the data size too much. Regarding computational complexity, the LSTM model falls somewhere in the middle compared to other models. While its complexity is lower than SVM, it is higher than traditional machine learning models due to the intricate nature of neural network architectures. However, the computational time for LSTM remains within an acceptable range, indicating that the fusion method, coupled with LSTM, provides a good balance between performance and efficiency.

The four sets of comparative experiments and experimental results preliminarily show that the integration of sentiment lexicon and machine learning methods is not only feasible in CSI/CSA, but also shows some effectiveness in the improvement and enhancement of the identification effect.

## 5 Conclusion

By comparing and integrating existing CSI/CSA methods, this paper carries out a series of comparative experiments for seven traditional machine learning models and two deep learning models (LSTM and CNN) using Athar-Corpus and tries to use the BERT method in text feature extraction. The main experimental findings obtained are as follows: ① In the comparative experiment of sentiment lexicon, the extended lexicon exhibited superior performance compared to the original lexicon; ② In the comparative experiment of text feature extraction methods, TF-IDF outperformed Word2Vec and BERT; ③ In the comparative experiment of data resampling, the "SMOTE+undersampling " strategy can get better results than using undersampling or SMOTE alone; ④ In the fusion comparative experiment, fusion strategy can further optimize the identification effect compared with using machine learning methods alone. In short, our work improves the accuracy of citation sentiment recognition compared with previous studies, which helps to analyze the connotation of citation context in depth and has an auxiliary effect for academic evaluation. It also makes the construction of a large-scale citation context corpus possible, which can break through the bottleneck of CSI/CSA and accelerate its development.

In the process of literature research and experimentation, this paper also found that existing studies pay less attention to the data imbalance problem of citation corpus. Only a few papers adopted the SMOTE method, most of them resampled the entire dataset prior to splitting, thereby altering the authenticity of the test set, and although a high accuracy was eventually obtained, their experimental results are of little reference significance. In addition, current research rarely considers the integration of different methods. The study by Ghosh & Shah (2020) stands out as it considers data resampling of the training set and incorporates sentiment lexicon features into the final machine learning classification. However, it solely utilizes SMOTE without undersampling, and its final feature set is predominantly influenced by the frequency of sentiment words, overlooking the significance of sentiment scores. In terms of sentiment identification (classification) results, the overall accuracy of this paper is 89%, with a Macro-F1 value of 0.60, both of which are better than those of the study by Ghosh (80.61% and 0.52, respectively). This further shows the effectiveness and application value of the citation sentiment fusion identification method proposed in this paper.

The current study still has deficiencies, mainly in the quality of the corpus. Although Athar-Corpus is the most widely used public dataset, it contains only citation sentences that may lead to incomplete expression of citation sentiment without taking into account their contextual information. Moreover, it also has a relatively low percentage of subjective sentiment records (i.e. data imbalance). Although the data imbalance problem is dealt with by resampling in our study, the inadequacy of the dataset itself in terms of citation context collection is difficult to be solved in a short period of time. Our team is planning to construct a more complete high-quality citation content dataset to provide a better data foundation for subsequent empirical research. In addition, some AIGC tools (e.g. ChatGPTs) can be selected to generate new positive/negative data that helps to alleviate corpus' imbalance, but the detailed method still needs to be explored.

We will focus on the following two aspects to deepen our follow-up exploration: first, to build a high-quality citation corpus(including both citation sentences and their contextual information) in the field of Bibliometrics, thus to break through the constraints caused by the current corpus quality (low) and size (small) on data-driven research such as CSI/CSA; second, to introduce more deep learning models(such as RNN), and to carry out more diversified fusion experiments(expanding from the feature layer fusion used in this study to the decision layer fusion) in order for providing a better support oriented to downstream tasks based on CSA/CSI.

## References

Abu-Jbara, A., Ezra, J., & Radev, D. (2013). Purpose and Polarity of Citation: Towards NLP-based Bibliometrics. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 596–606. https://aclanthology.org/N13-1067

Amjad, Z., & Ihsan, I. (2020). VerbNet based citation sentiment class assignment using machine learning. *International Journal of Advanced Computer Science and Applications, 11*(9), 621–627. Scopus. https://doi.org/10.14569/IJACSA.2020.0110973

Athar, A. (2011). Sentiment Analysis of Citations using Sentence Structure-Based Features. *Proceedings of the ACL 2011 Student Session*, 81–87. https://aclanthology.org/P11-3015

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Cunha, W., Mangaravite, V., Gomes, C., Canuto, S., Resende, E., Nascimento, C., Viegas, F., França, C., Martins, W. S., Almeida, J. M., Rosa, T., Rocha, L., & Gonçalves, M. A. (2021). On the cost-effectiveness of neural and non-neural approaches and representations for text classification: A comprehensive comparative study. *Information Processing & Management, 58*(3), 102481. https://doi.org/10.1016/j.ipm.2020.102481

Dehdarirad, T., & Yaghtin, M. (2022). Gender differences in citation sentiment: A case study in life sciences and biomedicine. *Journal of Information Science*, 016555152210743. https://doi.org/10.1177/01655515221074327

Dong, C., & Schäfer, U. (2011). Ensemble-style Self-training on Citation Classification. *Proceedings of 5th International Joint Conference on Natural Language Processing*, 623–631. https://aclanthology.org/I11-1070

Dong, Z.D, & Dong, Q. (2001). 知网和汉语研究 [Hownet and Chinese Language Research]. *Contemporary Linguistics, 1*, 33-44+77.

Esuli, A., & Sebastiani, F. (2006). SENTIWORDNET: A publicly available lexical resource for opinion mining. *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, 417–422. https://go.exlibris.link/gBz214Pm

Ghosh, S., Das, D., & Chakraborty, T. (2017). *Determining sentiment in citation text and analyzing its impact on the proposed ranking index*. https://doi.org/10.48550/arXiv.1707.01425

Ghosh, S., & Shah, C. (2020). Identifying Citation Sentiment and its Influence while Indexing Scientific Papers. *Hawaii International Conference on System Sciences*. https://doi.org/10.24251/HICSS.2020.307

Goodarzi, M., Mahmoudi, M. T., & Zamani, R. (2014). A framework for sentiment analysis on schema-based research content via lexica analysis. *7'th International Symposium on Telecommunications (IST'2014)*, 405–411. https://doi.org/10.1109/ISTEL.2014.7000738

Hassan, S.-U., Aljohani, N. R., Idrees, N., Sarwar, R., Nawaz, R., Martínez-Cámara, E., Ventura, S., & Herrera, F. (2020). Predicting literature's early impact with sentiment analysis in Twitter. *Knowledge-Based Systems, 192*, 105383. https://doi.org/10.1016/j.knosys.2019.105383

Ikram, M. T., Afzal, M. T., & Butt, N. A. (2018). Automated citation sentiment analysis using high order n-grams: A preliminary investigation. *Turkish Journal of Electrical Engineering & Computer Sciences, 26*(4), 1922–1932. https://doi.org/10.3906/elk-1712-24

Jochim, C., & Schütze, H. (2012). Towards a Generic and Flexible Citation Classifier Based on a Faceted Classification Scheme. *Proceedings of COLING 2012*, 1343–1358. https://aclanthology.org/C12-1082

Lauscher, A., Glavaš, G., Ponzetto, S. P., & Eckert, K. (2017). Investigating Convolutional Networks and Domain-Specific Embeddings for Semantic Classification of Citations. *Proceedings of the 6th International Workshop on Mining Scientific Publications*, 24–28. https://doi.org/10.1145/3127526.3127531

Mehmood, K., Essam, D., & Shafi, K. (2019). Sentiment Analysis System for Roman Urdu. K. Arai, S. Kapoor, & R. Bhatia, *Intelligent Computing*, 29–42. https://doi.org/10.1007/978-3-030-01174-1_3

Munkhdalai, T., Lalor, J. P., & Yu, H. (2016). Citation Analysis with Neural Attention Models. *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, 69–77. https://doi.org/10.18653/v1/W16-6109

Muppidi, S., Kumar, B. S., & Kumar, K. P. (2021). Sentiment Analysis of Citation Sentences using

Machine Learning Techniques. *2021 Innovations in Power and Advanced Computing Technologies (i-PACT)*, 1–5. https://doi.org/10.1109/i-PACT52855.2021.9696703

Raza, H., Faizan, M., Hamza, A., Mushtaq, A., & Akhtar, N. (2019). Scientific Text Sentiment Analysis using Machine Learning Techniques. *International Journal of Advanced Computer Science and Applications (IJACSA), 10*(12), Article 12. https://doi.org/10.14569/IJACSA.2019.0101222

Sula, C. A., & Miller, M. (2014). Citations, contexts, and humanistic discourse: Toward automatic extraction and classification. *Literary and Linguistic Computing, 29*(3), 452–464. https://doi.org/10.1093/llc/fqu019

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology, 61*(12), 2544–2558. https://doi.org/10.1002/asi.21416

Wang, X.Y, & Zhao, D.Q. (2024). 引文情感识别研究进展及评述[Review on Progress of Citation Sentiment Identification]. *Information Studies:Theory & Application, 47*(1):173-181+189.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *HLT/EMNLP 2005 - Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 347–354. https://doi.org/10.3115/1220575.1220619

Xu, J., Zhang, Y., Wu, Y., Wang, J., Dong, X., & Xu, H. (2015). Citation Sentiment Analysis in Clinical Trial Papers. *AMIA Annual Symposium Proceedings, 2015*, 1334–1341.

Yousif, A., Niu, Z., Tarus, J. K., & Ahmad, A. (2019). A survey on sentiment analysis of scientific citations. *Artificial Intelligence Review, 52*(3), 1805–1838. https://doi.org/10.1007/s10462-017-9597-8

Zuo, R.X, Tang, Z.H, Huang, X, & Wu, J. (2022). 基于情感词典的引文文本情感识别研究[Research on Sentiment Recognition of Citation Text Based on Sentiment Lexicon]. *Digital Library Forum, 02*, 10–17.