Pattern Recognition Letters 150 (2021) 250–257

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Scientific papers citation analysis using textual features and SMOTE resampling techniques

Muhammad Umer[a,b,1,1,*,*], Saima Sadiq[a,a,2], Malik Muhammad Saad Missen[b,b,**], Zahid Hameed[d], Zahid Aslam[b], Muhammad Abubakar Siddique[a], Michele NAPPI[c]

[a] Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan, Pakistan
[b] Department of Computer Science & Information Technology, The Islamia University of Bahawalpur, Bahawalpur, 63100, Pakistan
[c] Department of Computer Science, University of Salerno, Via Giovanni Paolo II, 132, Fisciano 84084, Salerno, Italy
[d] Department of Management Sciences, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan, Pakistan

ARTICLE INFO

Article history:
Received 3 March 2021
Revised 29 May 2021
Accepted 7 July 2021
Available online 24 July 2021

Edited by: P.S. Conti

Keywords:
Citation sentiment analysis
Machine learning
Feature engineering
TF-IDF
SMOTE

ABSTRACT

Ascertaining the impact of research is significant for the research community and academia of all disciplines. The only prevalent measure associated with the quantification of research quality is the citation-count. Although a number of citations play a significant role in academic research, sometimes citations can be biased or made to discuss only the weaknesses and shortcomings of the research. By considering the sentiment of citations and recognizing patterns in text can aid in understanding the opinion of the peer research community and will also help in quantifying the quality of research articles. Efficient feature representation combined with machine learning classifiers has yielded significant improvement in text classification. However, the effectiveness of such combinations has not been analyzed for citation sentiment analysis. This study aims to investigate pattern recognition using machine learning models in combination with frequency-based and prediction-based feature representation techniques with and without using Synthetic Minority Oversampling Technique (SMOTE) on publicly available citation sentiment dataset. Sentiment of citation instances are classified into positive, negative or neutral. Results indicate that the Extra tree classifier in combination with Term Frequency-Inverse Document Frequency achieved 98.26% accuracy on the SMOTE-balanced dataset.

1. Introduction

Pattern recognition is the process of extracting patterns from a dataset and classifying them into different classes. Among various fields where pattern recognition has been applied previously, machine learning has been extensively studied and practiced. The applications of pattern recognition include various scientific disciplines such as computer vision and natural language applications. A suitable feature representation technique and decision-making model is selected on the basis of the problem domain.

Scientific publications are linked with the other state-of-the-art scientific studies in the literature. In the scientific context, a citation is an expression used to cite other scientific work and reference is the identifier of the cited work [27]. An example of citation and its corresponding reference is presented in Fig. 1. Scientists acknowledge the contribution of other researchers by citing their studies which establish a link between cited and citing paper [8]. Citation is very important and valuable for the qualitative analysis of the paper but the size of citation text makes information retrieval a challenging job. Expansion in scientific literature is another challenge in evaluating the impact of publication. Citation not only presents the impact of publication but also the importance of an author. However, counting the number of citations is a quantitative measure but it does not bring forth a qualitative aspect to the citation.

Hitherto the only way to evaluate the quality of research work has been the citation count. Citation count that measures the frequency of research paper being cited by other researchers can mislead in assessing the strength of research paper [28]. Sometimes ci-

* Corresponding author.
** Principal corresponding author.
E-mail addresses: umersabir1996@gmail.com (M. Umer), s.kamrran@gmail.com (S. Sadiq), Saad.missen@iub.edu.pk (M.M.S. Missen), zahid.hameed@kfueit.edu.pk (Z. Hameed), zahid.aslam@iub.edu.pk (Z. Aslam), abubakar.ahmadani@gmail.com (M.A. Siddique), mnappi@unisa.it (M. NAPPI).
[1] orcid=0000-0002-6015-9326
[2] orcid=0000-0002-2611-3738

https://doi.org/10.1016/j.patrec.2021.07.009

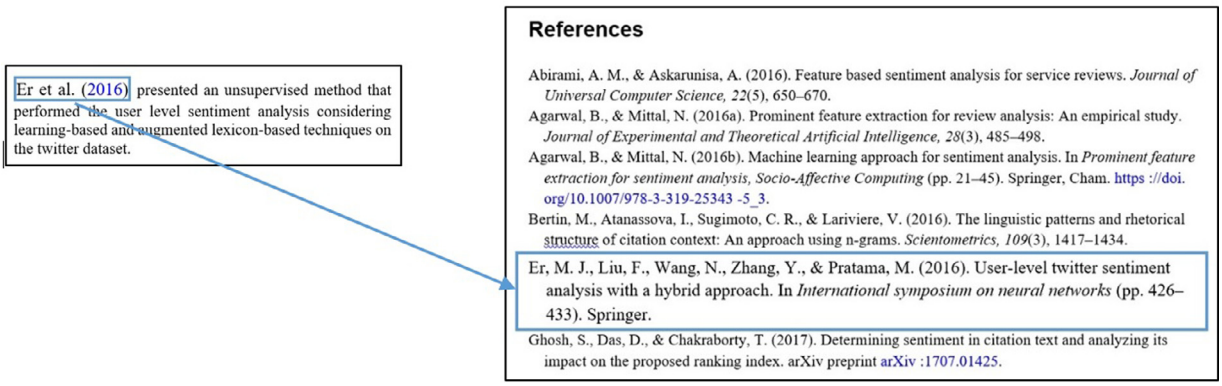

Fig. 1. Example of citation and its reference.

tations can be biased and intentionally made to increase the number of references because of co-authorship [11]. Cited paper sometimes is discussed by citing paper just to discuss shortcomings and improvement suggestions. Such types of citations are also counted in measuring citation indices [5]. Ranking systems based on citation count often fail to recognize productive research work.

Emotion recognition involves sentiment analysis of citation considers its qualitative aspect, which refers to the citing author's opinion about cited research paper. Sentiment classification also describes the context of citation. Sentiment classification has been extensively applied in product reviews [53], movie reviews [18], and citations [55]. Sentiments can be classified as objective and subjective or in a more fine-grained approach as positive or negative depending upon the domain of the text. Different approaches have been applied to classify text strings based on their hidden or ambiguous sentiment. However Sentiment analysis of citations is a relatively less explored area of research. Most of the citations in the research paper describe the findings of the research without showing any opinion, it is assumed that most of these citations are positive [5]. The authors also discussed the social aspect of citation, in which criticism is often shown in a polite way. Negative citations are often implicit and hidden in contrasting terms and cause big challenges for its detection. Cited paper can be positively discussed in terms of experimental design but may be critically analyzed in terms of evaluation metrics. The overall polarity of the text could be measured by assigning weights to the individual sentiments.

Traditionally Scientometrics evaluation schemes consider the quantitative aspect of citation. Most of the high-quality papers are never cited while low-quality papers are frequently cited mostly for criticism. Most of the existing indices are based on the frequency of the citation and can be biased as they are made just to increase the h-index of co-authors. To measure the effectiveness and influence of the research, there is a need to explore the qualitative aspect of citation. Qualitative aspects include the sentiment polarity and context of the citing paper. In recent years, many advancements have been achieved in the field of NLP and machine learning by including efficient text representation techniques such as word embedding and effective classification algorithms. The present work investigates whether these effective word embedding techniques that have shown improved results in text classification can be used to improve the result of sentiment analysis of citation through experiments and analysis. The major contributions of this study are as follows.

- Proposed an effective feature representation technique in combination with supervised machine learning models to determine the sentiment of citation instances into positive, negative, or neutral.
- Performance of seven supervised machine learning models namely Decision Tree (DT), Adaptive Boosting (AdaBoost), Logistic Regression (LR), Stochastic Gradient Decent (SGD), Random Forest (RF), Extra Tree Classifier (ETC), and Support Vector Machine (SVM) is compared in combination with frequency-based and prediction-based feature representation technique to explore pattern recognition of citations.
- Efficacy of Synthetic Minority Oversampling Technique (SMOTE) is analyzed in balancing citation sentiment dataset.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 presents an overview of material and methods adopted for the current research. Section 3 also gives the summary of the proposed methodology for citation sentiment analysis. Results are discussed in Section 4. The conclusion and future work are discussed in Section 5.

2. Related work

A number of citation count-based metrics have been developed to rank authors, articles, and journals. For journals' ranking citation count based metrics are: Source Normalized Impact per Paper (SNIP) [32], Journal Impact Factor (JIF) [22], Eigenfactor [9] and SCImago journal rank (SJR) Indicator [19]. For authors' ranking citation count based metrics are: h-index [26], Author Impact Factor [40] and Author Eigenfactor [51]. Different citation count based ranking metrics to check the influence of the articles are: Article Level Metric by Public Library Of Science (PLOS) [10], Altmetrics [36] and Plum Analytics metric [34]. However, the use of these metrics needs to be manipulated carefully as it impacts end users.

Sentiment analysis of citations is also an upcoming research field in bibliometrics [31]. Polarity (Positive, negative, or neutral) was assigned to the sentiments applying machine learning models [6]. As expressed sentiments are hidden in the citation text which makes the task of assigning sentiment more complex. Some researchers utilized SentiWordNet to assign polarity [7]. It is significant to find a scientific-literature-specific lexicon for sentiment analysis of scientific papers.

Machine learning is the combination of pattern recognition and learning theory and deals with the application that can make predictions after learning patterns from the dataset. Whereas supervised machine learning models search patterns within assigned labels to data points. Machine learning model has been widely used in text classification [29]. Researchers have tried to explore various citation behavior in scientometrics using machine learning models. They analyzed co-citation to link two literature [48], find the relationship between citation location and its type [15] and explored the patterns in citation to find the influence of the editor [47]. Shi et al. [46] explored the relationship between citing and cited paper

and their influence on citing publication. Spiegel-Rosing [49] was the first one who pointed out lack of content (appraisal/critical) in evaluating components of citations. The growing trend towards sentiment analysis of citation can be observed over the last decade [6,23].

Natural language applications are very effective in converting unstructured text into structured text. Active learning for sentiment lexicon extraction for better sentiment analysis is proposed by authors [41]. Researchers utilized fuzzy logic for partial emotion modeling [16]. Authors predicted personality using textual sentiment features [56]. Behavior analysis on textual data was performed by Sadiq et al. [42]. Ref. [37] applied supervised learning along with linguistic features for classification of the citations. Abu-Jbara et al. [1] explored the existing bibliometric measures and criticized that there is no difference between positive and negative citations. They applied NLP techniques to assign sentiment and identified the purpose and the polarity of the sentiment in citations. Athar [3] used n-grams, scientific-literature-based lexicons, sentence splitting, the negation of features, and dependency based features in sentiment analysis. They suggested that n-grams and dependency features outperformed in the classification of citations. Athar and Teufel [6] explored the context-based detection of sentiment of citations.

To get more deep insight into citation literature authors identified the intent of citation in [38]. Intent refers to the citation purpose of the existing work as authors cite published work for many reasons such as to describe their work or to contradict their claim. They claimed that the position of citation plays a significant role in the polarity of citation. The citation present in the result and discussion part is likely to be negative as in this part citation is presented just for comparison purposes to show superiority of the proposed method. Authors deal with the dataset imbalance problem using different techniques such as SMOTE and focal loss. They also performed experiments using CNN, RNN, and LSTM. Their proposed model namely ImpactCite outperformed other models in finding the impact of a citation by sentiment classification.

Authors calculated the quality of the article by citation sentiment with the help of SentiWordNet [44]. They extracted context from citation sentences and then identified adjectives by part of speech (POS) tagging. Authors assigned polarity to each adjective using SentiWordNet analyzer. The overall sentiment is calculated by aggregating scores of all adjectives. In [50] authors framed positive and negative polarity with respect to association and dissociation respectively between citing and cited authors. Positive sentiments refer to supportive arguments and claims for other's work. While negative sentiments refer to disagreement in claims and arguments for other's work. They also presented connections between citing and cited research through network graphs.

A new annotation methodology is established for citation sentiment classification by Hernández-Alvarez and Gómez [25]. They labeled citation sentences into three classes that are positive, negative, and neutral. The authors built a new corpus that is annotated with sentiments and then they applied SVM for citation sentiment classification. A rule-based approach was developed for citation extraction and more than thousands of citations were annotated manually [52]. They utilized n-grams and sentiment-based lexicon with SVM for citation sentiment classification. Their results proved that a combination of features achieved better results than individual ones. In [35] authors suggested using different features including unigram, p_index, author_id, and affiliation_id in order to improve analysis methods to measure the quality of the research work. They highlighted the need for improvement in H_index by considering negative polarity. But their work still requires feature engineering techniques to improve citation sentiment classification.

Sentiment classification of review text is different from the citation sentiment classification because citation text is quite formal. The intent of classification can be explored by exploring the section of the paper where that text appears. Authors in [2] applied different machine learning models to perform binary classification of in-text citations.They used cosine similarity of citation paper pairs and sentiment as features to perform classification. Researchers proposed Impactcite based on XLNet for citation impact analysis [39]. Their proposed Impactcite model outperformed for intent and sentiment classification[30]. performed opinion mining and qualitative analysis of citation text. They performed medical data analysis by Spearman correlation. Sentiment analysis of in-text citations is significant due to the unavailability of the appropriate dataset in this domain. Finding a sentiment from formal and analytical text is different from the informal and subjective text such as reviews or Twitter data.

3. Proposed methodology for citation sentiment analysis

This section provides the details of the dataset, SMOTE, evaluation parameters, and basic model of the proposed methodology for the classification of citation sentiment. The complete architecture of the proposed approach is presented in Fig. 2. A manually annotated dataset containing citation sentences is collected for this purpose. Experiments have been performed on various machine learning models such as DT, AdaBoost, LR, SGD, RF, ETC, and SVM discussed in Table 1. The pipeline of the proposed approach consists of several steps. Figure 2 demonstrates the proposed methodology of data and workflow of this research work.

3.1. Citation sentiment dataset

This research uses a dataset namely "citation sentiment corpus" [4] derived from ACL Anthology Network corpus. Dataset consists of 8736 citation sentences that have been assigned sentiment by manual annotation. The dataset consists of four attributes that are: Source_Paper ID, Target_Paper ID, Sentiment, and Citation_Text. Source_Paper ID is the ID of the citing paper from where the text has been taken. Target_Paper ID is the ID of the cited paper.

In Citation_Text, sentences contained citation text to the target paper and were assigned classes in the Sentiment column as positive, negative, or neutral according to the intention of the citing author.

3.2. SMOTE

Citation sentiment corpus was used to train the machine learning classifiers for citation sentiment analysis. The distribution of polarity of the positive, negative, and neutral instances in the corpus is 619, 206, and 5644 respectively.

As Data instances are biased toward the Neutral class, SMOTE is applied to deal with the problem of class imbalance. As biased class increases the wrong predictions and causes overfitting of machine learning models. SMOTE is a technique that performs upsampling and is being widely used to deal with class imbalance problems [17].

SMOTE uses Euclidean distance to generate synthetic data for minority class from its nearest neighbors. Newly generated data is similar to the original data on the basis of features. In this research, new data instances are created using SMOTE. Data instances of minority class have been increased according to the size of the majority class which is a neutral class in our case.

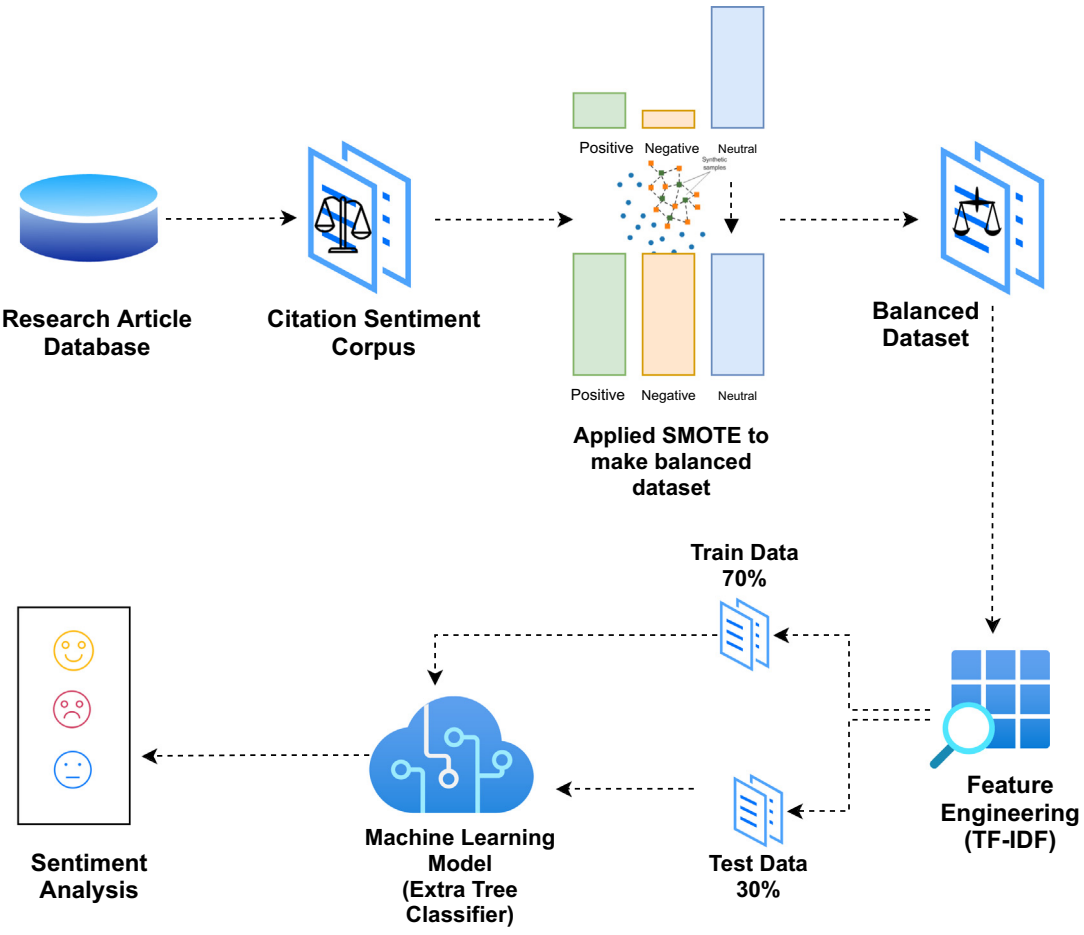

Fig. 2. Proposed Architecture for In-text Citation Sentiment Classification.

Table 1
Description of Machine Learning Models.

Reference	Model	Description
[14]	DT	A decision Tree is a classification algorithm that works well on both forms of data i.e., categorical and numerical. Generally speaking, a decision tree is utilized for the creation of tree-like structures. A decision tree is simple and easy to implement so it is widely used for medical data analysis.
[20]	AdaBoost	AdaBoost is the abbreviation of adaptive boosting. AdaBoost is normally used in conjunction with the other algorithms to enhance their performance. It uses boosting technique and transforms weak learners into strong learners. Every tree in the AdaBoost classifier is depending on the result that is the error rate of the last tree.
[12]	LR	Logistic regression usually deals with the classification problems. It is a predictive analysis algorithm and statistical model and based on the concept of probability. It is usually used to investigate binary data in which one or more variables work to determine output.
[21]	SGD	Stochastic Gradient Descent integrates many binary classifiers in the one-versus-all method. SGD has been widely used for large datasets because in each iteration it uses all the samples. The working principle is quite similar to the regression technique so it is quite easy to implement and understand. Its hyper parameters need to be accurately valued to obtain correct results. In terms of feature scaling sensitivity of SGD is high.
[13]	RF	Random forest is a tree-based ensemble learning model, which generates accurate predictions by merging numerous weak learners. The bagging technique is used by this model to train a variety of decision trees using various bootstrap samples. In a random forest, a bootstrap sample is derived by subsampling the original dataset with substitution, where the sample size is the same as that of the training data set.
[45]	ETC	Extra trees classifier working is quite similar to the random forest and only different from it in a method of construction of trees in the forest. Every decision tree in the extra tree classifier is made from the original training sample. Random subsamples of the k-best feature are utilized for decision. Information gain and Gini index are used to choose the top feature to split the node in the tree. These random samples of feature indication to the creation of multiple de-correlated decision trees.
[43]	SVM	Support Vector Machine is a supervised learning technique and based on mathematical models. It is applied for regression and classification problems. It performs classification by constructing high dimensional hyper-planes also called decision planes. Hyper-plane distinguishes one type of data from another type.

Table 2
Performance Measures used for Evaluation in this study.

Evaluation Metric	Formula
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1-Score	$2^* \frac{precision.recall}{precision+recall}$

3.3. Evaluation metrics

The performance of machine learning models is evaluated using several performance evaluation parameters. These evaluation parameters are expected to endorse the development of analytical research [33]. In this research, four evaluation metrics are Accuracy, Precision, Recall and F1-score will be examined for comparison of the machine learning-based algorithms. Confusion matrix allows visualization of the performance of machine learning models to calculate all four metrics. The elements of the confusion matrix are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The measures of the performance are given in Table 2.

3.4. Feature extraction

In order to apply machine learning models effectively, text representation is converted into numerical form. Selection of the best feature representation technique is essential in natural language processing-based classification tasks. Feature representation techniques are mainly categorized into frequency-based and prediction-based approaches. The strengths and weaknesses of each of these approaches is shown in Table 3.

4. Results and discussion

Comprehensive experiments have been performed in different settings for citation sentiment analysis. In each setting, machine learning models are trained using two different feature representation techniques on a balanced and Imbalanced dataset. For feature extraction, TF-IDF and Word2Vec have been selected as both techniques have been widely used for text classification and showing robust results. Likewise, we selected the most suitable machine learning model for citation text classification. Experiments have been designed to find the effectiveness of combining feature representation methods with supervised machine learning classifiers to predict citation sentiments into positive, negative, or neutral classes.

Table 4
Classification result of seven machine learning models using TF-IDF without SMOTE.

Models	Accuracy	Precision	Recall	F1-score
DT	84.73%	84%	85%	84%
AdaBoost	87.52%	85%	88%	85%
LR	87.14%	84%	87%	82%
SGD	88.70%	87%	89%	86%
RF	87.60%	84%	88%	84%
ETC	87.75%	85%	88%	84%
SVM	89.61%	87%	89%	87%

4.1. Performance of classification models using frequency-based feature engineering

Three settings have been considered for training classification models discussed in Table 1 using TF-IDF from frequency-based feature engineering technique for citation sentiment analysis. Performance comparison using TF-IDF is presented in Fig. 3.

Setting I: In order to compare the performance of classifiers for citation sentiment classification, at first experiment has been performed on an imbalanced dataset using TF-IDF as a feature set. Results shown in Table 4 indicates that SVM surpassed other models and achieved reasonable classification results with 89.61% Accuracy, 87% Precision, 89% Recall and 87% F1-score. Moreover, DT showed the worst results with 84.73% accuracy using TF-IDF on the imbalanced dataset. DT often cannot generalize data well in the case of an imbalanced class and can overfit. In the case of a high proportion of minority class instances having multiple features in the sample space, the trees can recognize the patterns but there is a probability that unstable deep trees will be prone to overfitting. It will affect the performance of the model. Generalization of DT as RF, AdaBoost, and ETC provide better alternatives in terms of better performance and stability.

Setting II: Then dataset is balanced by considering equal instances from each class. Data instances of the majority class are under-sampled according to the ratio of minority class and 280 instances from each class are selected randomly for the training of the model. Results of the supervised machine learning model after training on 280 data instances of each class using TF-TDF are presented in Table 5. In the case of a small-sized dataset, ETC achieved the highest accuracy which is 77.77%, the highest precision which is 78% and highest recall which is also 78%. The highest F1-score is achieved by ETC and SGD with 77%. DT showed the lowest results after training on 280 instances of each class. It seems clear from Fig. 3 that the performance of all models degraded analogously after reducing dataset instances. The principal cause is the

Table 3
Description of Feature Representation Techniques.

Technique	Type	Strengths	Weaknesses
TF-IDF	Frequency based approach	• Can find similarity between documents easily • Count the occurrence of the word in a document as well as in a whole corpus • Weight is directly proportional to document's word frequency and inversely proportional to words' frequency within documents. • Words like is, a, the, but, while are not significant as compared to words having rare occurrence.	• large vector size • Position and co-occurrence has no importance • Do not consider semantics and context. • Sparsity problem • Unable to find similarity between synonyms and differentiate in polysemy words.
Word2Vec	Prediction based approach	• Give probability of words • Consider word similarities • Map words to target vectors • CBOW predict the probability of words and skip-gram predicts context of words.	• Difficult to train models on Word2Vec because of large size of the vocabulary. • Consider word similarities • Take average of polysemy words by CBOW, skip-gram represents by separate vectors.

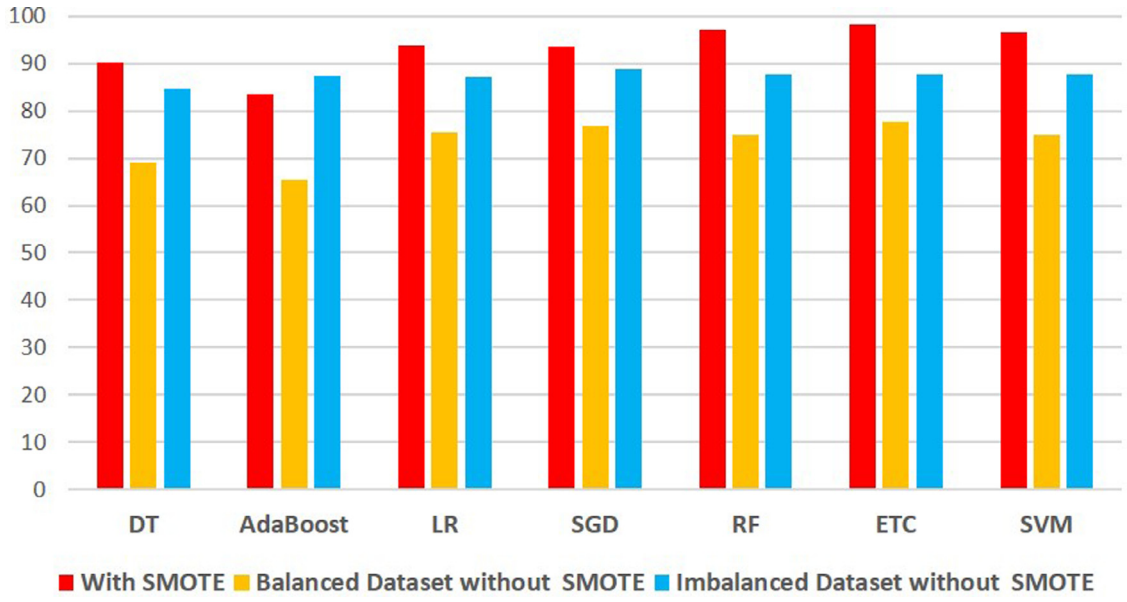

Fig. 3. Accuracy comparison of seven classifiers using TF-IDF.

Table 5
Classification result of seven machine learning models using TF-IDF with balanced data (280 records of each class).

Models	Accuracy	Precision	Recall	F1-score
DT	69.04%	69%	69%	69%
AdaBoost	65.47%	66%	65%	66%
LR	75.39%	75%	75%	75%
SGD	76.98%	77%	77%	77%
RF	75.00%	75%	75%	75%
ETC	77.77%	78%	78%	77%
SVM	75.00%	75%	75%	75%

Table 6
Classification result of seven machine learning models using TF-IDF with SMOTE.

Models	Accuracy	Precision	Recall	F1-score
DT	90.10%	90%	90%	90%
AdaBoost	83.61%	84%	79%	82%
LR	93.88%	94%	94%	94%
SGD	93.61%	96%	96%	96%
RF	97.29%	98%	96%	97%
ETC	98.26%	98%	98%	98%
SVM	96.69%	97%	97%	97%

small number of instances in each class. This fact not only poses a problem of less training instance for the model but also the reason for poor performance. In other words, the size of the dataset significantly affects the training and performance of the models for comparing the effective classification methods.

Setting III: After that class imbalance problem is solved by applying SMOTE. Data instances of minority class are over-sampled to make the dataset balanced. Then experiments have been performed on the balanced dataset and results are presented in Table 6. The results shown in Fig. 3 indicate that using SMOTE consistently improved the performance, as all experiments achieved higher results than their results on the imbalanced and under-sampled dataset. This contributes to the understanding that using SMOTE avoids the chance of misclassification and 6 out of 7 models achieved more than 90% results. While it is crystal clear that some algorithms take more advantage than others as ETC and RF achieved more than 97% accuracy. SMOTE generates synthetic data samples by joining k nearest neighbors of minority class chosen randomly according to the requirement. New synthetic data examples are generated by taking the difference between the sample feature vector and its nearest neighbor and then multiply this number to a random number and then added to the feature vector. This method makes minority class more general by adding new data examples. An extremely randomized tree classifier that is ETC totally selects random cut-point from random subspace. If the randomization level in ETC is adjusted properly then variance vanishes while bias increases according to the trees. Bias is the main problem with ETC. Hence, SMOTE has been applied in order to deal with the bias. In this way, the tree-based model generalized better and avoids overfitting. ETC outperformed all other classifiers and can predict citation sentiment with 98.26% accuracy and 98% F1-score.

Table 7
Classification result of seven machine learning models using Word2Vec.

Models	Accuracy	Precision	Recall	F1-score
DT	77.29%	80%	77%	78%
AdaBoost	85.54%	80%	86%	82%
LR	87.60%	84%	88%	83%
SGD	87.56%	85%	88%	83%
RF	87.52%	83%	88%	83%
ETC	87.52%	84%	88%	83%
SVM	87.29%	76%	87%	81%

4.2. Significance of the proposed model

Finally, the performance of classifiers has been evaluated and compared using Word2Vec from the prediction-based feature engineering method for citation sentiment analysis. Results presented in Table 7 indicate that classifiers using Word2Vec did not achieve robust results. Figure 4 shown a considerable difference between the performance of models using Word2Vec and using TF-IDF. Moreover, despite the reasonable results shown by all classifiers, word2Vec is not contributing to improving the performance of any classifiers used for the experiment. LR, SGD, RF, and ETC achieved the highest F1-score of 83% using Word2Vec for citation sentiment analysis which is 10% lower than that achieved by using TF-IDF

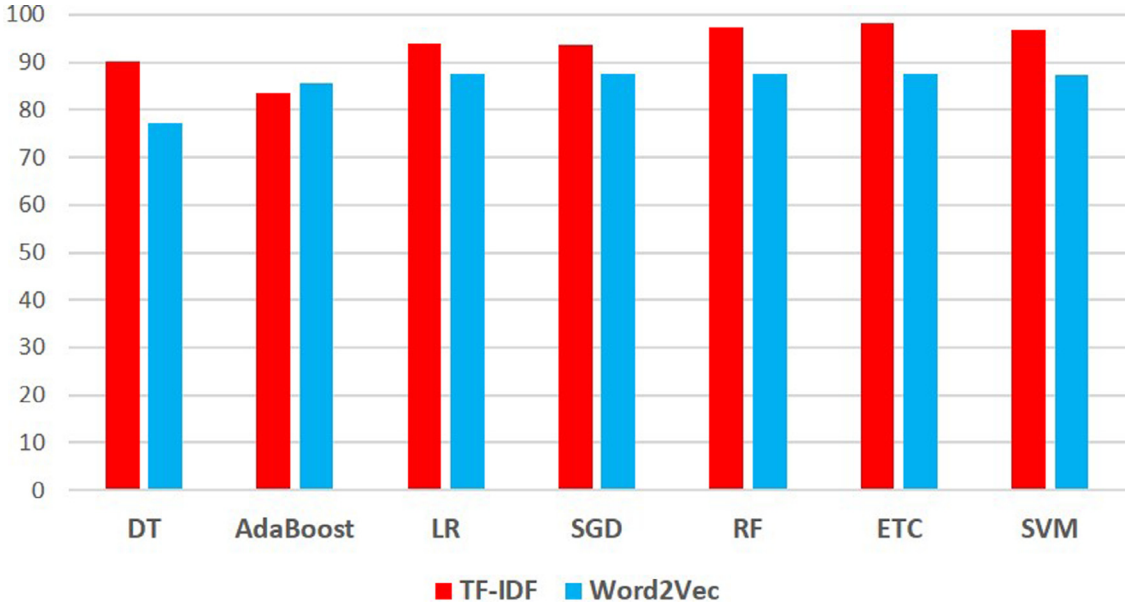

Fig. 4. Accuracy comparison of seven classifiers using TF-IDF and Word2Vec.

Table 8
Accuracy Comparison of seven machine learning models using TF-IDF with SMOTE AND ADASYN.

Models	With SMOTE	With ADASYN
DT	90.10%	88.79%
AdaBoost	83.61%	83.61%
LR	93.88%	91.47%
SGD	93.61%	93.61%
RF	97.29%	94.29%
ETC	98.26%	96.26%
SVM	96.69%	96.01%

with SMOTE. These consequences contribute to the understanding that Word2Vec is not adequate to improve the effectiveness of classifiers.

In order to prove the effectiveness of the proposed model, we also performed an experiment using ADASYN (Adaptive Synthetic Sampling) [24], an improved version of SMOTE, with seven machine learning models using TF-IDF. Comparison of seven machine learning models using TF-IDF with SMOTE AND ADASYN is presented in Table 8. ADASYN has shown good results especially AdaBoost with 83.61% accuracy and SGD with 93.61% Accuracy showed similar Accuracy results as with SMOTE. But all other models such as DT, LR, RF, ETC, and SVM performed better with SMOTE. It can be clearly observed from Table 8 that SMOTE surpassed ADASYN in sentiment analysis of in-text citations.

If we compare some previous sentiment analysis research works with the proposed approach, one thing is quite clear that previous research did not bother about the dataset class imbalance problems. They performed sentiment classification using the machine [54] and deep learning [29] with all kinds of features but totally neglected the biases of the training models in terms of dataset imbalance records. In this research work, we prefer to make the dataset balanced to get more reliable results in terms of training and testing. Furthermore, the aforementioned researchers worked on the single line reviews obtained from the user but this research work makes use of citation paragraphs to express the sentiment of the authors in citing other research works.

We also compared the result of our proposed approach with the state-of-the-art research work [2]. They also trained their models on the same dataset [4] as ours and achieved 84% precision value by RF. Our proposed ETC model using TF-IDF with SMOTE achieved higher precision with 97% value.

5. Conclusion

In order to assess the effectiveness and suitability of feature representation techniques, models, and their combinations, experiments have been performed with 4 different settings on citation sentiment corpus. By evaluating the classification model on performance evaluation metrics, it is concluded that Word2Vec is not appropriate for citation sentiment analysis. TF-IDF has a considerable advantage over Word2Vec in the prediction of citation sentiment. The effectiveness of SMOTE in improving the performance of machine learning classifiers has also been observed.

Experimental results confirm the superiority of ETC with 98.26% accuracy and 98% F1-score using TF-IDF on SMOTE balanced dataset for citation sentiment analysis. It seems ETC owes most of its dominance in terms of performance because of its combination with an appropriate feature representation technique. These investigations increase understanding of the task-specific combination of a feature engineering technique with dataset balancing method for classification performed by supervised machine learning models. Nevertheless, 98.26% accuracy achieved by ETC is remarkable to predict citation sentiment. Although, other classifiers used in the experiment did not show robust results so their flexibility could be explored further to get promising results. In the future, we are planning to use pre-trained word embedding in combination with deep learning models for citation sentiment analysis.

Declaration of Competing Interest

Manuscript has not been published elsewhere and that it has not been submitted simultaneously for publication elsewhere. Authors wish to confirm that there are no known conflicts of interest associated with this publication.

CRediT authorship contribution statement

Muhammad Umer: Writing – original draft, Methodology, Software. **Saima Sadiq:** Writing – original draft, Conceptualization.

Malik Muhammad Saad Missen: Writing – review & editing, Supervision. **Zahid Hameed:** Conceptualization. **Zahid Aslam:** Writing – review & editing. **Muhammad Abubakar Siddique:** Supervision. **Michele NAPPI:** Writing – review & editing, Methodology, Funding acquisition.

References

[1] A. Abu-Jbara, J. Ezra, D. Radev, 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, 2013, pp. 596–606.

[2] H. Aljuaid, R. Iftikhar, S. Ahmad, M. Asif, M.T. Afzal, Important citation identification using sentiment analysis of in-text citations, Telematics Inf. 56 (2021) 101492.

[3] A. Athar, Sentiment analysis of citations using sentence structure-based features, in: Proceedings of the ACL 2011 Student Session, 2011, pp. 81–87.

[4] A. Athar, Sentiment analysis of citations using sentence structure-based features, in: Proceedings of the ACL 2011 Student Session, Association for Computational Linguistics, Portland, OR, USA, 2011, pp. 81–87.

[5] A. Athar, S. Teufel, Context-enhanced citation sentiment detection, in: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2012, pp. 597–601.

[6] A. Athar, S. Teufel, Detection of implicit citations for sentiment detection, in: Proceedings of the Workshop on Detecting Structure in Scholarly Discourse, 2012, pp. 18–26.

[7] S. Baccianella, A. Esuli, F. Sebastiani, SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining., in: Lrec, vol. 10, 2010, pp. 2200–2204.

[8] J. Bar-Ilan, G. Halevi, Post retraction citations in context: a case study, Scientometrics 113 (1) (2017) 547–565.

[9] C.T. Bergstrom, J.D. West, Assessing citations with the eigenfactor™ metrics, 2008,

[10] L. Bornmann, Which kind of papers has higher or lower altmetric counts? A study using article-level metrics from PLOS and F1000Prime, 2014,

[11] L. Bornmann, H.-D. Daniel, What do citation counts measure? A review of studies on citing behavior, J. Doc. (2008).

[12] C.R. Boyd, M.A. Tolson, W.S. Copes, Evaluating trauma care: the TRISS method, J. Trauma Acute Care Surg. 27 (4) (1987) 370–378.

[13] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[14] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and regression trees. statistics/probability series, 1984,

[15] V. Cano, Citation behavior: classification, utility, and location, J. Am. Soc. Inf.Sci. 40 (4) (1989) 284–290.

[16] I. Chaturvedi, R. Satapathy, S. Cavallari, E. Cambria, Fuzzy commonsense reasoning for multimodal sentiment analysis, Pattern Recognit. Lett. 125 (2019) 264–270.

[17] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.

[18] B.L. Devi, V.V. Bai, S. Ramasubbareddy, K. Govinda, Sentiment analysis on movie reviews, in: Emerging Research in Data Engineering Systems and Computer Communications, Springer, 2020, pp. 321–328.

[19] M.E. Falagas, V.D. Kouranos, R. Arencibia-Jorge, D.E. Karageorgopoulos, Comparison of SCimago journal rank indicator with journal impact factor, FASEB J. 22 (8) (2008) 2623–2628.

[20] Y. Freund, R. Schapire, N. Abe, A short introduction to boosting, J.-Jpn. Soc. Artif.Intell. 14 (771–780) (1999) 1612.

[21] W.A. Gardner, Learning characteristics of stochastic-gradient-descent algorithms: a general study, analysis, and critique, Signal Process. 6 (2) (1984) 113–133.

[22] E. Garfield, The history and meaning of the journal impact factor, JAMA 295 (1) (2006) 90–93.

[23] S. Ghosh, D. Das, T. Chakraborty, Determining sentiment in citation text and analyzing its impact on the proposed ranking index, in: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, 2016, pp. 292–306.

[24] H. He, Y. Bai, E.A. Garcia, S. Li, ADASYN: adaptive synthetic sampling approach for imbalanced learning, in: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), IEEE, 2008, pp. 1322–1328.

[25] M. Hernández-Alvarez, J.M. Gómez, Citation impact categorization: for scientific literature, in: 2015 IEEE 18th International Conference on Computational Science and Engineering, IEEE, 2015, pp. 307–313.

[26] J.E. Hirsch, An index to quantify an individual's scientific research output, Proc. Natl. Acad. Sci. 102 (46) (2005) 16569–16572.

[27] R. Hjerppe, A bibliography of bibliometrics and citation indexing and analysis (1980).

[28] S. Huggett, Journal bibliometrics indicators and citation ethics: a discussion of current issues, Atherosclerosis 230 (2) (2013) 275–277.

[29] A. Ishaq, M. Umer, M.F. Mushtaq, C. Medaglia, H.U.R. Siddiqui, A. Mehmood, G.S. Choi, Extensive hotel reviews classification using long short term memory, J. Ambient Intell. Humaniz. Comput. (2020) 1–11.

[30] K. Jokar, M. Yaghtin, H. Sotudeh, M. Mirzabeigi, Correlation between quantitative citation analysis and opinion mining of citation contexts, Scientometrics Res. J. (2021).

[31] S.K. Kochhar, U. Ojha, Index for objective measurement of a research paper based on sentiment analysis, ICT Express 6 (3) (2020) 253–257.

[32] L. Leydesdorff, T. Opthof, Scopus's source normalized impact per paper (snip) versus a journal impact factor based on fractional counting of citations, J. Am. Soc. Inf.Sci. Technol. 61 (11) (2010) 2365–2369.

[33] T.-S. Lim, W.-Y. Loh, Y.-S. Shih, A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms, Mach. Learn. 40 (3) (2000) 203–228.

[34] J.M. Lindsay, Plumx from plum analytics: not just altmetrics, J. Electron. Resour. Med.Libraries 13 (1) (2016) 8–17.

[35] Z. Ma, J. Nam, K. Weihe, Improve sentiment analysis of citations with author modelling, in: Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2016, pp. 122–127.

[36] R. Melero, Altmetrics–a complement to conventional metrics, Biochem. Med. 25 (2) (2015) 152–160.

[37] R. Meng, W. Lu, Y. Chi, S. Han, Automatic classification of citation function by new linguistic features, iConference 2017 Proceedings(2017).

[38] D. Mercier, S.T.R. Rizvi, V. Rajashekar, A. Dengel, S. Ahmed, ImpactCite: an XLNet-based method for citation impact analysis, (2020) arXiv preprint arXiv: 2005.06611

[39] D. Mercier, S.T.R. Rizvi, V. Rajashekar, A. Dengel, S. Ahmed, ImpactCite: an XLNet-based solution enabling qualitative citation impact analysis utilizing sentiment and intent(2021).

[40] R.K. Pan, S. Fortunato, Author impact factor: tracking the dynamics of individual scientific impact, Sci. Rep. 4 (1) (2014) 1–7.

[41] S. Park, W. Lee, I.-C. Moon, Efficient extraction of domain specific sentiment lexicon with active learning, Pattern Recognit. Lett. 56 (2015) 38–44.

[42] S. Sadiq, A. Mehmood, S. Ullah, M. Ahmad, G.S. Choi, B.-W. On, Aggression detection through deep neural model on twitter, Future Gener. Comput. Syst. 114 (2021) 120–129.

[43] B. Schölkopf, C. Burges, V. Vapnik, Incorporating invariances in support vector learning machines, in: International Conference on Artificial Neural Networks, Springer, 1996, pp. 47–52.

[44] S. Sendhilkumar, E. Elakkiya, G. Mahalakshmi, Citation semantic based approaches to identify article quality, in: Proceedings of International Conference ICCSEA, 2013, pp. 411–420.

[45] A. Sharaff, H. Gupta, Extra-tree classifier with metaheuristics approach for email classification, in: Advances in Computer Communication and Computational Sciences, Springer, 2019, pp. 189–197.

[46] X. Shi, L.A. Adamic, B.L. Tseng, G.S. Clarkson, The impact of boundary spanning scholarly publications and patents, PLoS ONE 4 (8) (2009) e6547.

[47] M. Sievert, M. Haughawout, An editor's influence on citation patterns: a case study of elementary school journal, J. Am. Soc. Inf.Sci. 40 (5) (1989) 334–341.

[48] H. Small, E. Garfield, Analysis of scientific literature to assist in problem solving, J. Am. Soc. Inf.Sci. 40 (3) (1989). 152–152

[49] I. Spiegel-Rosing, Science studies: bibliometric and content analysis, Soc Stud Sci 7 (1) (1977) 97–113.

[50] C.A. Sula, M. Miller, Citations, contexts, and humanistic discourse: toward automatic extraction and classification, Literary Linguist. Comput. 29 (3) (2014) 452–464.

[51] J.D. West, M.C. Jensen, R.J. Dandrea, G.J. Gordon, C.T. Bergstrom, Author-level eigenfactor metrics: evaluating the influence of authors, institutions, and countries within the social science research network community, J. Am. Soc. Inf.Sci. Technol. 64 (4) (2013) 787–801.

[52] J. Xu, Y. Zhang, Y. Wu, J. Wang, X. Dong, H. Xu, Citation sentiment analysis in clinical trial papers, in: AMIA Annual Symposium Proceedings, vol. 2015, American Medical Informatics Association, 2015, p. 1334.

[53] L. Yang, Y. Li, J. Wang, R.S. Sherratt, Sentiment analysis for e-commerce product reviews in chinese based on sentiment lexicon and deep learning, IEEE Access 8 (2020) 23522–23530.

[54] A. Yousaf, M. Umer, S. Sadiq, S. Ullah, S. Mirjalili, V. Rupapara, M. Nappi, Emotion recognition by textual tweets classification using voting classifier (LR-SGD), IEEE Access (2020).

[55] A. Yousif, Z. Niu, J.K. Tarus, A. Ahmad, A survey on sentiment analysis of scientific citations, Artif. Intell. Rev. 52 (3) (2019) 1805–1838.

[56] J. Zhao, D. Zeng, Y. Xiao, L. Che, M. Wang, User personality prediction based on topic preference and sentiment analysis using LSTM model, Pattern Recognit. Lett. 138 (2020) 397–402.