Scientific Text Sentiment Analysis using Machine Learning Techniques

Hassan Raza[1], M. Faizan[2], Ahsan Hamza[3], Ahmed Mushtaq[4], Naeem Akhtar[5]
School of Computer Sciences
National College of Business Administration and Economics
Lahore, Pakistan

***Abstract*—Over time, textual information on the World Wide Web (WWW) has increased exponentially, leading to potential research in the field of machine learning (ML) and natural language processing (NLP). Sentiment analysis of scientific domain articles is a very trendy and interesting topic nowadays. The main purpose of this research is to facilitate researchers to identify quality research papers based on their sentiment analysis. In this research, sentiment analysis of scientific articles using citation sentences is carried out using an existing constructed annotated corpus. This corpus is consisted of 8736 citation sentences. The noise was removed from data using different data normalization rules in order to clean the data corpus. To perform classification on this data set we developed a system in which six different machine learning algorithms including Naïve-Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbor (KNN) and Random Forest (RF) are implemented. Then the accuracy of the system is evaluated using different evaluation metrics e.g. F-score and Accuracy score. To improve the system' accuracy additional features selection techniques, such as lemmatization, n-graming, tokenization, and stop word removal are applied and found that our system provided significant performance in every case compared to the base system. Our method achieved a maximum of about 9% improved results as compared to the base system.**

Keywords—Sentimental analysis; scientific citations; machine learning; scientific literature; classification

I. Introduction

Sentiment analysis of scientific citation is very well discussed and interesting topic in this era where WWW is excessively loaded with an enormous amount of text data [62]. This data contains tons of important information inside itself that can be very beneficial after being analyzed based on requirements. Sentimental analysis is also known as opinion mining that means to find out or identify the positive, negative, neutral opinions, views, attitudes, impressions, emotions and feelings indicated in the text [8]. Opinion mining from the citations is of prime importance because citations from the papers reinforce arguments and connect it to intellectual [25][47][84]. From the last decade, the importance opinion mining or sentimental analysis is mentioned by [86][91][93][94] as a research on a citation function.

We choose more specifically scientific domain as a problem statement in order to analyze the sentiments of citation sentences extracted from different scientific research papers just because of linguistic differences in this domain. A lot of work has been done previously in other genres like English and Chinese as compared to the scientific domain. Following are the problems related to scientific text has been mentioned in literature: usually, sentiments in scientific citations are hidden and not well expressed, scientific citations are often neutral, prevalent style of writing [9], objective style and personal biased of authors have to be hedged [35][64]. Conventionally scientific citations are written in dual-mode and that is also a problem in scientific literature. Some authors apply the strategy of dual-mode like prefacing some criticism after a light appraisement[55]. Identifying such opinions is a challenging task. Such kind of expressions are also found in other types of literature as well [96].

In this work citation sentence refers to the reference of other's papers in the text of a given scholarly work, the former will be known as cited and the latter is called citing paper as well. Along with the citation sentences usually, the citation references are mentioned using different styles and standards. One such famous standard of writing citations' references is "Harvard Style" that uses the author's last name followed by the year of publication [9].

In this research work, we have done sentimental analysis of scientific citations by using an annotated corpus consists of citation sentences developed by [9]. The corpus is made up of 8736 citation sentences constructed from the scientific domain related research papers extracted from ACL (Association for Computational Linguistics) Anthology. The corpus is annotated using some rules to assign the polarity to citation sentences. We have developed a system based on six different machine learning algorithms including Naïve-Bayes, Support Vector Machine, Logistic Regression, Decision Tree, K-Nearest Neighbor and Random Forest. Accuracy of the classification algorithms has been evaluated using different evaluations measures e.g., F-Score and Accuracy score to evaluate the classification system' correctness. To improve our system' performance, we have used different features selection techniques like lemmatization, n-gaming, tokenization, stop words and punctuation removal. After successful experimentation we have found that our system outperforms in each case as compare to the method adopted by the[9]. The maximum outperformance we achieved is 87% F-score as compare to 78% F-score reported by [9] which results in 9% improvement than [9].

II. Literature Review

To carry out this study, literature review is held out to analyze the current state of the domain. In the last couple of years, the interest towards the research in sentimental analysis has been increased for different domains but less work has been done on the scientific literature due to some problems mentioned above.

Sentimental classifiers can be developed using two machine learning approaches named Supervised and Unsupervised learning [50]. The most famous approach to build a classifier is supervised learning. In supervised approach, the classifier requires labeled training data. Yet the training data is to be annotated using manual or automatic approaches on the basis of some predefined rules. Using predefined rules the citation sentences are annotated as positive, negative and neutral. For the purpose of annotations, human annotators are required. While in the case of an unsupervised approach, there is no need for labeled training data. Instead, there is a need for sentiment lexicon to assign polarities to citation sentences. This approach is very difficult because it requires different varieties of a lexicon for different genres. From the literature review, we found that many researchers used supervised while others relyied on unsupervised approach.

Author in [23] worked on automatic citataion classification. Another work has been done in order to analyze the behaviors of authors, readers of research papers in the scientific field by [39]. The analysis was regarding the authors of research papers, how they frame their citations, how readers become interested in the citations of authors and how these processes contribute towards the maturity of Natural Language Processing (NLP). For the sake of analyzing the contributions and purpose of citations based on behavioral analysis, authors classify their data using two different schemes [20]. They prepared the data set of citation sentences extracted from 52 papers from ACR (Anthology Reference Corpus) and annotated the data set using some guidelines based on ACR [16]. They used core aspects of prior citation based on annotation schemes mentioned by [30][31][94][95]. BRAT Tool was used to perform the annotations [85]. They used the Random Forest classifier implemented using Sickit-learn [101]. The major reason of choosing this classifier is its ability to perform effecient for larger feature sets [21][98]. Structural features and grammatical features were used for the classification [28][30][31]. They compared their evaluation results with [89]. Their method with different features shows salient behavior of writers, readers and domain. We have also used a supervised learning approach and developed a system in which multiple machine learning classifiers are implemented. As input, our system takes corpus consisted of labeled citation sentences, to performs classification and also evaluate the classification accuracy of the system. To Increase the efficiency and system accuracy, we have applied different features selection techniques in the data pre-processing phase. Our data set consists of a huge set of citation sentences. Our system performed better than [9].

While processing the citations, finding the implicit citation is also a problem and this problem was addressed by [63]. In their work, the major goal was to identify the implicit citations with the help of improving citation context detection methods [63]. As the research work by [10] was restricted up to the detection of the author's sentiments towards citation reference. In order to create a summary of reference citations, text data may also helpful [72]. Mostly the negative opinions are appeared in explicit citations [10]. [81] claimed that researchers only read 20% of the papers to get the desired information. Different authors have different intentions towards the citations. The intentions of the author were classified using manually constructed and compared cue-phrases against citation context [90]. Citations' context detection is also helpful for creating summaries of different research topics that can support researchers to get a detailed and convenient view of papers [72]. The importance of citation context can also be felt from the fact that all the information retrieval systems that incorporate the concept of citation context have better retrieval effectiveness [73]. The authors developed a system in which data set prepared by [12] and [10] were tested using the method from the work of [72]. The data set consisted of 852 papers from ACL Anthology. To boost up the efficiency and accuracy results of the system authors applied some new classification features like sentence features and sentence similarity measures. The accuracy of the system was evaluated and find out that the system performed better. Our approach is totally different in which we have used new and different classification algorithms and different features have been used by us to improve the system accuracy scores. And our system outperformed the state of the art.

Another work is done in the domain of sentimental analysis which is not in English, Chinese or scientific domain but specifically in the Urdu domain [59], as very less work was done in the Urdu language [56]. They used the data set based on Urdu reviews related to movies, politics, mobile, dramas and miscellaneous domains extracted using scrapers as well as manual. The data set was then classified using different types of supervised learning classifiers and compare their results with each other.

Author in [67] used labeled data for the purpose of classification, they preferred the supervised learning approach. For the purpose of classification, the Naïve Bayes classifier is used. In this work, they have used a dataset of movie reviews. The reviews were classified as positive or negative based on their ratings. After the experimental evaluations, the system achieved an accuracy score of 83%. We have followed a totally different approach in this work as our method is based on the scientific domain. Our approach is comprised of not only one classifier as well as our system accuracy score is better than [67].

Another work following the supervised approach is done by [96]. They have developed a system that distinguishes between sentence-level as well as contextual polarity. In this work, their data set was comprised of 8984 sentences extracted from 425 documents. Their method gave 76 % accuracy.

In sentimental analysis, researchers used semi-supervised and unsupervised learning approaches. The importance of the ML approaches is based on the need and specific scenario. In the sentiment analysis of English text, the impact of an adjective in sentence is of potential effect. First, there should be the identification of adjective orientation in a sentence. The orientation of adjectives will decide the state of a sentence whether should be positive or negative. One such work that contributed towards the identification of adjective orientation is done by [28] and they followed the unsupervised approach. They presented a method for identifying adjectives' semantic orientation in a sentence. They suggested that orientation' information depends upon the conjunction between adjectives, where AND refers to similar conjunction e.g. "Fair and Honest" while BUT refers to different orientation e.g. "Simple but popular". They used a well-known lexicon named Wall Street Journal Corpus for extracting the conjunctions of adjectives. For the sake of determining orientation, they used a log-linear model and achieved 78% accuracy of the system. Similar work was done by [92] following the unsupervised mechanism, also worked on the orientation of words. They found out the estimated Point Mutual Information of each phrase to calculate semantic orientation and the system achieved an accuracy score of 74%. [92] extracted the sentences that contain adjectives using the POS tag pattern lexicon. They found that the large size of lexicon can be better to achieve outperform classification results.

One such work is presented by [87] in which expansion towards the lexicon is considered by using the concept of text position. They used the position of text in order to expand lexicon to get better-classified results. They used the concept of assigning weights to the parts of the text. They tend to assign more weight to more subject-oriented part while less weight to the less subject-oriented part. This method achieved an accuracy score of 65%, later they found that this technique is not as efficient as they expected. The authors expanded the lexicon by measuring the co-occurrence of words inside the sentence. For the classification of data semantic orientation of each sentence is calculated, and by applying the density estimation positive and negative polarities are assigned to sentences. This system achieved an accuracy score of about 90%.

III. Methodology

The purpose of the methodology is defined in this section. Our methodology is depicted in Fig. 1. First of all, we used the annotated dataset prepared by [12] mentioned in section V. We used python based machine learning library named Scikit-Learn [68] for implementing the system**.** Scikit-Learn is a well-known machine learning library tightly integrated with Python language and provides easy-to-interact interface [68]**.** First of all our system reads the data stored in the file having(Tab Separated Values) format. After reading, preprocessing phase is applied to clean and prepare the data for the use of machine learning algorithms. Directly text data cannot be given to machine learning algorithms, it should be converted into a suitable type. Using Scikit-Learn module named "count vectorizer", the text data firstly convert into numeric format and prepare the matrix of tokens count, now the data is ready for machine learning algorithms. Then 60% of data is splitted randomly to train the classifier and 40% for testing the classifier' accuracy. We perform our experiments in two phases, firstly we just apply N-grams (Length 1-3) features on data and compute accuracies using equation (1) and equation (2). Secondly, in order to improve the accuracy scores, we apply other features like(stop words & punctuation removal, lemmatizationnm, etc.) along with n-grams and then again compute the accuracies. The latter approach helps to reduce the noise and complexity of the data. Thirty iterations of each experiment were conducted to compute average results and a total of six experiments were performed. After computing the accuracies of each phase, we then select the best feature which is giving the best result and which classifier is better in a specific scenario.

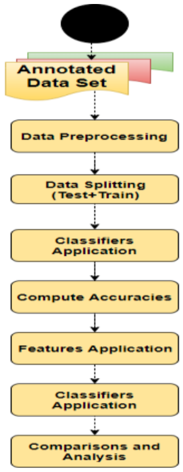

Fig. 1. Step by Step Flow of System Working.

IV. Evaluation Metrics

The evaluation of any research product decides the status and quality of that specific research work. This section briefly describes the metrics used to evaluate the sentimental analysis system we developed. The performance of sentimental analysis system is evaluated by computing the accuracy of the classification results given by the system. Accuracy of the system is to be mentioned in the form of some units that include F-score and Accuracy score. In our evaluation phase, we have calculated both Macro-F Score as well as Micro-F Score. Where FP is considered an error of type-1 and FN is considered an error of type-2. F-score is commonly used, a harmonic mean between precision and recall.

$$\text{Fscore} = (1+\beta^2)\frac{\text{Precision}*\text{Recall}}{\beta^2*\text{Precision}+\text{Recall}} \quad (1)$$

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FN}+\text{FP}+\text{TN}} \quad (2)$$

V. Corpus Construction

As mentioned earlier, we have used the data set prepared by [9]. However, we are going to highlight the process of corpus construction. As the authors restricted themselves to the field of computational logistics**,** they preferred to use the ACL (Association for Computational Linguistics) anthology mentioned by [16][29]. This digital archive contains journal and conference papers in PDF format since 1965 [15]. At the time of work done by the [9] the archive contains about 21,800 papers. The ACL anthology neither provides fully machine-readable text nor citation information that was a problem so this problem was solved by the resource of ACL anthology that provides the paper text converted from PDF using automated tools. The detailed corpus data is consists of 8736 citation sentences.

A. *Citation Sentiment Annotations*

Data annotation was done according to some defined rules. Citation sentences are classified into 3-classes positive, negative and neutral. The guidelines used for the annotation are as follows:

B. *Annotation Guidelines Negative*

- If direct mention of the problem or shortcoming of cited paper.
- If citing paper improves upon the cited paper.
- If citing paper gives outperform evaluation than cited paper.

Positive:

- If direct mention of the positive attribute of the cited paper.
- If citing paper not improves upon the cited paper.
- If citing paper gives not outperform evaluation than cited paper.

Neutral:

- If neither positive nor negative sentiment regarding the citation sentence is mentioned it will be tagged as neutral.

C. *Total Annotated Corpus Statistics*

The final sentiment corpus consists of 8736 sentences which were annotated using the above-mentioned rules. Here is the statistics of the sentiment annotated sentiment corpus in Table I.

TABLE. I. CORPUS ANNOTATION STATISTICS

Class	**Count**	**Percentage**
Positive	829	9.5 %
Negative	280	3.2 %
Neutral	7,627	87.3 %
Total	8,736	100 %

VI. CLASSIFICATION PROCESSING

This section gives brief details about the classification process used in this paper. The classification process is comprised of multiple processes like Data Pre-processing, Features Selections and classification classifiers used are discussed in details.

A. *Data Pre-Processing*

As mentioned in section V that corpus used for sentimental analysis classification is prepared or constructed by the [9]. This data set is comprised of a total of 8,736 citation sentences annotated as positive, negative, and neutral after applying rules. From total citation sentences, 60% of sentences were chosen randomly for training the classifier and the rest of 40% data was used for classifier' testing. The data set was cleaned to get the highest accuracy of the system. The following mentioned rules are used for normalization as shown in Table II.

TABLE. II. DATA NORMALIZATION RULES

No	**Original Sign**	**Convert into**
1	,	No change
2	:	No change
3	^	No Change
4	(jing , 2008)	<CIT>
5	(	-LRB-
6	)	-RRB-
7	[	-LRB-
8	]	-RRB-
9	{	-LCB-
10	}	-RCB-
11	.	Eliminate
12	\	Eliminate
13	|	Eliminate
14	**%**	**\%**
15	**Successive Citation**	**<OTH>**
16	‘	\’
17	*	*

B. *Features Selection*

For the sake of developing a system for sentiment analysis, different features are provided by ML framework [67][96][54]. We have used various features e.g. lemmatization, n-grams, stop words and term-document frequency to evaluate the classifier' accuracy. Later the evaluation results will be displayed.

C. *Lemmatization*

Lemmatization is a process of normalizing the inflected forms of words [70]. Homographic words cause ambiguity that disturbs searching accuracy and this ambiguity may also occur due to inflectional word forms [44]. For instance, words like “Talking”, “Talks” and “Talked” are the inflected forms of the word “Talk”. The process of lemmatization and stemming is similar with minor changes [70], while the benefits of both approaches are the same. We have applied only lemmatization and avoid stemming due to the problems of stemming process. The stemming process is worthwhile for short retrieval lists [11][27][34], while our system has to deal with large data set and processing lists so we did not apply stemming. Stemming performs normalization of inflected words by keeping different variations of words along with their derivation process [4][46][69]. The stemming process produces more potential results for the languages other than English – for example, Slovenian [71], French, modern Greek [41], Arabic [1] and Swedish [74], because other languages include less inflected form of words than English. In our case, we are dealing with the data set containing the Citation Sentences written in the English language.

D. N-Grams

N-grams refer to the combination of sequenced words in a text, where n means the number of words in that combination. If the N = 1, then it means a single word in a text if N = 2 then it leads to the combination of two sequenced words. We used 3 different kinds of N-grams in our classification that generate different results. The example of these N-grams with different values of N based on the sentence “I like to do research” is given in Table III.

Author in [67] claimed that uni-grams and bi-grams performed well for movie reviews data. In our work, we have applied tri-grams because tri-grams play a substantial part in scientific text [40].

E. Stop Words and Punctuation

English text contains a lot of meaningless and non-informative words [52] called stop words. These are not required in classification because their presence just increase the size of data. So we applied stop words removal technique in order to cleanse the data for better and efficient classification [80]. Some research works support the stop words removal from the data set to reduce the dimensions of data [13][7][66][24][45][83], while some researchers are against the removal of stop words because these words contain sentiment information [75][57][32][33]. The earliest work that contributes to the removal of stop words by [53], in which they advised that words can be categorized into two types (i) keywords and (ii) non-keywords and the latter were called as stop words. There are also pre-compiled stop word lists such as Van, Brown [22], called classic stop or standard stop lists. Later these stop word lists are criticized for being out dated [82][52]. We have used the latest and up to date, NLTK stop words list that provides 180 plus stopwords.

F. Term Document Frequency

Term document frequency refers to the count of specific words in the document [99]. We also used the concept of finding the term document using vectorizer.

G. Classification Classifiers

After preprocessing and features selection the very next step is to apply classification algorithms. Many text classifiers have been purposed in literature [19][36]. We have used 6 algorithms of machine learning including Naïve-Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbor (KNN), and Random Forest (RF).

a) Naïve Bayes: Naïve- Bayes is the most popular classification algorithm due to its simplicity and effectiveness [42][76]. This classifier works according to the concept of Bayes theorem [26]. It’s a kind of module classifier [102] that follows the idea of probabilities for the purpose of classification. Bernoulli and multinomial are the models of naïve Bayes classifier [49][2][58], Binarized Naïve Bayes model is described by [26].

b) Support Vector Machine: In the world of machine learning one such supervised learning algorithm that achieves enough improvements on a variety of tasks is a Support vector machine classifier [37]. Particularly in the case of analyzing the sentiments, SVM has demonstrated good results [67][96][54][43]. In-text classification the SVM contributes towards excellent precision scores while poor recall scores while adjusting the thresholds recall scores can be adjusted [36]. Adjustment of thresholds is of vital importance, a study by [78] described the mechanism of automatically adjusting the thresholds of SVM.

TABLE. III. N-GRAM EXAMPLE

N values	Called	Example
N=1	Unigram	I, like, to, do, research
N=2	Bi-gram	I like, like to, to do, do research
N=3	Trigram	I like to, like to do, to do research

c) Decision Tree: In various fields of text classification the use of decision tree classifier can be seen and analyzed [61]. Its popularity is based on the nature of classification rules that make it interesting for NLP researchers [14]. The decision is constructed by selecting the data from the data set randomly [3]. The information gain is calculated for all values and the feature with the highest information gain value becomes the tree's root [59] and the whole tree is constructed by finding the features for the next level again and again. The fast decision tree algorithm is developed by [38]. So the solution for such kind of a scenario is presented by [60].

d) Random Forest: [18] mentioned the importance of a random forest classifier and compared its performance with the other classifiers. [5][18] claimed that the random forest algorithm provides efficient and discriminative classification, as a result, it is considered an interesting classifier. [48][65] were the first who discussed the importance of random forest classifier in the field of computer vision. [97][79] Introduced class recognition based on random forest. [100][101] used random forest for bi-layer video segmentation, [17] used it for image classification, and [6] used it for personal identification.

e) K-th Nearest Neighbour: KNN is a simple and efficient classifier [88]. Called lazy learner because its training phase contains nothing but storing all the training examples as classifiers [77]. KNN requires a lot of memory while storing the training values [59]. The performance issue of KNN can also be solved by efficient estimations of parameters [51].

VII. RESULTS

For the sake of performing the experimental task, we have used the data set mentioned in Section V. The data is labeled using positive, negative, and neutral classes using annotated rules mentioned in Section V. Different machine learning algorithms used for the classification discussed in Section VI. The evaluation metrics mentioned in Section IV (Equation 1, 2) were used to validate the system. The detailed description of the experimental results using evaluation metrics is defined in Table IV, and Table V. In these tables terms, A1, B1, C1 denotes simply unigram, bigram, trigram features while A2, B2, and C2 denote the application of unigram, bigram, trigram along with other features. Table IV shows that Overall DT using n-grams gives the best F-score in macro while RF is best in case of micro average. LR is also overall best in the micro average without applying extra features. Uni-gram plays support in better performance of LR and DT, uni-gram along with other features plays significant performance in NB, KNN, and RF. DT gives better performance in the case of uni-grams, bi-grams, and tri-grams. LR performance is significant in case of uni-grams only, k-th nearest neighbor outperforms in case of n-grams along with other features and give worst performs without other features while RF performs best as same as KNN. The overall discussion describes that uni-gram, bi-gram, and tri-gram without other features perform best where uni-gram is at first position.

Table V shows that Overall SVM, LR, and RF performed very best with the highest accuracy scores. N-grams play significant performance in NB, SVM gives the best accuracy using uni-gram, LR performance is significant in case of bi-grams and tri-grams, KNN outperforms in case of n-grams without other features and gives worst performs with other features. The overall discussion describes that uni-gram, bi-grams, and tri-grams without other features performs best and give significant accuracy scores.

TABLE. IV. F-SCORES AFTER THIRTY ITTERATIONS

Features	NB		SVM		LR		DT		KNN		RF	
	Macro Scores %	Micro Scores %	Macro Scores %	Micro Scores %	Macro Scores %	Micro Scores %	Macro Scores %	Micro Scores %	Macro Scores %	Micro Scores %	Macro Scores %	Micro Scores %
A1	36	87	37	88	49	88	49	85	33	87	44	88
A2	49	83	48	87	46	87	48	85	34	87	46	88
B1	34	87	31	87	46	88	49	86	32	87	44	88
B2	46	79	47	87	46	87	48	85	34	87	46	88
C1	36	87	31	87	44	88	49	86	32	87	42	88
C2	45	77	46	87	46	87	48	85	34	87	46	88

TABLE. V. ACCURACY SCORES AFTER THIRTY ITTERATIONS

Features	NB %	SVM %	LR %	DT %	KNN %	RF %
A1	87	87	87	87	87	88
A2	83	88	84	87	86	88
B1	87	88	88	87	87	88
B2	79	88	85	87	86	88
C1	87	88	88	87	87	88
C2	77	88	86	87	86	88

VIII. CONCLUSION

In this research work, we presented a sentiment analysis system for scientific text. We have used different machine learning classifiers namely NB, SVM, DT, LR, KNN and RF along with different features to process the data and optimize the classification results. Experiments are performed on the data set prepared by the [9]. Data set is partitioned into training and testing sets according to the ratio of 60:40 and. Accuracies of the classifiers are computed by using various evaluation metrics like F-score, and Accuracy score. The results show that SVM performs better than other classifiers. After SVM Naïve Bayes performs well. In the case of the macro average, the performance of SVM classifier is best while computing F-score, and accuracy measures while the random forest is best in case of micro average. Uni-grams, bi-grams, and tri-gram features performed very well and support the classifiers to achieve highest accuracy scores.

We compared our findings with [9] in the experimental phase based on different features. We used the n-grams approach together with the lemmatization process to reduce the data dimensions as the latter approach was not applied by the [9]. Table VI describes the comparative analysis of our work and the work of [9].

The author of [9] used NB and SVM classifier and compute the accuracies of the system using an F-score. In this paper, we have implemented six classifiers LR, DT, KNN, and RF including NB, SVM used by [9]. We computed the accuracies by increasing the number of evaluation metrics F-score and accuracy including F-score used by [9] to evaluate the accuracies with the base system. Our results showed significant improvement like in the case of Naïve Bayes using uni-gram feature we achieved micro-F 87% while the base system described the result of micro-F = 78% and our results are approximately 9 % better.

Macro-F scores using uni-gram mentioned in the research work of [9] is 48% and we achieved the macro-F = 49% by reducing the data dimensions by using the lemmatization process and stop words removal mechanism. Based on bi-gram and tri-gram features our system achieved the same result of micro-F = 87%. The micro-F of [9] based on bi-gran and tri-gram features decreased from 78% to 76%. In our case, the micro-F based on bi-gram and tri-gram features increased by 11 %. While in the case of bi-gram and trigram research work of [9] showed the macro-F score of 47%, where our method achieved a macro-F score of 46% using bi-gram and 45% using tri-gram. Overall using Naïve Bayes classifier [9] work achieved maximum of (micro-F score = 78%, macro-F score = 48) while we improved our results to extant and achieved maximum of (micro-F = 87%, macro-F = 49%) that shows the significant improvement of our work.

TABLE. VI. COMPARATIVE ANALYSIS OF OUR WORK WITH [9]

Characteristics	[9]	Our Method
Classification algorithms used	2	6
Evaluation metrics used	1	2
Naïve Bayes F-scores	78 %	87 %
SVM F-scores	86 %	88 %

The second classifier used by the [9] is SVM. We also implemented SVM based on the same features and our results outperform [9]. In the case of using uni-gram, bi-gram and tri-gram features in SVM classifier the author [9] reported the micro-F = 86% and our result with micro-F = 88% for uni-gram and shows the significant improvement.

We also implemented other classifiers like DT, LR, KNN, and RF and achieved significant results. LR and DT performed well. In the case LR, we achieved the micro-F score = 88% and for DT micro-F score = 87% and for both macro-F score = 49%. If we compare this result with [9] results than the F-scores show the improvement of approximately 2% for macro-F scores for both LR and DT and approximately 10% for micro-F-score of LR. KNN and RF classifiers give improved results than [9] in the case of an F-score micro average with the improvement of approximately 10 %.

ACKNOWLEDGMENT

First of all, authors would like to thank Almighty Allah for His grace and blessings. Authors would also like to pay thanks their parents, friends and colleagues for their support and encouragement.

REFERENCES

[1] Abu-Salem, H., Al-Omari, M., & Evens, M. W. (1999). Stemming methodologies over individual query words for an Arabic information retrieval system. Journal of the American Society for Information Science, 50(6), 524-529.

[2] Aghila, G. (2010). A Survey of Na\" ive Bayes Machine Learning approach in Text Document Classification. arXiv preprint arXiv:1003.1795.

[3] Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision trees. International Journal of Computer Science Issues (IJCSI), 9(5), 272.

[4] Alkula, R. (2001). From plain character strings to meaningful words: Producing better full text databases for inflectional and compounding languages with morphological analysis software. Information Retrieval, 4(3-4), 195-208.

[5] Amit, Y., & Geman, D. (1997). Shape quantization and recognition with randomized trees. Neural computation, 9(7), 1545-1588.

[6] Apostolof, N., & Zisserman, A. (2007, September). Who Are You?-Real-time Person Identification. In BMVC (pp. 1-10).

[7] Asiaee T, A., Tepper, M., Banerjee, A., & Sapiro, G. (2012, October). If you are happy and you know it... tweet. In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 1602-1606). ACM.

[8] Athar, A. (2011, June). Sentiment analysis of citations using sentence structure-based features. In Proceedings of the ACL 2011 student session (pp. 81-87). Association for Computational Linguistics.

[9] Athar, A. (2014). Sentiment analysis of scientific citations (No. UCAM-CL-TR-856). University of Cambridge, Computer Laboratory.

[10] Athar, A., & Teufel, S. (2012, July). Detection of implicit citations for sentiment detection. In Proceedings of the Workshop on Detecting Structure in Scholarly Discourse (pp. 18-26). Association for Computational Linguistics.

[11] Avoy, J. (1999). A stemming procedure and stopword list for general French corpora. Journal of the American Society for Information Science, 50(10), 944-952.

[12] Awais Athar. Citation Context Corpus. http://www.cl.cam.ac.uk/ ~aa496/citation-context-corpus/. Accessed: 2015-05-13.2,5).

[13] Bakliwal, A., Arora, P., Madhappan, S., Kapre, N., Singh, M., & Varma, V. (2012). Mining sentiments from tweets. In Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis (pp. 11-18).

[14] Ben-Haim, Y., & Tom-Tov, E. (2010). A streaming parallel decision tree algorithm. Journal of Machine Learning Research, 11(Feb), 849-872.

[15] Bienz, T., Cohn, R., & Adobe Systems (Mountain View, Calif.). (1993). Portable document format reference manual (p. 214). Boston^ eMA MA: Addison-Wesley.

[16] Bird, S., Dale, R., Dorr, B. J., Gibson, B., Joseph, M. T., Kan, M. Y., & Tan, Y. F. (2008). The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics.

[17] Bosch, A., Zisserman, A., & Munoz, X. (2007, October). Image classification using random forests and ferns. In 2007 IEEE 11th international conference on computer vision (pp. 1-8). Ieee.

[18] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[19] Breiman, L. (2017). Classification and regression trees. Routledge.

[20] Chubin, D. E., & Moitra, S. D. (1975). Content analysis of references: Adjunct or alternative to citation counting?. Social studies of science, 5(4), 423-441.conference on Empirical methods in natural language processing (pp. 129-136). Association for Computational Linguistics.

[21] Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?. The Journal of Machine Learning Research, 15(1), 3133-3181.

[22] Fox, C. J. (1992). Lexical Analysis and Stoplists.

[23] Garzone, M., & Mercer, R. E. (2000, May). Towards an automated citation classifier. In Conference of the canadian society for computational studies of intelligence (pp. 337-346). Springer, Berlin, Heidelberg.

[24] Gokulakrishnan, B., Priyanthan, P., Ragavan, T., Prasath, N., and Perera, A. (2012). Opinion mining and sentiment analysis on a twitter data stream. In Advances in ICT for Emerging Regions (ICTer), 2012 International Conference on, pages 182–188. IEEE.

[25] Goodwin, J. (1980). Eugene Garfield, Citation Indexing-Its Theory and Application in Science, Technology, and Humanities (Book Review). Technology and Culture, 21(4), 714.

[26] Gupte, A., Joshi, S., Gadgul, P., Kadam, A., & Gupte, A. (2014). Comparative study of classification algorithms used in sentiment analysis. International Journal of Computer Science and Information Technologies, 5(5), 6261-6264.

[27] Harman, D. (1991). How effective is suffixing?. Journal of the american society for information science, 42(1), 7-15.

[28] Hatzivassiloglou, V., & McKeown, K. R. (1997, July). Predicting the semantic orientation of adjectives.

[29] In Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics (pp. 174-181). Association for Computational Linguistics.

[30] Hernández-Alvarez, M., & Gomez, J. M. (2016). Survey about citation context analysis: Tasks, techniques, and resources. Natural Language Engineering, 22(3), 327-349.

[31] Hernández-Alvarez, M., & Gómez, J. M. (2015, October). Citation impact categorization: for scientific literature. In 2015 IEEE 18th International Conference on Computational Science and Engineering (pp. 307-313). IEEE.

[32] Hu, X., Tang, J., Gao, H., & Liu, H. (2013, May). Unsupervised sentiment analysis with emotional signals. In Proceedings of the 22nd international conference on World Wide Web (pp. 607-618). ACM.

[33] Hu, X., Tang, L., Tang, J., & Liu, H. (2013, February). Exploiting social relations for sentiment analysis in microblogging. In Proceedings of the

sixth ACM international conference on Web search and data mining (pp. 537-546). ACM.

[34] Hull, D. A. (1996). Stemming algorithms: A case study for detailed evaluation. Journal of the American Society for Information Science, 47(1), 70-84.

[35] Hyland, K. (1995). The Author in the Text: Hedging Scientific Writing. Hong Kong papers in linguistics and language teaching, 18, 33-42.

[36] Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. WSEAS transactions on computers, 4(8), 966-974.

[37] Joachims, T. (1998, April). Text categorization with support vector machines: Learning with many relevant features. In European conference on machine learning (pp. 137-142). Springer, Berlin, Heidelberg.

[38] Johnson, D. E., Oles, F. J., Zhang, T., & Goetz, T. (2002). A decision-tree-based symbolic rule induction system for text categorization. IBM Systems Journal, 41(3), 428-437.

[39] Jurgens, D., Kumar, S., Hoover, R., McFarland, D., & Jurafsky, D. (2016). Citation classification for behavioral analysis of a scientific field. arXiv preprint arXiv:1609.00435.

[40] Justeson, J. S., & Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. Natural language engineering, 1(1), 9-27.

[41] Kalamboukis, T. Z. (1995). Suffix stripping with modern Greek. Program, 29(3), 313-321.

[42] Kim, S. B., Rim, H. C., Yook, D., & Lim, H. S. (2002, August). Effective methods for improving naive bayes text classifiers. In Pacific Rim International Conference on Artificial Intelligence (pp. 414-423). Springer, Berlin, Heidelberg.

[43] Korde, V., & Mahender, C. N. (2012). Text classification and classifiers: A survey. International Journal of Artificial Intelligence & Applications, 3(2), 85.

[44] Korenius, T., Laurikkala, J., Järvelin, K., & Juhola, M. (2004, November). Stemming and lemmatization in the clustering of finnish text documents. In Proceedings of the thirteenth ACM international conference on Information and knowledge management (pp. 625-633). ACM.

[45] Kouloumpis, E., Wilson, T., & Moore, J. (2011, July). Twitter sentiment analysis: The good the bad and the omg! In Fifth International AAAI conference on weblogs and social media.

[46] Krovetz, R. (1993, July). Viewing morphology as an inference process. In Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 191-202). ACM.

[47] Latour, B. (1987). Science in action: How to follow scientists and engineers through society. Harvard university press.

[48] Lepetit, V., & Fua, P. (2006). Keypoint recognition using randomized trees. IEEE transactions on pattern analysis and machine intelligence, 28(9), 1465-1479.

[49] Lewis, D. D. (1998, April). Naive (Bayes) at forty: The independence assumption in information retrieval.

[50] In European conference on machine learning (pp. 4-15). Springer, Berlin, Heidelberg.

[51] Lim, H. S. (2004, November). Improving kNN based text classification with well estimated parameters. In International Conference on Neural Information Processing (pp. 516-523). Springer, Berlin, Heidelberg.

[52] Lo, R. T. W., He, B., & Ounis, I. (2005, January). Automatically building a stopword list for an information retrieval system. In Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR)(Vol. 5, pp. 17-24).

[53] Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. IBM Journal of research and development, 1(4), 309-317.

[54] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1 (pp. 142-150). Association for Computational Linguistics.

[55] MacRoberts, M. H., & MacRoberts, B. R. (1984). The negational reference: Or the art of dissembling. Social Studies of Science, 14(1), 91-94.

[56] Malik, M. K. (2017). Urdu named entity recognition and classification system using artificial neural network. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 17(1), 2.

[57] Martínez-Cámara, E., Montejo-Ráez, A., Martín-Valdivia, M. T., & Ureña-López, L. A. (2013). Sinai: Machine learning and emotion of the crowd for sentiment analysis in microblogs. In Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013) (Vol. 2, pp. 402-407).

[58] McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification. In AAAI-98 workshop on learning for text categorization (Vol. 752, No. 1, pp. 41-48).

[59] Mehmood, K., Essam, D., & Shafi, K. (2018, July). Sentiment Analysis System for Roman Urdu. In Science and Information Conference (pp. 29-42). Springer, Cham.

[60] Mehta, M., Agrawal, R., & Rissanen, J. (1996, March). SLIQ: A fast scalable classifier for data mining. In International conference on extending database technology (pp. 18-32). Springer, Berlin, Heidelberg.

[61] Mitchell, T. M. (1997). Does machine learning really work?. AI magazine, 18(3), 11.

[62] Moravcsik, M. J., & Murugesan, P. (1988). Some Results on the Function and Quality of Citations: Social Studies of Science. 研究 技術 計画, 3(4), 538.

[63] Murray, J. (2015). Finding Implicit Citations in Scientific Publications: Improvements to Citation Context Detection Methods.

[64] Oppenheim, C., & Renn, S. P. (1978). Highly cited old papers and the reasons why they continue to be cited. Journal of the American Society for Information Science, 29(5), 225-231.

[65] Ozuysal, M., Fua, P., & Lepetit, V. (2007, June). Fast keypoint recognition in ten lines of code. In 2007 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-8). Ieee.

[66] Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In LREc (Vol. 10, No. 2010, pp. 1320-1326).

[67] Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up? sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10(pp. 79-86). Association for Computational Linguistics.

[68] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.

[69] Pirkola, A. (2001). Morphological typology of languages for IR. Journal of Documentation, 57(3), 330-348.

[70] Plisson, J., Lavrac, N., & Mladenic, D. (2004). A rule based approach to word lemmatization. Proceedings of IS-2004, 83-86.

[71] Popovič, M., & Willett, P. (1992). The effectiveness of stemming for natural-language access to Slovene textual data. Journal of the American Society for Information Science, 43(5), 384-390.

[72] Qazvinian, V., & Radev, D. R. (2010, July). Identifying non-explicit citing sentences for citation-based summarization. In Proceedings of the 48th annual meeting of the association for computational linguistics (pp. 555-564). Association for Computational Linguistics.

[73] Ritchie, A. (2009). Citation context analysis for information retrieval (No. UCAM-CL-TR-744). University of Cambridge, Computer Laboratory.

[74] Rosell, M. (2003). Improving clustering of Swedish newspaper articles using stemming and compound splitting. In NoDaLiDa 2003, Reykjavik, Iceland 2003 (pp. 1-7).

[75] Saif, H., He, Y., & Alani, H. (2012, November). Semantic sentiment analysis of twitter. In International semantic web conference (pp. 508-524). Springer, Berlin, Heidelberg.

[76] Schneider, K. M. (2005, February). Techniques for improving the performance of naive bayes for text classification. In International

Conference on Intelligent Text Processing and Computational Linguistics (pp. 682-693). Springer, Berlin, Heidelberg.

[77] Sebastiani, F. (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), 1-47.

[78] Shanahan, J. G., & Roma, N. (2003, September). Improving SVM text classification performance through threshold adjustment. In European Conference on Machine Learning (pp. 361-372). Springer, Berlin, Heidelberg.

[79] Shotton, J., Johnson, M., & Cipolla, R. (2008, June). Semantic texton forests for image categorization and segmentation. In 2008 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-8). IEEE.

[80] Silva, C., & Ribeiro, B. (2003, July). The importance of stop word removal on recall values in text categorization. In Proceedings of the International Joint Conference on Neural Networks, 2003. (Vol. 3, pp. 1661-1666). IEEE.

[81] Simkin, M. V., & Roychowdhury, V. P. (2002). Read before you cite! arXiv preprint cond-mat/0212043.

[82] Sinka, M. P., & Corne, D. W. (2003, October). Towards modernised and web-specific stoplists for web document analysis. In Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003) (pp. 396-402). IEEE.

[83] Speriosu, M., Sudan, N., Upadhyay, S., & Baldridge, J. (2011, July). Twitter polarity classification with label propagation over lexical links and the follower graph. In Proceedings of the First workshop on Unsupervised Learning in NLP (pp. 53-63). Association for Computational Linguistics.

[84] Spiegel-Rosing, I. (1977). Science studies: Bibliometric and content analysis. Social Studies of Science, 7(1), 97-113.

[85] Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. I. (2012, April). BRAT: a web-based tool for NLP-assisted text annotation. In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (pp. 102-107). Association for Computational Linguistics.

[86] Swales, J. (1986). Citation analysis and discourse analysis. Applied linguistics, 7(1), 39-56.

[87] Taboada, M., & Grieve, J. (2004, March). Analyzing appraisal automatically. In Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (AAAI Technical Re# port SS# 04# 07), Stanford University, CA, pp. 158q161. AAAI Press.

[88] Tan, S. (2006). An effective refinement strategy for KNN text classifier. Expert Systems with Applications, 30(2), 290-298.

[89] Teufel, S. (1999). Argumentative zoning: Information extraction from scientific text (Doctoral dissertation, University of Edinburgh).

[90] Teufel, S., Siddharthan, A., & Tidhar, D. (2006). An annotation scheme for citation function.

[91] Teufel, S., Siddharthan, A., & Tidhar, D. (2006, July). Automatic classification of citation function. In Proceedings of the 2006 conference on empirical methods in natural language processing (pp. 103-110). Association for Computational Linguistics.

[92] Turney, P. D. (2002, July). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 417-424).Association for Computational Linguistics.

[93] Valenzuela, M., Ha, V., & Etzioni, O. (2015, April). Identifying meaningful citations. In Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence.

[94] White, H. D. (2004). Citation analysis and discourse analysis revisited. Applied linguistics, 25(1), 89-116.

[95] White, H. D., Wellman, B., & Nazer, N. (2004). Does citation reflect social structure?: Longitudinal evidence from the “Globenet” interdisciplinary research group. Journal of the American Society for information Science and Technology, 55(2), 111-126.

[96] Wilson, T., Wiebe, J., & Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. Computational linguistics, 35(3), 399-433.

[97] Winn, J., & Criminisi, A. (2006). Object class recognition at a glance. In Video Proc. CVPR.

[98] Xia, R., Zong, C., & Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. Information Sciences, 181(6), 1138-1152.

[99] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In International Conference on Machine Learning, pages 412–420, 1997.

[100] Yanjun Qi., “Random Forest for Bioinformatics”. www.cs. cmu.edu/~qyj/papersA08/11-rfbook.pdf.

[101] Yin, P., Criminisi, A., Winn, J., & Essa, I. (2007, June). Tree-based classifiers for bilayer video segmentation. In 2007 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-8). IEEE.

[102] Yong-feng, S., & Yan-ping, Z. (2004). Comparison of text categorization algorithms. Wuhan university Journal of natural sciences, 9(5), 798-804.