Automated citation sentiment analysis using high order n-grams: a preliminary investigation

MUHAMMAD TOUSEEF IKRAM

MUHAMMAD TANVIR AFZAL

NAVEED ANWER BUTT

Recommended Citation

IKRAM, MUHAMMAD TOUSEEF; AFZAL, MUHAMMAD TANVIR; and BUTT, NAVEED ANWER (2018) "Automated citation sentiment analysis using high order n-grams: a preliminary investigation," *Turkish Journal of Electrical Engineering and Computer Sciences*: Vol. 26: No. 4, Article 20. https://doi.org/10.3906/elk-1712-24
Available at: https://journals.tubitak.gov.tr/elektrik/vol26/iss4/20

Automated citation sentiment analysis using high order n-grams: a preliminary investigation

Muhammad Touseef IKRAM[1],[*] , **Muhammad Tanvir AFZAL**[1], **Naveed Anwer BUTT**[2]
[1]Department of Computer Science, Capital University of Science and Technology, Islamabad, Pakistan
[2]Department of Computer Science, University of Gujrat, Gujrat, Pakistan

Abstract: Scientific papers hold an association with previous research contributions (i.e. books, journals or conference papers, and web resources) in the form of citations. Citations are deemed as a link or relatedness of the previous work to the cited work. The nature of the cited material could be supportive (positive), contrastive (negative), or objective (neutral). Extraction of the author's sentiment towards the cited scientific articles is an emerging research discipline due to various linguistic differences between the citation sentences and other domains of sentiment analysis. In this paper, we propose a technique for the identification of the sentiment of the citing author towards the cited paper by extracting unigram, bigram, trigram, and pentagram adjective and adverb patterns from the citation text. After POS tagging of the citation text, we use the sentence parser for the extraction of linguistic features comprising adjectives, adverbs, and n-grams from the citation text. A sentiment score is then assigned to distinguish them as positive, negative, and neutral. In addition, the proposed technique is compared with manually classified citation text and 2 commercial tools, namely SEMANTRIA and THEYSAY, to determine their applicability to the citation corpus. These tools are based on different techniques for determining the sentiment orientation of the sentence. Analysis of the results shows that our proposed approach has achieved results comparable to the commercial counterparts with average precision, recall, and accuracy of 90%, 81.82%, and 85.91% respectively.

Key words: Citation sentiment analysis, n-gram analysis, citation classification, SEMANTRIA, THEYSAY

1. Introduction

Due to the exponential growth of research publications on the Internet, the frequency of citations to scientific texts has become very high [1,2]. Citation text is valuable and of utmost importance for the qualitative assessment of a paper, but the sheer size of it makes access prohibitive [3–6]. To be able to leverage the available information, there is a need for some automatic mechanism for summarizing and processing scholarly big data [7,8]. In the current research repositories, there is no existing technique that summarizes research citations on the basis of sentiments expressed in them [9]. However, applying sentiment analysis to citation contexts is a challenging process as the criticism is often hidden, negative citations are hardly explicit, and most research papers are considered positive in general [10,11]. Similarly, additional academic pressure for more research publications coupled with increased difficulty of publishing also led to citation cartels in the form of unclean citations [12]. In addition, bibliometric studies have found a high increase in the citation of controversial

*Correspondence: touseefgrw@hotmail.com

papers. Moreover, the linguistic differences in scientific texts, such as use of technical terms, variation in lexical terms, hidden sentiment, and contrastive expressions, urge the scientific community to undertake new methods for citation sentiment classification [13,14].

In the domain of natural language processing and probability, n-grams are widely being utilized as features for opinion mining [15,16]. N-grams are the contiguous and continuous sequences of n-items, terms, or objects from a given document, sequence of text, or speech, which are used as a feature to get sentiment cues from the text [17,18]. Most contemporary studies have harnessed bigrams and trigrams [19–22]. However, the presence of technical and lexical terms in the citation context creates an extreme challenge for citation sentiment identification, which can be addressed with higher order n-grams that assist in capturing short-term contextual and positional information [23–25]. Therefore, in the proposed technique, we assume that higher order n-gram phrases, part of speech (POS) tagging, dependency relations, bag-of-words (BOW) models, and sentiment lexicons can play significant roles in improving classification accuracy.

Based on the above arguments, the following are the main contributions of this paper:

- The focal point is the identification of the sentiment used in context with reference to a particular citation by using different n-gram techniques: unigrams, bigrams, trigrams, and pentagrams comprising (adjective, adverbs, and adjective + adverb) patterns. For this, we have combined text mining and lexical analysis techniques to derive the opinion of the citing author towards the cited research paper from the citation text. First the sentiment-bearing words are extracted from the citation sentences using the linguistic phrase patterns, synonyms, and heuristic rule-based approach. Afterwards, we determine sentiment orientation of extracted sentences using SentiWordNet by contemplating the words around the linguistic expression. These sentiment orientations are further classified as positive, negative, and neutral based on the sentiment score.

- Secondly, we intend to scrutinize the applicability of the existing commercial sentiment analysis tools (SEMANTRIA and THEYSAY) in addressing the complexity associated with the citation genre. The attained results of the proposed approach are compared with 2 commercial tools to determine the effectiveness of our technique. The experimental evaluation has confirmed the effectiveness of using higher order n-grams (i.e. $n \geq 3$) in predicting the sentiment of scientific literature and demonstrated that the results are comparable to the commercially available tools. Overall, the results are evaluated by utilizing standard evaluation measures such as precision, recall, accuracy, and f-measure.

The rest of the paper is organized as follows: in Section 2, we present the state-of-the-art work done to evaluate the author's attitude towards scientific publications. Section 3 summarizes the proposed methodology for sentiment analysis of citation text. Results of experiments are presented in Section 4. In Section 5 of the paper, we conclude the research work based on the analysis of our experimental results.

2. Related work

In the literature, there exist various approaches that address different issues of citation sentiment analysis. In this section, we present the relevant literature for the sentiment classification of citations for scientific papers. In [26], the authors focused on the automatic sentiment polarity identification in the citation text using different features like n-grams ($n = 3$), dependency relations, negation features, scientific lexicon, and sentence splitting in the SVM framework. They categorized the citations into 3 different classes, i.e. positive, negative, and neutral. The results of their study revealed that trigrams and dependency relations provide robust results in this regard

and outperform the scientific lexicon and sentence splitting features. In [27], Kim and Thoma developed an automated sentiment detection method using machine learning techniques and linguistic clues. As a preliminary work, the authors presented a technique for citation text classification based on support vector machines using n-gram (unigrams and bigrams) word statistics as a feature vector. The proposed citation sentiment classification technique categorizes the text into 2 categories, i.e. positive and others. As a future work, they planned to improve their methodology by using a denser ground-truth dataset and enriching the feature set by considering more input features. A technique was proposed by Tandon and Jain in [28] for the generation of a structured summary of a research paper based on the citation text in citing papers. They classified the citation text using multilabel classification into one or more of 5 different classes: summary, strengths, limitations, related work, and applications. As a baseline, they used the naive Bayes algorithm while considering the combinations of adjectives, verbs, and n-grams in each class. As per the experimental results, the combination of adjectives, verbs, and bigrams achieved an average precision of 68.54%. Yu presented a technique in [29] for citation bias detection using manual citation sentiment analysis of biomedical research publications. He highlighted a number of differences between the approaches used by biomedical researchers and the automated citation sentiment analysis methods. According to the findings, researchers pay a lot of attention to citation sentiment aspects like strength and validity and have not vastly emphasized simple polarity-based sentiment. Similarly, in [30], a technique for context-based citation summary was presented based on the semantic similarity between the citing and cited articles. They utilized the surrounding text around the citation and the PageRank algorithm for calculating the qualitative citation index of the research papers. Butt et al. [31] found that many of the existing studies have been carried for citation sentiment classification in domain-specific areas like medicine, biomedicine, computer science, French humanities articles, etc. In their work, they classified the citations based on the naive Bayes classifier by selecting a window of 5 sentences around the cited text with an accuracy of 80%. For experimental purposes, they used generalized lexica that incorporate the citations from the multiple domains, which ascertained the demonstrability of their approach in multiple disciplines. Parthasarathy and Tomar [23,24] reported a literature survey related to journal citation sentiment analysis. They presented a framework for citation sentiment analysis, which consists of citation extraction from the paper, preprocessing of the extracted citations, feature extraction, and sentiment classification by application of different machine learning techniques like support vector machines, naive Bayes, and decision trees. According to the survey study, different sentiment classification techniques are being employed for term frequency, n-grams, negations, and lexica. In [32] an exploration of sentiment, polarity, and function analysis of the citations was performed. As per the findings, there is still plenty of room for further research and development in the domain of citation sentiment analysis. In [33] the authors proposed a method for the elucidation of an author's attitude towards the cited work by combining text mining and lexical analysis techniques. They encapsulated the extracted opinions in an objective measure for qualifying the impact of scientific publications. One of the key findings of their research work is that the majority of the citations were neutral in nature with considerable agreement in the utilization of the terminology for verbalizing the sentiments towards the cited work both in terms of positive and negative opinions. As a future study, they wanted to extend the model by considering different lexical elements and their different combinations such as conjunctions, adjectives, verbs, adverbs, and verb-adverb combinations.

3. Proposed system

In this section we describe the proposed research methodology, which is based on the following tasks: 1) extraction of the n-grams from the citation text; 2) determination of the sentiment score; 3) classification of the

citation text as positive, negative, or neutral; and 4) evaluation and comparison of the higher order n-grams' citation classification with the commercial counterparts.

3.1. Extraction of n-grams

We split the citation text into sentences and performed part-of-speech tagging on the gold-standard dataset [26]. A central and important aspect of sentiment analysis is the selection of good feature representation and determining the likelihood of tag occurrence on the basis of previous tag patterns. From the citation sentence, n-grams of different sizes (unigrams, bigrams, trigrams, and pentagrams) along with their associated tag patterns are extracted as an ordered sequence of words on the basis of criteria specified in Eq. (1). Unigrams are the BOWs separated by the spaces, whereas bigrams, trigrams, and pentagrams are features consisting of 2, 3, and 5 consecutive words. For example, in the sentence '*The proposed method outperformed the class based model*', the words 'the', 'proposed', 'method', 'outperformed', 'the', 'class', 'based', and 'model' are all distinct unigrams. Similarly, in the sentence *'the results are not competitive to the state-of-the-art systems'*, after removing the stop words 'the' and 'to', 'results are', 'are not', 'not competitive', 'competitive state-of-the-art', 'state-of-the-art systems' are distinct bigram features. The advantage of using these features is that they are capable of containing some contextual information. Considering contextual information in terms of higher order n-grams helps in capturing negations and subtle meanings in the form of implicit negations. Therefore, we have focused on using higher order n-gram features for citation sentiment classification. For higher order n-grams, the probability of a POS tag sequence is calculated as the product of conditional probabilities of its trigrams and pentagrams. If we denote the tag sequence as t_1, t_2, t_3, ...t_n and the corresponding word sequence as w_1, w_2, w_3, ...w_n, the above stated fact can be explained with Eq. (2).

$$L_{\min} \leq \text{Length (pos\_sequence)} \leq L_{\max} \tag{1}$$

$$P(t_i|w_i) = P(w_i|t_i).P(t_i|t_{i-4}, t_{i-3}, t_{i-2}, t_{i-1}) \tag{2}$$

This provides the transition between the tags and helps in capturing the context of the citation sentence in terms of higher order n-grams. The probabilities are computed with the following formula:

$$P(t_i/t_{i-4}, t_{i-3}, t_{i-2}, t_{i-1}) = f(t_{i-4}, t_{i-3}, t_{i-2}, t_{i-1}, t_i)/f(t_{i-4}, t_{i-3}, t_{i-2}, t_{i-1}) \tag{3}$$

For the sentiment analysis of the citation corpus, we have not only contemplated the adjectives and adverbs independently but have also evaluated their continuous and consecutive word sequences in the form of n-grams. We have calculated the degree of resemblance of each word pattern in the citation data d to the citation text t:

$$resp(d,t) = \begin{cases} 1, & \text{if the citation vector contains the pattern as it is,} \\ & \text{in the same order,} \\ \alpha.n/N, & \text{if } n \text{words out of the N words of the pattern appear} \\ & \text{in the citation text in the correct order,} \\ 0, & \text{if no words of the pattern appear in the citation text.} \end{cases} \tag{4}$$

3.2. Sentiment classification

In the sentiment classification module, we have used the unsupervised sentiment analysis technique for determining the sentiment polarity of the citation text based on the SentiWordNet 3.0 dictionary [26]. After

targeting the whole term profile of the citation sentence, only those terms are extracted that were labeled as adjectives, adverbs, and their n-grams. Computational linguists have suggested that adjectives and adverbs mostly represent sentiments in a sentence. The sentiment score of each tokenized string labeled as adjectives, adverbs, and their n-grams is calculated from the SentiWordNet lexical analyzer using the following formulas:

$$0 \leq \text{SwnPosScore, SwnNegScore, SwnNeuScore} \leq 1 \tag{5}$$

$$0 \leq (\text{SwnPosScore} + \text{SwnNegScore}) \leq 1 \tag{6}$$

$$\text{SwnNeuScore} = 1 - (\text{SwnPosScore} + \text{SwnNegScore}) \tag{7}$$

The frequency of occurrence of each n-gram is also aggregated for each of the citation texts. All the corresponding scores of adjectives, adverbs, and their combinations are aggregated to obtain the overall sentiment score of each citation text with the help of the following formulas:

$$\text{SynsetScore}\,(w) = \sum \frac{(\text{SwnPosScore}\,(w,s) - \text{SwnNegScore}\,(w,s))}{s} \tag{8}$$

$$wi = \langle \text{POS, SWN\_ID, SwnPosScore, SwnNegScore, SYNSETTERMS}, GLOSS \rangle \tag{9}$$

Document-level sentiment classification is performed to classify the entire citation text into 'positive', 'negative', or 'neutral' classes. In return, a sentiment polarity and overall sentiment score is assigned to each citation based on the average SentiScore of each opinionated phrase in the citation text, which is computed by using the following formulas:

$$\text{class\_label} = \begin{cases} POS \text{ if max(SwnPosScore, SwnNegScore, SwnNeuScore)} = POS \\ NEG \text{ if max(SwnPosScore, SwnNegScore, SwnNeuScore)} = NEG \\ else \end{cases} \tag{10}$$
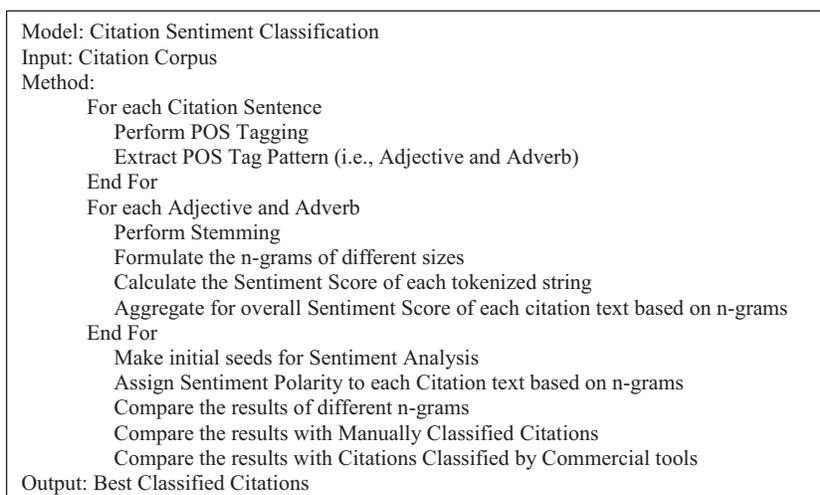
$$\text{SentiScore}\,(w) = \frac{\sum_{\text{for each term } w} \text{SynetScore}\,(w)}{N} \tag{11}$$

The proposed algorithm based on SentiWordNet incorporating different steps for document-level citation sentiment classification is shown in Figure 1.

However, there are some citation sentences that do not determine any sentiment polarity. The reason for this is that they do not have enough clearly opinionated words. An example of determining the sentiment score and classifying the citation text is presented in Table 1. SentiWordNet scores are calculated for positive and negative phrases found in the citation text and used these for determining the sentiment orientation by classifying the citation text into the class with the highest score. The analysis of the results depicted a clear tendency of improved and strengthened sentiment scores against the higher order n-grams as compared to lower order n-grams. A pattern of change in the sentiment polarity assignment from the objective class to the positive or negative class can also be seen in Table 2.

3.3. Sentiment classification using commercial tools

We have not only classified the citation text based on the adjective and adverb n-grams, but have also found the polarity of the citation sentences using commercial tools. For this purpose, 2 commercial tools are employed:

```
Model: Citation Sentiment Classification
Input: Citation Corpus
Method:
        For each Citation Sentence
            Perform POS Tagging
            Extract POS Tag Pattern (i.e., Adjective and Adverb)
        End For
        For each Adjective and Adverb
            Perform Stemming
            Formulate the n-grams of different sizes
            Calculate the Sentiment Score of each tokenized string
            Aggregate for overall Sentiment Score of each citation text based on n-grams
        End For
            Make initial seeds for Sentiment Analysis
            Assign Sentiment Polarity to each Citation text based on n-grams
            Compare the results of different n-grams
            Compare the results with Manually Classified Citations
            Compare the results with Citations Classified by Commercial tools
Output: Best Classified Citations
```

Figure 1. Proposed model for citation sentiment classification.

Table 1. Sentiment score of bigram features.

ID	Bigram	SENT_VAL	SENT_TYPE
4	the same	0	Neutral
4	same cut	0	Neutral
4	a reasonable	0.45	Positive
4	reasonable starting	0.45	Positive
4	subsequent research	0.30	Positive
10	neglected notable	−0.49	Negative

Table 2. A comparison of sentiment classification of n-grams.

ID	Score (n = 1)	Type (n = 1)	Score (n = 2)	Type (n = 2)	Score (n = 3)	Type (n = 3)	Score (n = 5)	Type (n = 5)
1	0.37	P	0.73	P	1.1	P	1.83	P
2	0	O	0	O	0	O	0	O
4	0.44	P	1.18	P	1.63	P	2.51	P
10	0	O	−0.49	N	−0.98	N	−1.96	N
11	0	O	0	O	0.41	P	0.41	P
17	0	O	0	O	−0.49	N	−1.96	N
51	0	O	−0.5	N	−1	N	−1.5	N

SEMANTRIA (which is a Microsoft Excel add-in) and THEYSAY (which is an online sentiment analysis tool). The primary objective of using these commercial tools is to validate the results obtained by the proposed model and to determine the reliability of sentiment analysis tools. There are many commercially available sentiment analysis tools, but we have utilized SEMANTRIA and THEYSAY because they provide feasibility in terms of usage and configuration. Moreover, they are applicable to all kinds of documents, databases, sentences, and phrases. SEMANTRIA performs the multilevel analysis of sentences incorporating parts of speech, assignment

of a sentiment score from dictionaries, application of intensifiers, and determination of the final sentiment score based on machine learning techniques. Each citation text is assigned a numerical sentiment score ranging from −2.0 to +2.0 along with a polarity of positive, negative, or neutral. THEYSAY assigns a sentiment score to the positive, negative, and neutral classes in percentages. It offers an in-depth analysis of the text considering different features like POS recognition, humor detection, gender detection, comparison detection, language detection, risk detection, and text summarization.

4. Dataset

For our experiments, we have utilized the gold-standard citation summary data consisting of 8736 citations manually annotated as positive, negative, and objective from 310 research papers [26]. Among the collected set of citation sentences, 829 are manually annotated and labeled for the positive sentiment class, 280 for the negative sentiment class, and the rest as neutral or others. This dataset is further expanded into 2 sets, one for training and the other for testing. The manually classified citation data are used as a baseline to compare the results with the proposed approach and SEMANTRIA and THEYSAY. We evaluated the selected citations based on 2 conditions: 1) the number of words in the citation must be above 20; 2) the selected citation text must contain at least one adjective or adverb. The average number of words in a citation sentence is 35 and most of the sentences are long, i.e. 81.22% of citations consisted of more than 20 words.

5. Results and discussion

We have used a prelabeled and manually classified corpus for the experiments to serve as a baseline for comparing and evaluating the results of the technique. Moreover, we have also labeled and classified the dataset consisting of the citation text using 2 commercial sentiment analysis tools, SEMANTRIA and THEYSAY. We compared the results of classification by our technique with the manually classified text and with the commercial tools. The results of the classification performed by commercial tools are different, which establishes that these tools are implemented based on different internal algorithms. Different percentages of positivity, negativity, and neutrality exhibit that both tools handle the negativity and neutrality in divergent ways. According to SEMANTRIA, among these 8730 citation texts, 2264 are in the positive sentiment class, 1266 have negative sentiment orientation, and the remaining ones are labeled as neutral. According to THEYSAY this distribution is 4802 positive, 2130 negative, and 1799 neutral. In the manually labeled dataset, the majority of the citations are labeled as neutral or objective, i.e. 87%, whereas the percentage of neutrality or objectivity is quite lower with the commercial tools, at 59.5% by SEMANTRIA and 20.5% by THEYSAY, respectively. The classification performed by manual annotation is highly skewed towards objectivity and the same pattern is depicted by SEMANTRIA.

This shows that SEMANTRIA has a clear tendency of assigning objective labels to the citation dataset as compared to THEYSAY, which demonstrates a biased behavior towards the positive class. This supports the findings of [21], which evaluated the utilization of SEMANTRIA and THEYSAY for the sentiment analysis of healthcare survey data. When comparing the tendency of class labels produced by commercial tools with the baseline manually annotated dataset, it is found that results produced by SEMANTRIA are closer to the benchmark as far as all class labels are concerned.

The presence of adjectives and adverbs in a citation text is utilized to classify it as positive, negative, or neutral. In experimentation, we extracted the BOW consisting of adjectives and adverbs based on n-grams. For this purpose, we have not only considered unigrams (n = 1) and bigrams (n = 2), but have also exploited the

use of trigrams (n = 3) and pentagrams (n = 5). The most specific reason for using the higher order n-grams is because most state-of-the-art approaches claim that they can play more significant roles than lower order n-grams in sentiment detection. Therefore, in this study, we have investigated how differently high order pairs of words behave in determining the sentiment orientation for the citation dataset. We have not only considered the adjectives and adverbs individually in determining the sentiment polarity of the citation text, but have also evaluated them in their different combinations. The extracted BOWs are sorted as per the frequency of their occurrence in the dataset to select the most important ones. The reason for this was to identify the frequent terms in the scientific literature for sentiment detection.

In this section, we present the results of citation sentiment classification by exploiting adjectives, adverbs, and their combinations by means of high order n-grams at the document level. We have not only compared the results obtained by our approach with the manually annotated corpora, but have also juxtaposed the results with classified citations using 2 commercial tools. In Table 3, the values of evaluation metrics (precision, recall, and recognition rate) are presented and a comparison is made with manual annotation and classification results against annotations performed by the commercial tools.

Table 3. Precision, recall, and accuracy of classified citations.

	Manual			SEMANTRIA			THEYSAY		
Feature	P	R	F-score	P	R	F-score	P	R	F-score
Adj. (1-g)	70.54	40	55.27	75.5	29.17	52.34	44.45	80	62.23
Adj. + Adv. (1-g)	75.54	55.56	65.55	94.11	55.56	74.84	53.35	78.17	65.76
Adj. (2-g)	75	50	62.5	70	54.17	62.09	55.56	76.92	66.24
Adj. + Adv. (2-g)	88.89	80	84.45	85.5	92.30	88.9	58.06	85.71	71.89
Adj. (3-g)	65.55	50	57.78	96	58.33	77.17	70.54	80	75.27
Adj. + Adv. (3-g)	88.89	72.72	80.81	96	68.57	82.29	59.38	82.60	70.99
Adj. (5-g)	75.54	60	67.77	94.11	66.67	80.39	54.17	81.26	67.72
Adj. + Adv. (5-g)	90	81.82	85.91	96.15	96.15	96.15	55.88	82.60	69.24

Let us now discuss the accuracy of the findings of classification results. When we select n-grams as features then semantic information is partially lost or neglected, more specifically in the case of BOWs, i.e. unigrams. The outcomes of the proposed study show an accuracy of above 80% against the manually annotated corpus using higher order n-gram adjective and adverb combinations, as explained in Figure 2. It can be observed from the results that with an increase in the value of n, the classification accuracy increases more specifically when adjectives and adverbs are used in a combinatory fashion. This further affirms that higher order n-gram adjective and adverb combinations are more precise and deterministic expressions than the lower order n-grams. With an accuracy of above 90% for adjective and adverb combinations, SEMANTRIA has produced the most
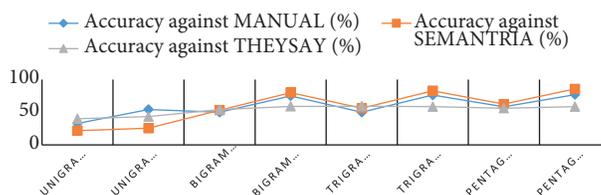
Figure 2. Comparison of accuracy against baseline.

precise predictions, following the same trend line as depicted against the manual baseline. This shows that these features (high order n-gram adjective and adverb combinations) produce better results as compared to unigrams and bigrams. The analysis of the results has revealed that the accuracy of the lower order n-grams (n = 1) remained the same for individual features (adjectives and adverbs) and both in a combinatorial fashion. The results also assert better accuracy for average positive citation texts as compared to negative and neutral ones. Furthermore, objective citations were misclassified as either positive or negative. The possible reason for this is that it is difficult to predict neutrality and negation. The analysis of the results has also shown that the use of higher order n-grams might solve the problem of compositionality (understanding a complex expression through the meanings of its constituent expressions).

It is again worth mentioning that the accuracy depicted against THEYSAY is persistently lower as compared to the SEMANTRIA and manual annotation of all the features. This is mainly because of the biased predictive behavior of THEYSAY towards the positive class. However, it is also evident for the accuracy trend line of THEYSAY that adjective and adverb combinations obtained a high accuracy rate as compared to isolated features. This also confirms and strengthens the accuracy of classification results. Though SEMANTRIA may not be considered as the best tool for sentiment analysis, it has shown a persistent behavior for different datasets, i.e. healthcare survey data [21] and a multilingual dataset [28]. THEYSAY has made incorrect classifications for large explanatory sentences. As the average sentence size of the experimental dataset was 35 words, it caused THEYSAY to make a majority of false predictions.

In Figure 3, for adjective and adverb combinations and high order n-grams, the proposed technique has achieved an average precision of about 90% against manual classification and 96.15% against SEMANTRIA, respectively. Thus, the effectiveness of the proposed technique is in an acceptable range when compared against manual classification and annotation results against SEMANTRIA. It is also worth mentioning that our proposed technique has acquired substantial improvement in precision as compared to recall. The precision and recall against THEYSAY are consistently low for all features. The high precision and recall values can be observed from Figures 3 and 4 for both individual and adjective and adverb combinations against SEMANTRIA, which
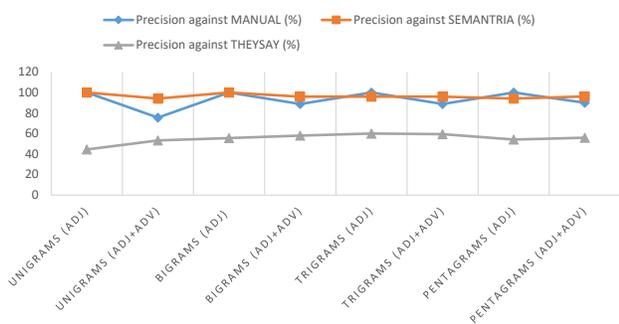

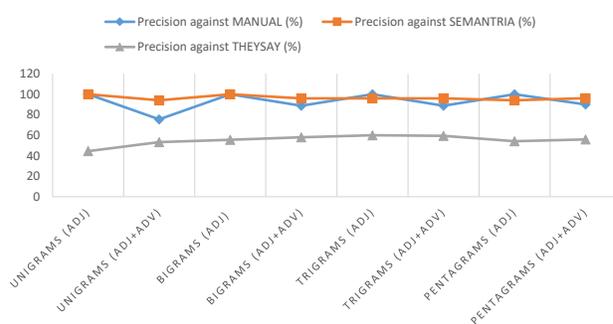**Figure 3**. Comparison of precision against baseline.

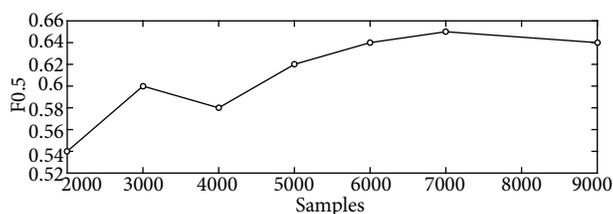**Figure 4**. Comparison of recall against baseline.

Figure 5. Impact of varied size of citation corpus on $F_{0.5}$ measure.

further ensures the reliability of the said tool. The combinations of adjectives and adverbs based on high order n-grams are a better choice in predicting the document-level sentiment for larger sentences. Further, to examine the computational load, we have examined the impact of the size of citation corpora on the performance of the proposed system. The results are presented in Figure 5. As seen from the graph, as the instances in the sample size are increased, there is improvement in the performance of the system. However, at a certain point in time, improvement may not be achieved by merely increasing the size of the training dataset when the size of the corpus is large.

6. Conclusion

The contribution of this study is twofold. Initially, it has explored the effectiveness of using adjectives, adverbs, and their combinations for document-level sentiment classification of citation text using a gold-standard citation sentiment corpus. Afterwards, it investigated the efficacy of applying commercial tools to the citation corpus for sentiment detection. Analysis of the results revealed that higher order n-grams (n = 5) for adjective and adverb combinations play a major role in improving the accuracy of the sentiment classification. The experimental results have also revealed that current sentiment analysis tools, and more specifically THEYSAY, are not efficient enough to detect sentiments accurately as the majority of the citations comprised multiple sentences. A possible direction for future work could be the contemplation of more features and dependency relationships for citation sentiment classification using different machine learning algorithms.

References

[1] Liu H. Sentiment Analysis of Citations Using Word2vec. arXiv preprint 1704.00177.

[2] Ma Z, Nam J, Weihe K. Improve sentiment analysis of citations with author modelling. In: Proceedings of WASSA 2016; San Diego, CA, USA. pp. 122-127.

[3] Klavans R, Boyack KW. Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? J Assoc Inf Sci Technol 2017; 68: 984-998.

[4] Oh J, Chang H, Kim JA, Choi M, Park Z, Cho Y, Lee EG. Citation analysis for biomedical and health sciences journals published in Korea. Healthc Inform Res 2017; 23: 218-225.

[5] Gurzki H, Woisetschläger DM. Mapping the luxury research landscape: a bibliometric citation analysis. J Bus Res 2017; 77: 147-166.

[6] Alam H, Kumar A, Werner T, Vyas M. Are cited references meaningful? Measuring semantic relatedness in citation analysis. In: ACM Proceedings of BIRNDL 2017; Tokyo, Japan.

[7] Meng R, Lu W, Chi Y, Han S. Automatic classification of citation function by new linguistic features. In: iConf Proceedings 2017; Wuhan, China.

[8] Nobre GC, Tavares E. Scientific literature analysis on big data and internet of things applications on circular economy: a bibliometric study. Scientometrics 2017; 111: 463-492.

[9] Wang Y, Bowers AJ, Fikis DJ. Automated text data mining analysis of five decades of educational leadership research literature: probabilistic topic modeling of EAQ articles from 1965 to 2014. Edu Adm Qua 2017; 53: 289-323.

[10] Hernández-Alvarez M, Gómez JM. Survey about citation context analysis: tasks, techniques, and resources. Nat Lang Eng 2016; 22: 327-349.

[11] Thelwall M. Data science altmetrics. J Data Inf Sci 2016; 1: 7-12.

[12] Fister JRI, Fister I, Perc M. Toward the discovery of citation cartels in citation networks. Front Phys 2016; 4: 49.

[13] Saggion H, Ronzano F. Scholarly data mining: making sense of scientific literature. In: ACM/IEEE Proceedings of JCDL 2017; June 2017; Toronto, Canada. pp. 1-2

[14] Jha R, Jbara AA, Qazvinian V, Radev DR. NLP-driven citation analysis for scientometrics. Nat Lang Eng 2017; 23: 93-130.

[15] Kuhn T, Perc M, Helbing D. Inheritance patterns in citation networks reveal scientific memes. Phys Rev X 2014; 4: 041036.

[16] Perc M. Self-organization of progress across the century of physics. Sci Rep 2013; 3: 1720.

[17] Michel JB, Shen YK, Aiden AP, Veres A, Gray MK, Pickett JP, Pinker S. Quantitative analysis of culture using millions of digitized books. Science 2011; 331: 176-182.

[18] Gao J, Hu J, Mao X, Perc M. Culturomics meets random fractal theory: insights into long-range correlations of social and natural phenomena over the past two centuries. J R Soc Interface 2012; 2012: rsif20110846.

[19] Li B, Liu T, Du X, Zhang D, Zhao Z. Learning document embeddings by predicting n-grams for sentiment classification of long movie reviews. arXiv preprint 1512.08183.

[20] Tembhurnikar SD, Patil NN. Topic detection using BNgram method and sentiment analysis on twitter dataset. In: IEEE Proceedings of ICRITO; September 2015; Noida, India. pp. 1-6.

[21] Georgiou D, MacFarlane A, Russell-Rose T. Extracting sentiment from healthcare survey data: An evaluation of sentiment analysis tools. In: IEEE Proceedings of SAI; July 2015; London, UK. pp. 352-361.

[22] Perc M. Evolution of the most common English words and phrases over the centuries. J R Soc Interface 2012; 2012: rsif20120491.

[23] Parthasarathy G, Tomar DC. Sentiment analyzer: analysis of journal citations from citation databases. In: IEEE Proceedings of Confluence; September 2014; Noida, India. pp. 923-928.

[24] Parthasarathy G, Tomar DC. A survey of sentiment analysis for journal citation. Indian Journal of Science and Technology 2015; 8: 35.

[25] Ikram MT, Butt NA, Afzal MT. Open source software adoption evaluation through feature level sentiment analysis using Twitter data. Turk J Elec Eng & Comp Sci 2016; 24: 4481-4496.

[26] Athar A. Sentiment analysis of citations using sentence structure-based features. In: ACM Proceedings of ACL; June 2011. pp. 81-87.

[27] Kim IC, Thoma GR. Automated classification of author's sentiments in citation using machine learning techniques: a preliminary study. In: IEEE Proceedings of CIBCB; 12 August 2015; Manchester, UK. pp. 1-7.

[28] Tandon N, Jain A. Citation context sentiment analysis for structured summarization of research papers. In: Springer Proceedings of KI; 24 September 2012; Saarbrücken, Germany. pp. 98-102.

[29] Yu B. Automated citation sentiment analysis: What can we learn from biomedical researchers? In: Proceedings of ASIST; 1 January 2013. pp. 1-9.

[30] Athar A, Teufel S. Context-enhanced citation sentiment detection. In: ACL Proceedings of HLT; 3 June 2012; Montreal, Canada. pp. 597-601.

[31] Butt BH, Rafi M, Jamal A, Rehman RSU, Alam SMZ, Alam MB. Classification of research citations (CRC). arXiv preprint 1506.08966.

[32] Hernández M, Gómez JM. Sentiment, polarity and function analysis in bibliometrics: a review. In: Natural Language Processing and Cognitive Science; 10 March 2015. p. 149.

[33] Stamou S, Mpouloumpasis N, Kozanidis L. Deriving the impact of scientific publications by mining citation opinion terms. Journal of Digital Information Management 2009; 7: 283-289.