CONTENT-BASED CITATION ANALYSIS USING MACHINE LEARNING ALGORITHMS

E.SUGANYA
Ph.D Research Scholar
Department of Computer Science
Bharathiar University
Coimbatore – 641046
elasugan1992@gmail.com

S.VIJAYARANI
Assistant Professor
Department of Computer Science
Bharathiar University
Coimbatore – 641046
vijimohan_2000@yahoo.com

***Abstract*— A citation is a reference to the source of information used in research. It is a way of knowing and learning new things from another source. It is used to find the information about the author, contents and journal. A citation analysis is defined as the analysis of information from citation indexes to determine the acclaim and the impact of distinguish articles, authors, and publications. Nowadays, Content-based Citation analysis is most require for measuring the quality of the research paper. In previous years, citations are evaluated using the number of counts cited by others for which paper, author, and journal. However, the quantity of the citations are not declared the quality which means if the paper has more citations and any one knows that has a positive citation or negative citations. To overcome this problem, content-based citation analysis is used for this research work to identify the citations' sentiments using sentiment analysis. Sentiment analysis of citations in scientific research papers and articles is a novel and interesting field to identify the sentiments. The main objective of this work is used to identify and analyze the meaning of the citation using Machine Learning algorithms such as KNN (K Nearest Neighbor), Random Forest, and Support Vector Machine. The performance measures used are precision, recall, and F-Measure. The sentiment analysis shows positive, negative, and neutral classes, which stated the meaning of the citations. It is very useful for the authors to know whether their publication may reach the researchers and academicians in a positive or negative way.**

Keywords — *Content based Citation Analysis, Sentiment Analysis, KNN, Random Forest, Support Vector Machine.*

I. INTRODUCTION

The sentiment is an emotional stance, idea, or assessment. Sentiment analysis, often defined as opinion mining, is the technique of identifying and extracting emotional information from various documents using natural language processing (NLP), text analysis, and computational linguistics (Haddi, 2013). Its goal is to predict an author's perspective toward a particular topic or a paper's discovering the best polarity (Aljuaid, 2020). When it comes to sentiment data, online is a valuable resource. People can upload their individual content on social media platforms, such as forums, microblogs, and online social networking sites. Many social media companies provide their Application Programming Interfaces (APIs), which encourage research and analysis by practitioners and researchers (Athar, 2011). As a result, sentiment analysis performs to have a strong base predicated on huge web data.

However, various limitations in this form of internet data could make sentiment analysis difficult. The first problem is people can publish whatever they want, and the quality of their thoughts cannot be recognized. Online spammers, for example, rather than provide topic-related opinions and post spam on forums. Sometimes spam is completely pointless, although some contain inappropriate or fraudulent opinions (Parthasarathy, 2014). The second problem is that such digital information does not necessarily have relevant data. Reality is more similar to a label for a particular opinion, specifying if it is positive, negative, or neutral (Kumar, 2018).

Sentiment analysis is a heuristic method for determining which emotions an author conveys in a research paper. Sentiment analysis has earned a lot of interest in recent decades (Yousif, 2019). The quality of a citation is determined by evaluating its meaning, which is a difficult task given that research papers are written in an unstructured format, with each work having its own format style (Xu, 2015). The citation and reference in the paper are unique. Some papers' citations are formatted numerically, while others are formatted in the APA style. As a result, to address the existing problem, content-based analysis is required (Muppidi, 2020). To solve this problem, semantic analysis of citations is used to determine the meaning of the citation with the use of sentiment analysis. In sentiment analysis, the Machine Learning algorithm is very important. THE REST OF the paper organized as follows: Section 2 described the previous studies related to content based citation analysis, sentiment analysis, and machine learning algorithms. Section 3 illustrates the methodology and machine learning algorithms such as KNN, Random Forest, and Support Vector Machine. Section 4 discussed the results and discussion of machine learning algorithms and conclusion is given in section 5.

II. EXISTING SYSTEM

Seth Porter (Porter, 2020) has analyzed the resources of Woodrow Wilson School Princeton Universities MPA/MPP (Master's in Public Affairs/ Master's in Public Policy) curriculum and the policy workshop reports from 2014-2018. Based on the analysis, 5.56% of the resources were inappropriate. Giovanni et al. (Giovanni Colavizza, 2017) The authors investigated the closeness of sets of articles that were co-cited at divergent co-citation steps of the journal, section, article, sentence, bracket, and paragraph. Their results indicated that the textual closeness shared references,

shared authorship, proximity at publication time rise at co-citation level gets lower (from journal to bracket). The authors compared results from four journals over the years 2010-2015. This study established that the similarity was the result of the journal to article co-citation; change at every level involves a rise in closeness. This is clearly seen from section to paragraph and paragraph to sentence/bracket levels.

Jeong et al. (Yoo Kyung Jeong, 2016) authors introduced a new approach to ACA by the proposed content-based similarity measure. The study adopted Word2Vec as the measure of author similarity. The authors also conducted an in-depth network analysis of author maps. Although the dataset is limited to JASIST, their method can be applied to other disciplines. As a follow-up study, the authors planned to extend network analysis of all cited authors and construct the author map with all cited authors. They also planned to identify the relationships between authors by analyzing the sources and conducting various statistical analyses such as factor analysis to verify their results. Zehra Taşkın et al. (Al, 2018) have analyzed the content-based citations for Turkish citation styles. They considered 423 peer-reviewed articles, 101,019 sentences, and 12881 related references, which are published in library and information science literature in Turkey. They divided the citations into four different categories i.e. meaning, purpose, shape, and array-based on the content. They achieved a 90% success rate for the citation classes.

Abdallah Yousif et al. (Zhendong Niu, 2019) presented a detailed study of scientific citation sentiment analysis. The process of scientific citation sentiment analysis is described, and recently developed approaches were given, analyzed, and criticized. Finally, they discovered that the majority of the articles employed traditional machine learning techniques. However, they have predicted that, in the future, hybrid and deep learning approaches will be able to tackle the difficulties of scientific citation sentiment analysis more effectively and consistently. Awais Athar (Athar, 2011) concentrated on automatic sentiment polarity detection in citations. The author assessed the usefulness of current and unique characteristics such as n-grams, scientific lexicon, dependency relations, and sentence splitting using a recently created and labeled citation sentiment corpus. Their findings revealed that 3-grams and dependencies dominate the scientific lexicon and sentence splitting features in this challenge.

Hanan Aljuaid (Aljuaid, H, 2020) proposed a content-based strategy for binary citation categorization by examining the similarities between research papers, sentiments of in-text citations, and finding the appropriate sentiment analysis model. Classification models have been used to categories the citations into binary categories. The proposed method used cosine similarity to generate the similarity score, and it was discovered that the linear SVC model is more appropriate for analyzing in-text citation sentiments than the other state-of-the-art models. Bahrainian et al. (S.A. Bahrainian, 2013) suggested an approach for detecting sentiment that is both unsupervised and hybrid. To classify tweets, they employed a binary SVM classifier. They demonstrated that by combining several features, the hybrid technique outperforms state-of-the-art algorithms.

Sugiyama et al. (Sugiyama, 2010) used the SVM classification technique to categorize citations into citing and non-citing groups utilizing data such as n-grams, previous and next sentences, proper nouns, orthographic properties, and location. They employed maximum entropy and discovered that proper noun and context categorization were the most effective characteristics for training the exact model. Agarwal et al. (S. Agarwal, 2010) used the Support vector machine and Multinomial Naive Based models to categorize the citations into eight groups. The experiment used a dataset of 43 open access papers from the field of medical science. The annotation was completed by discovering cue words/phrases in the citation context. Among 1710 sentences, there are 2,977 annotations. The average F-measure score for the developed system is 0.76.

S.Vijayarani et al. (Vijayarani, 2015) have compared two types of classification algorithms Support Vector Machine and Naïve Bayes to classify four types of kidney diseases. The performance was based on the performance factors such as classification accuracy and execution time. It was concluded that the SVM algorithm outperforms well than Naïve Bayes. E. Suganya et al. (Suganya, 2017) have evaluated and investigated classification algorithms for Road Accident Dataset. For performance factors they consider different parameters of accuracy and the error rate, it was found out that the KNN classification algorithm was the best algorithm with a maximum accuracy than linear regression, logistic regression, decision tree, SVM, Naïve Bayes, Random Forest, and gradient boosting algorithm. Surabhisaxena et al. (Surabhisaxena, 2016) have compared two different algorithms in image classification they were KNN classification and SVM classification. Although the performance of the KNN classification was very low compared to SVM. This was because the KNN classifier could not discriminate buildings where as SVM was very operative in classifying buildings.

III. METHODOLOGY

The primary aim of this work is used to identify the meaning of the citations using existing Machine Learning algorithms like K-Nearest Neighbor, Random Forest, and Support Vector Machine. The performance measures used are precision, recall, F-Measure, accuracy, and execution time. It shows positive, negative, and neutral classes, which stated the meaning of the citations. Figure1 is depicted the system architecture of Content-based Citation Analysis.

A. DATA COLLECTION

The dataset is collected from the Google Scholar web citation database. The dataset contains 41160 highly cited papers, which are in pdf files for computer science domains such as Artificial Intelligence, Big Data, Cloud

Computing, Data Mining, Image Processing, Internet of Things, Machine Learning, Soft Computing, Text Mining, and Web Mining.

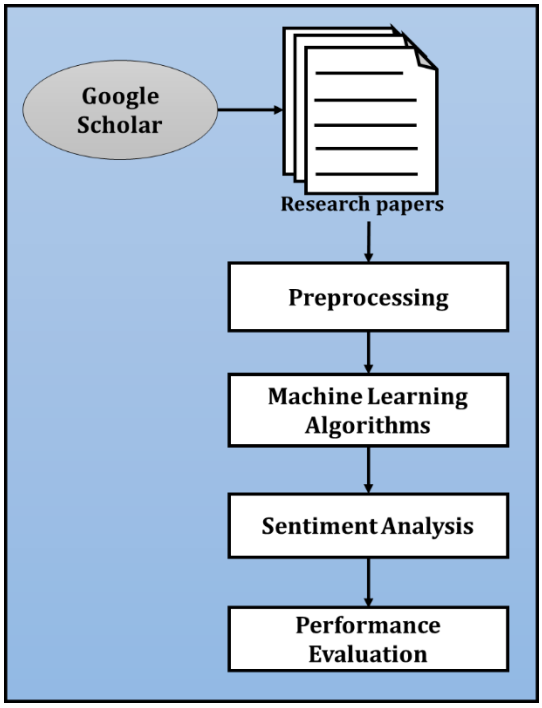

Fig. 1. System Architecture for Content-based Citation Analysis

B. PDF FILE CONVERSION

This research work required the content of the research papers. Pdf files are converted into plain text files because it is easy to identify the citations and citation sentences. The uncited papers are removed in this stage.

C. PREPROCESSING

Three strategies are employed in preprocessing to extract features from the content. These are stemming, stop word elimination, and tokenization. Stemming is a technique for deleting the suffix of a phrase from a research article. For instance, high, higher, more success, success, and so forth. However, there are fewer tokens implicated in the training, removing stop words minimizes the dataset volume and hence minimizes cost and time. These are the most generic terms in any language, including prepositions, conjunctions, pronouns, and so on, that do not add anything to the content.

In this research, select the citation sentences 3 to 4 lines across the citation. To select the sentence before the citation and after the citation. If a citation is present, select only one sentence in which citation is existing. In this work, 7368 citation sentences are identified for sentiment analysis.

D. MACHINE LEARNING ALGORITHMS

K-Nearest Neighbor Algorithm:

The K-NN algorithm is a distance-based metric for comparing the similarity scores between testing and training citation sentences. This method is used to evaluate the type of test citation sentences. This algorithm is a real-time machine learning technique for identifying sentiments based on the training sentiment's closest attributes (Huq, 2017). The strong feature space describes the training emotions. This is classified into three groups based on training sentiments: positive, negative, and neutral. The major goal of this method is to find sentiment similarity based on its neighbors (Hota, 2018). The features of the citation sentences are exclusively retrieved using this method from the high-dimensional feature space (Tyagi, 2018). The distance metrics can be computed and the k nearest feature determined in the classification stage. As a result, the closest value is used to represent the document's determined category (Damarta, 2021). This classification technique produces improved classification results as the number of categorized sentiments.

Pseudocode for KNN Algorithm:

Input: Citation Sentences

Output: Positive, Negative, and Neutral classes

Procedure:

Step 1. Let (X_i, C_i) represent data points with i = 1,2.....,n. For each i X_i denotes feature values and C_i denotes labels for X_i.

Step 2. Define C for the number of classes, where C_i $\in$ {1,2,3,…….,c} for all i values.

Step 3. Find d (x, x_i) i =1,2,.....,n, where d is the Euclidean distance between the points.

Step 4. Prepare the n Euclidean distances measured in a non-decreasing order.

Step 5. Take the first k distances from this sorted list, where k is a positive integer.

Step 6. Obtain the k-points that correspond to the provided k-distances.

Step 7. Let K_i signify the number of points in the i-th class out of a total of k points, i.e. k≥0

Step 8. Insert x in class i If ki >kj $\forall$ i ≠ j.

Step 9. Stop the process

Support Vector Machine:

Support vector machine (SVM) is the most powerful learning algorithm for classifying the sentiments. It is established on the structural risk minimization principle from the computational learning theory. The main aim of this principle is to find the hypothesis values at the lowest error rate (Jung HG, et al. 2014). The linear kernel threshold function is used in this research work. This classifier necessitates both positive and negative training sets, but the other classifiers are not. These positive and negative training sets are used to seek out the decision surface. It will divide the positive from the negative and negative from positive data in the high dimensional feature space, denoted as a hyperplane (Dong J-X, et al. 2005). The citation sentence representatives, which are closest to the decision surface, are called the support vector.

Pseudocode for SVM Algorithm:

Input: Classify the research papers, Training T={ (x_1,y_1),…..(x_n,y_n)}, Threshold t

Output: y ϵ {-1,1}

Procedure:

Step 1: Find the samples with minimum value of K

Step 2: Train SVM model on the k selected samples

Step 3: Classify the sentiments based on SVM probability

$$min\frac{1}{2} \parallel w \parallel^2$$

Subject to $y_i(w^T x_i + b) - 1 \geq 0 \; \forall i = 1,2,\dots,N$

Step 4: Make the Decision using t

Step 5: Stop the process

IV. RESULTS AND DISCUSSIONS

All the experiments are executed on a 2.20 GHz Intel CPU with 2 GB of memory and running on windows 8.1 pro.

A. DATASET DESCRIPTION

This research work focuses on ten different fields of highly cited research papers, which are under computer science domain such as Artificial Intelligence, Big Data, Cloud Computing, Data Mining, Image Processing, Internet of Things, Machine Learning, Soft Computing, Text Mining, and Web Mining, which is shown in table 1. The dataset contains 41160 pdf files, 75% of the data are considered as training set, and the remaining 25% of the data are considered as testing set.

Table 1. Dataset Description

S. No	Dataset - Topics	PDF Files
1.	Artificial Intelligence	4040
2.	Big Data	3932
3.	Cloud Computing	4020
4.	Data Mining	3680
5.	Image Processing	4160
6.	Internet of Things	4844
7.	Machine Learning	4092
8.	Soft Computing	3936
9.	Text Mining	4452
10.	Web Mining	4004

Evaluation of the Machine Learning algorithms using three significant performance factors such as precision, recall, and F-measure. These performance measures are evaluated the sentiments like positive, negative, and neutral classified citations. Based on the observation the SVM algorithm was evaluated and obtain better results after comparison. The classifiers used in this research are KNN, Random Forest, and SVM Machine Learning algorithms. In Table 2, three performance factors are used to evaluate the dataset using Machine Learning algorithms. From the experiments, the SVM outperforms well than existing algorithms.

Table 2. Performance Analysis of Machine Learning Algorithms

Topics	Algorithms	Precision	Recall	F-Measure
AI	KNN	68.1	68.9	68.5
	RF	69.09	69.89	69.49
	SVM	73.89	74.69	74.29
BD	KNN	66.4	67.8	67.1
	RF	67.39	68.69	68.09
	SVM	71.49	72.29	71.89
CC	KNN	64.6	65.9	65.3
	RF	65.59	66.89	66.29
	SVM	70.39	71.69	71.09
DM	KNN	67.2	67.9	67.6
	RF	68.19	68.89	68.59
	SVM	72.99	73.69	73.39
IP	KNN	66.6	67.7	67.2
	RF	67.69	68.79	68.29
	SVM	74.29	75.39	74.89
IoT	KNN	65.7	66.6	66.2
	RF	66.69	67.59	67.19
	SVM	71.49	72.39	71.99
ML	KNN	67.1	67.4	67.3
	RF	68.09	68.39	68.29
	SVM	72.89	73.19	73.09
SC	KNN	68.4	68.8	68.6
	RF	69.49	69.89	69.69
	SVM	76.09	76.49	76.29
TM	KNN	63.6	65.3	64.5
	RF	64.59	66.29	65.49
	SVM	69.39	71.09	70.29
WM	KNN	68.1	68.9	68.5
	RF	69.09	69.89	69.49
	SVM	73.89	74.69	74.29

Figure 2 depicts precision value, figure 3 illustrates recall value, and figure 4 shows F-Measure of Machine Learning algorithms for ten different topics of the Computer Science domain. From the experiments, SVM outperforms well than existing algorithms.

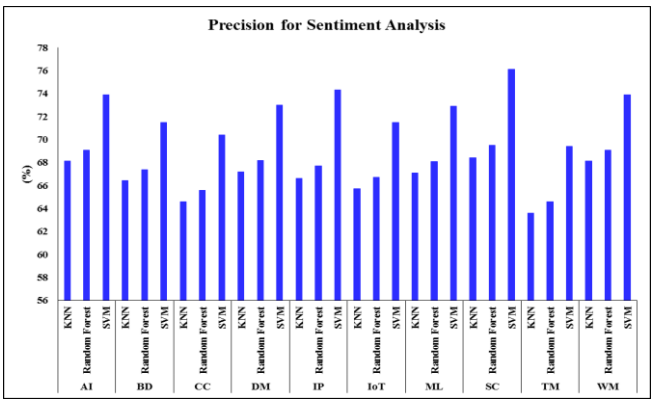

Fig. 2. Precision of Machine Learning Algorithms for Sentiment Analysis

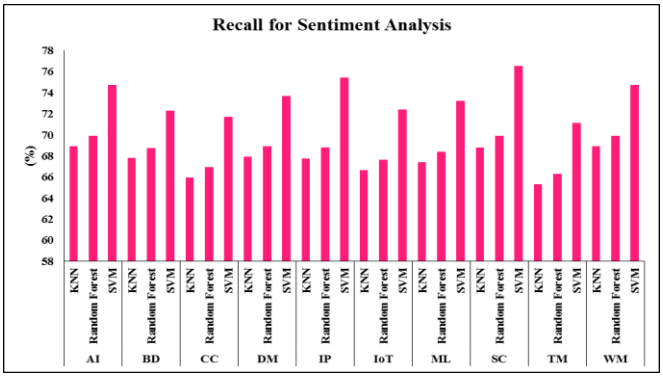

Fig. 3. Recall of Machine Learning Algorithms for Sentiment Analysis

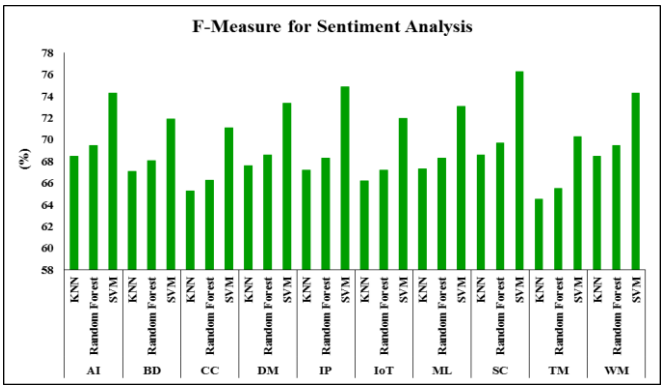

Fig. 4. F-Measure of Machine Learning Algorithms for Sentiment Analysis

V. CONCLUSION

The primary aim of this work is used to identify the meaning of the citations. This work consists of two sub-phases: Preprocessing and Sentiment Analysis. Before preprocessing, the pdf files are converted into a normal text file. After that, NLP techniques are used to perform stop words, stemming, and tokenization tasks. Then, to select the citations and identify the citation sentences for sentiment analysis. In order to do that, the existing Machine Learning algorithms are applied for sentiment analysis such as K-Nearest Neighbor, Random Forest, and Support Vector Machine algorithms. The performance measures used are precision, recall, and F-Measure. It shows positive, negative, and neutral classes, which stated the meaning of the citations. Based on the observation, SVM algorithm performs well than existing algorithms with greater accuracy. It is very useful for the authors to know whether their publication may reach the researchers and academicians positively or negatively.

References

[1] Abhilasha Tyagi, Naresh Sharma, Sentiments Analysis of Twitter Data using K- Nearest Neighbour Classifier, International Journal of Engineering Science and Computing, Vol 8, Issue 4, April 2018.

[2] Alatas, S. A. (2019). Sentiment Classification Within Online Social Media Using Whale Optimization Algorithm and Social Impact Theory Based Optimization. Physica A. doi:https://doi.org/10.1016/j.physa.2019.123094.

[3] Aljuaid, H., Iftikhar, R., Ahmad, S., Asif, M., Tanvir Afzal, M., Important citation Identification using Sentiment Analysis of In-text citations, Telematics and Informatics (2020), doi: https://doi.org/10.1016/j.tele.2020.101492

[4] Arnav Munshi, Sanchit Sapra, M.Arvindhan, A Novel Random Forest Implementation of Sentiment Analysis, International Research Journal of Engineering and Technology, Volume: 07 Issue: 06 | June 2020, e-ISSN: 2395-0056.

[5] Awais Athar. 2011. Sentiment analysis of citations using sentence structure-based features. In Proceedings of the ACL 2011 Student Session (HLT-SS '11). Association for Computational Linguistics, USA, 81–87.

[6] Bahrawi, (2019). Sentiment Analysis using Random Forest Algorithm Online Social Media Based. Journal of Information Technology and Its Utilization, Volume 2, Issue 2, December-2019: 29-33.

[7] Damarta, R & Hidayat, A & Abdullah, A. (2021). The application of k -nearest neighbors classifier for sentiment analysis of PT PLN (Persero) twitter account service quality. Journal of Physics: Conference Series. 1722. 012002. 10.1088/1742-6596/1722/1/012002.

[8] Emma Haddi, X. L. (2013). The Role of Text Pre-processing in Sentiment Analysis. Procedia Computer Science (pp. 26-32). Elsevier.

[9] Esha Tyagi and Arvind Kumar Sharma, Sentiment Analysis of Product Reviews using Support Vector Machine Learning Algorithm, Indian Journal of Science and Technology, 2017, Volume: 10, Issue: 35, Pages: 1-9, DOI: 10.17485 /ijst/2017/v10i35/118965

[10] G. Parthasarathy and D. C. Tomar, Sentiment analyzer: Analysis of journal citations from citation databases, 2014 5th International Conference - Confluence The Next Generation Information Technology Summit (Confluence), 2014, pp. 923-928, doi: 10.1109/CONFLUENCE.2014.6949321.

[11] Imamah, Husni, Eka Malasari Rachman, Ika Oktavia Suzanti, and Fifin Ayu Mufarroha, Text Mining and Support Vector Machine for Sentiment Analysis of Tourist Reviews in Bangkalan Regency, Journal of Physics, 2020, doi:10.1088/1742-6596/1477/2/022023

[12] Jeong, Y. K. (2014). Content -Based Author Co-Citation Analysis. Journal of Informetrics, 197-211.

[13] Kumar, H. &. (2020). A New Feature Selection Method for Sentiment Analysis in Short Text. Journal of Inteligent System, 29(1), 1122 - 1134. Retrieved https://doi.org/10.1515/jisys-2018-0171

[14] M. Naz, K. Zafar, and A. Khan, "Ensemble Based Classification of Sentiments Using Forest Optimization Algorithm," Data, vol. 4, no. 2, p. 76, May 2019.

[15] Mehreen Naz, K. Z. (2019). Ensemble Based Classification of Sentiments Using Forest Optimization Algorithm. Data, 1-13.

[16] Mohammad Rezwanul Huq, Ahmad Ali, Anika Rahman, Sentiment Analysis on Twitter Data using KNN and SVM, International Journal of Advanced Computer Science and Applications, Vol. 8, No. 6, 2017.

[17] Munshi, A., Arvindhan, M. and Thirunavukkarasu, K. (2021). Random Forest Application of Twitter Data Sentiment Analysis in Online Social Network Prediction. In Emerging Technologies for Healthcare (eds M. Mangla, N. Sharma, P. Mittal, V.M. Wadhwa, K. Thirunavukkarasu and S. Khan). https://doi.org/10.1002/9781119792345.ch12

[18] Muppidi, Satish, Gorripati, Satya Keerthi, and Kishore, B. An Approach for Bibliographic Citation Sentiment Analysis Using Deep Learning'. 1 Jan. 2020 : 353 – 362.

[19] Nazir S, Asif M, Ahmad S, Bukhari F, Afzal MT, Aljuaid H (2020), Important citation identification by exploiting content and section-wise in-text citation count. PLoS ONE 15(3):e0228885. https://doi.org/10.1371/journal.pone.0228885

[20] P. Karthika, R. Murugeswari and R. Manoranjithem, "Sentiment Analysis of Social Media Network Using Random Forest Algorithm," 2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), 2019, pp. 1-5, doi: 10.1109/INCOS45849.2019.8951367.

[21] Soudamini Hota, Sudhir Pathak, KNN Classifier Based Approach for Multi-Class Sentiment Analysis of Twitter Data, International Journal of Engineering & Technology, International Journal of Engineering & Technology, 7 (3) (2018) 1372-1375

[22] Souvick Ghosh and Dipankar Das and Tanmoy Chakraborty, Determining sentiment in citation text and analyzing its impact on the proposed ranking index, 2017.

[23] Spiegel Rosing I. Science Studies: Bibliometric and Content Analysis. Social Studies of Science. 1977;7(1):97–113. https://doi.org/10.1177/030631277700700111

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

[25] Yousif, A., Niu, Z., Tarus, J.K. et al. A Survey on Sentiment Analysis of Scientific Citations. Artif Intell Rev 52, 1805–1838 (2019). https://doi.org/10.1007/s10462-017-9597-8.

[26] Zainuddin, Nurulhuda & Selamat, Ali. (2014). Sentiment analysis using Support Vector Machine. I4CT 2014 - 1st International Conference on Computer, Communications, and Control Technology, Proceedings. 333-337. 10.1109/I4CT.2014.6914200 .