IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 25 July 2022, accepted 13 August 2022, date of publication 17 August 2022, date of current version 26 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3199420

RESEARCH ARTICLE

Important Citation Identification by Exploding the Sentiment Analysis and Section-Wise In-Text Citation Weights

SHAHZAD NAZIR[1], MUHAMMAD ASIF[1], SHAHBAZ AHMAD[1], HANAN ALJUAID[2], RIMSHA IFTIKHAR[1], ZUBAIR NAWAZ[3], AND YAZEED YASIN GHADI[4]

[1]Department of Computer Science, National Textile University, Faisalabad 37610, Pakistan
[2]Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University (PNU), Riyadh 11671, Saudi Arabia
[3]Department of Data Science, University of the Punjab, Lahore 54590, Pakistan
[4]Department of Computer Science/Software Engineering, Al Ain University, Abu Dhabi, United Arab Emirates

Corresponding author: Muhammad Asif (asif@ntu.edu.pk)

This work was supported by the Princess Nourah bint Abdulrahman University Researchers Supporting Project, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia, under Grant PNURSP2022R54.

ABSTRACT A massive research corpus is generated in this epoch based on some previously established concepts or findings. For the acknowledgment of the base knowledge, researchers perform citations. Citations are the key considerations used in finding the different research measures, such as ranking the institutions, researchers, countries, computing the impact factor of journals, allocating research funds, etc. But in calculating these critical measures, citations are treated equally. However, researchers have argued that all citations can never be equally influential. Therefore, researchers have proposed other techniques to identify the important content-based, meta-data-based, and bibliographic-based citations. However, the produced results by the state-of-the-art still need to be improved. In this research work, we proposed an approach based on two primary modules, 1) The section-wise citation count and 2) Sentiment based analysis of citation sentences. The first technique is based on extracting the different sections of the research articles and performing citation count. We applied Neural Network and Multiple Regression on section-wise citations for automatic weight assignment. The citation sentences were extracted in the second approach, and sentiment analysis was used for sentences. Citations were classified with Support Vector Machine, Multilayer Perceptron, and Random Forest. F-measure, Recall, and Precision were considered to evaluate the results, compared with the state-of-the-art results. The value of precision with the proposed approach was enhanced to 0.94.

INDEX TERMS Important citation identification, sentiment analysis, weight assignment, machine learning.

I. INTRODUCTION

Scientific research always has its roots in the literature of the domain [1]. Citation specifies the relationship between the citing and cited articles. In the research community, citations act as an acknowledgment of the stat-of-the-art work and the researcher. Therefore, the citation is deemed as a gauge to measure the different research aspects such as the impact factor of journals [2], H-index, I-index, research grants, and funds [3], awards, ranking of researchers [4], institutions, etc. To compute such parameters, all citations are given equal weightage. In this era, the researchers have asserted that each citation is not influential [5], and the importance of citations varies for reasons such as researchers can cite an article to provide technical background, enhance the results, or compare the findings. To analyze the citations, qualitative features should be accompanied by quantitative aspects. The research community suggests that a citation reflecting only literature knowledge and a citation that enhances the work can never be of equal importance. In research articles, citations

The associate editor coordinating the review of this manuscript and approving it for publication was Vivek Kumar Sehgal.

are primarily made to provide the general background of the research work [6]. Therefore, researchers have adopted multidisciplinary approaches to discriminate between important and non-important citations. If a citation enhances the work, it is considered important and non-important in the case of providing only background knowledge [7].

The researchers have developed multiple models and approaches to classify the citations concerning their reasons. This classification was converged to automatic categorization by manually asking the reasons from the authors. Finney [8] was the first researcher who proposed an automatic model to classify citations into seven categories. The different groups of citations were merged, forming two classes such as important and non-important citations. The key approaches for the classification of citations are 1) Content-based [7], [9], 2) Mata data-based [10], 3) Count based [11], 4) Sentiment based [5], 5) Hybrid approaches [12], etc. In Meta-data based and Content-based techniques, the similarity of the corpus is calculated while the frequency of citation is considered in the count-based approach. Zhu *et al.* [9] performed the pioneer binary classification of citations. This work was enhanced by Valenzuela *et al.* [7]. The author utilized contextual features and categorized the citations into non-important and important categories. Qayyum and Afzal *et al.* [10] used the Meta-data approach and enhanced the results further. Wang *et al.* [12] introduced the syntactic and contextual-based approach. The author produced a 0.85 value of the F-measure. The produced results by the state-of-the-art need to be enhanced for potential decisions.

This research presents a hybrid approach to identifying the credible citations of research articles. To experiment, two annotated datasets were used. The first dataset was collected by Valenzuela *et al.* [7], and the domain experts annotated this dataset. The second dataset was compiled by [10] and annotated by a Faculty member of the Central University of Science and Technology Islamabad. To classify the citations, different modules were considered, such as 1) Citation Count, 2) Similarity of research articles, 3) Section-wise weights for in-text citation, and 4) Sentiment analysis of citations. In citation count, the direct and indirect frequency of citations was considered. Furthermore, a cosine similarity algorithm was utilized to calculate the text similarity of citing and cited research articles.

Further, the sentiment analysis on citation sentences was performed, and the citation was categorized as positive, negative, or neutral. Finally, a section-wise citation count was performed to assign the automatic weights to sections. Considering the section-wise citation count Neural Network and Multiple Regression algorithms were utilized to produce appropriate weights for sections. Support Vector Machine, Multilayer Perceptron, and Random Forest were considered to classify the citations. The performance of the approach was measured with Precision, Recall, and F-measure values. The produced results were compared with stat-of-the-art. The outcomes of the experiments enhanced the state-of-the-art results from 0.9 to 0.94 value of the F-measure.

This research considered potential features for identifying important citations, such as section-wise in-text citation weights, sentiment analysis, and similarity of research articles. These features effectively classified the citations producing a significant value of the F-measure. As a result, the proposed approach outperformed as compared to the state-of-the-art making a considerable contribution to the literature.

II. LITERATURE REVIEW

Citations are the key factors in effectively estimating the different technical aspects, such as the impact factor of journals, H-index, and I-index. Citations represent the bond between the citing and cited research articles. The esteem citation analysis is used to acquire scientific information about the author's research work. The pioneer of the domain citation analysis was Garfield [13]. The author worked on the correlation between citations and the Prize winners. Furthermore, Inhaber and Przednowek [3] developed the idea of considering the relationship between citations and research fund winners. Garfield [13] performed the research and extracted the 15 reasons for citations to find out why the authors do citations. These reasons were investigated by Bornmann and Daniel [2]. Moraviscki and Murugesan [6] performed the citation classification based on the reasons. The author claimed that the citations are performed considering different reasons. Therefore, citations are not equally influential. The classification categories of citations were reduced to 13 by Spiegel-Rosing *et al.* [14]. In the early era of citation analysis, the reasons for citations were manually asked by authors. The manual citation reason finding was unfeasible for a massive corpus. Therefore, the need of the hour was to classify the citations automatically.

Roger Mayer *et al.* [15] explained that specific words or phrases with citations could justify citation category. Moravisks and Murugesan [6] introduced the citation classification technique and reported that a single citation could belong to different categories. Finney performed the first semi-automatic citation classification [8]. The author classified the citations into seven categories. The fully automated approach for citation classification was introduced by Garzone and Mercer [16]. The author highlighted the shortcomings of the Finny model. The citations were categorized into 35 categories using 195 lexical and 14 parsing rules for documents. The approach was implemented on a dataset of 20 research articles. The experiments showed better results on the known dataset, but for the unknown dataset, the results were averaged. Giles developed the first automatic citation indexing engine [17], later named CiteSeer. This engine is a digital library consisting of literature on computer science. Pham and Hoffmann [18] categorized the citations into four categories. Bi *et al. et al.* [19] proposed a similar approach. The author considered the direct and indirect citations. The author stated that the proposed system achieved higher results than a state-of-the-art method like SCi and PageRank. Another automated approach was introduced

by Teufel *et al.* [20] based on a supervised machine learning model.

The author considered different linguistic rules and classified the citations into four groups. The citation groups were further divided into 11 subgroups. The dataset consisted of 548 citations, and these citations were categorized considering 892 linguistic phrases. A 90% dataset was utilized for training the model, and the model was tested on the remaining 10%. The results depicted that 65% of citations were neutral with a 0.71 value of the F-measure. Sugiyama *et al.* enhanced this idea [21]. The author classified the citation into citing and non-citing categories. The Support Vector Machine (SVM) model was considered to implement the approach, utilizing different features such as nouns, position phrases, following sentences, n-grams, and previous sentences. It was reported that context and proper nouns were significant for training purposes.

Agarwal *et al.* [22] used SVM and Naïve Based approaches to classify citations considering eight categories. The dataset used by the author consisted of 43 research articles from the domain of medical science. The annotation was performed with phrases from the context of citations. The results were presented in the form of an F-measure value of 0.76. Next, Small [23] performed the sentiment analysis of citations to understand the social process. The dataset of 20 research articles was used, consisting of words and phrases depicting the sentiments of citations. The author reported the correlation of sentiments with social and cognitive reasons. Finally, Shahid *et al.* [24] developed an approach to find the relevant research articles. The author used a dataset of 16404 reference pairs and stated that the articles would be relevant if the citation frequency were five or more. This approach was further enhanced by Hou *et al.* [25]. The author claimed that if the in-text citation frequency is more than 10, there would be vital relevancy between the citing and cited research articles. For this experiment, the dataset of 651 articles was used, and the results showed closely related references more often.

To classify the citations, Balaban [26] introduced the approach of assigning more weightage to the citations of famous authors. The author also stated that a research article would be significant if cited by the high impact factor article. Dong and Schäfer [27] reduced the classes to three, considering that more classes can produce a conflict for citations. The classes were 1) Positive, 2) Negative, and 3) Neutral. Athar [28] also classified the citations into three categories. Next, the author performed the sentiment analysis on citations. Citation analysis was further implemented by Jochim and Schütze [29]. Finally, the author proposed an approach to find the citations having more impact in the research domain. For this experiment, different lexical features from context were utilized, and the dataset was collected from ACL Anthology. Classification of citations into two categories was performed by Roger Mayers [15]. The author used a dataset of 20 articles. Another classification technique based on keywords was introduced by Kumar [30]. The citations were categorized into 1) Positive and 2) negative classes after performing sentiment analysis. The dataset was collected from Association for Computational Linguistics (ACL) Anthology.

Lee *et al.* [31] classified the citations into three categories 1) Positive, 2) Negative, and 3) Neutral. These categories were further distributed into 12 subcategories. The dataset consisted of 6,355 citations. To perform the experiment n-gram technique was used. The model achieved a 0.67 value of the F-measure. Butt *et al.* [32] proposed extracting the five sentences with citations. The author implemented sentiment analysis using Naïve Bayes to classify the citations. The accuracy of the model was 80%. The same research was conducted by Sula and Miller [33], and the author used the Naïve Bayes model to classify the citations. They manually extracted the citation sentences and annotated them as positive and negative citations. But the proposed model could not extract the multiple citations in a sentence. Another approach for the classification of citations was proposed by Kumar [30]. The approach was keyword-based. The citations were categorized as positive and negative citations. The dataset was collected from ACL Anthology.

Zhu *et al.* [9] classified the citations into two categories and termed them categories influential and non-influential. The author used the machine learning algorithm Support Vector Machine for the classification. The dataset consisted of hundred research articles collected from ACL Anthology. Five features were used: citation frequency, similarity, position-based, context-based, and miscellaneous. This research was further enhanced by Zhu *et al.* [9]; the author classified the citations into important and non-important citations. The dataset collected from ACL Anthology consisted of 465 pairs of citation articles. Field experts annotated the citations. The experts 93.6% agreed on classification. For classification, twelve features were utilized, such as citation count, similarity, direct, indirect, etc. The authors utilized SVM and Random Forest models while computing the value of precision as 0.65 and Recall as 0.90.

The results were further boosted by Qayyum and Afzal *et al.* [10], and citations were classified into two categories. For this experiment, two datasets were used. The first dataset was collected from the research work of Valenzuela *et al.* [7], and the other dataset consisted of 324 citation pairs. The Faculty of Computer Science, CUST Islamabad, collected and annotated this dataset. The research work was performed on metadata [34] of research articles. Eight different features were used, such as title similarity, abstract similarity, keywords similarity, etc. The author used three machine learning models SVM, KLR, and Random Forest. The author claimed the best results with Random Forest by achieving a 0.72 value of precision. Aljuaid *et al.* [5] enhanced this idea by considering the sentiment analysis and achieved a 0.83 precision value. After that, we [11] contributed to the citation classification domain and increased the results to 0.84. Currently, the state-of-the-art approach [12] has achieved a 0.9 value of precision, but considering the citation important is still not optimal.

TABLE 1. Dataset D1 statistics.

Annotation	Label	Category	Citations
Related Work	0	Non-important	398
Comparing Work			
Utilizing Work	1	Important	67
Enhancing Work			

TABLE 2. Dataset D1.

Annotator	Papers	Cited by	Follow up
A	J02-3001	E03-1040	0
A	J02-3001	P04-1043	0
A	J02-3001	P07-1071	1
B	P05-1045	D13-1191	0
B	P05-1045	E09-1007	1

III. METHODOLOGY

To classify the citations into two categories, such as important and non-important citations, the overall approach is given in Figure 1. This experiment consisted of four key modules 1) Similarity calculation, 2) Citation Count, 3) Section-wise weights assignment, and 4) Sentiment Analysis of citation sentences. Machine learning algorithms Random Forest, Multilayer Perceptron, and Support Vector Machine were used. The results were evaluated using Precision, Recall, and F-measure.

A. DATASET

This research was conducted considering two datasets. The first dataset was collected by Valenzuela *et al.* [7], and the other dataset was composed by Faiza Qayyum and Afzal [10]. Valenzuela's dataset consisted of 20,527 scientific research papers. This dataset was obtained from ACL Anthology, and two experts in the domain performed the annotation of the dataset. The extracted number of citations was 106,509, and due to difficulty in labeling a massive number of citations, the annotators only considered 465 citation sets. The domain experts categorized the citations into four classes concerning their importance in articles. Further, the four groups converged into binary classes. In dataset, 0 represents non-important and important citations are reflected by 1. 14.6% of citations were annotated as non-important, and 85.4% as important, as presented in Table 1.

In Valenzuela's dataset, the IDs of research articles are placed, and by using these IDs, the pdf files can be downloaded from http://www.aclweb.org/anthology/. The IDs will be linked at the end of the Anthology URL to extract the Pdf files. The annotated dataset is shown in Table 2. Unfortunately, while performing the scraping, IDs 1) L08-1584, 2) W07-2058, 3) L08-1267, and 4) L08-1584 were unavailable, and four IDs were unable to be scrapped. Therefore, 457 research articles were considered from the first dataset for the experiment.

In the first column of Table 2, A and B represent the annotators. The following field presents IDs of cited research papers; the third column consists of citing research papers.

TABLE 3. Dataset D2 statistics.

Annotation	Label	Category	Citations
Related Work	0	Non-important	216
Comparing Work			
Utilizing Work	1	Important	95
Enhancing Work			

TABLE 4. Dataset D2 titles.

Sr #	Title
11	Single-ISA Heterogeneous Multi-Core Architectures
12	JEOPARD – Java Environment for Parallel Real-Time Development
13	Experiments with Cholesky Factorization on Clusters of SMPs
14	A Parallel Java Grande Benchmark Suite

TABLE 5. Dataset D2 citation pairs.

PaperID	CitedBy	Fine Grained Value
16	35	0
16	36	0
37	16	1
38	16	0

Finally, the fourth column describes whether the citation is important or not. Here, 0 is presenting a non-important citation, and 1 is for an important citation. To increase the citation pairs, we considered another dataset consisting of 324 research articles and 311 citation pairs.

The research papers were from several publishers such as IEEE, Elsevier, Science Direct, etc. This dataset was gathered by [10] and annotated by the members of the Faculty of Central University of Science and Technology (CUST) Islamabad. As described in Table 3, most of the citations were from non-important citation classes, and important citations were more minor in number. This dataset contained two spreadsheets, the first sheet comprised the titles of research articles and their IDs and the second sheet had follow-up of citation pairs, describing if the citation is important or non-important. The dataset D2 is presented in Table 4 and Table 5. These tables consist of Titles and citation pairs.

B. PDF TO TEXT CONVERSION

The research articles of dataset D1 were automatically downloaded from Anthology, considering their IDs. on the other hand, for dataset D2, we manually downloaded the articles from different publishing sites. All the research articles were in PDF format. PDF files store the text in the form of a content stream, and parsing the PDF files is a difficult task to perform. In comparison, Text files are the simplest document form and can be easily parsed. Therefore, both datasets' PDF files were converted into Text files. To perform this conversion XPDF tool was used, which is openly available on GitHub. This tool implements the R language and is considered efficient for such conversions. Therefore, using XPDF[1] the PDF files were automatically converted into Text files.

[1] https://github.com/kermitt2/xpdf-4.00

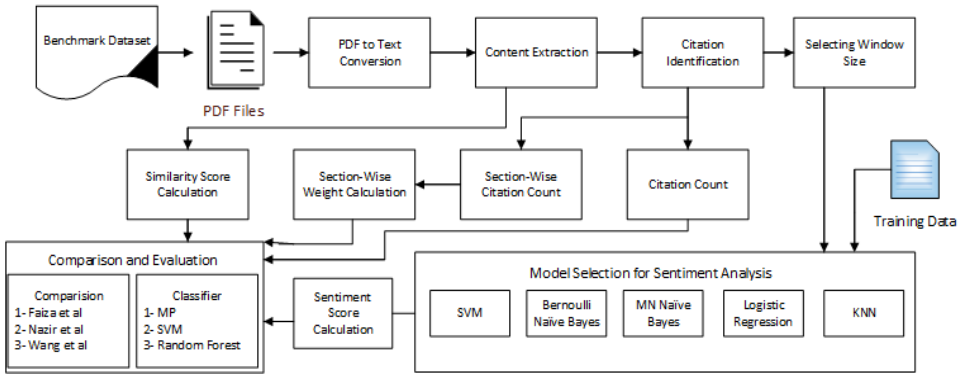

FIGURE 1. Overall methodology.

C. CONTENT EXTRACTION

After converting the PDF files into Text files, the content extraction from Text files was performed. Therefore, text files were parsed using an openly available tool ParCit [35], and can be downloaded from the GitHub site. This tool uses a Conditional Random Field Model and is already trained. Therefore, we do not need to train it. It performs tokenization at the sentence level and labels the tokens. It considers different research elements such as (1) Title, (2) Authors, (3) Abstracts, (4) Different Sections, and (5) citations. This tool parses the files considering their structure and can extract bibliographic portions. The text files of datasets D1 and D2 were parsed, and different elements of files were extracted.

D. CITATION IDENTIFICATION

It is a complex task to identify the citation location as the citation styles vary concerning publishers, but in both datasets, the citation pattern was similar. Therefore, the identification process became feasible. For citation identification, the sections were parsed, and the program automatically located the citations. First, for each citation pair, the Authors name and the publishing year were the main elements for locating the citations. Next, the structure of citations was considered as it starts and ends with round brackets. Then, the publishing year and name of the author in citation were compared with articles publishing year and author of articles. Finally, this pattern matching was executed for each section, and the citation locations were identified. The identified citations were further utilized in calculating overall citations of manuscript, section-wise citations and in citation sentence extraction for performing sentiment analysis.

E. SECTION-WISE WEIGHT CALCULATION FOR IN-TEXT CITATION

The section-wise weight calculation phase consists of two sub-modules, 1) calculating section-wise citation count and 2) Assigning weights to sections. In this step, we identified the section-wise citation and counted the citations concerning the sections. After that, we used machine learning algorithms to assign weights to sections. Four sections were considered that are a standard part of most of the research articles such as 1) Introduction, 2) Literature Review, 3) Methodology, and 4) Results and Discussions.

1) CALCULATING SECTION-WISE CITATION COUNT

The frequency of citation can be termed as its occurrences in a specific research article. This approach is quantitative and simple, where the frequency of citation is considered. For example, if a research article is cited three times, the citation frequency will be three. To calculate the section-wise citations, we considered four key sections [36] 1) Introduction, 2) Literature Review, 3) Methodology, and 4) Results and Discussions. Then, considering citation pairs, the citations were computed. Table 6 presents the section-wise citation count for dataset D1.

2) WEIGHT ASSIGNMENT TO SECTIONS

For automatic weight assignment, we used supervised machine learning models 1) Neural Network and 2) Multiple Regression. Many researchers used these models to calculate the weights. For example, Karakaya and Awasthi [37] proposed an approach for relative weight calculation using Multiple Regression. Multiple Regression considers a single dependent feature and multiple independent features. Similarly, Neural Network was utilized by Choi *et al.* [38] for landslide susceptibility analysis. Therefore, we focused on Multiple Regression and Neural networks for weight calculation.

F. SENTIMENT ANALYSIS

The sentiment analysis provides the intent of citation, whether the document is positively or negatively cited. The positively cited citations would be more probably important ones. In this phase, we extracted the sentences with citations and performed sentiment analysis to explore the essence of citation. This step consists of three modules. In the first module, we select the window size for citation sentences.

TABLE 6. Section-wise frequencies of dataset D1.

Paper	Cited-by	Follow-up	Frequency	Introduction	Literature	Methodology	Results
A97-1011	P01-1006	1	1	0	0	1	0
A97-1011	E12-1072	0	1	0	0	1	0
C00-1072	C02-1130	1	2	0	0	2	0
C00-1072	C04-1077	0	1	0	0	1	0

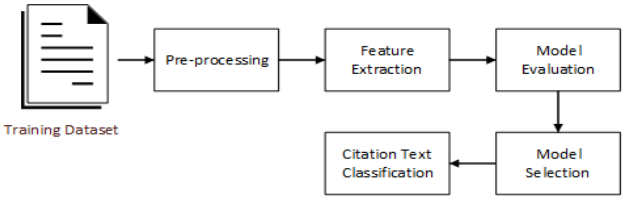

FIGURE 2. Classification of citation text.

In the second step, we performed the sentiment analysis, and thirdly, the sentiment score was calculated.

1) SELECTING WINDOW SIZE

For sentiment analysis, different citation windows can be considered, such as 2-3 sentences across the citation, a single sentence after, and a single sentence before the citations, or it can be one sentence where the citation is present. We selected the window size 1, considering the sentence where citation occurred as this approach is termed better than other approaches [39]. Next, the citation sentence was extracted using in-text citation identification. The citation sentence extraction consisted of the following steps:

- Identification of the citation using in-text citation identification.
- The existing text before opening brace was picked till the full stop of the last sentence appeared.
- The after the text of closing brace was picked till the full stop of this sentence.
- Storing the picked sentence in comma-separated values (CSV)file.

2) CITATION SENTIMENT ANALYSIS

To classify the text, sentiment analysis was used. Different machine learning algorithms exist for text classification, but the model produces better results. Therefore, the commonly used classifiers were evaluated, such as Multinomial Naïve Bayes, Support Vector Machine, Random Forest, K-Nearest Neighbour, and Logistic Regression. Finally, the model that produced high accuracy and a high average macro score was selected for the classification of citations into Negative, Positive, and Neutral categories. The process is represented in Figure 2.

3) TRAINING DATASET

To train the models, a training dataset was used that was collected by [23]. This dataset consisted of four features, 1) Source_Paper_ Id, 2) Target_Paper_ID, 3) Sentiment and 4) Citation_Text.

TABLE 7. Training dataset statistics.

Sentences	Quantity
Positive	829
Negative	280
Neutral	7,627
Total Sentences	8,736

The first feature contains the ID of the cited research articles, while the second feature shows the IDs of citing research articles. In the Sentiment feature, 'p' reflects the positive, 'n' represents negative, and 'o' represents neutral or objective sentiments. The last feature contained the text having the citation sentences. This dataset was given as input to the state-of-the-art machine learning models. The summary of the training dataset is presented in Table 7.

4) PRE-PROCESSING

Preprocessing is important while manipulating the text or performing text classification. In this step, we tokenized the sentences and then converted those tokens into lower case. After that, the tokens were labeled using parts of speech (POS) tagging. Finally, the stop words were removed from the text, and lemmatization was performed. This preprocessing was conducted using Wordnet and Natural Language tool kit.

5) FEATURE EXTRACTION AND SIMILARITY CALCULATION

To extract the features for sentiment analysis, we used the Term Frequency – inverse document frequency. This technique is statistical and measures the relevancy of a word to a document. Thus, it multiplies two metrics, 1) the occurrence of a word in a document and 2) the inverse document frequency concerning a set of documents.

$$tf - idf(t, d, D) = tf(t, d) * idf(t, D) \tag{1}$$

Equation (1) is used for extracting the features. In this experiment top, 30% [5] of features were considered. Using this technique, Unigram, bigram, and trigram features were evaluated. To calculate the similarity between the features of citation sentences, Cosine similarity [40] was used. Cosine similarity can reflect the relatedness of a text corpus. The high value of cosine similarity, the higher the relatedness of input text.

6) MODEL SELECTION AND SCORE COMPUTATION

Different performance measures were used to evaluate the models, such as F-measure, Precision, Recall, and mean

TABLE 8. Characteristics of D1.

Total Citation Pairs	Important Citations Pairs	Non-Important Citations Pairs	Sections of Articles	Citations Frequency
457	69	388	Introduction	155
			Literature Review	131
			Methodology	404
			Results and Discussions	77

TABLE 9. Characteristics of D2.

Total Citation Pairs	Important Citations Pairs	Non-Important Citations Pairs	Sections of Articles	Citations Frequency
282	89	193	Introduction	157
			Literature Review	122
			Methodology	116
			Results and Discussions	69

accuracy values. For validation, we utilized the 10-fold technique. Six algorithms were investigated for sentiment analysis: Multinomial NB, Linear SVC, Bernoulli NB, Logistic Regression, and K-Neighbors. First, the model that produced high f-measure value than other models was selected. Furthermore, the linear Support Vector Classifier observed a high macro score compared to other models. Therefore, a linear support vector classifier was used to classify the citations into negative, positive, and neutral classes. The selected machine learning algorithm calculated the sentiment scores. The sentences were classified into three categories such as positive, negative, and neutral. The frequency of citations was also considered, and for each citation, the citation sentence was extracted, and we calculated the sentiment score.

G. CITATION COUNT

In this step, the frequency of the citations is calculated [41]. It is the identification of citation occurrences in a research document. For example, if a research article is cited five times in another article, the citation count would be 5. This experiment counts all the citation of citation pairs irrespective of their sections. For citation count calculation, ParCit [35] was used, which is an openly available tool. This tool considers the structure of the research article for citation extraction and calculation.

H. DOCUMENT SIMILARITY SCORE CALCULATION

The similarity of citation pairs such as cited and cited articles can be deemed important in verifying the important citation. To calculate the similarity, we considered the whole content of the research papers. For this purpose, the content was extracted from the files. Further, the preprocessing was performed for removing stop words were removed, and words were converted to their base by applying stemming. Further, the key terms were identified using cosine similarity and term frequency-inverse document frequency [40] on extracted vital terms. The similarity value indicated the relatedness of the research articles.

IV. RESULTS AND DISCUSSIONS

A. SECTION-WISE WEIGHT ESTIMATION

To conduct this research work, we utilized two datasets, D1 and D2. The dataset D1 was compiled by Valenzuela and two experts of the domain, who annotated this dataset. This dataset consisted of 457 citation pairs, out of which 388 citation pairs were non-important, and 69 citation pairs were annotated as important. We categorized these citations concerning their sections, and 155 citation pairs were observed in the Introduction section; similarly, 131 citations were in the Literature review, and 404 and 77 citations were found in the Methodology and Results and Discussion sections, respectively. Overall the citation found in dataset D1 was 767. The citations in the Methodology sections were more in number than in the other sections, and the minimum citations were in the Results and Discussions sections. For dataset D1, the characteristics are described in following Table 8.

While the dataset D2 consisted of 311 citation pairs but for ID's 32, 71, 135, 152, 156, 157, 163, 164, 175, 180, 187, 191, 192, 195, 198, 199,216,222, 228, 230, 235, 244, 246, 262, 266, 290, 303, 316 and 317, we were unable to download the articles. Therefore, we researched with 282 citation pairs, out of which 193 citation pairs were non-important, and 89 were important. Further, we considered the sections for citations; 157 citations were observed in Introduction, 122 citations in Literature Review, 116 were found in Methodology, and similarly, 69 citations were in the Results and Discussion sections of the citation pairs. The total citation frequency for all citation pairs was 464 because a citation can be made multiple times. The maximum citations were made in the Introduction sections, and the minimum ones were in the Results and Discussion sections. The statistics of dataset D2 are presented in following Table 9.

To automatically assign the weights to sections, we utilized two machine learning algorithms 1) Neural Network and Multiple Regression. However, both datasets were imbalanced as there were more non-important citations. Therefore, we applied Smote Filter while considering five neighboring instances. This filter creates virtual instances by considering the neighbor instances. These algorithms were trained on 60% data and tested on the remaining 40% [5].

In Multiple Regression, we kept the Y-intercept as 0, so there would be no need to add the constant value to the obtained weights. As a result, the obtained weights by both machine learning algorithms had a sum equal to 1. Table 10 presents the weights obtained by the Neural Network. The algorithm Neural Network assigned a maximum

TABLE 10. Section-wise weights by neural network.

Algorithm	Sections	Assigned Weights	Order
Neural Network	Introduction	0.19095378	3
	Literature Review	0.17626289	4
	Methodology	0.28763501	2
	Results and Discussions	0.34514832	1

TABLE 11. Section-wise weights by multiple regression.

Algorithm	Sections	Assigned Weights	Order
Multiple Regression	Introduction	0.1891921316	3
	Literature Review	0.1470393226	4
	Methodology	0.3663496373	1
	Results and Discussions	0.2974189085	2

TABLE 12. Multiplication of weights with section-wise citation count.

Paper	Cited-by	Follow-up	Frequency	Introduction	Literature	Methodology	Results
A00-1043	C00-2140	0	1	0	0	0.28763501	0
A00-1043	P02-1057	0	1	0.19095378	0	0	0
H05-1079	D10-1074	0	1	0	0.17626289	0	0
H05-1079	J07-3004	0	1	0	0	0	0.34514832
H05-1079	N06-1005	1	4	0.76381512	0	0	0

weight to the Results and Discussion sections, and the least weight was assigned to the Literature Review sections. On the other hand, The Multiple Regression assigned more weight to the Methodology sections and less to the Results and Discussions sections, as mentioned in Table 11. At the same time, the research community focuses more on Results and Discussion citations. Therefore, we utilized the weights obtained by Neural Network for further manipulation. The weights were further multiplied to the section-wise in-text citations. For dataset D1, the multiplication of weights to the section-wise citation count is presented in Table 12.

B. SENTIMENT SCORE

To compute the sentiment score, we extracted the citation sentences and classified the sentences into positive, negative, and neutral categories. We considered six algorithms, such as Linear SVC, Multinomial NB, Bernoulli NB, K-Neighbors, and Logistic Regression, for extracting the essence of citations. These algorithms were trained on the separate training dataset, and we utilized Unigram, Bi-gram, and Tri-gram measures. These algorithms were evaluated based on Accuracy, Precision, Recall, and F-measure parameters. The following table presents the macro F-measure values with Unigram, Bi-gram, and Tri-gram features.

As described in Table 13, on average, the considered models produced better results with Unigram than Bigram and Trigram. Therefore, Unigram was selected for feature evaluation. The Weighted average, Micro average, and Macro average values for Accuracy, Precision, Recall, and F-measure are presented in the following Figure 3.

The overall results produced by Linear SVC were better than the other models. Therefore, for further computations, we utilized Linear SVC. This model was applied to citation sentences, and the sentences were categorized as positive, negative, and neutral sentences. The following table describes the results obtained with the Linear SVC algorithm.

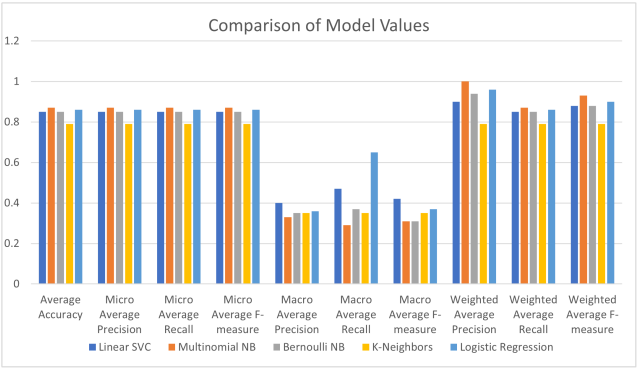

FIGURE 3. Performance comparison of models.

In Table 14, we computed the sentiment of citation sentences concerning their number of occurrences. As in the first row, a paper is cited a single time and was classified as Neutral. While the last row explains that a paper was cited twice, and both times, it had different sentiments: one time it was positively cited, and one time it was cited as neutral.

C. CLASSIFICATION OF CITATIONS WITH ALL COMBINED FEATURES

To classify the citations, we combined all the considered features consisting of 1) Citation Count, 2) Content Similarity, 3) Section-wise in-text citation count, and 4) Citation Sentence Sentiment Score.

The Section-wise in-text citation count feature was further divided into four features such as 1) Introduction Citation Frequency, 2) Literature Review Frequency, 3) Methodology Citation Frequency, and 4) Results and Discussions Citation Frequency. Similarly, the Sentiment Scores were divided into three sub-features 1) Positive, 2) Negative, and 3) Neutral citation frequencies. Finally, we combined all these features and classified the citations using machine learning algorithms Support Vector Machine, Random Forest, and Multilayer

TABLE 13. Macro F-measure values.

Features	Linear SVC	Multinomial NB	Bernoulli NB	K-Neighbors	Logistic Regression
Unigram	0.43	0.31	0.31	0.35	0.37
Bigram	0.42	0.31	0.40	0.36	0.36
Trigram	0.41	0.32	0.39	0.35	0.36

TABLE 14. Sentiment of citation sentences.

Paper	Cited-by	Follow-up	Frequency	Positive Frequency	Negative Frequency	Neutral Frequency
A00-1043	C00-2140	0	1	0	0	1
A00-1043	P02-1057	0	1	0	0	1
A97-1011	W09-1118	0	1	0	0	1
I05-2038	P06-4005	1	2	1	0	1

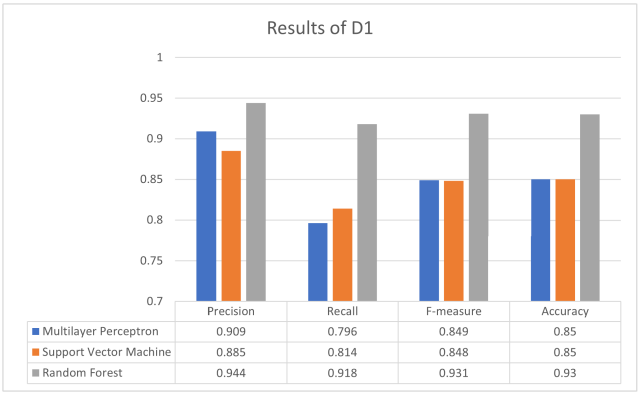

FIGURE 4. Results of dataset D1 with all features.

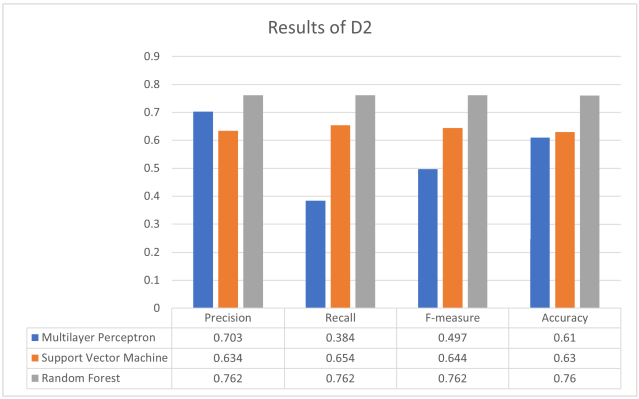

FIGURE 5. Results of dataset D2 with all features.

Perceptron. We utilized Smote filter to balance important and non-important classes while considering five neighbors to create virtual instances. For dataset D1, the results are presented in Figure 4. The Random Forest produced better results than other models. This model achieved a 0.94 value for precision, 0.918 value for Recall, and 0.93 for F-measure. The highest accuracy achieved was 0.93. at the same time, the performance of other models, such as Multilayer Perceptrons and Support Vector Machine, was comparable. The values of F-measure and Accuracy achieved by both models were 0.84 and 0.85, respectively. At the same time, Multilayer Perceptron gained a 0.9 value of precision and a 0.79 value of Recall. SVM produced 0.88 precision and 0.814 Recall value.

While for dataset D2 Random Forest was observed with high results compared to other models. Random forest achieved 0.76 value of Precision, Recall, and F-measure. While Multilayer Perceptron gained 0.73 Precision, 0.384 Recall, and 0.497 F-measure. While SVM achieved 0.634 Precision value, 0.654 Recall value and 0.644 value of F-measure. While Random Forest achieved the maximum accuracy. The results of D2 are presented in Figure 5. On both datasets, the Random Forest model outperformed other candidate models. However, for dataset D1, the proposed approach achieved a value of precision of 0.94, and for D2, it was 0.76. There exists a difference in the results of D1 and D2. The reason for achieving higher results for D1 is that the articles in this dataset mostly contain multiple citations against a single article. Conversely, for D2, most articles contain single citations against a single article.

D. COMPARISON

Using dataset D1, the research work of Valenzuela *et al.* [7] achieved 0.68 precision values by utilizing metadata and content-based features. This work was further enhanced by Qayyum and Afzal *et al.* [10]. The author enhanced the precision value up to 0.72 by using metadata-based features only. Similar research was conducted by Aljuaid *et al.* [5], and the researcher achieved a 0.83 value of precision with the Random Forest algorithm. We have also contributed to important citation identification in our previous research work, and the precision value was enhanced to 0.85. Finally, the state-of-the-art approach of Wang *et al.* [12] further improved the results and gained a 0.91 precision value. The graphical comparison of different approaches till now is presented graphically in the following Figure 6.

While the dataset D2 was utilized by Faiza *et al.*, Aljuaid *et al.*, and in our previous research work. The research work of Faiza *et al.* gained a 0.62 F-measure, Aljuaid *et al.* achieved a 0.69 value F-measure, and in our previous contribution, we were able to enhance the value up to 0.72. In our proposed approach, we have added sentiment features that improved the results further. The comparison for dataset D2 is presented in Figure 7. The proposed approach utilized section-wise in-text citation weights, sentiment analysis,

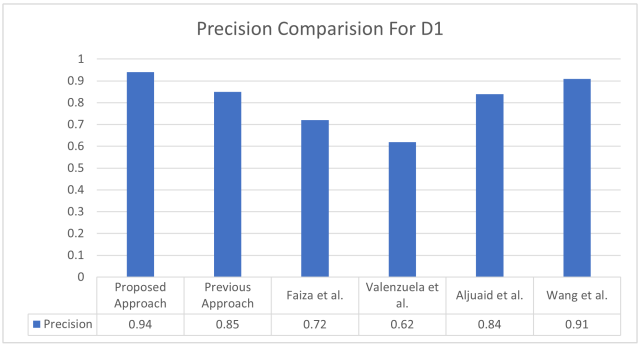

FIGURE 6. Results comparison for D1.

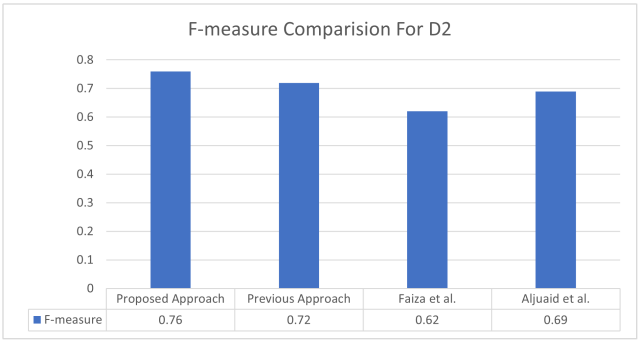

FIGURE 7. Results comparison for D2.

and article similarity features which were not combinedly considered by any approach. Therefore this approach performed better than previous state-of-the-art approaches and achieved a higher F-measure value.

V. CONCLUSION

To conduct a research study, a citation is considered a scientific measure to evaluate the significance of the work. Citations are used for computing multiple aspects of research such as impact factor of journals, ranking of researchers, ranking of institutions, etc. While the criteria of computing these important measures, all the citations are counted with equal importance. The research community concluded that all citations are not equally important. The reasons for citations should be incorporated, as a citation providing background and the other one extending the work cannot be of the same worth. Therefore, researchers have developed different approaches to distinguish important citations from non-important citations. These state-of-the-art approaches are content-based, bibliographic based, or meta-data based. However, the achieved accuracy of these state-of-the-art approaches is insufficient for making potential decisions. In this work, we have introduced an approach based on four submodules such as 1) automatically assigning the appropriate weights to sections where the citations were made, using Neural Network, 2) sentiment analysis on citation sentences, 3) calculating the similarity of research articles, and 4) utilizing overall count of citations. We used two datasets, D1 and D2, collected by Valenzuela *et al.* and Faiza *et al.* to perform the citation classification. These were earlier used in state-of-the-art approaches. Multilayer Perceptron, Support Vector Machine, and Random Forest were utilized for classification. The Random Forest algorithm achieved the highest value. The results revealed that the proposed approach achieved a 0.94 value, comparatively higher than any other approach.

CONFLICT OF INTEREST

The authors have no conflict of interest to report regarding the present study.

REFERENCES

[1] F. Narin, *Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity*. Hill, NJ, USA: Computer Horizons Cherry, 1976.
[2] L. Bornmann and H. Daniel, "What do citation counts measure? A review of studies on citing behavior," *J. Document.*, vol. 64, no. 1, pp. 45–80, Jan. 2008.
[3] H. Inhaber and K. Przednowek, "Quality of research and the Nobel prizes," *Social Studies Sci.*, vol. 6, no. 1, pp. 33–50, 1976.
[4] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 46, pp. 16569–16572, 2005.
[5] H. Aljuaid, R. Iftikhar, S. Ahmad, M. Asif, and M. T. Afzal, "Important citation identification using sentiment analysis of in-text citations," *Telematics Informat.*, vol. 56, Jan. 2021, Art. no. 101492.
[6] M. J. Moravcsik and P. Murugesan, "Some results on the function and quality of citations," *Social Stud. Sci.*, vol. 5, no. 1, pp. 86–92, 1975.
[7] M. Valenzuela, V. Ha, and O. Etzioni, "Identifying meaningful citations," in *Proc. Workshops 29th AAAI Conf. Artif. Intell.*, 2015, pp. 1–6.
[8] B. Finney, "The reference characteristics of scientific texts," Ph.D. dissertation, City Univ., London, U.K., 1979.
[9] X. Zhu, P. Turney, D. Lemire, and A. Vellino, "Measuring academic influence: Not all citations are equal," *J. Assoc. Inf. Sci. Technol.*, vol. 66, no. 2, pp. 408–427, 2015.
[10] F. Qayyum and M. T. Afzal, "Identification of important citations by exploiting research articles' metadata and cue-terms from content," *Scientometrics*, vol. 118, no. 1, pp. 21–43, Jan. 2019.
[11] S. Nazir, M. Asif, S. Ahmad, F. Bukhari, M. T. Afzal, and H. Aljuaid, "Important citation identification by exploiting content and section-wise in-text citation count," *PLoS ONE*, vol. 15, no. 3, Mar. 2020, Art. no. e0228885.
[12] M. Wang, J. Zhang, S. Jiao, X. Zhang, N. Zhu, and G. Chen, "Important citation identification by exploiting the syntactic and contextual information of citations," *Scientometrics*, vol. 125, no. 3, pp. 2109–2129, Dec. 2020.
[13] E. Garfield, "Is citation analysis a legitimate evaluation tool?" *Scientometrics*, vol. 1, no. 4, pp. 359–375, May 1979.
[14] I. Spiegel-Rosing, "Science studies: Bibliometric and content analysis," *Social Stud. Sci.*, vol. 7, no. 1, pp. 97–113, 1977.
[15] C. R. Myers, "Journal citations and scientific eminence in contemporary psychology," *Amer. Psycholog.*, vol. 25, no. 11, p. 1041, 1970.
[16] M. Garzone and R. E. Mercer, "Towards an automated citation classifier," in *Proc. Conf. Can. Soc. Comput. Stud. Intell.* Cham, Switzerland: Springer, 2000, pp. 337–346.
[17] C. L. Giles, K. D. Bollacker, and S. Lawrence, "CiteSeer: An automatic citation indexing system," in *Proc. 3rd ACM Conf. Digit. libraries (DL)*, 1998, pp. 89–98.
[18] S. B. Pham and A. Hoffmann, "A new approach for scientific citation classification using cue phrases," in *Proc. Australas. Joint Conf. Artif. Intell.* Cham, Switzerland: Springer, 2003, pp. 759–771.
[19] H. H. Bi, J. Wang, and D. K. J. Lin, "Comprehensive citation index for research networks," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 8, pp. 1274–1278, Aug. 2011.
[20] S. Teufel, A. Siddharthan, and D. Tidhar, "Automatic classification of citation function," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2006, pp. 103–110.

[21] K. Sugiyama, T. Kumar, M.-Y. Kan, and R. C. Tripathi, "Identifying citing sentences in research papers using supervised learning," in *Proc. Int. Conf. Inf. Retr. Knowl. Manage. (CAMP)*, Mar. 2010, pp. 67–72.
[22] S. Agarwal, L. Choubey, and H. Yu, "Automatically classifying the role of citations in biomedical articles," in *Proc. AMIA Annu. Symp.*, 2010, p. 11.
[23] H. Small, "Interpreting maps of science using citation context sentiments: A preliminary investigation," *Scientometrics*, vol. 87, no. 2, pp. 373–388, May 2011.
[24] A. Shahid, M. Afzal, and M. Qadir, "Discovering semantic relatedness between scientific articles through citation frequency," *Austral. J. Basic Appl. Sci.*, vol. 5, no. 6, pp. 1599–1604, 2011.
[25] W.-R. Hou, M. Li, and D.-K. Niu, "Counting citations in texts rather than reference lists to improve the accuracy of assessing scientific contribution: Citation frequency of individual articles in other papers more fairly measures their scientific contribution than mere presence in reference lists," *BioEssays*, vol. 33, no. 10, pp. 724–727, Oct. 2011.
[26] A. T. Balaban, "Positive and negative aspects of citation indices and journal impact factors," *Scientometrics*, vol. 92, no. 2, pp. 241–247, Aug. 2012.
[27] C. Dong and U. Schäfer, "Ensemble-style self-training on citation classification," in *Proc. 5th Int. joint Conf. natural Lang. Process.*, 2011, pp. 623–631.
[28] A. Athar, "Sentiment analysis of scientific citations," Comput. Lab., Univ. Cambridge, Cambridge, U.K., 2014.
[29] C. Jochim and H. Schütze, "Towards a generic and flexible citation classifier based on a faceted classification scheme," in *Proc. COLING*, 2012, pp. 1343–1358.
[30] S. Kumar, "Structure and dynamics of signed citation networks," in *Proc. 25th Int. Conf. Companion World Wide Web (WWW) Companion*, 2016, pp. 63–64.
[31] S. Lee, J. Choi, U. Chwae, and B. Chang, "Landslide susceptibility analysis using weight of evidence," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jun. 2002, pp. 2865–2867.
[32] B. H. Butt, M. Rafi, A. Jamal, R. S. U. Rehman, S. M. Z. Alam, and M. B. Alam, "Classification of research citations (CRC)," 2015, *arXiv:1506.08966*.
[33] C. A. Sula and M. Miller, "Citations, contexts, and humanistic discourse: Toward automatic extraction and classification," *Literary Linguistic Comput.*, vol. 29, no. 3, pp. 452–464, Sep. 2014.
[34] S. Nazir, M. Asif, and S. Ahmad, "Exploring the proportion of content represented by the metadata of research articles," in *Proc. 3rd Int. Conf. Advancements Comput. Sci. (ICACS)*, Feb. 2020, pp. 1–7.
[35] I. G. Councill, C. L. Giles, and M.-Y. Kan, "ParsCit: An open-source CRF reference string parsing package," in *Proc. 6th Int. Conf. Lang. Resour. Eval.*, vol. 8, 2008, pp. 661–667.
[36] A. Y. Khan, A. K. Shahid, and M. T. Afzal, "Extending co-citation using sections of research articles," *Turkish J. Elect. Eng. Comput. Sci.*, vol. 26, no. 6, pp. 3345–3355, 2018.
[37] F. Karakaya and A. Awasthi, "Robustness and sensitivity of conjoint analysis versus multiple linear regression analysis," *Int. J. Data Anal. Techn. Strategies*, vol. 6, no. 2, pp. 121–136, 2014.
[38] J. Choi, H.-J. Oh, J.-S. Won, and S. Lee, "Validation of an artificial neural network model for landslide susceptibility mapping," *Environ. Earth Sci.*, vol. 60, no. 3, pp. 473–483, Apr. 2010.
[39] L. C. Smith, "Citation analysis," *Library Trends*, vol. 30, no. 1, pp. 83–106, 1981.
[40] A. R. Lahitani, A. E. Permanasari, and N. A. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment," in *Proc. 4th Int. Conf. Cyber IT Service Manage.*, Apr. 2016, pp. 1–6.
[41] S. Nazir, M. Asif, and S. Ahmad, "Important citation identification by exploiting the optimal in-text citation frequency," in *Proc. Int. Conf. Eng. Emerg. Technol. (ICEET)*, Feb. 2020, pp. 1–6.

SHAHZAD NAZIR received the M.S. degree in computer science from the National Textile University, Faisalabad, where he is currently pursuing the Ph.D. degree in computer science. His research interests include recommending relevant documents, information systems, deep learning, and natural language processing.

MUHAMMAD ASIF received the M.S. and Ph.D. degrees from Asian Institute of Technology (AIT), in 2009 and 2012, respectively, on HEC Foreign Scholarship. During the course of time, he was a Visiting Researcher with the National Institute of Information and Communications Technology, Tokyo, Japan. He has worked on some projects, including the Air Traffic Control System of the Pakistan Air Force. He is currently a Tenured Associate Professor of computer science with the National Textile University, Faisalabad. Before this, he was a Research Scholar with the Department of Computer Science and Information Management, Asian Institute of Technology, Thailand. He also serves as an Associate Editor for IEEE Access, the prestigious journal of IEEE. He is a reviewer of several reputed journals and authored several research papers in reputed journals and conferences. He is also a permanent member of the Punjab Public Service Commission (PPSC) as an Advisor, and a Program Evaluator at the National Computing Education Accreditation Council (NCEAC), Islamabad.

SHAHBAZ AHMAD received the M.S. degree in computer science from the National Textile University, Faisalabad. He is currently pursuing the Ph.D. degree in computer science with the Capital University of Science and Technology. He is currently working as a Lecturer with the Department of Computer Science, National Textile University. He has published many high class research papers in journals and conferences.

HANAN ALJUAID is currently working as an Associate Professor with the Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University (PNU), Riyadh, Saudi Arabia.

RIMSHA IFTIKHAR received the M.S. degree in computer science from the National Textile University, Faisalabad. She has research and teaching experience. Her research interests include recommending relevant documents, information systems, deep learning, and natural language processing.

ZUBAIR NAWAZ is working as an Assistant Professor of data science with the University of the Punjab. He has several years of teaching and research experience.

YAZEED YASIN GHADI is currently a Professor and the Director of the Software Engineering and Computer Science Programs, Al Ain University, Abu Dhabi Campus. He has an extensive experience of teaching research and publications. He has published his research in world top class journals and conferences.