# A Comprehensive Evaluation of Cue-Words based Features and In-text Citations based Features for Citation Classification

Syed Jawad Hussain[1], Sohail Maqsood[2], Usman Ahmed[6]
Department of Computer Science & IT
The University of Lahore, Islamabad Pakistan

Azeem Khan[4]
Americian Degree Program ADP
Taylor's University
Subang Jaya, Selangor, Malaysia

NZ Jhanjhi[3]
School of Computing & IT (SoCIT)
Taylor's University
Subang Jaya, Selangor, Malaysia

Mahadevan Supramaniam[5]
Research & Innovation Management Centre
SEGI University
Malaysia

***Abstract*—Citation plays a vital role in the scientific community of evaluating the contributions of scientific authors. Citing sources delivers a measurable way of evaluating the impact factor of journals and authors and allows for the recognition of new research issues. Different techniques for classifying citations have been proposed. Citations that provide background knowledge in the citing document have been classified as non-important or incidental by previous researchers. Citations that extend previous work in the citing document are classified as important. The accuracy achieved by existing citation models is not much higher. Better features need to be included for accurate predictions. A hybrid approach would present all possible combinations of cue-words and in-text citation-based features for citation classifications.**

***Keywords*—Cue-words; in-text citation; hybrid**

## I. Introduction

Why are citations important? This question has grabbed many authors' attention in recent decades, and a variety of different answers were found in the existing literature. Citations are important because journals base their impact factor on them. Author rankings and awards also rely on citation-based measurements. The quality of work is measured by how often it is cited by others and whether it extends the work of others. Highly cited works in the research community mean recognition and a greater impact on scholarly research. Researchers recognize diverse causes of citations; however, past analysis shows that not all citations are equal. Some citations extend current works using the same algorithms and comparing results, whereas others are used as background knowledge that are from the same domain but do not directly influence the new work. Different techniques have been proposed to classify citations; these are identified as important and non-important. Citations that provide background knowledge for the citing document have been classified as incidental by previous researchers. In the scholarly community, different researchers have claimed that some authors unintentionally read cited papers incorrectly [1], [2] merely to fill up the citations section. Other errors that occur include page and volume number not being the same in the cited section. Some authors claim that 40% of citations are copied to fill in the reference section without being cited in the paper [3], [4], [5]. This work describes references using multiple feature types that influence the research community on citing research articles. These are influential references which contribute to citing papers such as new ideas, research problems, methodologies, and experiments. An automated system was proposed by [6] to classify citations into two classes: negative and positive. Past analysis shows that not all citations are treated equally, such as those deemed essential vs. non-essential or non-important. Recently [7], [8] contributed to the research community by being the first to tackle the problem of identifying important citations. Citations that provide background knowledge in the citing document are classified as incidental by previous researchers. Citations that extend previous work in the citing document are classified as important. Their approach is to use different features to identify important and incidental citations. These include total number of direct citations, number of direct citations per section, total number of indirect citations, and number of indirect citations per section. Author overlap is also considered helpful, where citations appear in tables or captions. Features also include number of reference, number of paper citations or all citations, similarity between abstracts, page rank, number of total citing papers after transitive closure, and the field of cited paper. Based on these features, they identify important and incidental citations. Our work is similar to Valenzuela et al; we make a detailed comprehensive evaluation of all possible combinations of cue-words based features and in-text based features and create a hybrid approach to classifying citations. In the literature review, we identified a research gap in that no one has detailed all possible combinations of cue-words based, in-text based, and hybrid features to evaluate the best feature among them.

*1) Cue-words:* In this feature, we use cue-words from the sentence in which the citations occur. We selected the sentence where citations appear and one sentence before the citations and one sentence after the citations appear. We collected all citations in the respective paper and counted the words.

*2) Cue-words count*: In this feature, we counted cue-words and checked the occurrence of each cue-word in the whole paper.

Section cue-word count: In this feature, we counted cue-words and checked the occurrence of single cue-words in each section to identify the word's importance.

*3) In-text count:* In-text citations which the author uses to support background knowledge or to extend previous work within the whole paper.

Section in-text count: In this feature, we use an in-text count and check the occurrence of single in-text within each section to identify importance of each citation.

*4) Hybrid:* all possible combination of cue-words bansed features and in-text citations based features.

The rest of the paper is divided in five sections, including related work, methodology, experimental results and discussion, conclusion, and future work.

## II. Related Works

Considerable effort has been made to improve the citation classification system in the past decades. Many authors contribute different approaches to examine citation classification. Some reasons why authors cite a work include [9]:

- Criticized for weak work by the community.
- Appreciation for great effort by the research community.
- To provide background knowledge.
- To reference tools and techniques.
- To strongly disagree with those who claim others' work.

Garzone [10] introduced 35 different categories of classifying citation functions based on cue phrases. This work is based on different combinations of schemes, but the focus is on [11] the scheme with limitations applied. We further subdivided the information into 10 top level categories after all processing was complete. They were the first to claim a new fully automated classification scheme [12], [13]. The automatic scheme inputs entire articles. The results return different citations along with different sets of functions. The literature dealing with the important function of cue phrases in citation classification was given unique consideration. The study relies exclusively on cue-phrase features and content citation-based features for classifying as important and identical. According to CiTO citation typing ontology [14] [15], they identified 90 semantic relationships between papers and citations. In [5] author describes references using multiple types of features; they had a strong influence on the research community in the citing research article. These are some influential references which contribute in citing papers: new ideas, research problems, methodologies, and experiments. According to [5], the classification of these citations would be two broad categories: 1 important and 2 non-important.

Author in [16] presents a technique called a novel automated technique, which classified as sentiment positive or sentiment negative. In this technique, the citations appearing in citing papers are extracted using the sentiment lexicon then classified into positive and negative. The data set was 150 research papers and they used classifier Naive Bayes for sentiment analysis. The approach used in this research work for classifying citations is as follows:

- TYPE I: Positive
- TYPE II: Negative
- TYPE III: Neutral

In TYPE I: Accuracy achieved in precision 0.84, recall 0.94 and f-measure 0.89: Positive results against support of 109 research papers. In TYPE II: Negative precision 0.25, recall 0.10 and f-measure 0.14 results against support of 21 papers.

They used a dataset to classify citations for this research work consisting of 2829 citations against 116 research papers. Every citation in the above-mentioned categories was tagged manually for citation classification and used 10-fold cross validation. [17] proposed a technique that supervised the citation classification function. The system is categorized into four levels:

- Neutral category: didn't appreciate or criticize any approach in this category.
- Contrast Category: compared different approaches.
- Positive Category: made agreement or compatibility with existing techniques.
- Weak Category: criticized the cited weak work.

They used 548 citations of adopted supervised learning in these categories:

- CoCoGM: compared the methodology and goal of the research article.
- PMot: motivation for the work.
- PSim: similarity between both works citing and cited paper.
- PSup: current work is based on previous work.
- Coco: cite the superior state of author's work.

This experiment was applied on datasets of 116 research articles; precision 0.75, Kappa 0.59 and Micro f 0.68 were calculated.

Cue phrases were used to classify citation system in KAFTAN, presented by [18]. The four categories to classify citations are as follows:

- Basis: In this Category based on another work.
- Support: In this Category supported by other work.
- Limitation: In this Category criticizing the cited work on its weaknesses.
- Comparison: In this Category comparing different approaches.

The technique is to describe several features based on different types of polarity and context-level features such as 1) grouping referencing; 2) tagging referencing; and 3) polity and non-syntactic referencing removal [17][19]. Support vector machine (SVM) and Random forest (RF) classifier used for citation classification. Accuracy achieved precision 92%, recall 76.4% and f-measure 70.5% used 10-fold cross validation. [20] Hassan technique expanded [7] work and proposed 14 different features to classify citations. Grouped into three main categories such as context-based features, cue-words based features and textual features. These features are evaluated according to five different classifiers: K Nearest Neighbor (KNN), SVM, RF, Naive Bayes, and Decision Tree. The best classifier identified with the highest accuracy is RF with 91%.

Author in [21] compared and built a technique that classifies citations into important vs. non-important. They used four different state-of-the-art datasets of their work with 64 different features, 29 of which are for Extra Tree Classifier. They manually selected 450 annotated citations and classified the citations using RF and SVM classifiers. They used 20,527 research articles from a well-known dataset, with 106,509 citations chosen against the dataset. In [22][23], the author describes citation function and polarity to classify them using a scheme of eight categories, which helps show the importance of citations in the community. This paper is in the biomedical domain. The dataset used 640 biomedical research articles and collected 1,823 meaningful citations for polarity and experiments. Set of two main features used to automatic citation classification, such as Part-of-speech tags (POS) and word n-gram using a machine learning algorithm to classify citations into eight categories using Maximum Entropy and SVM classifiers against meaningful citations to generate results.

Positive: in this class, the author agrees with previous work or extends it; two categories belong to this class: Confirmation precision 0.822, recall 0.638 and f1-score 0.719. and Being-confirmed. precision 0.77, recall 0.42 and f1-score 0.54.

Negative: in this class, the author disagrees with previous work for these reasons: weakness of the previous work, data not satisfying. Two Categories belong to this class: Contrast/Conflict precision 0.77, recall 0.52 and f1-score 0.62. Unsolved precision 0.554, recall 0.463 and f1-score 0.504.

Neutral: in this category, the author was not criticized and previous work was not appreciated. Four categories belong to this class: Perfunctory/Background, precision 0.67, recall 0.792 and f1-score 0.736. Statement precision 0.802, recall 0.582 and f1-score 0.674. Comparison precision 0.557, recall 0.788 and f1-score 0.653. and Multi-comparison precision 0.552, recall 0.431 and f1-score 0.484.

Here are the results for detailed feature combinations on citation function classification: The SVM classifier with POS tags + 1-3 grams + dependencies features achieved the best result. An automated system was proposed by [6] to classify citations into two classes: negative and positive. Past analysis shows that all citations are not treated equally, such as essential or non-essential/non-important. Recently in the research community [7]. They became the first to tackle this problem by identifying important citations that are referred to for providing background knowledge in the citing document. These citations are categorized in the class of incidental by previous researchers. Citations which referred to previous work in the citing document were categorized in the class of Important. Their approach is to use different features to identify important and incidental citations. These features are: Total number of direct citations, Number of direct citations per section, Total number of indirect citations and number of indirect citations per section, Author overlap, is considered helpful, citations appear in the table/caption, 1/number of reference, number of paper citations/all citations, Similarity between abstract, PageRank, Number of total citing paper after transitive closure, field of cited paper. On the bases of these features they identify important citations. A new term is used in the research community by Valenzuela, who categorized citation into important vs. incidental. These categories were sub-divided into two further categories: Important, Using others' work, and expanding on others' work. Incidental: Related work and Comparison. These categories were evaluated using 12 different features to classify citation:

- (F1) Total number of direct citations:
- (F2) Total number of direct citations per section:
- (F3) Total number of indirect citations and number of indirect citations per
- section:
- (F4) Author overlap:
- (F5) Is considered helpful:
- (F6) Citation appears in table or caption:
- (F7) 1/number of references:
- (F8) Number of paper citation / all citations:
- (F9) Similarity between abstracts:
- (F10) PageRank:
- (F11) Number of total citing papers after transitive closure:
- (F12) Field of the cited paper:

Author in [24] proposed a technique that classifies binary citation. This scheme is based on metadata and content-based parameters to classify citations. Faiza is the first to classify

citation as important vs. non-important using metadata-based hybrid parameters. Two types of datasets were used in this work: D1 and D2. Dataset D1 consists of a standard dataset which means it is authentic, published in top of the conference and even latest in the current domain. The dataset used 20,575 research articles which have 106,509 citations; among those, 465 annotated pairs of paper datasets are used in D1. D2 consists of pairs of 488 papers with best source and annotated (citing papers) citations. Features were used to classify important vs. non-important citations. There are two different types of parameters: metadata based and content based.

- Metadata

Title, Author name, key-word, Category, Reference.

- Content-based Abstract,

Cue-phrase set of static cue-phases used in this work.

WEKA Machine learning tool is used for classifications with the help of these classifier to generate results against dataset.

- SVM
- KLR
- RF

The generated results are better than the benchmark precision achieved on SVM 0.68, KLR 0.62 and RF 0.72. The similarity of our work is based on identifying important citations by [7], [21] and [24]. Authors describe how not all citations are treated equally. They categorized citation into important and non-important classes. Citations were classified into important and incidental using different features. The classifiers used to evaluate citations were SVM, KNN, Naive Bayes and RF. The dataset used for this thesis contained 465 research articles from the computing domain form ACL anthology. Only 14% of the citations calculated as important and the rest were incidental citations. The evaluation criteria were Precision, Recall, and f-measure. After a comprehensive study of the literature, the state-of-the-art approaches were found in the same domain. We found that methods for citation classification are based on linguistic cue phrases and In-text citations. The concise sign of these approaches is described in Table 2.1 with references, methodologies, strengths and weaknesses. In this thesis, we begin with a literature review of the citation classification. The varieties of classification techniques and their automatic classification schemes with machine learning algorithms were closely observed. In [16] present a technique called a novel automated technique that classified sentiment positive or sentiment negative. In this technique, the citations appearing in citing papers are extracted using the sentiment lexicon; then they are classified into positive and negative. Author in [9] introduced 35 different categories of classifying citation functions based on cue phrases. This work is based on a different combination of schemes, where the focus is on [11] and the scheme applies limitations to it. Author in [17] proposed a technique that supervised for citation classification function; this system is categorized into four different levels: Neutral, Contrast, Positive, and Weak. The average SVM accuracy achieves precision 0.83, Kappa 0.84, and Micro f 0.83 and RF accuracy has Precision 0.83, Kappa 0.84, and Micro f 0.83. The literature reviews that deal with the important function of cue phrases in citation classification were given unique consideration. CiTO citation typing ontology [14]; according to [15], they identified 90 semantic relations between papers and citations. [16] presents a technique called a novel automated technique, which classifies sentiment positive or sentiment negative. Author in [5] describes references using multiple feature types; they had a strong influence on the research community in the citing research article. These are some influential references which contribute to citing papers, such as new idea, research problems, methodologies, and experiments. According to [5], the classification of these citations would be two broad categories: one important and two non-important. In [24] proposed a technique that classifies binary citation; this scheme is based on metadata and content-based parameters to classify citation. Faiza was first to classify citations as important and non-important using metadata-based hybrid parameters. Two types of dataset set were used in this work: D1 and D2. Dataset D1 consists of a standard dataset which means authentic, published in top of the conference, and even latest in the current domain. The generated results are better than the benchmark precision achieved on SVM 0.68, KLR 0.62 and RF 0.72. Authors [22] and [23] describe citation functions and polarity for classification. This helps us understand the importance of citations in the research community. To classify, eight categories are merged in three main categories. Maximum Entropy and SVM classifiers against meaningful citations generated the following results: Neutral precision - 0.806, recall 0.931, and F1 0.838. Positive precision - 0.806, recall 0.931, and F1 0.838; and negative precision 0.806, recall 0.931, and F1 0.838. Author in [21] compared and built a technique that classified citations into important and non-important. They used four different state-of-the-art datasets for their work. They used 64 different features, 29 of which are for Extra Tree Classifier. RF and SVM are the classifiers used for classification. The average SVM results were: precision 0.87, recall 0.89, and f-measure 0.84, and RF accuracy was precision 0.9, recall 0.91. and f-measure 0.91.

## III. METHODOLOGY

The literature review was critically examined to explore the different techniques proposed by different authors to classify citations. In citation related studies, the main purpose is identifying and classifying citations. Recently [7] published an article at the AAAI conference on the A1 category. Valenzuela classifies citations into two broad categories: important and non-important. In this thesis, to classify citations using the above-mentioned broad categories, we proposed a new technique using a different combination of cue-words based and in-text-based features to classify the citations that are evaluated in the proposed methodology in Fig. 1. An architecture diagram explains our proposed methodology. In this thesis we use Valenzuela et al.'s 2015 publicly available dataset as a benchmark and Scholarly Big data for experiments published in the AAAI workshop. To evaluate our features, we performed some experiments on the collected dataset to show which features perform better on which classifiers. Author in [25] used four standard label mapped sections: Introduction,

Related work, Methodology and Results. We performed preprocessing on the collected dataset applying the stemming and stop words removal algorithm. After preprocessing, we removed duplicate words and the remaining list is a list of unique words that allow us to evaluate all possible combinations of features. We used four different classifiers: Random Forest RF, Support Vector Machine SVM, Naive Bayes, and KNN. We computed each model on the same dataset and calculated precision, recall and f-measure. For the training and testing of data, samples were studied using the 10 cross fold validation method, and SMOTE was applied to balance the dataset with synthetic value and minority class on the collected dataset for the experiment performed. Python was used for evaluation and generated results and compared with benchmark.

### A. Experimental Dataset

We collected a data set of 416 papers downloaded from the ACL anthology. Approximately 21,500 words were extracted from the dataset to classify citations. 14,000 words were marked as incidental and 7500 words were marked as important. Next preprocessing was applied on the collected words. Onix Text Retrieval Toolkit stop word List was applied for stop word removal from the collected dataset, and a suffix-stripping algorithm was used for stem words in their root form. We removed duplicate words from data which was generated after stop word and stemming was applied. A total of 858 unique words were found and two different files were created. One file contains 631 unique words marked as incidental class and another file contains 227 unique words marked as important class. We used [7] dataset as benchmark which is publicly available for experiments. There are four fields in the dataset. Annotator [25] is used in first field; the Paper field contains the ID of the root paper; the second field contains the Cited By ID of the paper, which refers to the root paper, and the last field contains the labels which range from 0 to 3. 0 and 1 indicate the incidental class; 2 and 3 indicate the important class through papers and Cited by IDs. For further processing from ACL anthology, all papers were downloaded manually.

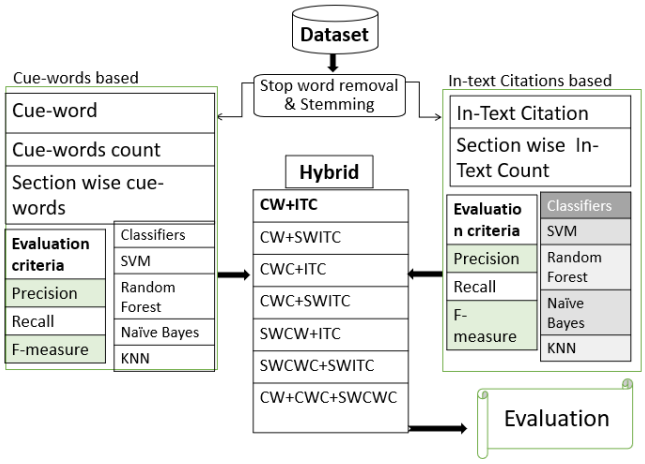



Fig. 1. Context Diagram of Proposed System.

### B. Feature Extraction

Our proposed methodology contains a number of different features. Based on these features, we classified citations as important and incidental. Different classifiers were used to evaluate these features: (1) Cue-words based features; (2) In-text based features; and (3) Hybrid (*H* indicates a Hybrid Feature), which includes possible combinations of cue-words based and in-text based features. These features are further subdivided into the following features:

*1) Cue-words*: In this feature we used Cue-words from the sentences in which citations occur. We picked the sentence where citations appear and one sentence before citations appear and one sentence after citations appear. We collected all citations in the respective paper and counted the words.

*2) Cue-words count:* In this feature, we used Cue-words count and checked the occurrence of each Cue-word in the whole paper.

*3) Section cue-words:* In this feature, we used the Cue-words count and checked the occurrence of single Cue-words in each section to identify the importance of that word.

*4) In-text citations count:* In this feature, In-text Citations were cited by the author to support background knowledge or to extend the previous work within whole paper.

*5) Section in-text count:* In this feature, we used In-text counts and checked the occurrence on single in-text citations in each section to identify the importance of that citation.

*6) Cue-words and in-text count H1:* In this feature is the combination of total occurrence of cue-words and in-text citations counted for a respective paper.

*7) Cue-words and section in-text count H2:* In this feature is the combination of total occurrence of cue-words and Section in-text citations counted for a respective paper to classify citations.

*8) Cue-words count and in-text count H3:* In this feature is the combination of total occurrence of cue-words count and in-text citations count for a respective paper to classify citations.

*9) Cue-words count and section in-text count H4:* In this feature is the combination of total occurrence of cue-words count and Section in-text citations count for a respective paper to classify citations.

*10)Section cue-words and in-text count H5:* In this feature is the combination of total occurrence of section cue-words count, and in-text citations count for a respective paper to classify citations.

*11)Section cue-words and section in-text count H6:* In this feature is the combination of total occurrence of Cue-words, Cue-words count and section cue-words count for a respective paper to classify citations.

*12)Cue-words and cue-words count and section cue-words H7:* In this feature cue-words based and in-text citations-based features were used to create all possible combinations.

### C. Model Selection

In the literature survey, different authors used different classifiers to classify citations. Each classifier performed according to their functionality and classification system. Our focus was to evaluate which classifier performed better using our proposed approach. The classifiers are as follows:

- SVM.
- RF.
- Naive Bayes (NB).
- KNN.

### D. Model Training and Testing

The evaluation criteria used for this research thesis is precision, recall and f-measure. These are widely used to evaluate results for classifying citations into important and incidental classes. For training and testing of citation classification, in this thesis, we are classifying citation using Python; we used four different classifiers with the Scikit learn library for citation classification, and we used the Seaborn Library for graph plots. We categorized datasets based on training models to classify citations into two categories. The dataset of citations, which categorize whether it belongs to the important or incidental class. The returns values of precision recall and f-measure. We evaluated the performance of each classifier on cue-words based, in-text citations based, and Hybrid features. A detailed evaluation of each feature using different classifiers will be discussed in the evaluation step.

## IV. Experimental Results and Discussion

### A. Experimental Setup

The data set collected for this research work was the most authentic and publicly updated available from [7]. It was published for the AAAI workshop. In the dataset, the author classified citations into two broad categories: important and non-important. The dataset contained 465 citations to be classified in the above-mentioned categories. In the results, 14% citations obtained were important and 86% citations were non-important. We downloaded 416 papers and 49 papers manually from the ACL anthology that were missed in the ACL. The new dataset of 416 is maintained for further processing**.**

### B. Performance

In the previous chapters, the literature review discussed the importance of citation classification. Different authors describe different ways of classifying citations. Different authors do not agree upon treating all citations equally and describe the importance of citations in sections, citations in introductions, and related work that belongs to incidental class. Citations that appear in methodology and results are marked as important class. We have collected a dataset of 416 papers by downloading it from the ACL anthology. Approximately 21,500 words were extracted from the dataset to classify citations. 14000 words were marked as incidental and 7500 words were marked as important. Then preprocessing is applied on the collected words, Onix Text Retrieval Toolkit stop word List apply for stop word removal from collected dataset and suffix-stripping algorithm used for stem words in their root form. We removed duplicate words from the data which were generated after the stop word and stemming were applied. A total of 858 unique words were found. Two different files were created. One file contains 631 unique words marked as incidental class and another file contains 227 unique words marked as important class. Then the dataset is ready for applying experiments on it. We classified citation into two broad categories: important and incidental, with the help of cue-words based and in-text based features. We used Python to classify citations. In Python, each feature is selected manually, and different sets of classifiers apply to each feature and generate results against experiments. Classifying citations into important and incidental experiments was performed on Python with the following classifiers: “RF,” “Naive Bayes,” “KNN,” and ”SVM” applied to all possible combinations of features.

### C. Features

*1) Cue-words based features:* In this feature, we concentrated on evaluating the total occurrence of cue-words only to classify citations. In this feature, we will use only the distinct cue-words attained from citations context and associate them with both important and incidental classes. In the cue-words count feature we concentrated on evaluating the total occurrence of cue-words in the paper. In the section cue-words feature, we evaluate the total occurrences of cue-words along with the section in which these cue-words appeared. In this feature, we will use both cue-words and their occurrences in each section.

*2) In-text citation-based features:* In this feature, we concentrated on evaluating the total occurrence of in-text citation counts alone to classify citations. In this feature, we will use only the count of a specific in-text citation. In the section cue-word, we concentrated on evaluating the total occurrence of section citation counts to classify citations. Here we will use the count of citations presented in logical sections of research papers.

*3) Hybrid features:* In this section, we are interested in evaluating the strength of cue-words and in-text citations, along with the sections and total number of occurrences, to classify citations. In this feature, we will use all possible combinations of cue-words based features and in-text citations-based features together to classify citations.

### D. Classifier and Results

In this thesis we are interested in classifying citations; the classifiers used in this work are commonly used in the research community. The following are the classifiers we used for this research work:

*1) SVM:* Support vector machines (SVM) are a particularly influential and flexible class of supervised algorithms for both regression and classification. The SVM classifier is mostly used in the literature for classifications. The SVM classification report was generated against hybrid features of cue-words based and in-text citations-based features as shown in Fig. 2. A total of 858 citations were

classified into two broad categories: important and incidental. The 629 citations that belong to the incidental class achieved precision 0.96, recall 0.94 and f1-score 0.97. The remaining 228 citations belonging to the important class achieved precision 0.97, recall 0.99 and f1-score 0.95. This is the best result achieved with the help of the SVM classifier using hybrid features as shown in Fig. 3.

This line chart evaluates the result of using hybrid features as shown in Fig. 4. We used the SVM classifier with evaluation criteria precision, recall and f-measure. In the Predicted Label figure, the blue line indicates incidental class and achieved a result against SVM on precision 0.96, recall 0.94, and f-measure 0.97. In the Predicted Label figure, the orange line indicates the important class result and achieved precision 0.97, recall 0.99, and f-measure 0.95 against SVM.

*2) Random forest:* Random Forest (RF) is one of the most popular and widely used machine learning algorithms. This classification report was generated against hybrid features of cue-words based and in-text citations-based features. The total 858 citations were classified into two broad categories: important and incidental. The 629 citations that belong to the incidental class achieved precision 0.93, recall 0.99, and f1-score 0.96. The remaining 228 citations belong to the important class and achieved precision 0.96, recall 0.90, and f1-score 0.92. This is the best result achieved with the help of the RF classifier using hybrid features as shown in Fig. 5 and 6.

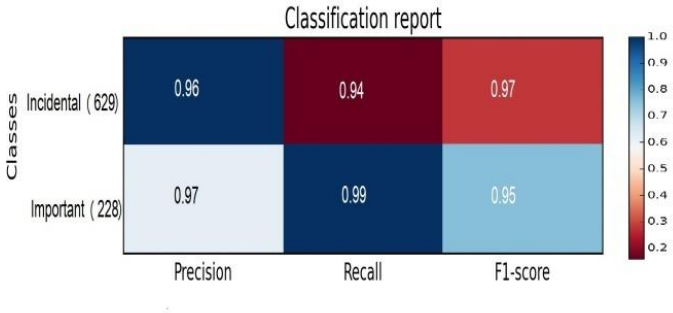



Fig. 2. Classification Report.

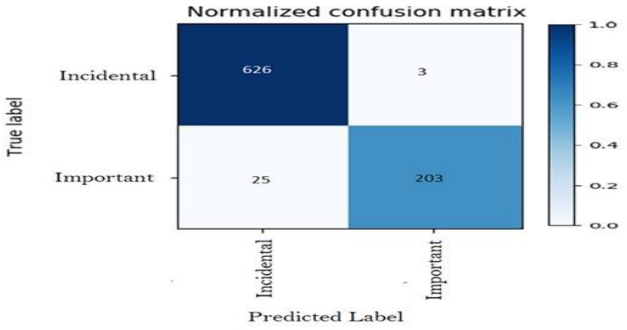



Fig. 3. Confusion Matrix.

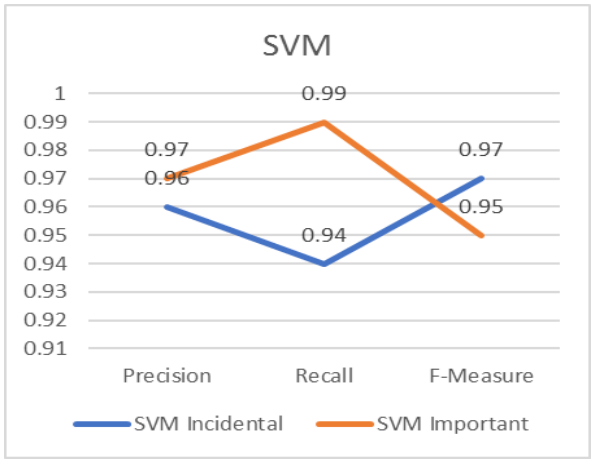



Fig. 4. Line Chart.

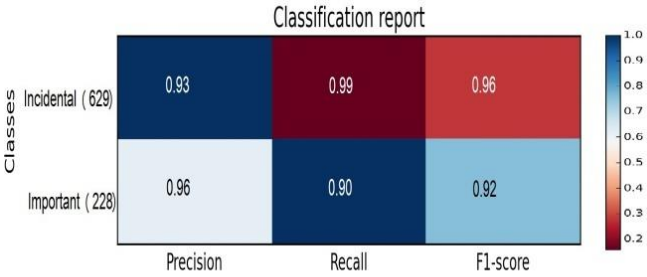



Fig. 5. Classification Report.

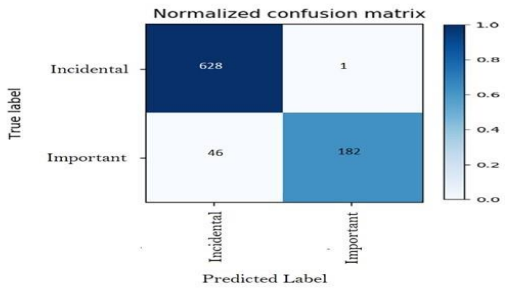



Fig. 6. Confusion Matrix.

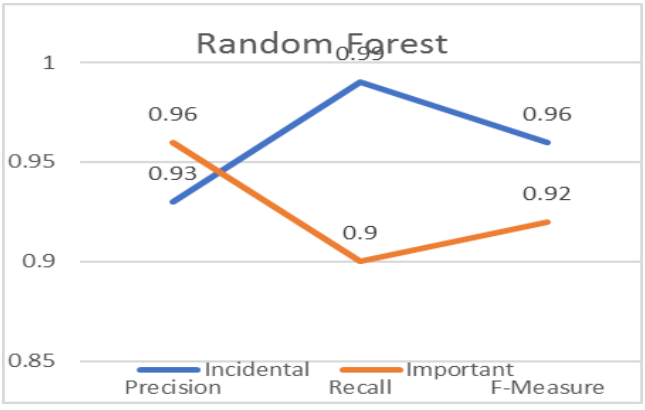



Fig. 7. Line Chart.

In this chart we evaluated the result against hybrid features using the RM classifier with evaluation criteria precision, recall and f-measure as shown in Fig. 7. In the figure, the blue line indicates incidental class results achieved against RM on precision 0.93, recall 0.99, and f-measure 0.96. In the figure, the orange line indicates important class results achieved against RM on precision 0.96, recall 0.90, and f-measure 0.92.

*3) Naive bayes:* Naive Bayes models are a set of supervised learning algorithms, Naive Bayes learners and classifiers can be exceptionally quick as compared to other sophisticated models. The Naive Bayes classifier was generated against hybrid features of cue-words based and in-text citations-based features as shown in Fig. 8. The total 858 citations were classified into two broad categories: important and incidental. The 629 citations that belonged to the incidental class achieved precision 0.93, recall 0.94, and f1-score 0.94. The remaining 228 citations belong to the important class and achieved precision 0.90, recall 0.88, and f1-score 0.89. This is the best result achieved with the help of the NB classifier using hybrid features.

In this chart, we evaluated the result against hybrid features using the NB classifier with evaluation criteria precision, recall and f-measure as shown in Fig. 9. In the figure, the blue line indicates incidental class result achieved against NB on precision 0.93, recall 0.94, and f-measure 0.94. The orange line indicates important class results achieved against NB at precision 0.90, recall 0.88, and f-measure 0.89.

*4) KNN:* K Nearest Neighbor (KNN) is a very simple, straightforward, adaptable and one of the topmost machine learning algorithms. The KNN classifier was generated against hybrid features of cue-words based and in-text citations-based features. The total 858 citations were classified into two broad categories: important and incidental as shown in Fig. 10 and 11. The 629 citations that belong to the incidental class achieved precision 0.93, recall 0.94, and f1-score 0.93. The remaining 228 citations belong to the important class and achieved precision 0.88, recall 0.86, and f1-score 0.87. This is the best result achieved with the help of the KNN classifier using hybrid features.

In this chart we evaluated the result against hybrid features using the KNN classifier with evaluation criteria precision, recall and f-measure as shown in Fig. 12. In the figure, the blue line which indicates an incidental class result, achieved precision 0.93, recall 0.94, and f-measure 0.93 against KNN. In the figure, the orange line indicates the important class result achieved precision 0.88, recall 0.86, and f-measure 0.87 against KNN.

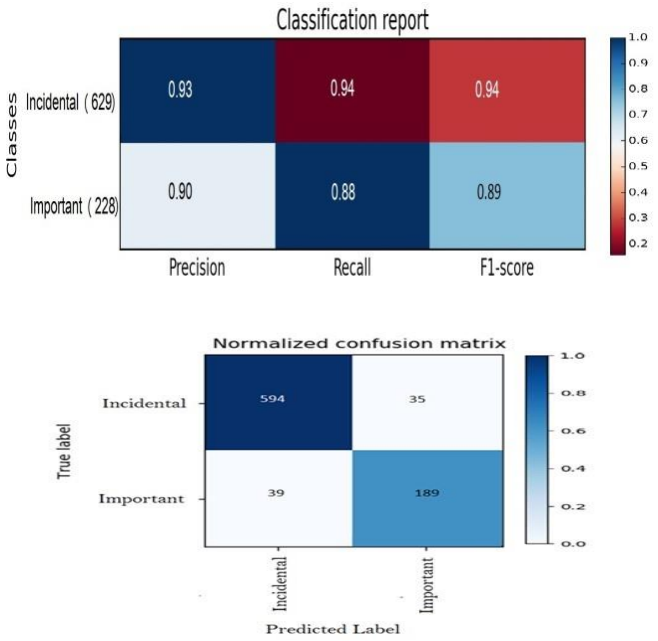



Fig. 8. Confusion Matrix.

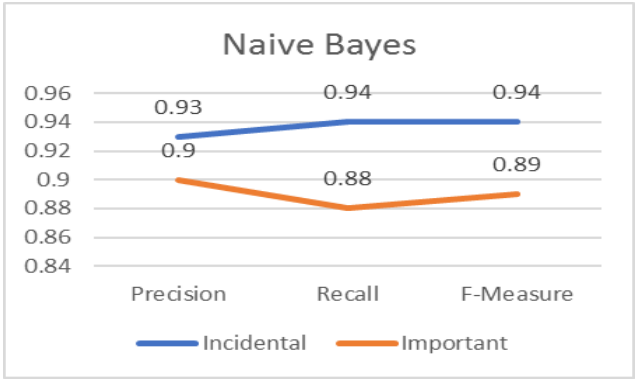



Fig. 9. Line Chart.

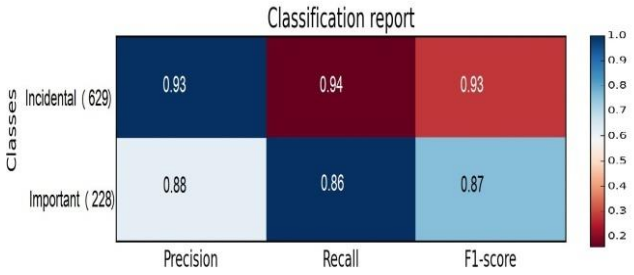



Fig. 10. Classification Report.

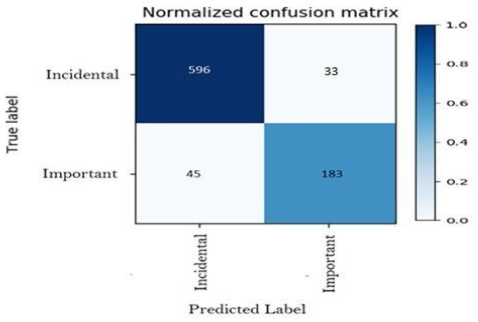



Fig. 11. Confusion Matrix.

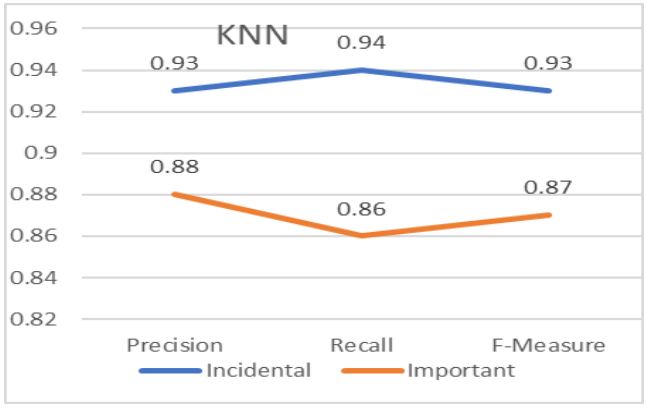
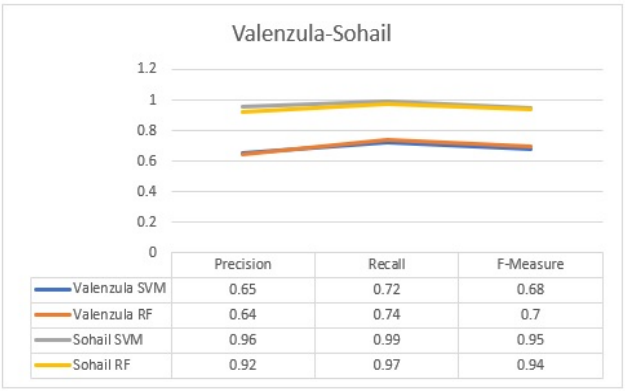



Fig. 12. Line Chart.



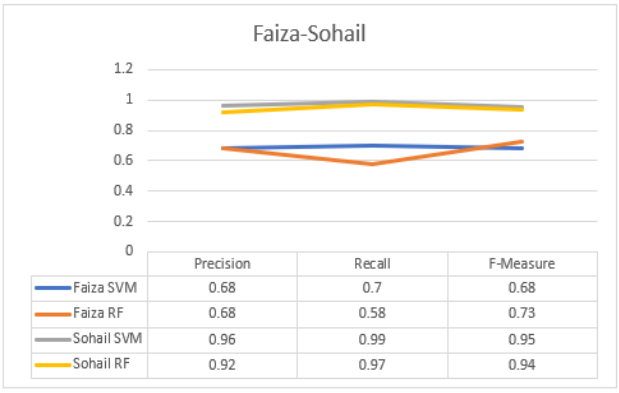




Fig. 13. Comperison with Existing Approaches.

## V. Comparison

In this section we compared of our approach generated results with the results of existing approaches by the research community, with same domains that use the same classifiers such as SVM and RF.

*1) Valenzuela's:* dataset is used in most of the articles. Valenzuela claimed that she is the first to in the research community to identify two broad categories for citation: important vs. incidental. Most of the approaches used different classifiers with Random Forest and Support Vector Machine being the most common classifier used in all the approaches as shown in Fig. 13.

*2) Faiza:* proposed classifying binary citation technique. This scheme is based on metadata and content-based parameters to classify citation. Faiza was first to classify citation as important vs. non-important using a permanent metadata-based hybrid. Features used to classify important vs. non-important citation included two different types: of parameter: metadata based and content-based. She used these classifiers to classify citations with precision achieved on SVM 0.68, KLR 0.62, and RF 0.72, Recall on SVM 0.7, KLR 0.59 and RF 0.69 and f-measure on SVM 0.68, KLR 0.58 and RF 0.73.

In this section we compare our results of classifying citations with Valenzuela et al. and Faiza et al.'s existing approaches in the research community with the same domain as shown in Fig. 13. We used four classifiers to classify citations: SVM (Support Vector Machine) and Random Forest (RF) are most commonly used classifiers and compare with other approaches. The average SVM accuracy achieved precision 0.96, recall 0.99, and f-measure 0.95, and RF accuracy achieved Precision 0.92, recall 0.97, and f-measure 0.94 as the best result.

## VI. Conclusion and Future Work

Citations are essential in the scientific community to assess the qualifications of the scientific authors. It is imperative in many situations because we must take direction from them in their subject. We proposed methodology to comprehensively compute all possible combination of cue-words based and in-text based features for classifying citations. We used the publicly available Valenzuela dataset as a benchmark and we used Scholarly Big data for experiments published in the AAAI workshop. We performed preprocessing on the collected dataset, applying the stemming and stop words removal algorithm. After preprocessing we removed duplicate words so that the remaining list of unique words totaled 858. Among these, 629 are incidental and 229 are important citations. When

the collected dataset was ready for experiment and evaluation, it was performed using Python. Our features are categorized into three main categories: (1) Cue-words based features; (2) In-text based features; and (3) Hybrid where all possible combinations of cue-words based and in-text based features were examined. These features are further subdivided. We evaluated all possible combinations of features and used five different classifiers: SVM, RF, Naive Bayes, and KNN. Each classifier performance was different for each feature. In the hybrid approach, all feature performances were better in SVM, and RF performed better than others. At the end, we made a grand comparison of results for our approach to classifying citation with the existing approaches by the research community in same domain. The average SVM accuracy achieved was precision 0.93, recall 0.99, and f-measure 0.96; for RF, accuracy is precision 0.97, recall 0.99, and f-measure 0.92, Naive Bayes accuracy achieved precision 0.90, recall 0.94, and f-measure 0.92, and average KNN accuracy achieved precision 0.93, recall 0.90, and f-measure 0.90, which are the best results as compared with other approaches.

## VII. Future Work

In this research work our focus was identifying important and incidental citation and comparing techniques. This is the publicly available dataset for this domain. To ensure the accuracy of the technique, studies must be performed on other domain as well with large datasets to ensure the validity of the technique.

## References

[1] M. V. Simkin and V. P. Roychowdhury, “Read before you cite!”arXivpreprint cond-mat/0212043, 2002.

[2] D. E. Chubin and S. D. Moitra, “Content analysis of references: Adjunct or alternative to citation counting?” Social studies of science, vol. 5, no. 4, pp. 423–441, 1975.

[3] M. J. Moravcsik and P. Murugesan, “Some results on the function and quality of citations,” Social studies of science, vol. 5, no. 1, pp. 86–92, 1975.

[4] D. Dubin, “The most influential paper gerardsalton never wrote,” 2004.

[5] X. Zhu, P. Turney, D. Lemire, and A. Vellino, “Measuring academic influence: Not all citations are equal,” Journal of the Association for Information Science and Technology, vol. 66, no. 2, pp. 408–427, 2015.

[6] B. H. Butt, M. Rafi, A. Jamal, R. S. U. Rehman, S. M. Z. Alam, and M. B. Alam, “Classification of research citations (crc),” arXiv preprint arXiv:1506.08966, 2015.

[7] M. Valenzuela, V. Ha, and O. Etzioni, “Identifying meaningful citations.” in AAAI Workshop: Scholarly Big Data, 2015.

[8] 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, 2011.47Bibliography

[9] E. Garfield and R. K. Merton, Citation indexing: Its theory and application in science, technology, and humanities. Wiley New York, 1979, vol. 8.

[10] M. Garzone and R. E. Mercer, “Towards an automated citation classifier,” in Conference of the Canadian Society for Computational Studies of Intelligence. Springer, 2000, pp. 337–346.

[11] B. Finney, “The reference characteristics of scientific texts,” Ph.D. dissertation, City University (London, England), 1979.

[12] C. L. Giles, K. D. Bollacker, and S. Lawrence, “Citeseer: An automatic citation indexing system,” in Proceedings of the third ACM conference on Digital libraries. ACM, 1998, pp. 89–98.

[13] L. Steve, G. Lee, and B. Kurt, “Digital libraries and autonomous citation indexing,” IEEE Computer, vol. 32, no. 6, pp. 67–71, 1999.

[14] S. Peroni and D. Shotton, “Fabio and cito: ontologies for describing bibliographic resources and citations,” Web Semantics: Science, Services and Agents on the World Wide Web, vol. 17, pp. 33–43, 2012.

[15] D. Shotton, “Cito, the citation typing ontology,” in Journal of biomedical semantics, vol. 1, no. 1. BioMed Central, 2010, p. S6.

[16] A. SHAHID, “Recommending relevant papers using in-text citation frequencies and patterns,” 2015.

[17] S. Teufel, A. Siddharthan, and D. Tidhar, “Automatic classification of citation function,” in Proceedings of the 2006 conference on empirical methods in natural language processing. Association for Computational Linguistics, 2006, pp. 103–110.

[18] S. B. Pham and A. Hoffmann, “A new approach for scientific citation classification using cue phrases,” in Australasian Joint Conference on Artificial Intelligence. Springer, 2003, pp. 759–771.

[19] A. Abu-Jbara, J. Ezra, and D. Radev, “Purpose and polarity of citation: Towards nlp-based bibliometrics,”in Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies, 2013, pp. 596–606.

[20] S.-U. Hassan, A. Akram, and P. Haddawy, “Identifying important citations using contextual information from full text,” in Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries. IEEE Press, 2017, pp. 41–48.

[21] S.-U. Hassan, I. Safder, A. Akram, and F. Kamiran, “A novel machinelearning approach to measuring scientific knowledge flows using citation context analysis,” Scientometrics, vol. 116, no. 2, pp. 973–996, 2018.

[22] M. Jia, “Citation function and polarity classification in biomedical papers,”2018.

[23] S. Agarwal, L. Choubey, and H. Yu, “Automatically classifying the role of citations in biomedical articles,” in AMIA Annual Symposium Proceedings, vol. 2010. American Medical Informatics Association, 2010, p. 11.

[24] F. Qayyum and M. T. Afzal, “Identification of important citations by exploiting research articles’ metadata and cue-terms from content,” Scientometrics, pp. 1–23, 2018.

[25] L. B. Sollaci and M. G. Pereira, “The introduction, methods, results, and discussion (imrad) structure: a fifty-year survey,” Journal of the medical library association, vol. 92, no. 3, p. 364, 2004.