CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Identification of Important Citations Using Section-wise Similarities

by

Abdul Rauf

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Computing
Department of Computer Science

2020

This thesis work is devoted to my beloved Teachers, Family and Friends. I have a special feeling of gratitude for my beloved parents, brothers, sisters and wife. I would like to thank my supervisor for his firm belief and confidence that enabled me to reach this milestone.

CERTIFICATE OF APPROVAL

Identifications of Important Citations Using Section-wise Similarities

by

Abdul Rauf

(MCS173015)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Ayyaz Hussain	QAU Islamabad
(b)	Internal Examiner	Dr. M. Shahid Iqbal Malik	CUST Islamabad
(c)	Supervisor	Dr. M. Tanvir Afzal	CUST Islamabad

Dr. Muhammad Tanvir Afzal
Thesis Supervisor
May, 2020

Dr. Nayyer Masood
Head
Dept. of Computer Science
May, 2020

Dr. Muhammad Abdul Qadir
Dean
Faculty of Computing
May, 2020

Author's Declaration

I, **Abdul Rauf** hereby state that my MS thesis titled "**Identification of Important Citations Using Section-wise Similarities**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Abdul Rauf)

Registration No: MCS173015

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Identification of Important Citations Using Section-wise Similarities**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been dully acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Abdul Rauf)

Registration No: MCS173015

Acknowledgements

I would like to express my gratitude and thanks to **Allah Almighty** for providing me abilities to accomplish this research work. Secondly, I would like to express my sincerest thanks to my esteemed supervisor **Dr. Muhammad Tanvir Afzal** for his guidance, support, inspiration and encouragement. He has taught me, both consciously and unconsciously, how good experimental work is carried out. I have learned not only research under his supervision, but also his attitude and behavior towards others. Sir, you will always be remembered as an inspirational teacher. May Allah Pak shower his countless blessings upon all of you, **Jazakallah.**

(Abdul Rauf)

Registration No: MCS173015

Abstract

In the scientific literature, citations count has been employed by different researchers to rank institutions, peer judgments and calculating impact factor of journals over the years. However, researchers have argued that all citations are not of equal significance and weights. They concluded that the only count of the citation is not enough and a proper reason behind the citations must be considered. The researchers have pointed out various reasons of citing a specific work which includes: (1) to present background knowledge, (2) to extend other researchers work, etc. Lately, different researchers have concentrated only on those reasons for different classifications of citations that help to achieve the reliability of citation count approach and divided citation reasons into only two broad types: (1) Important and (2) Non-Important citations. Important citations are those in which authors extend or adapt the existing work while Non-Important citations are just used to provide the background knowledge. We have critically reviewed more than 40 research articles and explored the research gap in the field. Different researchers have discussed different approaches that depend on the content and metadata of research articles. Their proposed metadata-based and contents features include title similarity, in-text citation count, cue phrases, etc. These features have their own limitations like in cue-phrases case, for every new dataset, there is a need to update the list of cue-phrases which is a time-consuming factor. In such cases, there should be some alternative modes available for citations classification. We have proposed a comprehensive approach to address the above-raised issue by classifying citations into (1) Important and (2) Non-Important citations. The proposed approach utilizes content of corresponding logical sections of citing and cited research articles such as Abstract, Introduction, Methodology, etc. The cosine similarity has been used to calculate the similarity scores of corresponding logical sections and different sections have been combined with average and weighted average formulas. The experiments have been performed on the comprehensive annotated dataset of Valenzuela. After comparing our results with metadata-based and content-based approaches, the proposed approach outperformed all state-of-the-art approaches by achieving the F-Measure score of 0.75.

Contents

List of Figures

List of Tables

Abbreviations

ACM	Association for Computing Machinery
Avg.	Average
Avg.(A, M)	Average (Abstract, Methodology)
Avg.(A, R)	Average (Abstract, Result)
Avg.(A, M, R)	Average (Abstract, Methodology, Result)
CUST	Capital University of Science and Technology
JUCS	Journal of Universal Computer Science
TF-IDF	Term-Frequency Inverse-Documents-Frequency
W.Avg.	Weighted Average
W.Avg.(A, M)	Weighted Average (Abstract, Methodology)
W.Avg.(A, R)	Weighted Average (Abstract, Result)
W.Avg.(A, M, R)	Weighted Average (Abstract, Methodology, Result)

Chapter 1

Introduction

This document demonstrates and reviews state-of-the-art approaches for important citation identifications for cited research articles of citing articles. Furthermore, it shows the proposed technique and its evaluations through different experiments. Section 1.1 describes the importance and background of the work associated to improve the identification of important citations. Section 1.2 gives the research objectives and Section 1.3 provides the scope of this research.

Researchers organize and carry out their research work by depending upon the very well-known work of their famous and expert predecessors in the domain. The above-mentioned sentence is also proved by Ziman [2], pointing out that a scientific paper does not stand alone; it is embedded in the literature of the subject. Narin [3] described that reference is the acknowledgment of a research article that it gives credit to another article and citation is the acknowledgment of research article which it receives from other research articles. Ziman [2] described the importance of examining citations of different research studies. According to the author, the high frequency of citation counts indicates the popularity and importance of that specific work. Different researchers used citation count to calculate and measure Institutions' and researchers' ranking [4], researchers Nobel prizes and awards [5], peer judgments[6], Research funds allocation and calculating impact factor of researchers and different Journals. According to the report of Wildson [7], citations can be used to measure and explain the quality of the research. Another latest report of Benedictus [8] inspected the role of citations that is used to check the quality of any individual research work.

Keeping in view the above discussion, it is significant to look for the answer that why researchers and authors refer or cite a specific work in their research articles? Garfield [9] explained 15 reasons of citing a specific work which includes: (1) Credit given on the basis of literature review, (2) Present background knowledge, (3) Criticize other researchers work, (4) Methodology identification, (5) correction of own work and (6) correction of other researchers work and so on. Different authors explained many reasons to refer to a particular research article after the publication of Garfield [9] report. Then the exploration of these citation reasons helped authors to critically examine citation count (i.e. quantitative citation approach).

Ziman [2] described that usage of citation count is not suitable and argued that many citation counts take place when the author criticizes someone else works in the research article. These criticized citation counts should not be given any importance and significance. Moravcsik [10] described that half of the citations (i.e. 40%) occurred in the citing research articles are those which discuss and provide the background knowledge. Garfield [11] investigated the citation count-based approaches critically and concluded that if there are high citations count then quality work becomes low which has received more criticism. Teufel [12] discussed that all citations in research articles are not of equal significance and weights. Benedictus [8] explored that when citation count is taken into account to check the quality and excellence of any researcher, then quantity is superior to the quality.

Different researchers found many citation reasons but the question arises now that automation of citation classification can be done or not? Brooks [13] described that manually citation classifications takes place on the basis of interviewing of author before or after the publication of research articles, and sometimes call the author that why he used someone else work; or interviewing of author at the time of writing research articles that why he is citing and using the work [14]. In 1979, Finney [15] was the first person who presented an idea in her master thesis that citation classification can be done automatically and she divided citations into 7 types. A different number of researchers follow her idea for automatic citation classification.

In 2000, Garzone [16] used the citation classification automation idea of Finney [15] and proposed a first fully automated citation classification system. Their system used 10 citation categories types which were further divided into 35 categories. They constructed a grammar for citation classification which contains 195 lexical matching rules and 14 parsing rules depending on section location and cue words. The research articles and citation sets are given to the proposed system as an input for generating a suitable category. However, Garzone [16] proposed system was upbraided by Radoulov [17] because it contained more number of categories types which can cause divergence with each other.

Lately, different researchers have concentrated only on those reasons for different classifications of citations that help to achieve the reliability of the citation count approach. For this objective, Valenzuela [1] presented the first approach for the identification of (1) Important and (2) Incidental citations. According to Valenzuela's definition of important citations, authors extend or adapt the proposed technique of cited article. While in incidental citations, authors just write the background information or some theory portion of the proposed technique of the cited article. They used 12 features for classifications which are: (1) Total number of direct citations, (2) Number of direct citations per section, (3) Total number of indirect citations and number of indirect citations per section (4) Author Overlap, (5) Similarity between abstracts and so on. Which depends on the content of research articles. They used 465 research articles having 48 tuples of root paper and cited paper from ACL anthology.

Another approach presented by Qayyum & Afzal [18] for the identification of important citations. They used a binary scheme which is composed of the metadata and the contents of research articles. The metadata consists of the paper title, author name, keywords, and references that are openly available on the internet. They picked abstract and cue terms from the content of articles. They used 05 features for classification which include title similarity, author overlap, references, abstract and cue terms. They used two benchmark datasets D1 and D2. D1 is the Valenzuela dataset from ACL anthology and D2 is CUST University Islamabad dataset consists of 488 paper and citation pairs. We will use the same dataset as proposed by Valenzuela [1] and was further used by a recent approach of Qayyum & Afzal [18]. This will help us to make a comparison with state of the art approaches on the same dataset.

1.1 Background of Problem

We have seen from the above literature that there are two major techniques of Qayyum & Afzal [18] and Valenzuela [1] used for the extraction of important citations. Qayyum & Afzal [18] criticize the Valenzuela [1] technique and comment that it depends mostly on the content of research articles which are not openly accessible by major journals like ACM, Elsevier, IEEE, Springer, etc. and they also used 12 different features to achieve the accuracy of 0.65. However, Qayyum & Afzal [18] used metadata-based features with the contents feature (i.e. abstract, references and cue words) and evaluated their approach by using unigram, bigram, and trigram. Their best combination achieved an accuracy of 0.72 by combining all the features.

Both of these approaches have given importance to metadata and in-text citation counts and their positions, however, research papers are written with the help of domain-specific terms and knowledge, the research gap we have identified is that the no one has tried to compare the important terms represented in different corresponding logical sections of the research papers. This is also supported by another factor of Valenzuela [1] important citations definition. They described that cited papers have important citations with citing papers if cited papers extend or adapt the presented idea of the citing papers. It is more probable that both papers might use similar vocabulary and terms as they belong or they are closely working on the same topic or extending once work in another work. So, there are more chances that both papers belong to a similar domain and citing paper uses domain-specific terms and knowledge in the Abstract, Introduction, Methodology, Literature review and Result section. We argue that no researcher utilized domain-specific terms of the corresponding logical sections for the identification of important citations.

1.2 Research Objectives

Based on the critical analysis presented in the previous section, this thesis evaluates the role of section-wise content similarity for the identification of important citations. The main objectives of this work are to enquire, search, identify and evaluate the role of sections to identify the important citations. In the end, we will be able to conclude that cited paper has an important citation with citing paper if the contents and vocabulary terms used in those papers corresponding logical sections are similar. The Important and Non-Important citations are describing as below:

Important: The citations in which authors extend or adapt the proposed technique of cited article.

Non-Important: The citations in which authors just write the background information or some theory portion of the proposed technique of the cited article.

1.3 Research Question

Based on the critical analysis of state-of-the-art approaches presented in the background of the problem section, this thesis identifies the important citations by considering the corresponding logical sections of research articles. The following research question has been devised in this research document.

RQ:

What is the role of section-based comparison of the content to classify citations as (1) Important citations and (2) Non-Important citations?

1.4 Scope

This thesis scope is to exploit the paper-citations pairs to quantify and classify the citations into just two classes as Important and Non-Important (Incidental) citations. The scope of this study is also limited to the available dataset provided by Valenzuela [1]. This annotated dataset has been used for experimentations that contain 465 tuples of root paper and cited paper from ACL anthology.

1.5 Applications of the Proposed Solution

This research thesis can help scientific society in the different fields which are as following:

- Educational Institutions Ranking
- Researchers and Authors Ranking
- Journals Ranking
- Countries Ranking
- Research funds allocation
- Researchers Nobel prizes and awards allocation
- Impact factor calculation of researchers
- Impact factor calculation of Journals
- Peer judgments

Chapter 2

Literature Review

According to a recent survey by Beel [19] more than 120 innovative approaches in 216 research publications were published for research paper recommender systems but even after 16 years of research, we were unable to tell that which approach is best for recommending academic literature. The main limitation in all approaches is the evaluation criteria. The authors emphasized the importance to know the best approach and to find its strengths and weakness so that it will be helpful for researchers to keep track of their research areas. The authors concluded in this survey that it is a more difficult task to investigate the best approach because all approaches do not follow all important aspects in order to recommend research papers.

Nascimento [20] proposed a research paper recommender system approach that used content-based algorithms to recommend relevant papers. This approach picks a research article as an input and generates several queries and keywords on the basis of words (contents) on that paper. Then these keywords and queries inserted in different digital libraries that contain sources of research articles that generate a set of candidate research articles. Content-based recommending algorithms applied to a set of candidate research articles to rank these papers. This technique used metadata of abstract and title to recommend the most relevant research articles which are publicly available.

Beel [21] discussed the Docear research recommender system that helped researchers for creating, organizing and searching research articles. This system used a unique feature of mind mapping for creating profiles of the researchers, where a tree data structure is used. Docear system continuously sensing the researcher activities like that for which topic a user is searching and reading, for which topic a researcher has already written a research article and so on. A user model was built from the users mind mapping and compared with Docears digital library to get relevant recommendations.

Ferrara [22] presented a content-based approach for research paper recommendations. They used the key phrase extraction module for creating the profiles of documents and users. This approach used KPEM (Key Phrase Extraction Module) for creation profiles. The user profile was created on the basis of tagging which allocated him earlier. So, the recommendation of research articles or papers took place on the basis of the profiles of users and documents. For checking similarity measures, uni-grams, bi-grams, and tri-grams used. For experimentation and evaluation of the proposed approach, the ACL Anthology Reference Corpus dataset was used which is publicly available. The dataset contains 597 papers of 28 researchers. The presented results showed that tri-grams give better results than uni-grams and bi-grams.

Pruitikanee [23] proposed an approach for research article recommendations depends on fuzzy clustering. This approach consists of four steps. In the first step, the system gets user queries and returns all research articles/papers from the database which contains at least one keyword from the query. Research articles grouping takes place on the basis of the same interests and topics by using fuzzy clustering in the second step. In the third step, representative research articles are computed to minimize user interaction from the large database. In the last step, research article representatives are ranked by using classical ranking models i.e. Page Rank.

Normalized Similarity Index (NSI) approach was presented by Nassiri [24] to calculate the similarity between two research articles. The root and cited research articles' similarity was measured by using a bibliography, co-citation, and longitudinal coupling. Bibliography coupling means that two research articles are relevant or similar if they share some references. Co-citation is used to find the relevancy among research articles. Suppose if Z cited both X and Y article then it means they may be relevant to each other. While longitudinal coupling means that research articles indirectly relevant to each other. NSI validation was calculated for five cited networks. NSI results were compared with the results of combined linkage (CL) and weighted direct citation (WDC) on the same data where NSI outperformed both CL and WDC.

Smith [6] discussed that citations play an important role to gauge different factors like the impact factor of different journals and authors, peer judgments, research grants, institution rankings, etc. Garfield [11] discussed that citations show a relationship between a part or whole of a cited research article and a part or whole of the cited research article. There are fifteen different reasons in which the researcher cites the work of other research articles. These approaches include criticizing previous work, Identifying methodology, providing background reading, etc.

Ziman [2] described that usage of citation count is not suitable and argued that many citations counts take place when the author criticized someone else works in the research article. These criticized citation counts should not be given any importance and significance. Moravcsik [10] described that half of the citations (i.e. 40%) occurred in the citing research articles are those which discuss and provide the background knowledge. Garfield [11] investigated the citation count-based approaches and concluded that if there are high citations count then quality work becomes low. Teufel [12] discussed that all citations in research articles are not of equal significance and weights.

The question arises now that automation of citation classification can be done or not? Garfield [9] citations reasons gave courage to researchers and authors to extract and calculate different aspects of referring a particular article, but there was not any automation of citation classification available. Brooks [13] described that manually citation classifications takes place on the basis of interviewing of author before or after the publication of research articles, and sometimes call the author that why he used someone else work [13], or interviewing of author at the time of writing research articles that why he is citing and using the work [14].

In 1979, Finney [15] presented an idea in her master thesis that citation classification can be done automatically and she divided citations into 7 types. These types or categories include (1) Background knowledge, (2) Tentative references, (3) Methodological references, (4) Conformational references, (5) Negational references, (6) Interpretational references and (7) Future research references. She developed a system where she combined citation function with cue words and citation location in a classification algorithm.

Bonzi [25] in 1982 found the parameters which help to calculate the closeness between citing and cited research articles. He found 13 parameters which includes: source of citation, date of citation, author self-citation, journal self-citation, types of journal, date of publication, sex of author, type of article, length of article, no. of citations, no. of citation in footnote, multiple mention of citations and placement of citations in text. They picked 31 research articles of library and information science published in different journals and contains 500 citations. Their results showed that citing and cited research articles relevancy depends upon the source of cited work, source of citing work, type of citing research article and how much time a work is cited in the text.

In 2000, Garzone [16] used the citation classification automation idea of Finney [15] and proposed a fully automated citation classification system. The research articles and citation sets are given to the system as an input for generating a suitable category. Their system used 10 citation categories types which include: Negational, Affirmational, Assumptive, Tentative, Methodological, Interpretational/Developmental, Future Research, Use of Conceptual Material Type, Contrastive Type Categories and Reader Alert Type Category. These 10 citation categories are further divided into 35 categories. They construct a grammar for citation classification which contains 195 lexical matching rules and 14 parsing rules depending on section location and cue words. They classified the results in three types which are completely right, partially right and completely wrong. The systems generate a better result on seen research articles and average results on unseen research articles. However, Garzone [16] proposed system was reproached by Radoulov [17] because it contains more number of categories types which can cause divergence with each other.

In 2006, Teufel [12] presented a supervised machine learning approach for classification of citations and they used linguistic rules for the purpose of making difference between citation categories. This annotation scheme for citation is the extension and adoption of Spiegel-Rusing [26] scheme. They used four citation categories taxonomies which include neutral, weakness, comparisons and compatibility. The above four categories taxonomies are further divided into 11 categories taxonomies. They used 26 annotated research articles which contain 548 citations. They made 892 linguistic cue phrases for citation classification into a specific category. The system trained on 90 % of the dataset and tested on 10% of the dataset and attained F-Measure of 0.71.

In 2011, Shahid [27] described in-text citation frequency of the cited research article for recommending the relevant papers to the user. When citation frequency is high in the cited research article then cited and citing research articles are more relevant to each other. They tested their approach on the Journal of Universal Computer Science (J. UCS) dataset and evaluated a new citation frequency approach by selecting a threshold value. The author concluded that cited paper is more relevant to citing paper if it has been cited five or more than five times in the citing paper.

In 2013, Livne [28] applied the machine learning technique (i.e. support vector regression machine (SVR)) on a diversified dataset for the identification of predictive features like the reputation of the venue. They used a dataset from Microsoft's academic search for experimentation. The dataset contains 38 million papers of 19 million authors from 15 different academic domains like physics, English, medicine, engineering, etc. Then future citation count was calculated from the extracted dataset. From the huge number of datasets, they calculated their results on the dataset from the year 2000 to the year 2005 on 7 academic domains. The used domains were CS, Physics, Biology, Mathematics, Chemistry, Engineering and Medicine. The presented approach gave better results in some domains and there is a need for consideration in their subdomains in the future.

Shahid [29] described that citations are an important factor to conclude the ranking of universities, impact factor of journals and recommend the relevant research articles to users on the basis of citation counts. Their research methodology consists of three steps. PDF to XML documents conversion takes place with the help of pdfbox in the first step. Then in-text citation frequencies of the dataset are calculated in the second step. And then clustered made on the basis of frequency. From 16000 citations only 3000 citations were extracted due to mathematical ambiguities, string variations, wrong allotment and commonality in content issues. Their results showed that 42% of in-text citations were un-identified and overall achieved accuracy of the system was 58%.

In 2015, Zhu [30] classified citations into two types which are: (1) influential and (2) non-influential. They used five features for citation classifications which include: In-text count-based, similarity-based, Context-based, Position-based and Miscellaneous features. The term Influential means that the author used the methodology, experiment and idea of a cited research article in the citing article. While in the non-influential term, the author just writes the background information of the proposed technique of the cited article in the citing research article. They construct 3143 paper-reference pairs from the 100 research articles for experimentation which were crawled from ACL anthology. These paper-reference pairs were annotated by the authors of the articles. Support vector machines (SVM) were used for the training of their features. Their system gained 0.35 precision on in-text citation count feature and outperformed other features.

In 2015, Valenzuela [1] proposed a supervised machine learning approach for the identification of important and incidental citations. In important citations, authors extend or adapt the proposed technique. While in incidental citations, authors just write the background information of the proposed technique of the cited paper. They used 12 features for classifications which are a Total number of direct citations, Number of direct citations per section, the total number of indirect citations per section, Author overlap, Similarity between abstracts and so on. Random forests and support vector machines were used for the training of their features. They used 465 tuples of root paper and cited a paper from ACL anthology. Their system gained 0.65 F-measure scores and in-text citation count feature outperformed all other 11 features with the precision of 0.37.

In 2019, Qayyum & Afzal [18] presented a supervised machine learning approach for the identification of important citations. They used a binary scheme which is composed of the metadata and the contents of research articles. The metadata consists of the paper title, author name, keywords, and references that are openly available on the internet. They picked abstract and cue terms from the content of articles. They used 05 features for classification which include: title similarity, author overlap, references, abstract and cue terms. These features were trained on Support Vector Machine (SVM), Random Forests (RF) and Kernel Logistic Regression (KLR). They used two benchmark datasets D1 and D2. D1 is the Valenzuela dataset consists of 465 tuples of root paper and cited a paper from ACL anthology and D2 is the CUST University Islamabad dataset consists of 488 paper and citation pairs. Their system gained overall 0.72 accuracy.

2.1 Critical Analysis

After the broad investigation of state-of-the-art approaches, we found that the citation classification techniques mainly rely on In-text citation count, Cue phrases and the content of the research article. The summarize sketch and overview of these techniques are discussed with features, limitations, and results in Table 2.1 below.

Table 2.1: Critical Analysis of different approaches containing Features, Results, and Limitations

Ref	Feature	Results	Limitations
[31]	They used Cue Phrases and cue words as features.	The system gained precision is 0.76.	With the change in the domain, the new extensive list of Cue-words and Cue-phrases needs to be developed.
[16]	They used Cue Phrases and cue words as features.	The system generated a better result on seen research articles and average results on unseen research articles.	With the change in the domain, the new extensive list of Cue-words and Cue-phrases needs to be developed. Construction of 195 lexical matching rules and 14 parsing rules needs expert human-level knowledge. Ten (10) citation categories and their further division into 35 categories can cause confliction with each other.
[12]	Cue Phrases and Cue words.	The system gained F-measure of 0.71.	With the change in the domain,

Table 2.1 – continued from previous page

Ref	Features	Results	Limitations
			the new extensive list of Cue-words and Cue-phrases needs to be developed and there is need to recreate list for every new dataset. Citations are annotated where manually selected words appear and not annotated where manually selected words don't appear e.g. better.
[30]	In-Text citations count, Count based features, Similarity-based features, Context-based features, Position based features.	In-Text citations count precision is 0.35 and it outclassed other features.	It ignores important cue phrases which occur immediately before and after the In-text citations.
[1]	In-Text citations count, Similarity between abstracts.	The f-measure score is 0.65 and in-text citation count feature outperformed all other 11 features	It ignores important cue phrases which occur immediately before and after the In-text citations.

Table 2.1 – continued from previous page

Ref	Features	Results	Limitations
		with the precision of 0.37.	The list of keywords needed to be updated for every new dataset.
[18]	They used 05 features for classification which includes title similarity, author overlap, references, abstract and cue terms.	The system gained 0.68 precision value just by depending on freely available metadata.	They used metadata for important citations extraction that was not domain-specific.

We have seen from the above literature that for the classifications of citations into (1) important and (2) non-important categories, two major approaches were proposed by Qayyum & Afzal [18] and Valenzuela [1]. Qayyum & Afzal [18] criticize the Valenzuela [1] technique and comment that it depends mostly on the content of research articles which are not openly accessible by major journals like ACM, Elsevier, IEEE, Springer, etc. and they also used 12 different features to achieve the accuracy of 0.65. However, Qayyum & Afzal [18] used metadata-based features (i.e. title similarity, author overlap, references) with the contents feature (i.e. abstract and cue words) and evaluated their approach by using unigram, bigram, and trigram. Both of these approaches have given importance to metadata and in-text citation counts and their positions, however, research papers are written with the help of domain-specific terms and knowledge. Therefore, we argue that no one (i.e. researchers and authors) has used the content of corresponding logical sections of research papers and tried to compare the important terms represented in them.

Chapter 3

Proposed Methodology

This chapter presents a comprehensive methodology of our proposed approach. The proposed methodology is divided into different chunks such as comprehensive dataset selection, PDF to text conversion, pre-processing, cosine similarity measure, content-based comparisons, section-wise content-based comparisons, techniques for the combination of different logical sections and evaluation steps. The description of each chunk is presented in the sections below. The geographical presentation of the proposed methodology is shown in Figure 3.1 with a detailed explanation of each step.

3.1 Comprehensive Data Set Selection Criteria

For the reliability of our proposed approach, we need a comprehensive diversified annotated dataset which should include different logical sections (i.e. Abstract, Introduction, Literature Review, Methodology and Result.) of research articles. To the best of our knowledge, there exist three studies such as: Zhu [30], Valenzuela [1] and Qayyum & Afzal [18] for important citations classifications. We used the comprehensive annotated dataset of Valenzuela [1] because they have publically released their data in 2015. Moreover, this dataset has been used by two recent state-of-the-art approaches i.e. 1)- Qayyum & Afzal [18] and 2)-Valenzuela [1].

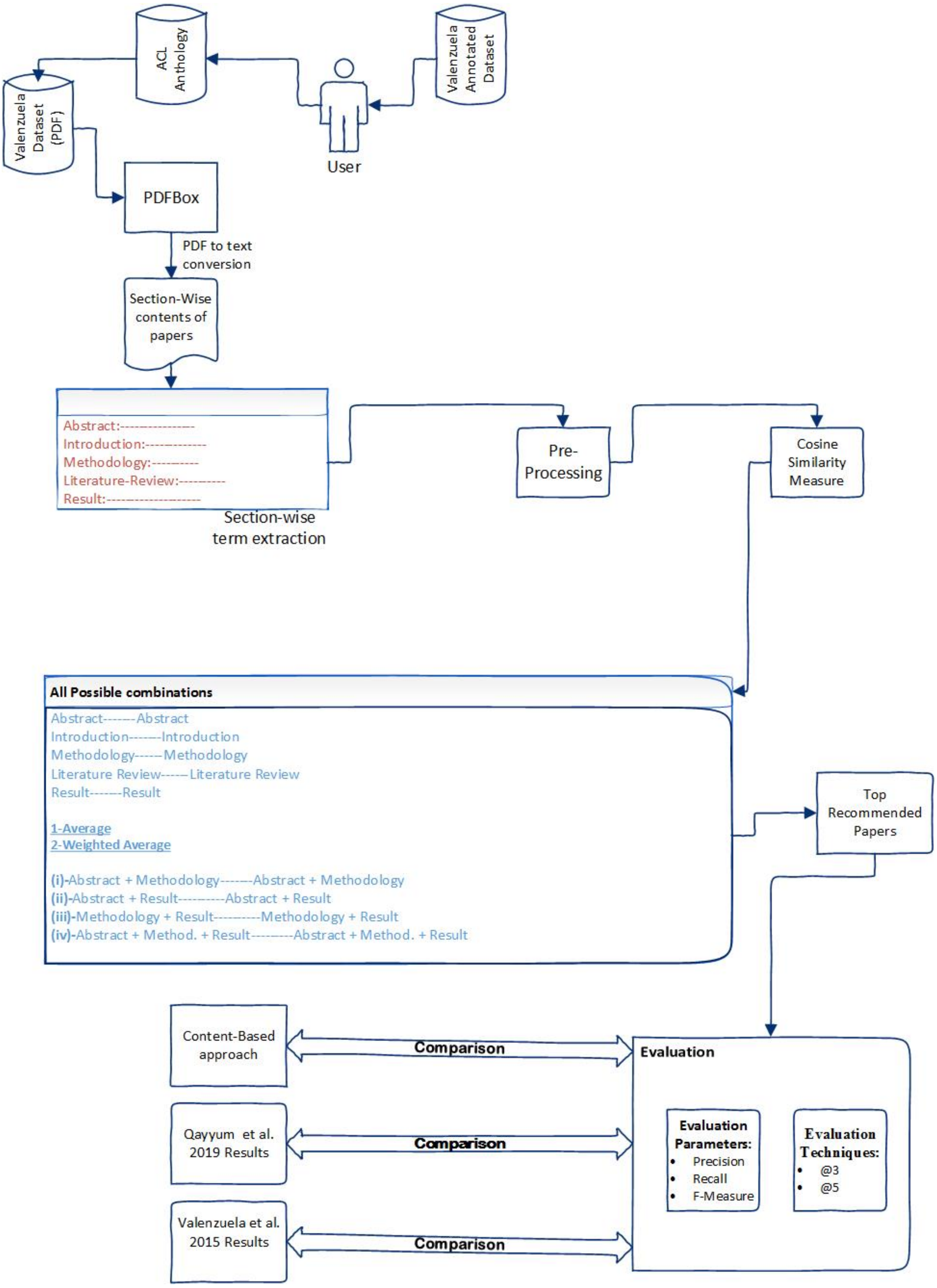


Figure 3.1: Proposed Methodology Diagram

Qayyum & Afzal [18] published their research work in the well reputed Scientometricsjournal in the year 2019. While Valenzuela [1] published their research work in AAAI conference on artificial intelligence which is a leading IEEE conference in the field of artificial intelligence.

Therefore, we have picked this diversified annotated dataset of Valenzuela. This dataset contains 465 research articles having 48 tuples of root papers and cited papers (i.e. cited paper - citing paper). Now based on the provided dataset by Valenzuela [1], we have extracted PDF files of the papers from ACL (Association of Computational and Linguistics) anthology. From these PDF files, we have extracted the logical sections such as Abstract, Introduction, Literature Review, Methodology and Result sections from research papers. The dataset description is shown in Figure 3.2. Column A is the annotator column in which annotation of research papers is done by two domain experts in the field. The second column (i.e. column B) represents the source IDs of research papers of ACL anthology. The third column (i.e. column C) represents the IDs of cited-by paper of source paper. Column D is the Follow-up column in which scores assigned from two annotators are displayed. The scores 0, 1 are for Non-Important and scores 2, 3 are for Important paper-citations pairs.

3.2 PDF to Text Conversion

The PDF research articles are available on ACL anthology. These articles contain all the logical sections of a research paper. After this, we used the PDFBox tool that converts PDFs into text files. These files contain the combination of all sections (e.g. Abstract, Introduction, Literature Review, Methodology, Results, etc.) of research articles as shown in Figure 3.3
The PDF file in Figure 3.3 contains different headings (i.e. Abstract, Introduction, etc.). These headings are treated as different logical sections. And we extracted these logical sections manually from this PDF. Then we generated another sub text files according to the different logical sections (e.g. Abstract, Introduction,

	A	B	C	D
1	Annotator	Paper	Cited-by	Follow-up
2	A	A00-1043	C00-2140	0
3	A	A00-1043	P02-1057	0
4	A	A97-1011	W09-1118	1
5	A	A97-1011	A00-2017	1
6	A	A97-1011	C00-2099	0
7	A	A97-1011	W04-1505	0
8	A	A97-1011	P99-1033	0
9	A	A97-1011	W06-0202	3
10	A	A97-1011	P01-1006	3
11	A	A97-1011	E12-1072	0
12	A	C00-1072	C02-1130	2
13	A	C00-1072	C04-1077	1

FIGURE 3.2: Benchmark Dataset

Literature Review, Methodology, Results, etc.) as shown in Figure 3.4 and Figure 3.5. All of the required five logical sections of research articles based on headings have been extracted on a Similar pattern as explained above.

3.3 Pre-Processing

Pre-processing is required to remove the noise from the dataset and achieving better results from the proposed methodology. We have divided the pre-processing phase into two steps. Initially, we removed stop words from the research papers and then the steaming process is applied to convert all remaining words in their root words for experimentation. The step by step description of these two approaches is as follows:

Abstract

This paper describes FERRET, an interactive question-answering (Q/A) system designed to address the challenges of integrating automatic Q/A applications into real-world environments. FERRET utilizes a novel approach to Q/A – known as *predictive questioning* – which attempts to identify the questions (and answers) that users need by analyzing how a user interacts with a system while gathering information related to a particular scenario.

1 Introduction

As the accuracy of today's best factoid question-answering (Q/A) systems (Harabagiu et al., 2005; Sun et al., 2005) approaches 70%, research has begun to address the challenges of integrating automatic Q/A systems into real-world environments. A new class of applications – known as interactive Q/A systems – are now being developed which allow users to ask questions in the context of extended dialogues in order to gather information related to any number of complex scenarios. In this paper, we describe our interactive Q/A system – known as FERRET – which uses an approach based on *predictive questioning* in order to meet the changing information needs of users over the course of a Q/A dialogue.

Answering questions in an interactive setting poses three new types of challenges for traditional than cooperatively answer a user's single question. Instead, in order to keep a user collaborating with the system, interactive Q/A systems need to provide access to new types of information that are somehow relevant to the user's stated – and unstated – information needs.

Second, we have found that users of Q/A systems in real-world settings often ask questions that are much more complex than the types of factoid questions that have been evaluated in the annual Text Retrieval Conference (TREC) evaluations. When faced with a limited period of time to gather information, even experienced users of Q/A may find it difficult to translate their information needs into the simpler types of questions that Q/A systems can answer. In order to provide effective answers to these questions, interactive question-answering systems need to include *question decomposition* techniques that can break down complex questions into the types of simpler factoid-like questions that traditional Q/A systems were designed to answer.

Finally, interactive Q/A systems must be sensitive not only to the content of a user's question – but also to the context that it is asked in. Like other types of task-oriented dialogue systems, interactive Q/A systems need to model both what a user knows – and what a user wants to know – over the course of a Q/A dialogue: systems that fail to represent a user's knowledge base run the risk of returning redundant information, while sys-

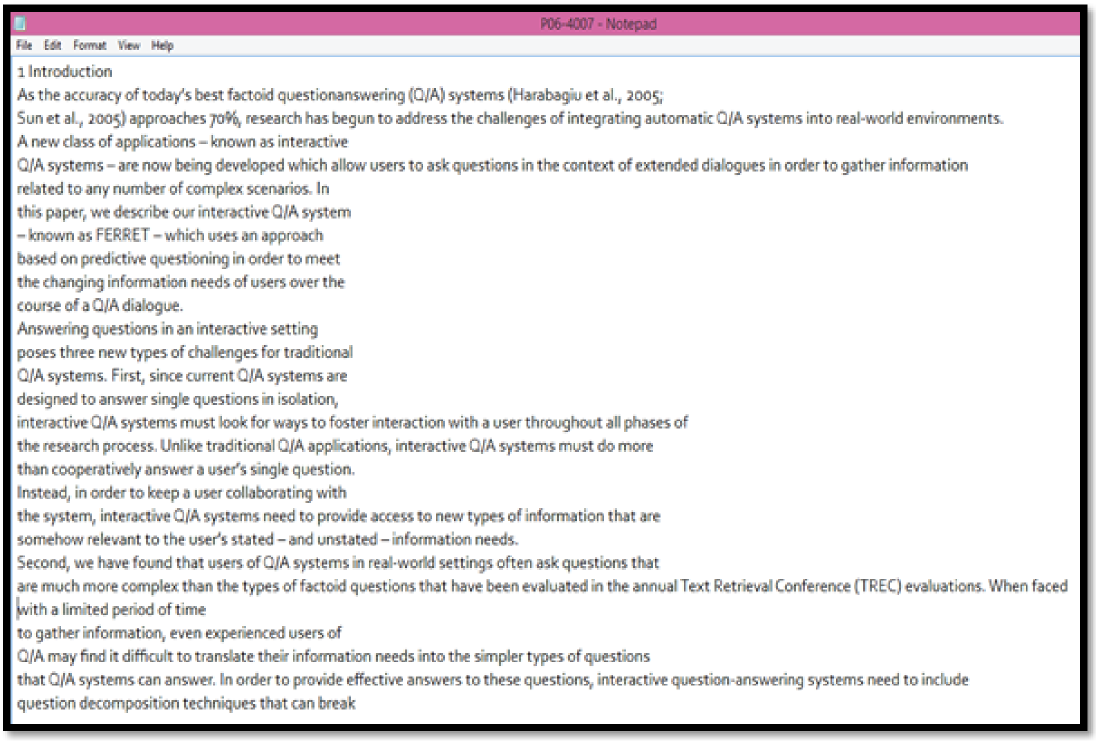
FIGURE 3.3: First page of a Typical Research Article

P06-4007 - Notepad

File Edit Format View Help

Abstract This paper describes FERRET, an interactive question-answering (Q/A) system designed to address the challenges of integrating
automatic Q/A applications into
real-world environments. FERRET utilizes
a novel approach to Q/A – known as predictive questioning – which attempts to
identify the questions (and answers) that
users need by analyzing how a user interacts with a system while gathering information related to a particular scenario.

FIGURE 3.4: Abstract of Research Article



P06-4007 - Notepad

File Edit Format View Help

1 Introduction
As the accuracy of today's best factoid questionanswering (Q/A) systems (Harabagiu et al., 2005;
Sun et al., 2005) approaches 70%, research has begun to address the challenges of integrating automatic Q/A systems into real-world environments.
A new class of applications – known as interactive
Q/A systems – are now being developed which allow users to ask questions in the context of extended dialogues in order to gather information
related to any number of complex scenarios. In
this paper, we describe our interactive Q/A system
– known as FERRET – which uses an approach
based on predictive questioning in order to meet
the changing information needs of users over the
course of a Q/A dialogue.
Answering questions in an interactive setting
poses three new types of challenges for traditional
Q/A systems. First, since current Q/A systems are
designed to answer single questions in isolation,
interactive Q/A systems must look for ways to foster interaction with a user throughout all phases of
the research process. Unlike traditional Q/A applications, interactive Q/A systems must do more
than cooperatively answer a user's single question.
Instead, in order to keep a user collaborating with
the system, interactive Q/A systems need to provide access to new types of information that are
somehow relevant to the user's stated – and unstated – information needs.
Second, we have found that users of Q/A systems in real-world settings often ask questions that
are much more complex than the types of factoid questions that have been evaluated in the annual Text Retrieval Conference (TREC) evaluations. When faced
with a limited period of time
to gather information, even experienced users of
Q/A may find it difficult to translate their information needs into the simpler types of questions
that Q/A systems can answer. In order to provide effective answers to these questions, interactive question-answering systems need to include
question decomposition techniques that can break

Figure 3.5: Introduction of Research Article

3.3.1 Stop Words Removal

The words like the, in, an, a, and, as, at, be, by, for, is, are, which, from, that, has, he, is, its, of, on, to, was, were, will, with occur frequently in English language sentences. So, there is a need to remove these stop words from the contents of different logical sections to obtain rare terms from the content. For removing the stop words, we picked the Onix Text Retrieval Toolkit Stop word List1[1] for removing stop words.

3.3.2 Stemming

In the stemming process, terms/words of contents of research articles are changed into their root words. For example, if we have two words like retrieval in the cited paper and retrieved in the citing paper, then these two words will not find their match when they are syntactically matched without the stemming process. The

[1] https://www.lextek.com/manuals/onix/stopwords1.html

Algorithm of Porter Stemming is applied to convert the terms/words into their root word [32]. For example, the words like retrieval, retrieved and retrieves will convert into their base word retriev. This algorithm was applied to the entire logical sections of the research articles.

3.4 Similarity Measures

Similarity measures techniques are used to check that how much content of two text documents (i.e corresponding logical sections) is similar. From the literature review, we have reviewed that there are different similarity measure techniques were used to examine the similarity between the content of two text documents. These similarity measures techniques were the cosine similarity measure, Jaccard similarity measure, Euclidean distance similarity measure etc. However, cosine similarity is one of the most common similarity measure technique used in literature to find similarity between text documents, for example in various applications for information retrieval and data mining. In this research work, we mainly focused on the cosine similarity measure for extracting our results. Because state-of-the-art studies revealed that it produces more accurate results than others like Jaccard similarity and Euclidean distance similarity measures [33][34].

Every research paper needs to be converted into a vector and each vector is helpful to identify the similarity between this vector and other vectors with the help of cosine similarity. Cosine similarity is used to find the normalized dot product of the two documents d1 and d2. The cosine similarity determines the cosine between the angles. If the two documents (d1, d2) are parallel then the cosine similarity among these documents is 1 while the documents which make an angle of 90 have the cosine similarity of 0. In this thesis, we will convert the documents (d1, d2) into the vectors and then cosine similarity will be calculated according to the presented equation by Korenius et al. 2007 [35] and Jain et al. 2017 [36] showing below.

$$CosineSimilarity(d1, d2) = \frac{d1.d2}{|d1| * |d2|} \tag{3.1}$$

$$|d1| = \sqrt{d1[0]^2 + d1[1]^2 + d1[2]^2 + ... + d1[n]^2} \tag{3.2}$$

$$|d2| = \sqrt{d2[0]^2 + d2[1]^2 + d2[2]^2 + ... + d2[n]^2} \tag{3.3}$$

3.5 Content-Based Comparisons

We are interested to evaluate the content of research papers by a standard approach of comparing the contents of different research papers. And additionally, we are also interested to evaluate the section-wise contents of research papers comparisons. We used the Apache Lucene tool kit for comparing the contents of research papers. This toolkit contains a variety of APIs to perform a different kind of text similarity tasks. One of the starting task is to calculate or identify the TF-IDF (term-frequency inverse- document-frequency) for rare terms/words retrieval and calculate cosine similarity between research articles which can be done. The Apache Lucene is frequently using for measuring research articles/text similarity [37] [38].

3.5.1 Extracting Important Terms

We need to extract important terms from the content of research articles. First, text files are given to Lucene API that extracts important terms by using the TF-IDF scheme. This uses the following equation (3.4) for the extraction of key terms from all text documents.

$$Tf * Idf(t, d, D) = tf(t, d) * idf(t, D) \tag{3.4}$$

TF-IDF identifies those terms (i.e. t) as important which occur frequently in a particular document (i.e. d) and do not give weightage those terms (i.e. t) which occur frequently in all other documents (i.e. D).

3.5.2 Ranking Research Papers Based on Content

There are many techniques and ways to find the similarity between different items. These techniques include Cosine similarity, Jaccard similarity, Euclidean distance, etc. However, Cosine similarity is widely used to measure text similarity between documents. The cosine similarity measure is used to find text similarity between two research articles. The key terms of document d1 are represented as a vector A and key terms of document d2 are represented as a vector B that is shown in equation (5). We calculated cosine similarity for each document against all documents/research articles in the dataset; all research articles got a similarity score, which is sorted and the rank list is obtained of similar documents.

$$Similarity = cos(\Theta) = \frac{A.B}{\|A\| \, \|B\|} = \frac{\sum_1^n (A_i * B_i)}{\sqrt{\sum_1^n A_i^2}\sqrt{\sum_1^n B_i^2}} \tag{3.5}$$

3.6 Section-wise Content-Based Comparisons

As highlighted in section 3.5, we are interested to evaluate the section-wise contents of research papers for the identification of important citations. Therefore, different logical sections of research articles are required to perform section-wise comparisons of content. The similar steps as discussed above in section 3.5 are required with some additional tasks to perform section-wise content-based comparisons. All text files have been converted into five logical sections sub text files (i.e. abstract, introduction, literature review etc.). These logical sections include Abstract, Introduction, Literature Review, Methodology, and Results sections.

3.6.1 Section-wise Extraction of Important Terms

Once the sections have been extracted, now we want to identify important terms within those sections for implementing the section-wise content-based similarity. Section-wise text files are given to Lucene API that extracts important terms by using the TF-IDF scheme that is needed for a proposed approach. This uses the following equation (3.6) for all the text documents for the extraction of key terms.

$$Tf * Idf(t, ds, Ds) = tf(t, ds) * idf(t, Ds) \tag{3.6}$$

TF-IDF identifies those terms (i.e. t) as important which occur frequently in a particular document section (i.e. ds) and do not give weightage those terms (i.e. t) which occur frequently in all other documents sections (i.e. Ds).

3.6.2 Ranking Research Papers Based on Section-wise Comparisons

Every research article is divided into five logical sections which include Abstract, Introduction, Literature Review, Methodology and Results sections. Each research article is donated by vector term. Lets we try to understand it with an example. If we have two research articles where d1 is a cited article donated by vector A and d2 is a citing article denoted by vector B. Now we want to compare vector A with vector B for calculating similarity. The cosine similarity approach is used for calculating the similarity between vector A and vector B. Similarly, cosine similarity of each corresponding logical section of a research article is calculated against all research articles sections in the dataset. All research articles got a similarity score, which is sorted in descending order in the form of a rank list. In the end, we used average(mean) and weighted average techniques to combine different logical sections of research articles which will be discussed in detail in the next section.

3.7 Combination Techniques of Logical Section

After an independent comparisons of each logical section Abstract Vs Abstract, Introduction Vs Introduction, Literature Review Vs Literature Review, Methodology Vs Methodology and Results Vs Results, we have formed combinations of different logical sections. There are following different possible combinations of the logical sections of research articles:

- Abstract + Introduction Vs Abstract + Introduction
- Abstract + Literature Review Vs Abstract + Literature Review
- Abstract + Methodology Vs Abstract + Methodology
- Abstract + Results Vs Abstract + Results
- Introduction + Literature Review Vs Introduction + Literature Review
- Introduction + Methodology Vs Introduction + Methodology
- Introduction + Results Vs Introduction + Results
- Literature Review + Methodology Vs Literature Review + Methodology
- Literature Review + Results Vs Literature Review + Results
- Methodology + Results Vs Methodology + Results
- Abstract + Introduction + Literature Review Vs Abstract + Introduction + Literature Review
- Abstract + Introduction + Methodology Vs Abstract + Introduction + Methodology
- Abstract + Introduction + Results Vs Abstract + Introduction + Results
- Abstract + Literature Review + Methodology Vs Abstract + Literature Review + Methodology

- Abstract + Literature Review + Results Vs Abstract + Literature Review + Results
- Abstract + Methodology + Results Vs Abstract + Methodology + Results
- Introduction + Literature Review + Methodology Vs Introduction + Literature Review + Methodology
- Introduction + Literature Review + Results Vs Introduction + Literature Review + Results
- Literature Review + Methodology + Results Vs Literature Review + Methodology + Results
- Abstract + Introduction + Literature Review + Methodology vs Abstract + Introduction + Literature Review + Methodology
- Abstract + Introduction + Literature Review + Results vs Abstract + Introduction + Literature Review + Results
- Introduction + Literature Review + Methodology + Results vs Introduction + Literature Review + Methodology + Results
- Abstract + Introduction + Literature Review + Methodology + Results vs Abstract + Introduction + Literature Review + Methodology + Results

All the above mentioned possible combinations of logical sections have been combined with each other by using the Average and Weighted Average techniques that are briefly discussed in the following sections 3.7.1 and 3.7.2.

3.7.1 Average

An average which is also referred as an arithmetic mean is calculated by taking the sum of values in a cluster and dividing it by the total number of values in the cluster. The average technique is used in this research to combine similarity score of different logical sections of research articles which are discussed in section 3.7 of

=AVERAGE(E3,F3)

C	D	E	F	G	H	I
Cited	Citing	Abstract Score	Methodology Score	Result Score	Average(Abstract, Methodology)	Average(Abstract, Result)
A97-1011.txt	A00-2017.txt	0.16842414	0.233827254	0.132761389	0.201125697	0.150592764
A97-1011.txt	P99-1033.txt	0.118466917	0.270121863	0.231294486	0.19429439	0.174880701
A97-1011.txt	W06-0202-Imp.txt	0.044599906	0.223285023	0.136228785	0.133942465	0.090414345
A97-1011.txt	P01-1006-imp.txt	0.026018326	0.226617417	0.199672283	0.126317871	0.112845304
A97-1011.txt	E12-1072.txt	0.060641246	0.186974236	0.176509497	0.123807741	0.118575371
A97-1011.txt	W04-1505.txt	0.045098124	0.201923258	0.117893593	0.123510691	0.081495858
A97-1011.txt	W09-1118.txt	0.022786525	0.15970037	0.125576084	0.091243447	0.074181305
A97-1011.txt	C00-2099.txt	0.016430754	0.152573448	0.073708485	0.084502101	0.04506962

FIGURE 3.6: Combination of Two Sections by Using Average

this chapter. The average technique used the following formula shown in equation (3.7).

$$Average = \frac{\sum_{i=1}^{n} X_i}{n} \tag{3.7}$$

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + X_3 + X_4 + ... + X_n \tag{3.8}$$

Where X1, X2,, Xn are the similarity scores of different sections of research articles and n is showing total number of sections of research articles.

Example: We used the average technique to combine the similarity score of two or more than two sections. It is shown in Figure 3.6 that Abstract and Methodology sections similarity scores are combined in column I and Abstract and Results sections similarity score are combined in column J by using average formula between them. Similarly, the average technique is used in this research to combine different sections of research articles discussed in section 3.7.

3.7.2 Weighted Average

We have also performed another experiment in which different weights have been assigned to the logical sections based on their importance. This research has used the weighted average technique to combine similarity score of different logical sections which are discussed in section 3.7 of this chapter. This technique used the formula shown in equation (3.9). Weights are assigned to every text files of logical sections on the basis of counting the words in that file through online wordcounter[2] .

$$WeightedAverage = \frac{(x_1 * \frac{1}{count(s_1)}) + (x_2 * \frac{1}{count(s_2)}) + ... + (x_n * \frac{1}{count(s_n)})}{SUM((\frac{1}{count(s1)}), (\frac{1}{count(s2)}), ..., (\frac{1}{count(sn)}))} \quad (3.9)$$

Where x1, x2,,xn are the similarity scores of an Abstract, Methodology and Result sections of research articles while count(s1),count(s2),...,count(sn) showing the total number of words of an Abstract, Methodology and Result sections respectively.

Example The weighted average technique is applied to combine the similarity scores of different logical sections of research articles. It is shown in Figure 3.7 that there are Abstract section similarity score and weights (i.e. count(abstract)) are presented in Column F and column G respectively. While Methodology section score and weights (i.e count(methodology)) are shown in column I and column J respectively. The weighted average of the Abstract and Methodology section is calculated in the last column (i.e. Column L) of Figure 3.7 by using the formula shown in equation (3.9). Similarly, the weighted average formula is used in this research to combine similarity score of different sections of research articles discussed in section 3.7.

3.8 Evaluation

For the evaluation of our results, we used formulas of Precision, Recall and F-Measure for calculation of the results of our proposed technique. The standard formulas of Precision, Recall, and F-Measure are as following:

[2]https://wordcounter.net/

=((F4*H4)+(I4*K4))/SUM(H4,K4)

Cited	Citing	Abstract Score	Count(Abstract)	1/Count(Abstract)	Methodology Score	Count(Methodology)	1/Count(Methodology)	W.Avg(A,M)
A97-1011.txt	A00-2017.txt	0.1684	179	0.005586592	0.2338	2974	0.000336247	0.172137162
A97-1011.txt	P99-1033.txt	0.1185	84	0.011904762	0.2701	3112	0.000321337	0.122452842
A97-1011.txt	E12-1072.txt	0.0606	124	0.008064516	0.1870	2178	0.000459137	0.067446324
A97-1011.txt	W06-0202-Imp.txt	0.0446	110	0.009090909	0.2233	1284	0.000778816	0.05869988
A97-1011.txt	W04-1505.txt	0.0451	84	0.011904762	0.2019	8465	0.000118133	0.046639042
A97-1011.txt	P01-1006-imp.txt	0.0260	59	0.016949153	0.2266	1608	0.000621891	0.033118113
A97-1011.txt	W09-1118.txt	0.0228	71	0.014084507	0.1597	1627	0.000614628	0.028511427
A97-1011.txt	C00-2099.txt	0.0164	95	0.010526316	0.1526	1790	0.000558659	0.023292057

FIGURE 3.7: Combination of Two Sections by Using Weighted Average

$$Precision = \frac{(TruePositive)}{(TruePositive + FalsePositive)} \tag{3.10}$$

$$Precision = \frac{(TruePositive)}{(TotalPredictedPositive)} \tag{3.11}$$

$$Recall = \frac{(TruePositive)}{(TruePositive + FalseNegative)} \tag{3.12}$$

$$Recall = \frac{(TruePositive)}{(TotalActualPositive)} \tag{3.13}$$

$$F - Measure = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \tag{3.14}$$

Our proposed approach results will be compared with the results of the techniques of Qayyum & Afzal [18], and Valenzuela [1]. We have used the comprehensive diversified annotated dataset of Valenzuela [1] which is being used by Qayyum & Afzal [18], and Valenzuela [1] approaches.

This chapter (i.e. Proposed Methodology Chapter) addressed the brief proposed methodology of the research. The proposed methodology was divided into various steps such as comprehensive dataset selection, PDF to text conversion, preprocessing, cosine similarity measure, content-based comparisons, section-wise content-based comparisons, combination techniques of different logical sections and evaluation steps. For the combination of different logical sections similarity scores of research articles, average and weighted average techniques were used. These all possible combinations are discussed in section 3.7. For the evaluation of our experiments and results, we used formulas of Precision, Recall and F-Measure for calculation of the results of our proposed technique by considering the top @3 and top @5 ranked research articles.

Chapter 4

Experiments and Results Discussion

In chapter 3 (i.e. Proposed Methodology), the precise methodology has been presented to address the identified gap from the literature. This chapter depicts the results that have been attained by applying the proposed methodology.

4.1 Dataset Collection

We performed our experiments on the comprehensive annotated dataset of Valenzuela [1]. This dataset contains 465 research articles (i.e. paper-citation pairs) having 48 tuples of root paper and cited paper, crawled from ACL (Association of Computational and Linguistics) anthology. While crawling 465 paper-citation pairs, 33 pairs couldnt be crawled due to unavailability on ACL. From 465 paper-citation pairs, domain experts have annotated 14.6% pairs as Important and remaining 85.4% pairs as Non-Important. The crawled 432 paper-citation pairs contain 375 research articles as Non-Important and remaining 57 research articles as Important.

4.2 PDF to Text Conversion

In order to access content, the PDF research articles crawled from ACL anthology need to be converted to text format (i.e.txt). We have used the PDFBox tool to convert PDFs to text files. There are various headings in the PDF file (i.e. Abstract, Introduction, etc.). Such headings are treated as different logical sections. And from this PDF we manually extracted these logical sections. We then generated another subtext files based on the different logical sections (e.g. Abstract, Introduction, Literature review, Methodology, Result, etc.). All the required five logical sections of research articles based on headings were extracted using similar patterns as explained above.

4.3 Pre-Processing

Pre-processing is required to remove the noise from the dataset and achieving better results from the proposed methodology. We have divided the pre-processing phase into two steps. Initially, we removed stop words from the research articles by using the Onix Text Retrieval Toolkit Stop word List1[1]. Then, the Algorithm of Porter Stemming is applied to convert all remaining words in their root words for experimentation [32].

4.4 Similarity Measures

Similarity measures techniques are used for the comparison of content between two text documents (i.e cited papers and cited by papers). We have employed various similarity measures to compare the text present in different logical sections between cited and cited by papers. In this research work, we mainly focused on the cosine similarity measure for extracting our results. Because state-of-the-art

[1] www.lextek.com/manuals/onix/stopwords1.html

Different possible combinations

1-{Abstract + Introduction Vs Abstract + Introduction}

2-{Abstract + Literature Review Vs Abstract + Literature Review}

3-{Abstract + Methodology Vs Abstract + Methodology}

4-{Abstract + Results Vs Abstract + Results}

5-{Introduction + Literature Review Vs Introduction + Literature Review}

6-{Introduction + Methodology Vs Introduction + Methodology}

7-{Introduction + Results Vs Introduction + Results}

8-{Literature Review + Methodology Vs Literature Review + Methodology}

9-{Literature Review + Results Vs Literature Review + Results}

10-{Methodology + Results Vs Methodology + Results}

11-{Abstract + Introduction + Literature Review Vs Abstract + Introduction + Literature Review}

12-{Abstract + Introduction + Methodology Vs Abstract + Introduction + Methodology}

13-{Abstract + Introduction + Results Vs Abstract + Introduction + Results}

14-{Abstract + Literature Review + Methodology Vs Abstract + Literature Review + Methodology}

15-{Abstract + Literature Review + Results Vs Abstract + Literature Review + Results}

FIGURE 4.1: All Possible Combinations

studies revealed that it produces more accurate results than others like Jaccard similarity and Euclidean distance similarity measures [33][34].

4.5 All Possible Combinations

After an independent comparisons of each logical section Abstract Vs Abstract, Introduction Vs Introduction, Literature Review Vs Literature Review, Methodology Vs Methodology and Results Vs Results, combinations of different logical sections have been formed. There are following different possible combinations of the logical sections of research articles shown in Figure 4.1 and Figure 4.2.

Different possible combinations

16-{Abstract + Methodology + Results Vs Abstract + Methodology + Results}

17-{Introduction + Literature Review + Methodology Vs Introduction + Literature Review + Methodology}

18-{Introduction + Literature Review + Results Vs Introduction + Literature Review + Results}

19-{Literature Review + Methodology + Results Vs Literature Review + Methodology + Results}

20-{Abstract + Introduction + Literature Review + Methodology vs Abstract + Introduction + Literature Review + Methodology}

21-{Abstract + Introduction + Literature Review + Results vs Abstract + Introduction + Literature Review + Results}

22-{Introduction + Literature Review + Methodology + Results vs Introduction + Literature Review + Methodology + Results}

23-{Abstract + Introduction + Literature Review + Methodology + Results vs Abstract + Introduction + Literature Review + Methodology + Results}

Figure 4.2: All Possible Combinations

4.6 Score Calculation

After applying the steps of preprocessing and division of research articles into five logical sections, we have calculated the results and analyzed the outcomes. All the formulas discussed in the Methodology chapter were applied to different sections and their possible combinations of research articles. The calculated score of every formula lies in the range from 0 to 1.

4.7 Evaluation

This thesis employs standard evaluation parameters used in this domain such as Precision, Recall and F-Measure for the evaluation process. The five corresponding logical sections and their all possible combinations of research papers are shown in Figure 4.1 and Figure 4.2. In this thesis, cosine similarity was used to measure the similarity score between papers. The similarity score of each section is ranked in descending order and then Precision, Recall and F-measure of top 3 and top 5 ranked research papers are considered for results calculation.

4.7.1 Single Parameter

In a single parameter, we are interested to just utilize one logical section from all extracted logical sections (i.e. Abstract, Introduction, Literature Review, Methodology, and Result) for the identification of important citations. As we have five logical sections, therefore, in this evaluation we will get five rankings. The first ranking will be achieved by comparing the abstract of citing paper with the abstract of the cited paper. This comparison will be based on the cosine similarity computation between the mentioned sections. This process will continue for all citing papers for a focused cited paper. The similarly scores will be ranked in descending order to achieve the ranking of documents. The ranking will then be evaluated based on the above mentioned evaluation parameters. In a similar way, the second ranking will be computed based on the comparisons of the Introduction of citing paper and the introduction of the cited paper. This process will continue for Literature review, Methodology, and result sections. All such five rankings will then be compared individually with the state-of-the-art ranking achieved by including all-content of the citing and cited-by papers. This process is shown in the algorithm listing in Figure 4.3

4.7.2 Abstract Vs Abstract Parameter

In the abstract-abstract parameter, the content of an abstract section of a cited paper is matched with the content of an abstract of all its citing papers. The achieved cosine similarity score of each (i.e., cited paper-citing paper) pair is ordered in descending order. After that, the Precision, Recall and F-Measure score of top 3 and top 5 ranked papers is calculated. The overall F-Measure score of the abstract-abstract Parameter is then compared to the F-Measure score of the

```
Algorithm: Single Parameter Extraction for Identification of Important Citations
Input: Research Papers in PDF
Output: Ranking of different sections of Research Papers
    Cited_Paper ← Assign All Cited Paper
    Citing_Paper ← Assign All Citing Paper Against Cited_Paper
    For x in Cited_Paper do
        For each section in x do // read all section of the paper x
            For each Citing in Citing_Paper[x] do // read all citing papers of the paper x
                For each section in Citing do // read all sections of the citing paper
                    If x[section]== Citing[section]
                        Test[I]= Cosine-Similarity(x[section], Citing[section])
                        Similarity_Score[section]← Cosine-Similarity(x[section], Citing[section])
                         I←I+1
                End for Loop
            End for Loop
            Sort_Score←Sort_in_Descending(Test[I])
            Average[section]← (Average[section] + Find_F-Measure(Sort_Score)) //Select top 3 and top 5
        End for Loop
    End for Loop
```

Figure 4.3: Single Parameter Extraction

All-Content (i.e., whole content) parameters of paper.

The top 3 ranked abstract-abstract parameter papers scored 0.70 against the top 3 ranked All-Content parameter papers with a score of 0.63. The top 5 ranked abstract-abstract parameter papers achieved 0.69 F-Measure score against the top 5 ranked All-Content parameter papers with a score of 0.65. The F-measure scores of the top 3 ranked papers show that the abstract-abstract parameter outperformed the All-Content parameter which is depicted in Figure 4.4 and 4.5. While the F-measure scores of the top 5 ranked papers show that the abstract-abstract parameter outperformed the All-Content parameter which is depicted in Figure 4.6 and 4.7.

4.7.3 Introduction Vs Introduction Parameter

In the Introduction-Introduction parameter, the content of the introduction section of a cited paper is matched with the content of an introduction of all its citing papers. The achieved similarity score of each (i.e., cited paper-citing paper) pair is ordered in descending order. After that, the Precision, Recall and F-Measure score of top 3 and top 5 ranked papers is calculated. The overall F-Measure score of the introduction-introduction Parameter is then compared to the F-Measure

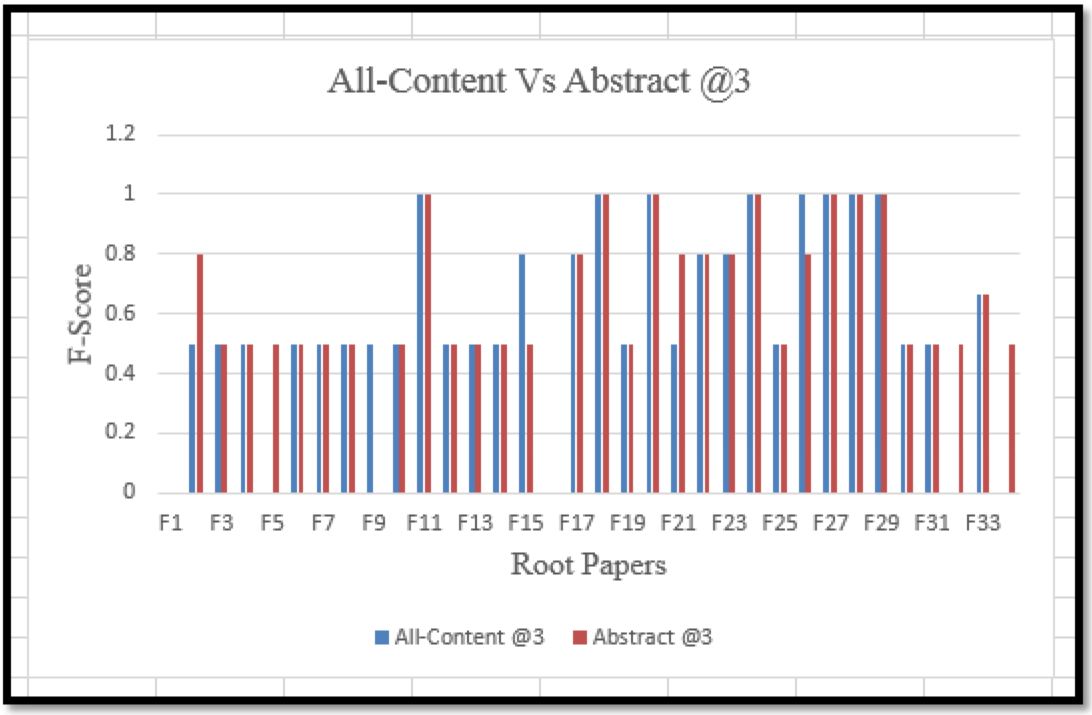


FIGURE 4.4: All-Content Vs Abstract @3

Abstract@3	All-Content@3
F2	F9
F5	F15
F21	F26
F32	
F34	

FIGURE 4.5: All-Content Vs Abstract @3

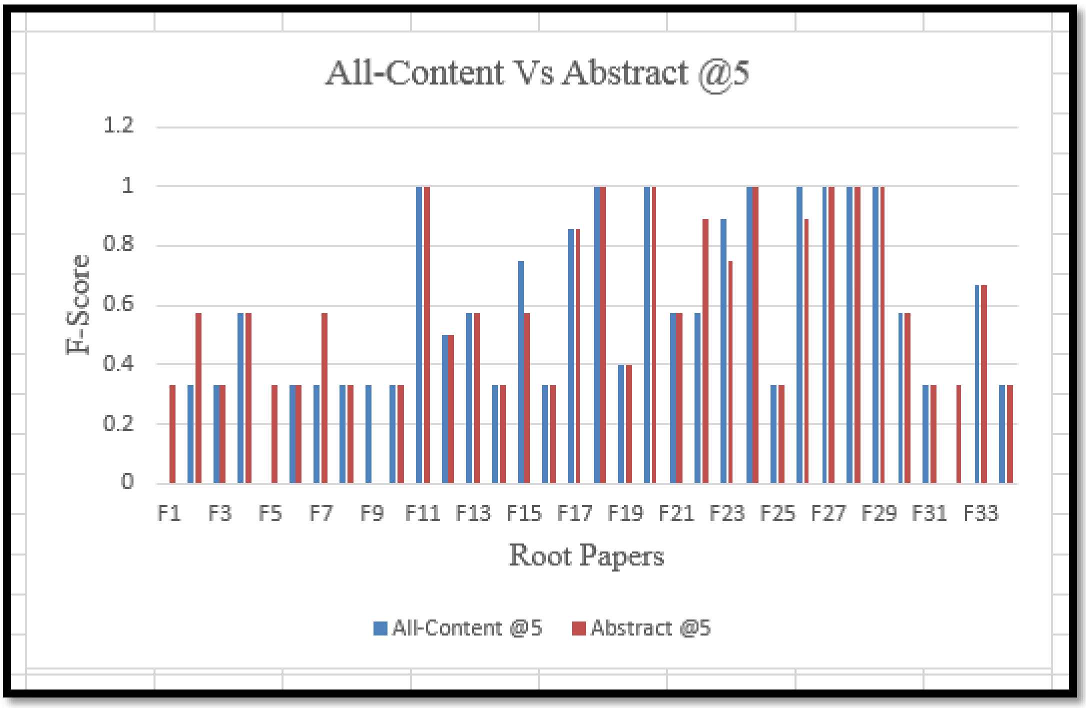


FIGURE 4.6: All-Content Vs Abstract @5

Abstract@5	All-Content@5
F1	F9
F2	F15
F5	F23
F7	F26
F22	
F32	

FIGURE 4.7: All-Content Vs Abstract @5

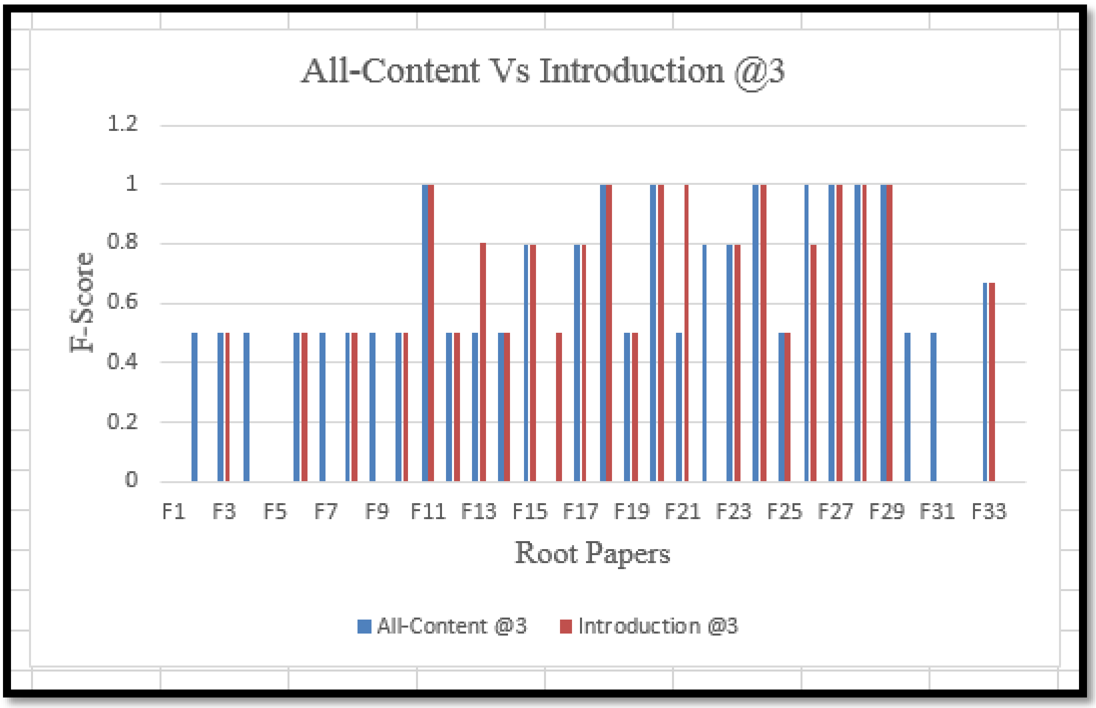


FIGURE 4.8: All-Content Vs Introduction @3

score of the All-Content (i.e., whole content) parameter of paper.

The top 3 ranked Introduction-Introduction parameter papers scored 0.57 against the top 3 ranked All-Content parameter papers with a score of 0.63. The top 5 ranked Introduction-Introduction parameter papers achieved 0.59 F-Measure score against the top 5 ranked All-Content parameter papers with a score of 0.65. The F-measure scores of the top 3 ranked papers show that the All-Content parameter scored better than the IntroductionIntroduction parameter which is demonstrated in Figure 4.8 and 4.9. While the top 5 ranked papers show that the All-Content parameter also scored better than the IntroductionIntroduction parameter which is demonstrated in Figure 4.10 and 4.11.

4.7.4 Literature Review Vs Literature Review Parameter

In the Literature Review-Literature Review parameter, the content of the Literature-Review section of a cited paper is matched with the content of Literature-Review of all its citing papers. The achieved cosine similarity score of each (i.e., cited paper-citing paper) pair is sorted in descending order. After that, the Precision,

Introduction@3	All-Content@3
F13	F2
F16	F4
F21	F7
	F9
	F22
	F26
	F30
	F31

FIGURE 4.9: All-Content Vs Introduction @3

Recall and F-Measure score of top 3 and top 5 ranked papers is calculated. The overall F-Measure score of the Literature Review-Literature Review parameter is then compared to the F-Measure score of the All-Content parameters (i.e., whole content) of paper.

The top 3 ranked Literature Review-Literature Review parameter papers scored 0.59 against the top 3 ranked All-Content parameter papers with a score of 0.63. The top 5 ranked Literature Review-Literature Review parameter papers achieved 0.62 F-measure score against the top 5 ranked All-Content parameter papers with a score of 0.65. The F-measure scores of the top 3 ranked papers show that All Content parameter scored better than the Literature Review-Literature Review parameter which is shown in Figure 4.12 and 4.13. While the F-measure scores of the top 5 ranked papers also show that All Content parameter scored better than the Literature Review-Literature Review parameter which is shown in Figure 4.14

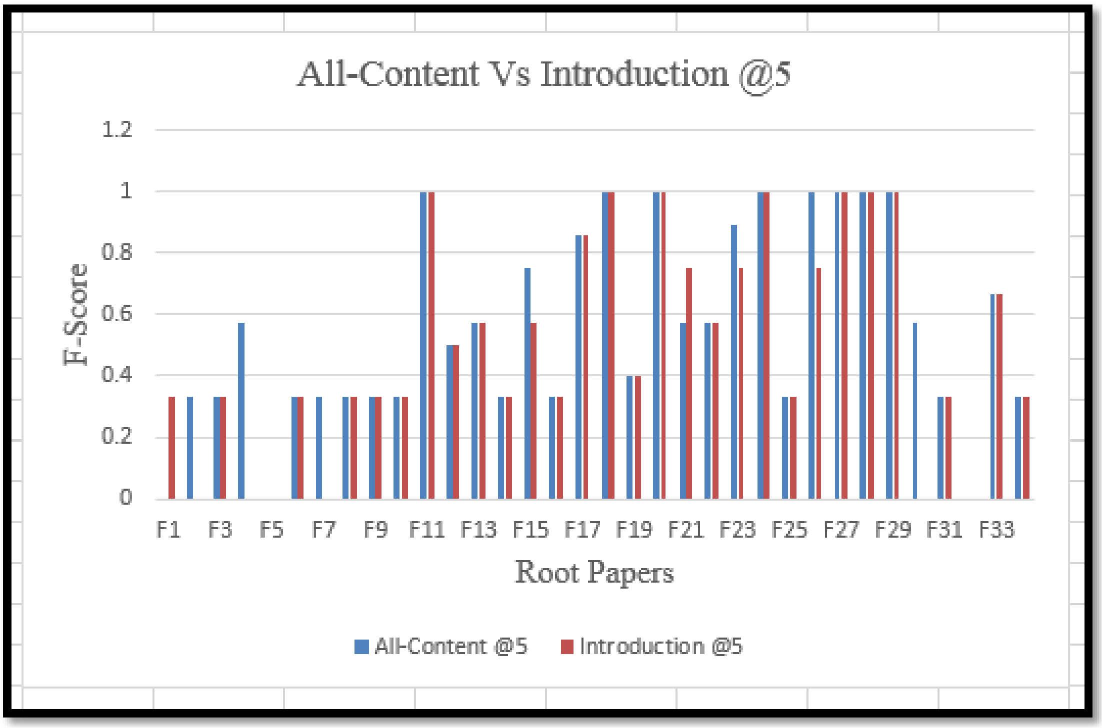


FIGURE 4.10: All-Content Vs Introduction @5

Introduction@5	All-Content@5
F1	F2
F21	F4
	F7
	F15
	F23
	F26
	F30

FIGURE 4.11: All-Content Vs Introduction @5

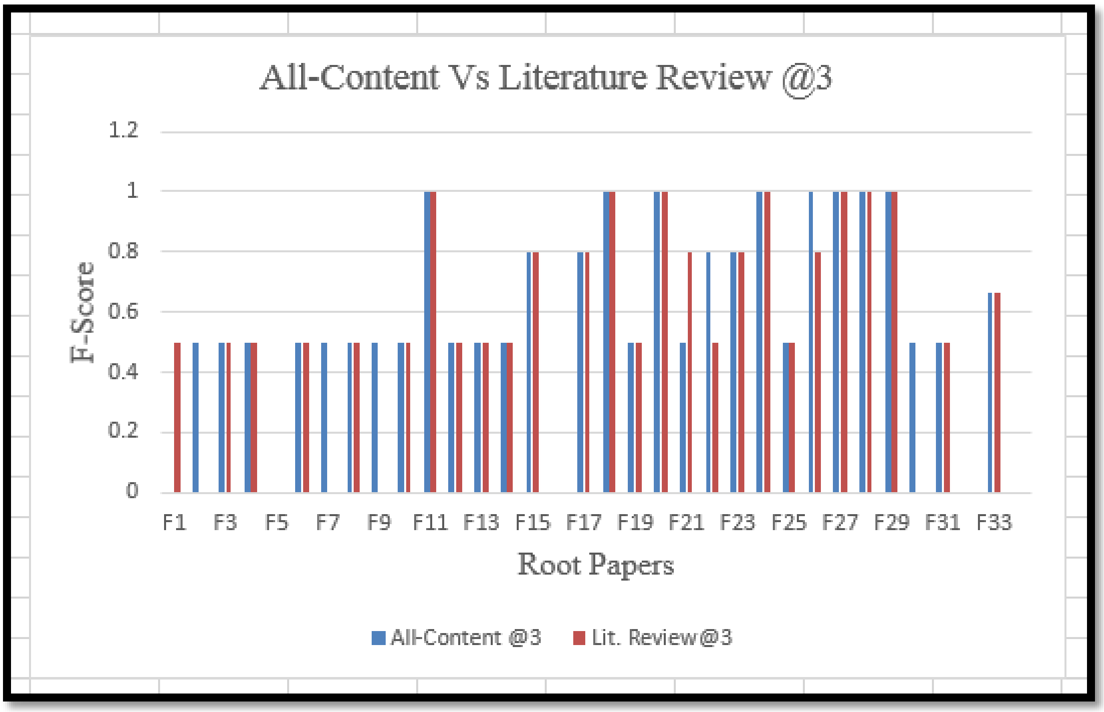


FIGURE 4.12: All-Content Vs Literature Review @3

and 4.15.

4.7.5 Methodology Vs Methodology Parameter

In the Methodology-Methodology parameter, the content of the Methodology section of a cited paper is matched with the content of the Methodology of all of its citing papers. The achieved cosine similarity score of each (i.e., cited paper-citing paper) pair is ordered in descending order. After that, the Precision, Recall and F-Measure score of top 3 and top 5 ranked papers is calculated. The overall F-Measure score of the Methodology-Methodology parameter is then compared to the F-Measure score of the All-Content parameters (i.e., whole content) of paper. The top 3 ranked Methodology-Methodology parameter papers scored 0.68 against the top 3 ranked All-Content parameter papers with a score of 0.63. The top 5 ranked Methodology-Methodology parameter papers achieved 0.66 F-Measure score against the top 5 ranked All-Contents parameter papers with a score of 0.65. The F-measure scores of the top 3 ranked papers show that the All-Content parameter was outperformed by the Methodology-Methodology parameter which is

L.R.@3	All-Content@3
F1	F2
F21	F7
	F9
	F22
	F26
	F30

FIGURE 4.13: All-Content Vs Literature Review @3

depicted in Figure 4.16 and 4.17. While the F-measure scores of the top 5 ranked papers show that the Methodology-Methodology parameter scored equally with the "All-Content parameter which is depicted in Figure 4.18 and 4.19.

4.7.6 Result Vs Result Parameter

In the Result-Result parameter, the content of the Result section of a cited paper is matched with the content of the Result of all its citing papers. The achieved cosine similarity score of each (i.e., cited paper-citing paper) pair is sorted in descending order. After that, the Precision, Recall and F-Measure score of top 3 and top 5 ranked papers is calculated. The overall F-Measure score of the Result-Result parameter is then compared to the F-Measure score of the All-Content parameters (i.e., whole content) of paper.

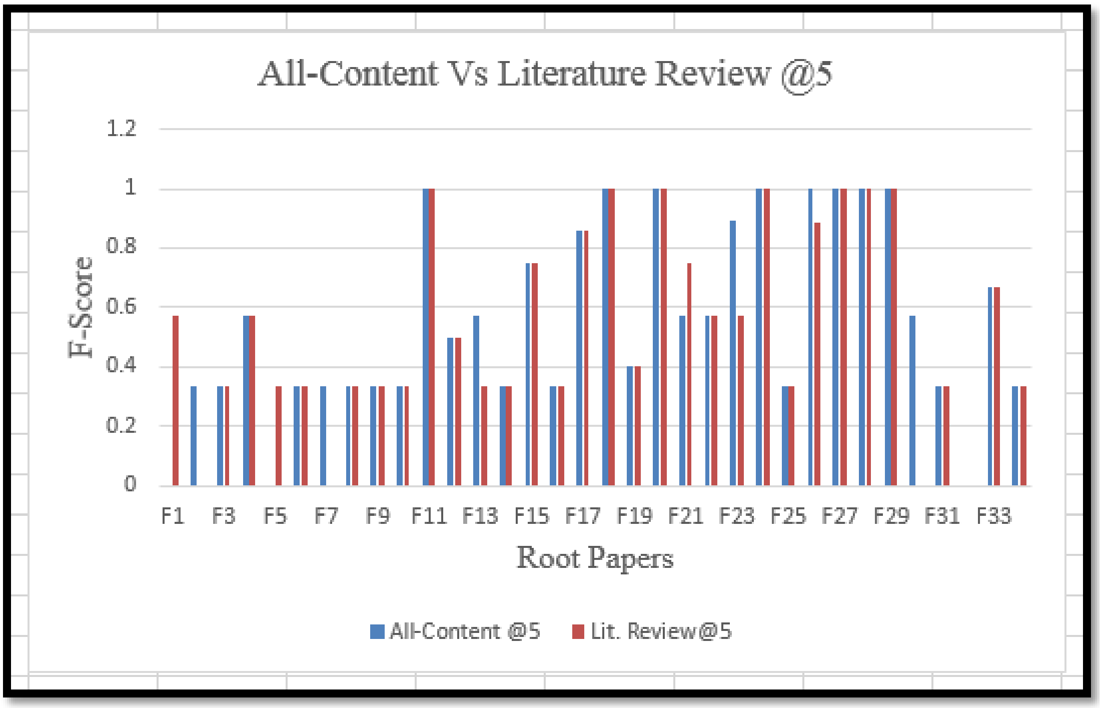


FIGURE 4.14: All-Content Vs Literature Review @5

The top 3 ranked Result-Result parameter papers scored 0.64 against the top 3 ranked All-Content parameter papers with a score of 0.63. The top 5 ranked Result-Result parameter papers achieved an F-Measure score of 0.63 against the top 5 ranked All-Content parameter papers with a score of 0.65. The F-measure scores of the top 3 ranked papers show that the All-Content parameter scored approximately equal with the Result-Result parameter which is demonstrated in Figure 4.20 and 4.21. While the F-measure scores of the top 5 ranked papers also show that the All-Content parameter scored approximately equal with the Result-Result parameter which is demonstrated in Figure 4.22 and 4.23.

4.7.7 Single Parameters Conclusion

The figure 4.24 concludes the results of single parameters in comparison of All-Content parameter. The results are deduced from the top 3 ranked and top 5 ranked research papers. We can clearly see that the highlighted lines single parameters like Abstract-Abstract, Methodology-Methodology, and Result-Result

L.R.@5	All-Content@5
F1	F2
F5	F7
F21	F13
	F23
	F26
	F30

FIGURE 4.15: All-Content Vs Literature Review @5

outperformed All-Content Parameter. If we take an example of an Abstract section and compare its result with All-Content parameter, then abstract section scored better than the All-Content parameter. As we know in abstract section, author has very limited space (i.e. 10 to 12 sentences) where he picks some points from introduction section, includes some methodology steps, discusses the results of proposed approach etc,. It means that there are more chances that an author uses the domain specific terms and knowledge in this section if cited paper is closely relevant to cited by papers. Similarly, in case of Methodology and Result sections, author uses domain specific words if both papers belong to similar domain or topic.

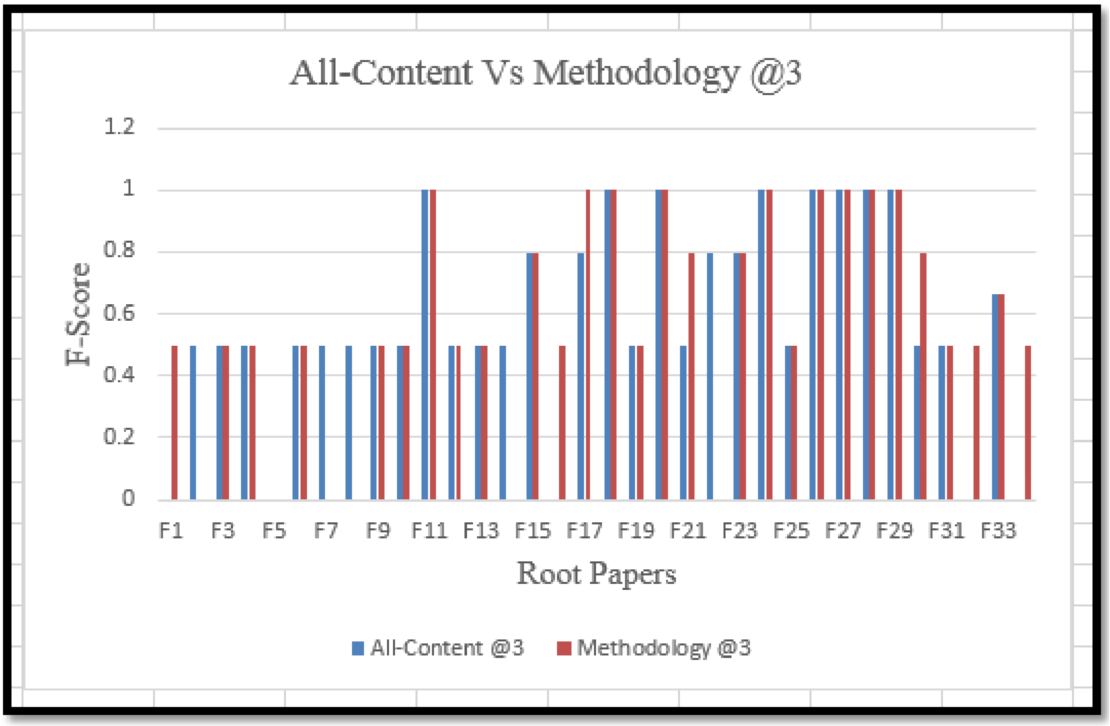


FIGURE 4.16: All-Content Vs Methodology @3

Methodology@3	All-Content@3
F1	F2
F16	F7
F17	F8
F21	F14
F30	F22
F32	
F34	

FIGURE 4.17: All-Content Vs Methodology @3

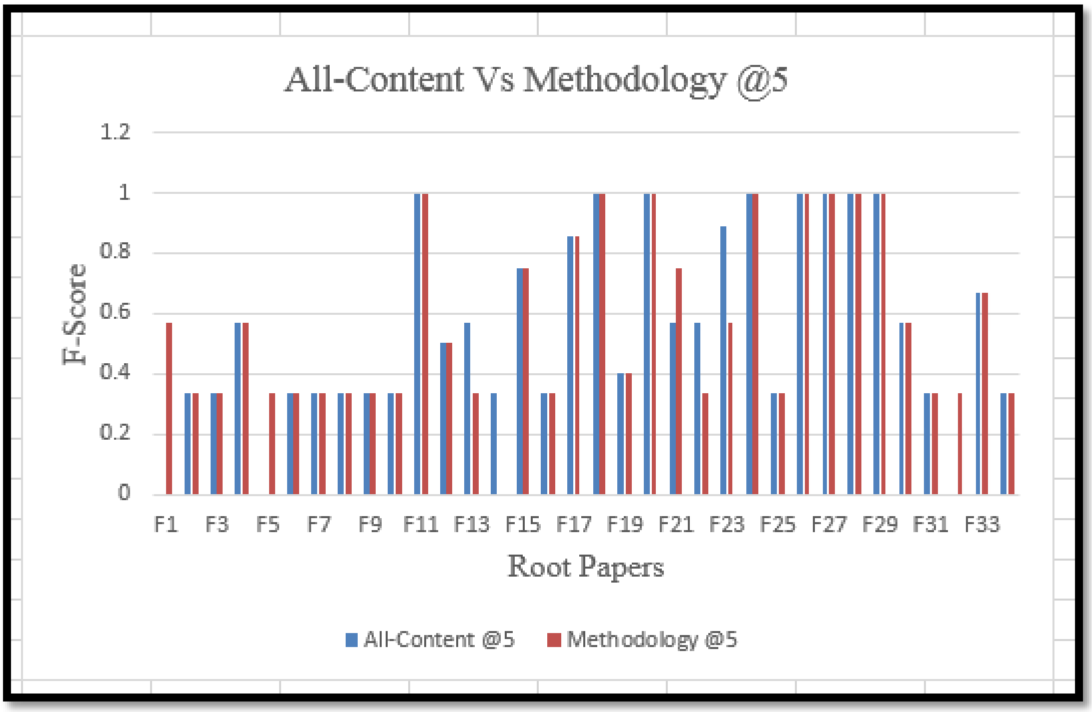


FIGURE 4.18: All-Content Vs Methodology @5

Methodology@5	All-Content@5
F1	F13
F5	F14
F21	F22
F32	F23

FIGURE 4.19: All-Content Vs Methodology @5

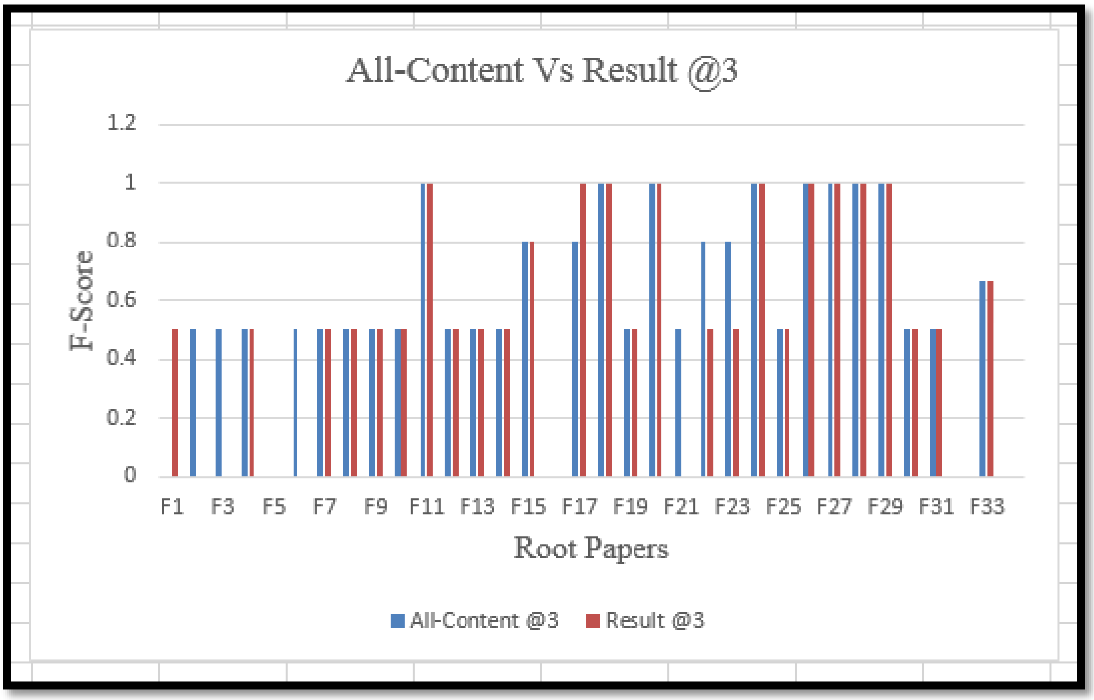


FIGURE 4.20: All-Content Vs Result @3

Result@3	All-Content@3
F1	F2
F17	F3
	F6
	F21

FIGURE 4.21: All-Content Vs Result @3

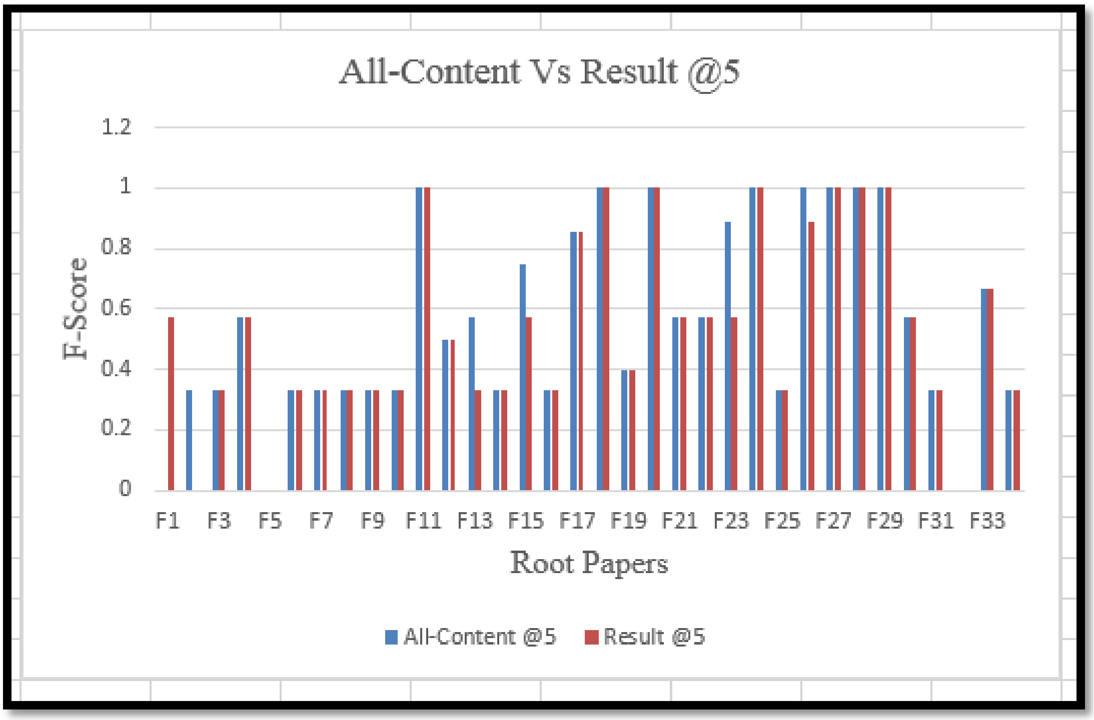


FIGURE 4.22: All-Content Vs Result @5

Result@5	All-Content@5
F1	F2
	F13
	F15
	F26

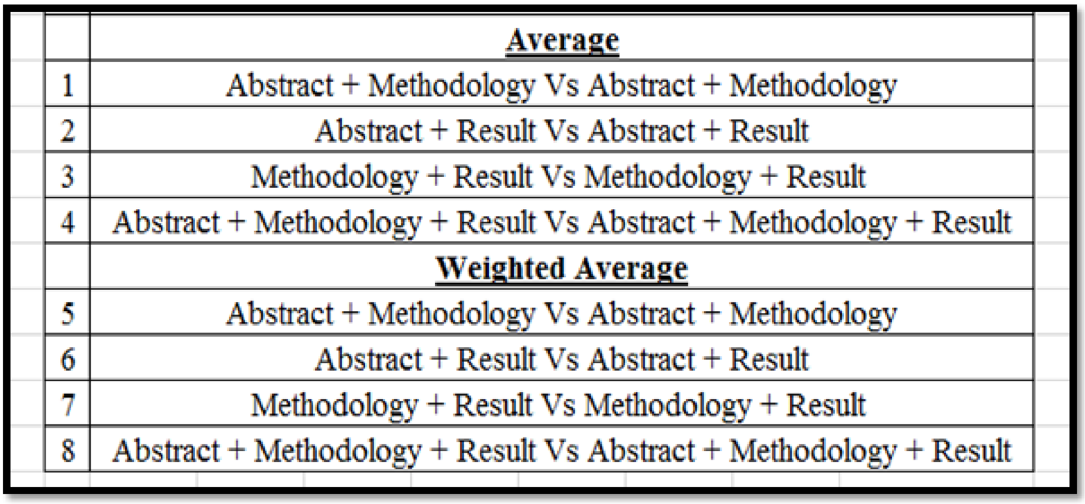
FIGURE 4.23: All-Content Vs Result @5

Single Parameters Vs All-Content			
Single Parameter	**R. Papers**	**Score**	**All-Content**
Abstract-Abstract	Top 3 Ranked	0.7	0.63
Abstract-Abstract	Top 5 Ranked	0.69	0.65
Introduction-Introduction	Top 3 Ranked	0.57	0.63
Introduction-Introduction	Top 5 Ranked	0.59	0.65
LR-LR	Top 3 Ranked	0.59	0.63
LR-LR	Top 5 Ranked	0.62	0.65
Methodology-Methodology	Top 3 Ranked	0.68	0.63
Methodology-Methodology	Top 5 Ranked	0.66	0.65
Result-Result	Top 3 Ranked	0.64	0.63
Result-Result	Top 5 Ranked	0.63	0.65

FIGURE 4.24: Single Parameters Conclusion

However, other two single parameters like Introduction-Introduction and Literature Review-Literature Review produced low F-measure scores than the All-Content parameter during the experimentation. In these sections, authors have more space where they make comparisons with existing approaches or sometime criticizing others work and these my be the reasons for producing lower results than All-Content parameter. Due to this reason, these two parameters (i.e. Introduction-Introduction and Literature Review-Literature Review) were neglected for further experimentation.

Therefore, we have used only three parameters (i.e. Abstract, Methodology and Result) with their all possible combinations for further experimentations. These all possible combinations are depicted in Figure 4.25. Each parameter is combined with other parameters with the help of Average and Weighted Average techniques to form double and triple parameters.



	Average
1	Abstract + Methodology Vs Abstract + Methodology
2	Abstract + Result Vs Abstract + Result
3	Methodology + Result Vs Methodology + Result
4	Abstract + Methodology + Result Vs Abstract + Methodology + Result
	Weighted Average
5	Abstract + Methodology Vs Abstract + Methodology
6	Abstract + Result Vs Abstract + Result
7	Methodology + Result Vs Methodology + Result
8	Abstract + Methodology + Result Vs Abstract + Methodology + Result

FIGURE 4.25: All Combinations of Abstract, Methodology and Result

4.7.8 Double Parameters Combination by Average

4.7.8.1 Avg.(Abstract, Methodology) Vs Avg.(Abstract, Methodology) Parameters

In this parameter combination, the content of an abstract and methodology sections of a cited paper is matched with the content of an abstract and methodology sections of all its citing papers. The comparison is made by measuring the cosine similarity. The abstract and methodology sections are combined with an average technique by using the formula depicted in equation 3.7 mentioned in section 3.7.1. The achieved cosine similarity score of each (i.e., cited paper-citing paper) pair is sorted in descending order. After that, the Precision, Recall and F-Measure score of top 3 and top 5 ranked papers is calculated. The overall F-Measure score of the Avg.(A,M)-Avg.(A,M) parameter is then compared to the F-Measure score of the All-Content parameter (i.e., whole content) of paper.

The top 3 ranked Avg.(A,M)-Avg.(A,M) parameter papers scored 0.74 against the top 3 ranked All-Content parameter papers with a score of 0.63. The top 5 ranked Avg.(A,M)-Avg.(A,M) parameter papers achieved an F-Measure score of 0.70 against the top 5 ranked All-Content parameter papers with a score of 0.65.

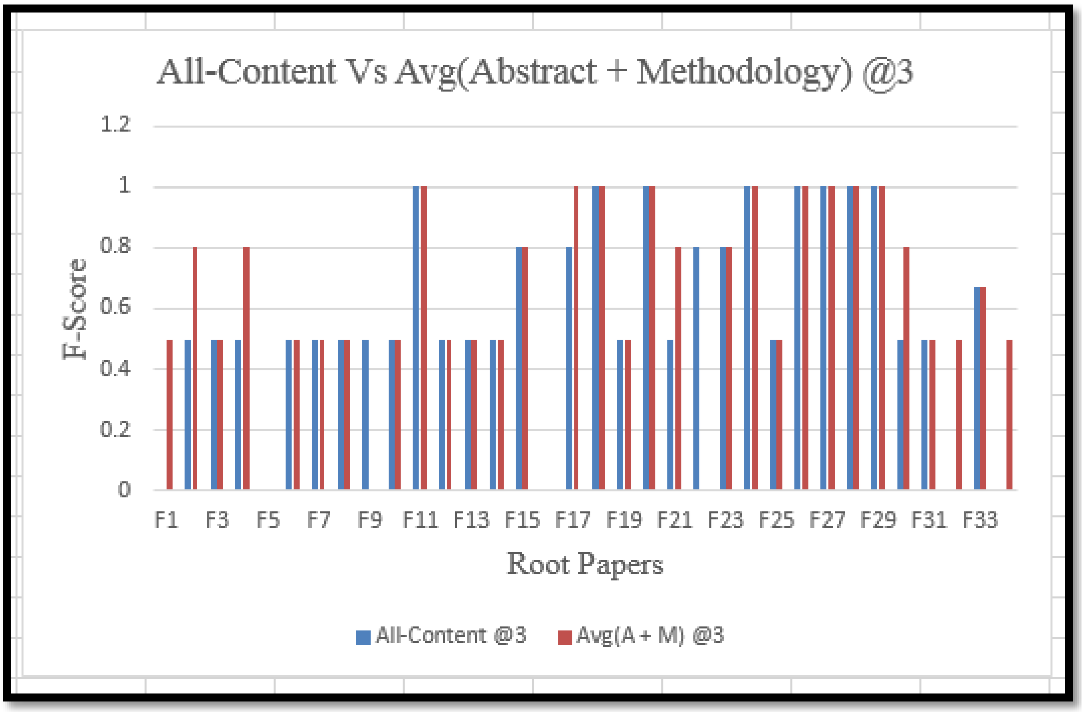


FIGURE 4.26: All-Content Vs Avg.(A,M) @3

The F-measure scores of the top 3 ranked papers show that the "Avg.(Abstract, Methodology) parameter outperformed the All-Content parameter which is depicted in Figure 4.26 and Figure 4.27. While the F-measure scores of the top 5 ranked papers show that the "Avg.(Abstract, Methodology) parameter also outperformed the All-Content parameter which is depicted in Figure 4.28 and Figure 4.29.

4.7.8.2 Avg.(Abstract, Result) Vs Avg.(Abstract, Result) Parameters

In this parameter combination, the content of an abstract and result sections of a cited paper is matched with the content of an abstract and result sections of all its citing papers. The comparison is made by measuring the cosine similarity. The abstract and results sections are combined with an average technique by using the formula depicted in equation (7) mentioned in section 3.7.1. The achieved cosine similarity score of each (i.e., cited paper-citing paper) pair is sorted in descending order. After that, the Precision, Recall and F-Measure score of top 3 and top

Avg.(A, M)@3	All-Content@3
F1	F9
F2	F22
F4	
F17	
F21	
F30	
F32	
F34	

FIGURE 4.27: All-Content Vs Avg.(A,M) @3

5 ranked papers is calculated. The overall F-Measure score of the Avg.(A,R)-Avg.(A,R) parameter is then compared to the F-Measure score of the All- Content parameter (i.e., whole content) of paper.

The top 3 ranked Avg.(A,R)-Avg.(A,R) parameter papers scored 0.63 against the top 3 ranked All-Content parameter papers with a score of 0.63. The top 5 ranked Avg.(A,R)-Avg.(A,R) parameter papers achieved F-Measure score of 0.67 against the top 5 ranked All- Content parameter papers with a score of 0.65. The F-measure scores of the top 3 ranked papers show that "Avg.(Abstract, Result) parameters generate equal results against the All-Content parameter shown in the Figure 4.30 and 4.31. While the top 5 ranked papers of the "Avg.(Abstract, Result) parameters outperformed the All-Content parameter shown in Figure 4.32 and Figure 4.33.

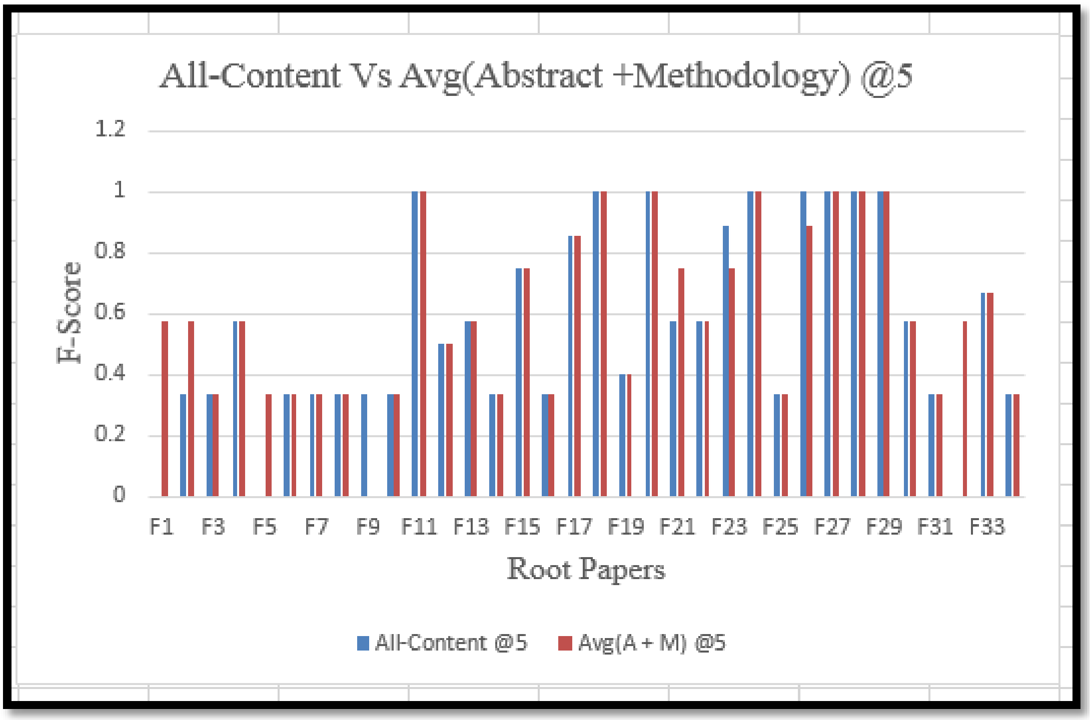


FIGURE 4.28: All-Content Vs Avg.(A,M) @5

Avg.(A, M)@5	All-Content@5
F1	F9
F2	F23
F5	F26
F21	
F32	

FIGURE 4.29: All-Content Vs Avg.(A,M) @5

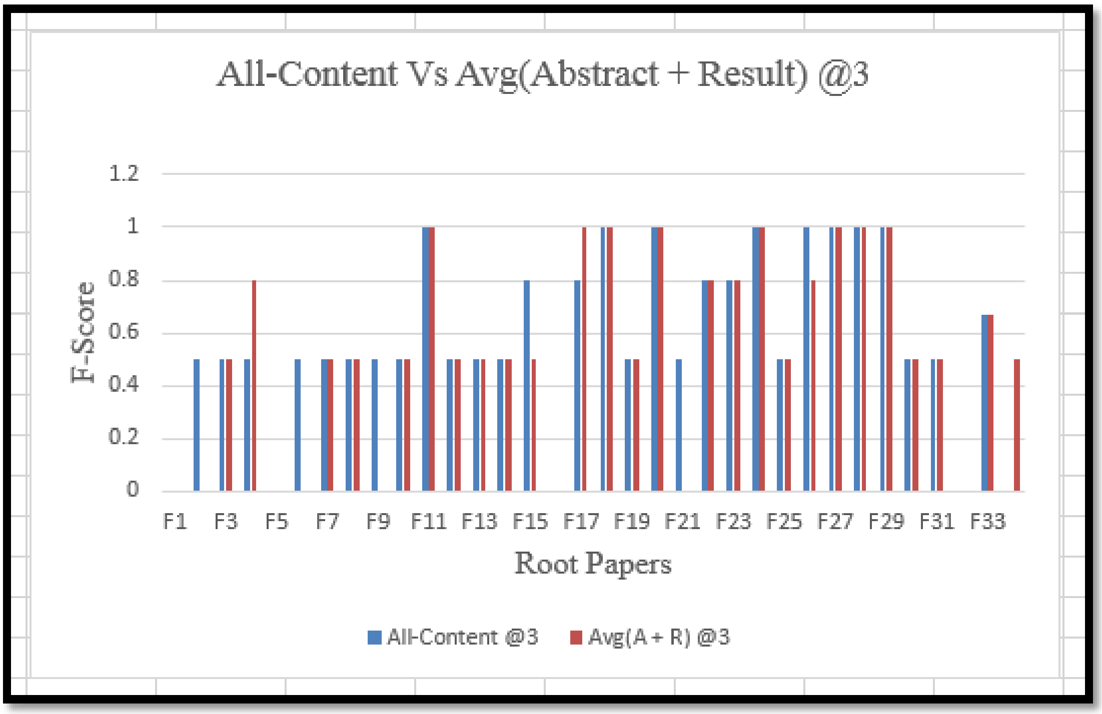


FIGURE 4.30: All-Content Vs Avg.(A,R) @3

Avg.(A, R)@3	All-Content@3
F4	F2
F17	F6
F34	F9
	F15
	F21
	F26

FIGURE 4.31: All-Content Vs Avg.(A,R) @3

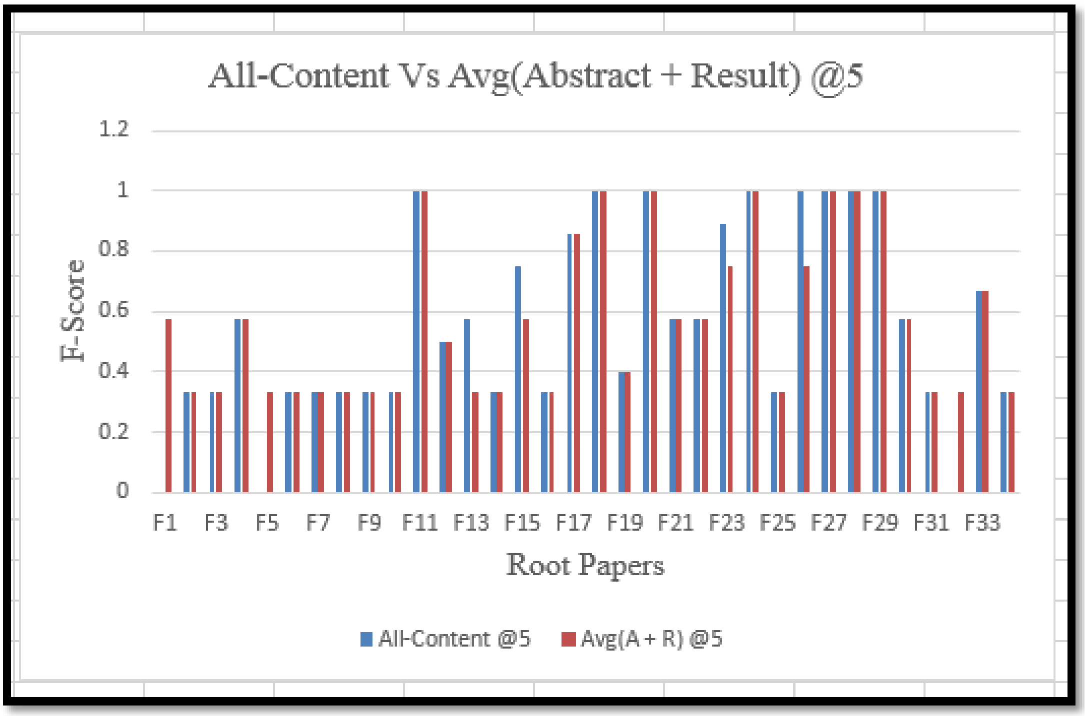


FIGURE 4.32: All-Content Vs Avg.(A,R) @5

Avg.(A, R)@5	All-Content@5
F1	F13
F5	F15
F32	F23
	F26

FIGURE 4.33: All-Content Vs Avg.(A,R) @5

4.7.8.3 Avg.(Methodology, Result) Vs Avg.(Methodology, Result) Parameters

In this parameter combination, the content of methodology and result sections of a cited paper is matched with the content of methodology and result sections of all its citing papers. The comparison is made by measuring the cosine similarity. The methodology and results sections are combined with an average technique by using the formula depicted in equation (7) mentioned in section 3.7.1. The achieved cosine similarity score of each (i.e., cited paper-citing paper) pair is sorted in descending order. After that, the Precision, Recall and F-Measure score of top 3 and top 5 ranked papers is calculated. The overall F-Measure score of the Avg.(M,R)-Avg.(M,R) parameter is then compared to the F-Measure score of the All-Content parameter (i.e., whole content) of paper.

The top 3 ranked Avg.(M,R)-Avg.(M,R) parameter papers scored 0.67 against the top 3 ranked All-Content parameter papers with a score of 0.63. The top 5 ranked Avg.(M,R)-Avg.(M,R) parameter papers achieved F-Measure score of 0.67 against the top 5 ranked All- Content parameter papers with a score of 0.65. The F-measure score of the top 3 ranked papers show that Avg.(Methodology, Result) parameters outperformed the All- Content parameter shown in Figure 4.34 and Figure 4.35. While the top 5 ranked papers scored equally against the All- Content parameter shown in Figure 4.36 and Figure 4.37.

4.7.9 Double Parameters Combination by Weighted Average

4.7.9.1 W.Avg.(Abstract, Methodology) Vs W.Avg.(Abstract, Methodology) Parameters:

In this parameter combination, the content of an abstract and methodology sections of a cited paper is matched with the content of an abstract and methodology sections of all its citing papers. The comparison is made by measuring the cosine

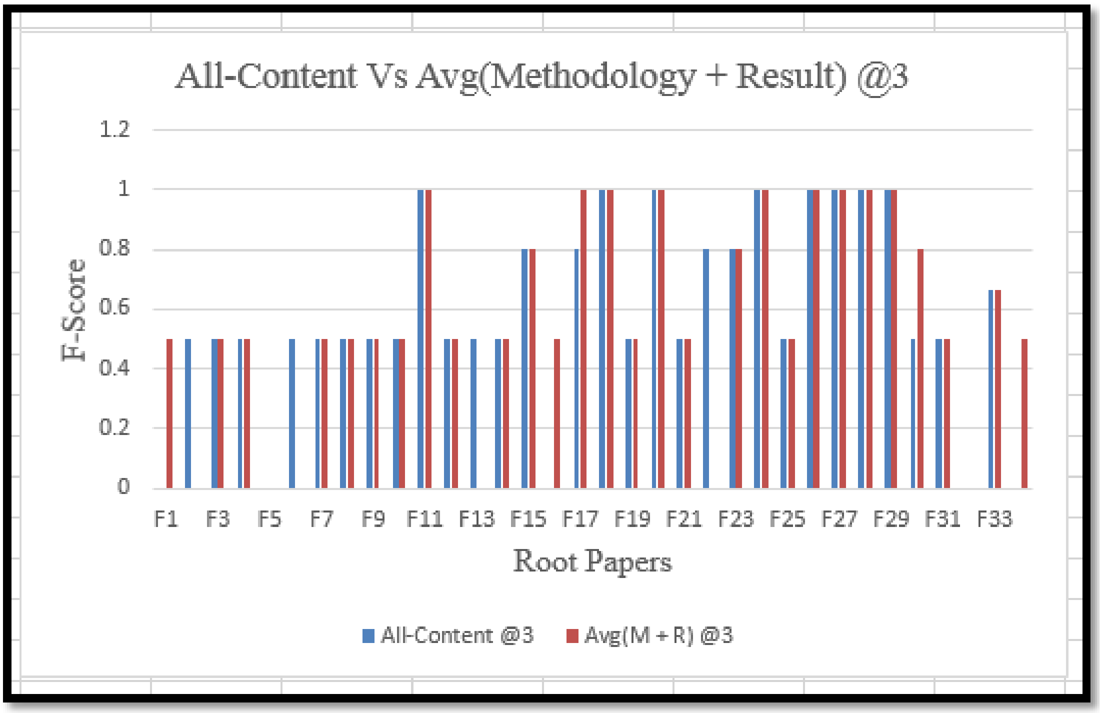


FIGURE 4.34: All-Content Vs Avg.(M,R) @3

Avg.(M, R)@3	All-Content@3
F1	F2
F16	F6
F17	F13
F30	F22
F34	

FIGURE 4.35: All-Content Vs Avg.(M,R) @3

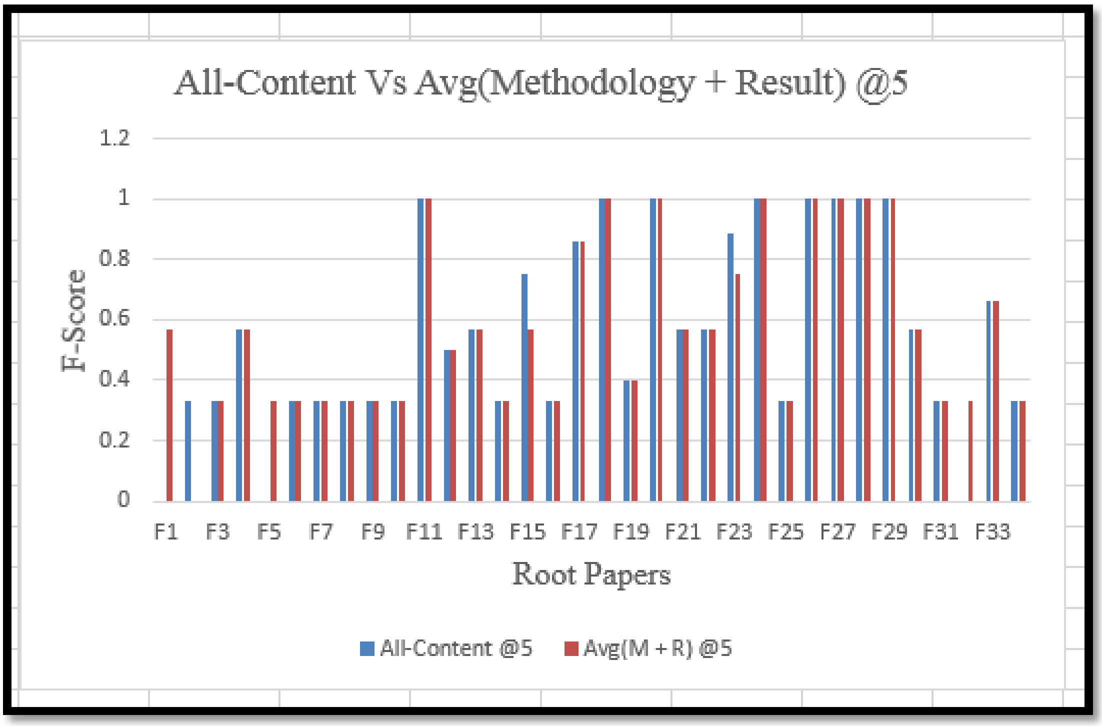


FIGURE 4.36: All-Content Vs Avg.(M,R) @5

Avg.(M, R)@5	All-Content@5
F1	F2
F5	F15
F32	F23

FIGURE 4.37: All-Content Vs Avg.(M,R) @5

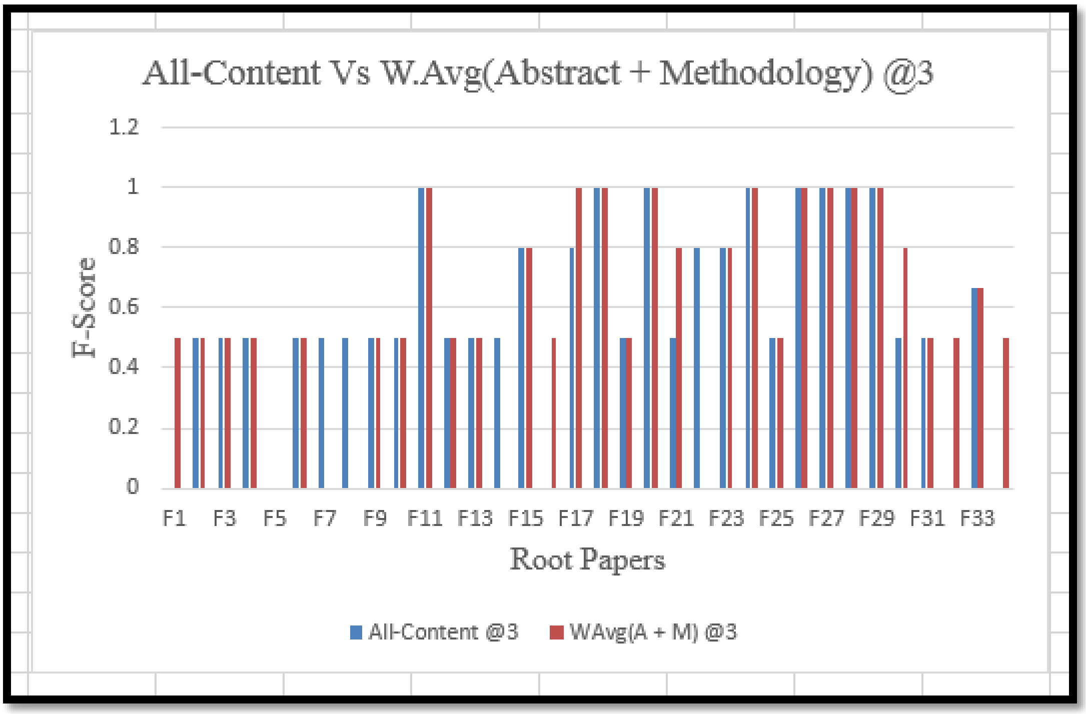


FIGURE 4.38: All-Content Vs W.Avg.(A,M) @3

similarity. The abstract and methodology sections are combined with the weighted average technique by using the formula depicted in equation (8) mentioned in section 3.7.2. The achieved cosine similarity score of each (i.e., cited paper-citing paper) pair is sorted in descending order. After that, the Precision, Recall and F-Measure score of top 3 and top 5 ranked papers is calculated. The overall F-Measure score of the W.Avg(A,M)-W.Avg(A,M) parameter is then compared to the F-Measure score of the All-Content parameter (i.e., whole content) of paper. The top 3 ranked W.Avg(A,M)-W.Avg(A,M) parameter papers scored 0.72 against the top 3 ranked All-Content parameter papers with a score of 0.63. The top 5 ranked W.Avg(A,M)-W.Avg(A,M) parameter papers achieved an F-Measure score of 0.68 against the top 5 ranked All-Content parameter papers with a score of 0.65. The F-measure scores of the top 3 ranked papers show that the W.Avg.(Abstract, Methodology) parameter outperformed the All-Content parameter which is demonstrated in Figure 4.38 and Figure 4.39. While the top 5 ranked papers of W.Avg.(Abstract, Methodology) parameter also outperformed the All-Content parameter shown in Figure 4.40 and Figure 4.41.

W.Avg.(A, M)@3	All-Content@3
F1	F7
F16	F8
F17	F14
F21	F22
F30	
F32	
F34	

FIGURE 4.39: All-Content Vs W.Avg.(A,M) @3

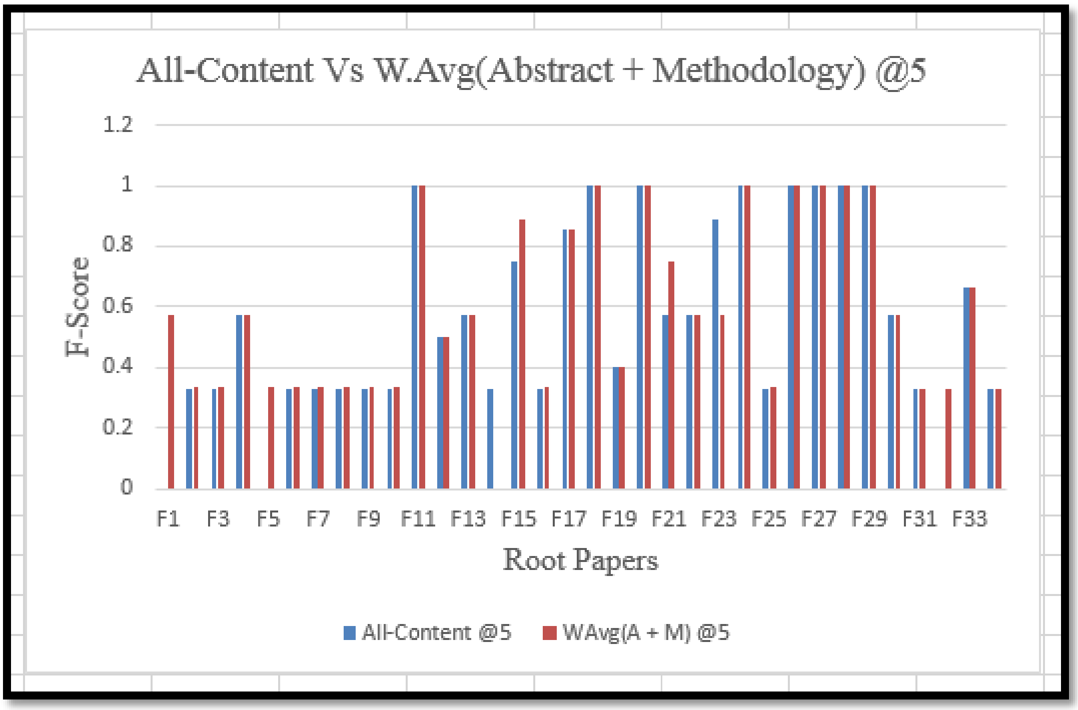


FIGURE 4.40: All-Content Vs W.Avg.(A,M) @5

W.Avg.(A, M)@5	All-Content@5
F1	F14
F5	F23
F15	
F21	
F32	

FIGURE 4.41: All-Content Vs W.Avg.(A,M) @5

4.7.9.2 W.Avg.(Abstract, Result) Vs W.Avg.(Abstract, Result) Parameters:

In this parameter combination, the content of an abstract and result sections of a cited paper is matched with the content of an abstract and result sections of all its citing papers. The comparison is made by measuring the cosine similarity. The abstract and results sections are combined with the weighted average technique by using the formula depicted in equation (8) mentioned in section 3.7.2. The achieved cosine similarity score of each (i.e., cited paper-citing paper) pair is sorted in descending order. After that, the Precision, Recall and F-Measure score of top 3 and top 5 ranked papers is calculated. The overall F-Measure score of the W.Avg.(A,R)-W.Avg.(A,R) parameter is then compared to the F-Measure score of the All-Content parameter (i.e., whole content) of paper.

The top 3 ranked W.Avg.(A,R)-W.Avg.(A,R) parameter papers scored 0.67 against the top 3 ranked All-Content parameter papers with a score of 0.63. The top 5 ranked W.Avg.(A,R)-W.Avg.(A,R) parameter papers achieved F-Measure score of 0.64 against the top 5 ranked All-Content parameter papers with a score of 0.65.

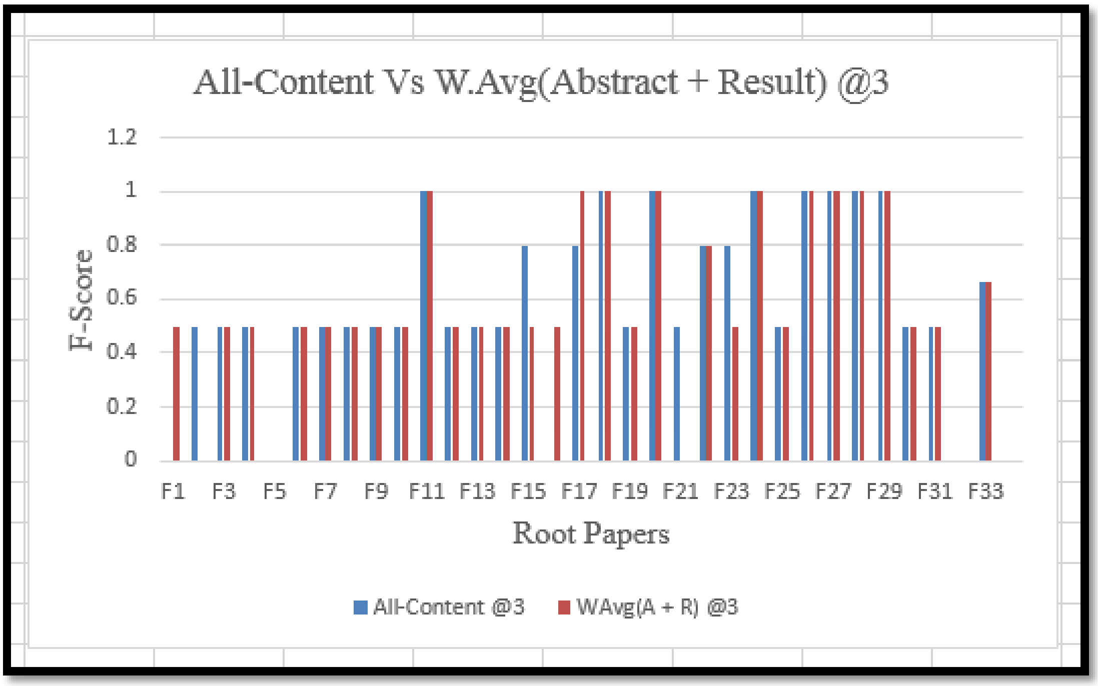


FIGURE 4.42: All-Content Vs W.Avg.(A,R) @3

W.Avg.(A, R)@3	All-Content@3
F1	F2
F16	F15
F17	F21
	F23

FIGURE 4.43: All-Content Vs W.Avg.(A,R) @3

The F-measure scores of the top 3 papers show that W.Avg.(Abstract, Result) parameters outperformed the All-Content parameter demonstrated in Figure 4.42 and Figure 4.43. While the top 5 ranked papers of the W.Avg.(Abstract, Result) parameters generate approximately equal results with the All-Content parameter shown in Figure 4.44 and Figure 4.45.

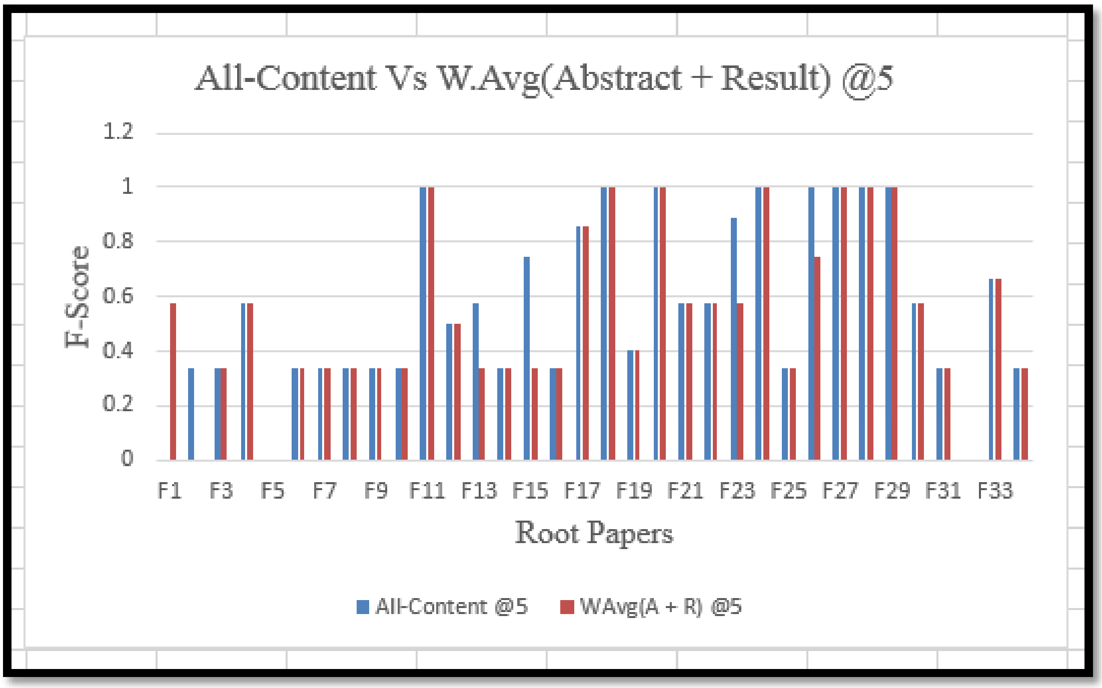


FIGURE 4.44: All-Content Vs W.Avg.(A,R) @5

W.Avg.(A, R)@5	All-Content@5
F1	F2
	F13
	F15
	F23
	F26

FIGURE 4.45: All-Content Vs W.Avg.(A,R) @5

4.7.9.3 W.Avg.(Methodology, Result) Vs W.Avg.(Methodology, Result) Parameters:

In this parameter combination, the content of methodology and result sections of a cited paper is matched with the content of methodology and result sections of all its citing papers. The comparison is made by measuring the cosine similarity. The methodology and results sections are combined with the weighted average technique by using the formula depicted in equation (8) mentioned in section 3.7.2. The achieved cosine similarity score of each (i.e., cited paper-citing paper) pair is sorted in descending order. After that, the Precision, Recall and F-Measure score of top 3 and top 5 ranked papers is calculated. The overall F-Measure score of the W.Avg.(M,R)-W.Avg(M,R) parameter is then compared to the F-Measure score of the All-Content parameter (i.e., whole content) of paper.
The top 3 ranked W.Avg.(M,R)-W.Avg.(M,R) parameter papers scored 0.68 against the top 3 ranked All-Content parameter papers with a score of 0.63. The top 5 ranked W.Avg.(M,R)-W.Avg.(M,R) parameter papers achieved F-Measure score of 0.65 against the top 5 ranked All-Content parameter papers with a score of 0.65. The F-measure score of the top 3 ranked papers shows that the W.Avg.(Methodology, Result) parameter outperformed the All-Content parameter which is demonstrated in Figure 4.46 and Figure 4.47. While the F-measure score of the top 5 ranked papers shows that the W.Avg.(Methodology, Result) parameter scored equally with the All- Content parameter shown in Figure 4.48 and Figure 4.49.

4.7.10 Double Parameters Conclusion

The figure 4.50 concludes the results of double parameters combination with Average technique in comparison of All-Content parameter. The results are deduced from the top 3 ranked and top 5 ranked research papers. While the figure 4.51 concludes the results of double parameters combination with Weighted Average technique in comparison of All-Content parameter. We can clearly see that the double parameter like (Abstract, Methodology) scored better than all other double

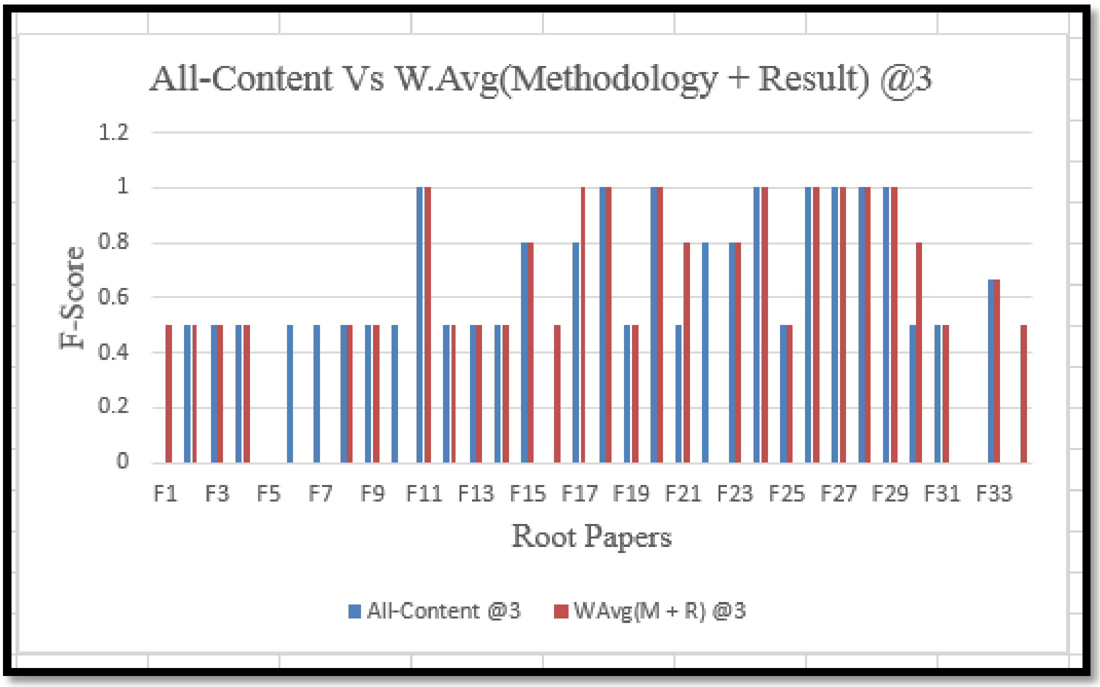


FIGURE 4.46: All-Content Vs W.Avg.(M,R) @3

W.Avg.(M, R)@3	All-Content@3
F1	F6
F16	F7
F17	F10
F21	F22
F30	
F34	

FIGURE 4.47: All-Content Vs W.Avg.(M,R) @3

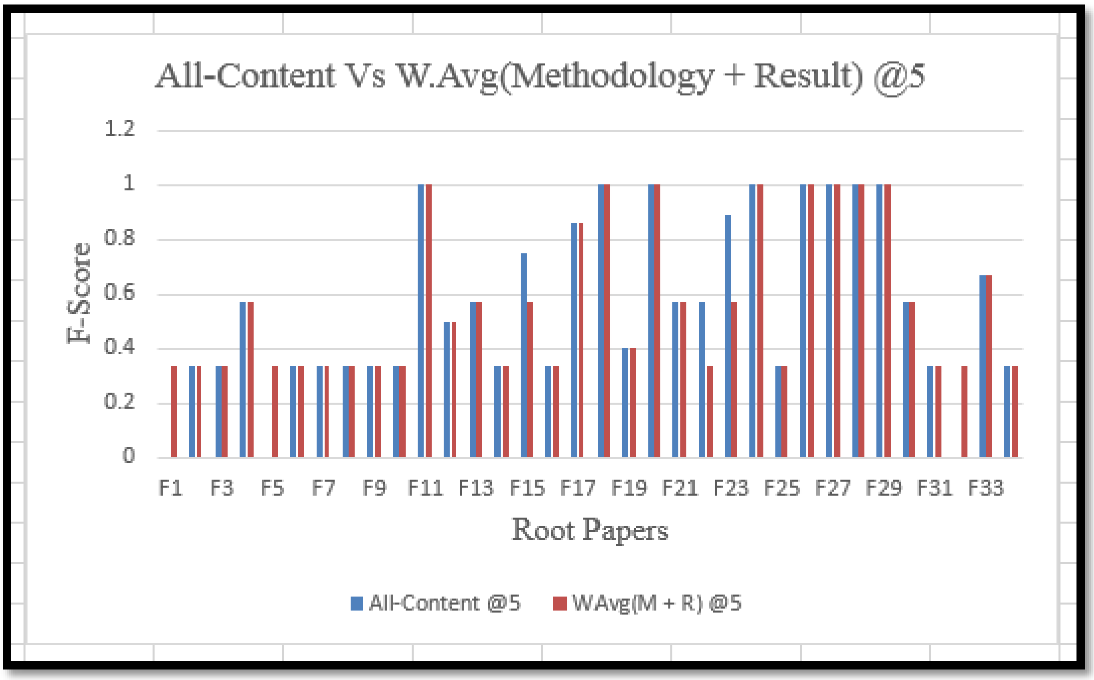


FIGURE 4.48: All-Content Vs W.Avg.(M,R) @5

W.Avg.(M, R)@5	All-Content@5
F1	F15
F5	F22
F32	F23

FIGURE 4.49: All-Content Vs W.Avg.(M,R) @5

Double Parameters Vs All-Content			
Double Parameters	**R. Papers**	**Score**	**All-Content**
Avg.(Abstract, Methodology)	Top 3 Ranked	0.74	0.63
Avg.(Abstract, Methodology)	Top 5 Ranked	0.7	0.65
Avg.(Abstract, Result)	Top 3 Ranked	0.63	0.63
Avg.(Abstract, Result)	Top 5 Ranked	0.67	0.65
Avg.(Methodology, Result)	Top 3 Ranked	0.67	0.63
Avg.(Methodology, Result)	Top 5 Ranked	0.67	0.65

FIGURE 4.50: Double Parameters Conclusion by Average Technique

parameters when combined with 1)-Average technique and 2)-Weighted Average technique. This parameter scored F-measure of 0.74 and 0.70 of top 3 and top 5 ranked research articles respectively when combined with average technique. While when it combined with Weighted Average technique then scored F-measure score of 0.72 and 0.68 of top 3 and top 5 ranked research articles respectively.

4.7.11 Triple Parameters Combining by Average

4.7.11.1 Avg.(Abstract, Methodology, Result) Vs Avg.(Abstract, Methodology, Result) Parameters:

In this parameter combination, the content of an abstract, methodology and result sections of a cited paper is matched with the content of an abstract, methodology and result sections of all its citing papers. The comparison is made by measuring the cosine similarity. The abstract, methodology and results sections are combined with an average technique by using the formula depicted in equation (7) mentioned in section 3.7.1. The achieved cosine similarity score of each (i.e., cited

Double Parameters Vs All-Content			
Double Parameters	**R. Papers**	**Score**	**All-Content**
W.Avg.(Abstract, Methodology)	Top 3 Ranked	0.72	0.63
W.Avg.(Abstract, Methodology)	Top 5 Ranked	0.68	0.65
W.Avg.(Abstract, Result)	Top 3 Ranked	0.67	0.63
W.Avg.(Abstract, Result)	Top 5 Ranked	0.64	0.65
W.Avg.(Methodology, Result)	Top 3 Ranked	0.68	0.63
W.Avg.(Methodology, Result)	Top 5 Ranked	0.65	0.65

FIGURE 4.51: Double Parameters Conclusion by Weighted Average Technique

paper-citing paper) pair is sorted in descending order. After that, the Precision, Recall and F-Measure score of top 3 and top 5 ranked papers is calculated. The overall F-Measure score of the Avg.(A,M,R)-Avg.(A,M,R) parameter is then compared to the F-Measure score of the All-Content parameter (i.e., whole content) of paper.

The top 3 ranked Avg.(A,M,R)-Avg.(A,M,R) parameter papers scored 0.75 against the top 3 ranked All-Content parameter papers with a score of 0.63. The top 5 ranked Avg.(A,M,R)-Avg.(A,M,R) parameter papers achieved F-Measure score of 0.71 against the top 5 ranked All-Content parameter papers with a score of 0.65. The F-measure score of the top 3 ranked papers shows that Avg.(Abstract, Methodology, Results) parameters outperformed the All-Content parameter demonstrated in Figure 4.52 and Figure 4.53. While the F-measure score of the top 5 ranked papers shows that Avg.(Abstract, Methodology, Results) parameters also outperformed the All-Content parameter demonstrated in Figure 4.54 and Figure 4.55.

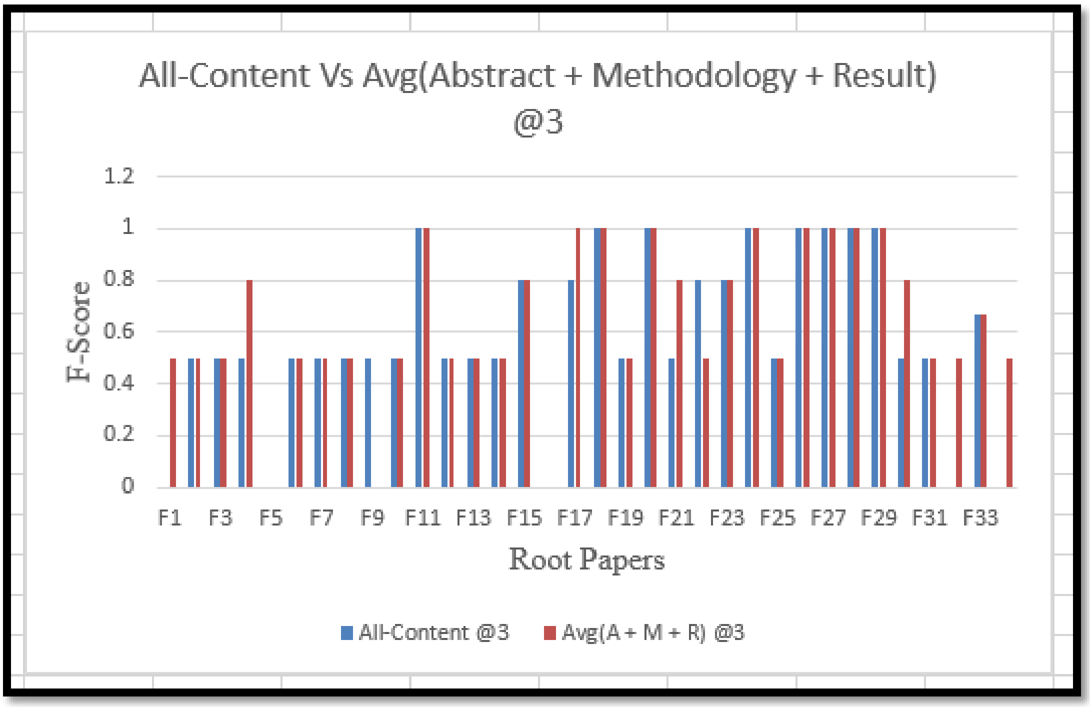


FIGURE 4.52: All-Content Vs Avg.(A,M,R) @3

Avg.(A, M, R)@3	All-Content@3
F1	F9
F4	F22
F17	
F21	
F30	
F32	
F34	

FIGURE 4.53: All-Content Vs Avg.(A,M,R) @3

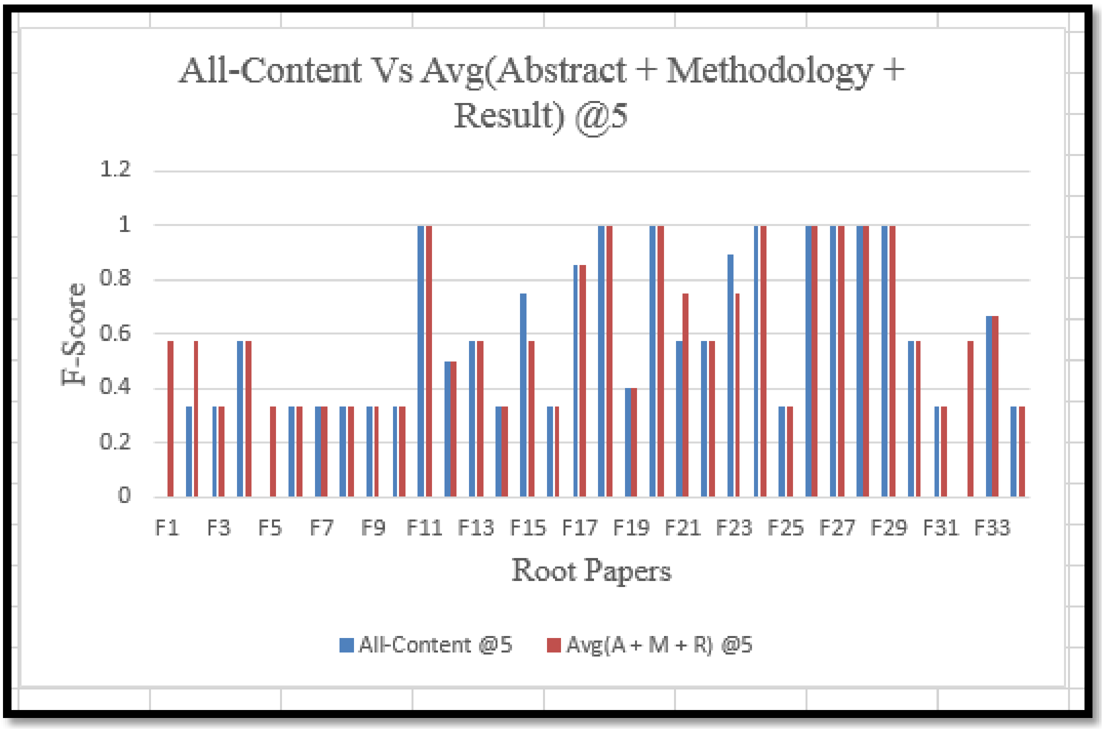


FIGURE 4.54: All-Content Vs Avg.(A,M,R) @5

Avg.(A, M, R)@5	All-Content@5
F1	F15
F2	F23
F5	
F21	
F32	

FIGURE 4.55: All-Content Vs Avg.(A,M,R) @5

4.7.12 Triple Parameters Combining by Weighted Average

4.7.12.1 W.Avg.(Abstract, Methodology, Result) Vs W.Avg.(Abstract, Methodology, Result) Parameters:

In this parameter combination, the content of an abstract, methodology and result sections of a cited paper is matched with the content of an abstract, methodology and result sections of all its citing papers. The comparison is made by measuring the cosine similarity. The abstract, methodology and result sections are combined with the weighted average technique by using the formula depicted in equation (8) mentioned in section 3.7.2. The achieved cosine similarity score of each (i.e., cited paper-citing paper) pair is sorted in descending order. After that, the Precision, Recall and F-Measure score of top 3 and top 5 ranked papers is calculated. The overall F-Measure score of the W.Avg.(A,M,R)-W.Avg.(A,M,R) parameter is then compared to the F-Measure score of the All-Content parameter (i.e., whole content) of paper.

The top 3 ranked W.Avg.(A,M,R)-W.Avg.(A,M,R) parameter papers scored 0.68 against the top 3 ranked All-Content parameter papers with a score of 0.63. The top 5 ranked W.Avg.(A,M,R)-W.Avg.(A,M,R) parameter papers achieved F-Measure score of 0.67 against the top 5 ranked All-Content parameter papers with a score of 0.65. The F-measure score of the top 3 ranked papers shows that W.Avg.(Abstract, Methodology, Results) parameters outperformed the All-Content parameter demonstrated in Figure 4.56 and Figure 4.57. While the F-measure score of the top 5 ranked papers shows that W.Avg.(Abstract, Methodology, Results) parameters scored equally against the All-Content parameter demonstrated in Figure 4.58 and Figure 4.59.

4.7.13 Triple Parameters Conclusion

The triple parameters are formed when combining the abstract, methodology, and result sections with the help of average and weighted average techniques. The results are deduced from the top 3 ranked and top 5 ranked research papers.

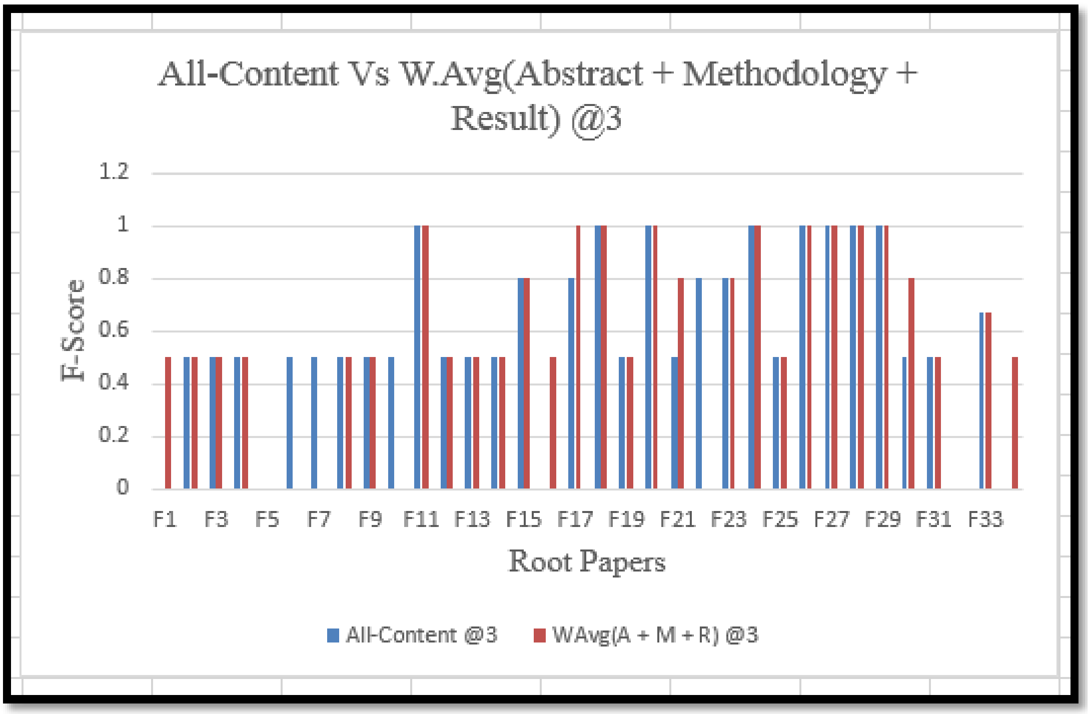


FIGURE 4.56: All-Content Vs W.Avg.(A,M,R) @3

W.Avg.(A, M, R)@3	All-Content@3
F1	F6
F16	F7
F17	F10
F21	F22
F30	
F34	

FIGURE 4.57: All-Content Vs W.Avg.(A,M,R) @3

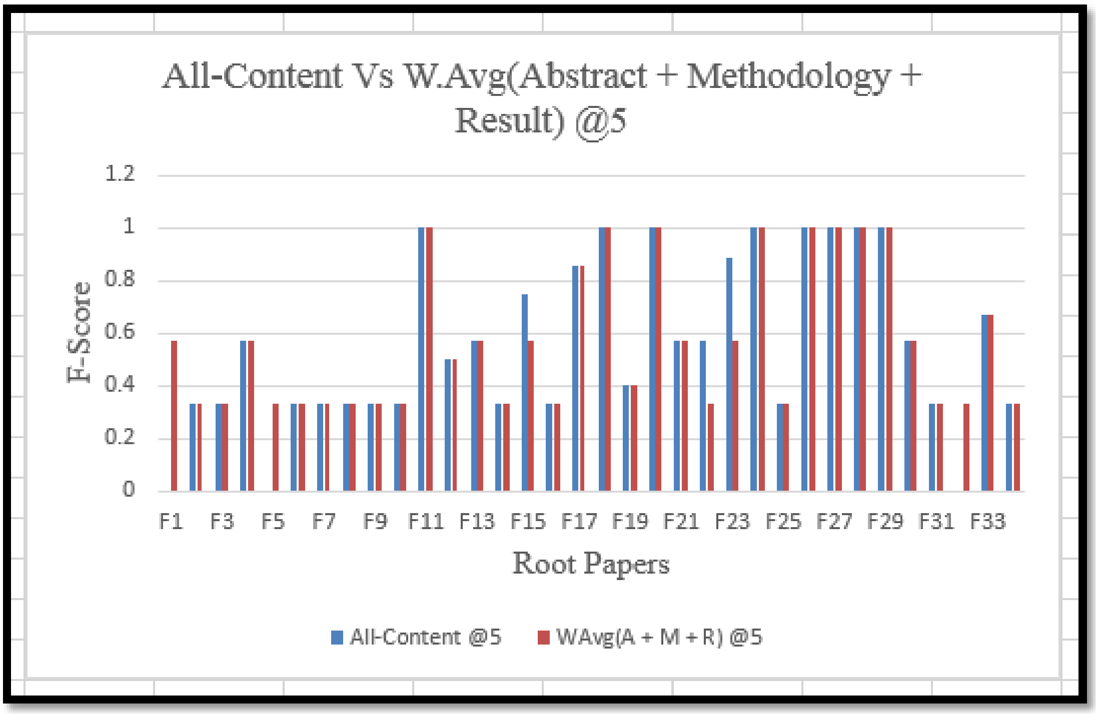


FIGURE 4.58: All-Content Vs W.Avg.(A,M,R) @5

W.Avg.(A, M, R)@5	All-Content@5
F1	F15
F5	F22
F32	F23

FIGURE 4.59: All-Content Vs W.Avg.(A,M,R) @5

The triple parameter Avg. (Abstract, Methodology, Result) scored F-measure of 0.75 and 0.71 of the top 3 and top 5 ranked research articles respectively. This triple parameter is the best scorer parameter from all the single, double and triple parameters. While the other triple parameter like W.Avg. (Abstract, Methodology, Result) scored an F-measure score of 0.68 and 0.67 of the top 3 and top 5 ranked research articles respectively.

4.8 Comparisons

The Valenzuela [1] and Qayyum & Afzal [18] have proposed approaches to identify important citations. Valenzuela [1] used 12 different features for classifications from which mostly depend on the content of research articles that is not openly accessible by major journals like ACM, Elsevier, IEEE, Springer, etc. While Qayyum & Afzal [18] performed a binary citation classification having 05 features based on metadata of research articles for important citations extraction. However, metadata is not domain-specific. Therefore, there is a need to propose an alternative mechanism for the identification of important citations by using domain-specific terms and knowledge. This mechanism can be constructed by using the content of research articles. Our proposed approach depends on the different logical sections of research papers. These logical sections include Abstract, Introduction, Literature Review, Methodology, Result, and all possible combinations. We have compared our proposed approach results with two recent state-of-the-art approaches of Valenzuela [1] and Qayyum & Afzal [18] due to the following two reasons.

1)-Both of the state-of-the-art approaches of Valenzuela [1] and Qayyum & Afzal [18] published their research work in well reputable journals and conferences. Valenzuela [1] published their work "in Workshops at the twenty-ninth AAAI conference on artificial intelligence with titled "Identifying meaningful citations." in the year 2015. While Qayyum & Afzal [18] published research work in Scientometrics journal with titled "Identification of important citations by exploiting research articles metadata and cue-terms from the content." in the year 2019.

2)- We have used the same data set (i.e. Valenzuela comprehensive annotated

dataset) as used by Valenzuela [1] and Qayyum & Afzal [18] for comparison.
This thesis has considered all five logical sections with all possible combinations. The similarity score of all logical sections has been calculated by using the cosine similarity approach and different logical sections have been combined by using average and weighted average techniques. The results have extracted by using every possible combination of logical sections from single combinations parameters to triple combinations. The extracted results have been sorted in descending order and then Precision, Recall, and F-measure scores have been calculated of the top 3 and top 5 ranked papers.
The Valenzuela's [1] proposed approach achieved an average precision score of 0.65 by combining all of their 12 features. While Qayyum & Afzal [18] approach achieved 0.73 average precision by combining all the features. In our proposed approach, the single parameter scored an F-Measure score of 0.63 of the top 3 ranked papers and 0.64 of the top 5 ranked papers. By using two parameters combined with a weighted average technique system scored an F-measure score of 0.69 of the top 3 ranked papers and 0.66 of the top 5 ranked papers. When two parameters combined with average technique then the system scored an F-measure score of 0.68 of the top 3 and top 5 ranked papers. When we combined three parameters through weighted average technique then the system achieved 0.68 F-Measure score of top 3 ranked papers and F-Measure score of 0.67 of the top 5 ranked papers. Similarly, when three parameters combined through an average technique then our system achieved 0.75 F-Measure scores of the top 3 ranked papers and 0.71 F-Measure scores of top 5 ranked papers. Our proposed section based approach best feature (i.e. triple parameter) yielded a precision of 0.75 when considering the top 3 ranked papers demonstrated in Figure 4.60. While when considering the top 5 ranked papers then the best feature (i.e. triple parameter) yielded a precision of 0.73 presented in Figure 4.61.

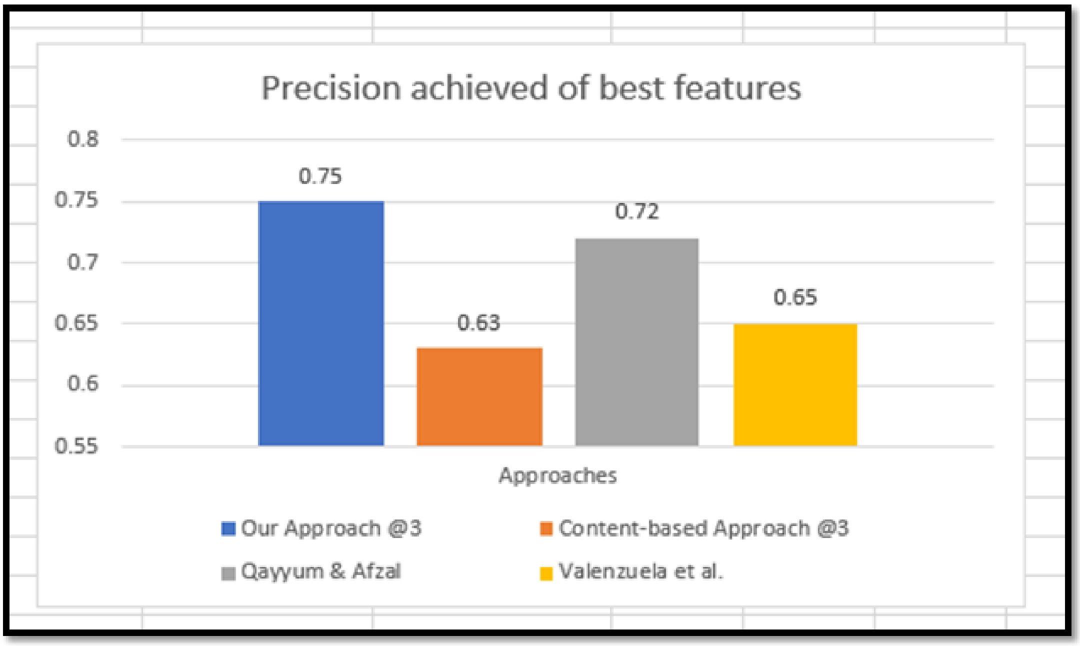


FIGURE 4.60: Comparisons of overall results

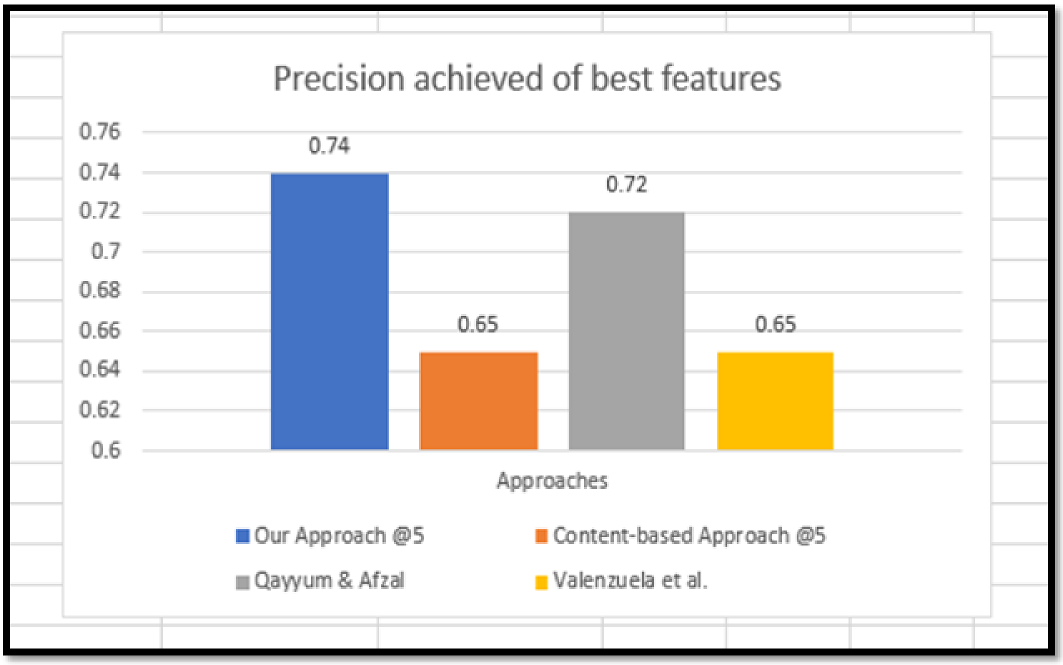


FIGURE 4.61: Comparisons of overall results

Chapter 5

Conclusion and Future Work

5.1 Conclusion

So far, researchers have utilized different citation analysis-based approaches to assist in the formulation of different scientific policies like institutions and researchers ranking [4], researchers Nobel prizes and awards [5], peer judgments [6], research funds allocation and calculating impact factor of researchers and different journals. In literature, researchers have critically analyzed the citations and argued that all citations in research articles are not of equal significance and weights. The latest approaches in citation classification community have combined and merged various citation reasons into two types. These types include (1) important and (2) non-important. In important citations, authors extend or adapt the proposed technique of the cited paper. While in non-important citations, authors just write the background information of the proposed technique of the cited paper. These two classes ensure the reliability of citation count by considering only important citations while counting. With the help of knowing important citations, new and emerging trends for research can be identified and beneficiate in finding the most relevant articles against some research topics.

We have critically reviewed more than 40 research articles in the field. Researchers have discussed and presented different approaches that depend on the content and metadata of research articles. Their proposed metadata-based and content features include title similarity, author overlap, references, in-text citation count, cue words, cue phrases, etc. These features have their own limitations like in cue-words or cue-phrases case, for every new dataset, there is a need to update the list of cue words and phrases which is a time-consuming factor. And in case of high in-text citation count of citing and cited research articles, it does not mean that cited article has important citations with the cited articles according to Ziman [2], Moravcsik [10] and Garfield [11].

All of the existing approaches have given importance to metadata, in-text citation counts, and their positions. Whereas metadata of the research paper does not hold domain-specific terms that can adversely affect accuracy. To the best of our knowledge, none of the existing approaches have compared the important terms represented in different corresponding logical sections of the research articles to find the important and non-important relation between two articles (i.e. citing and cited articles). We have proposed a comprehensive methodology to address the above-raised issue. Our methodology compares the content of corresponding logical sections of cited and citing research articles to trace whether we get better results or closer results than the existing approaches.

We have considered 5 logical sections of research articles which include Abstract, Introduction, Methodology, Literature Review, and Result sections. The cosine similarity approach has been used to calculate the similarity scores of corresponding logical sections and their all possible combinations depicted in Figure 4.25 of paper-citation pairs. The average and weighted average techniques are used to combine two or more sections of articles. The experiments are performed on the comprehensive annotated dataset of Valenzuela [1]. This dataset contains 465 research articles (i.e. paper-citation pairs) having 48 tuples of root paper and cited paper, crawled from ACL (Association of Computational and Linguistics) anthology. Pre-processing is applied to remove the stop words and conversion of words into their root words. It involved two steps, initially, we removed stop words by using the Onix Text Retrieval Toolkit Stop word List1[1]. Then the Algorithm of

[1] https://www.lextek.com/manuals/onix/stopwords1.html

Porter Stemming is applied to change the terms/words into their root word [32]. The F-measure scores of top 3 ranked and top 5 ranked of five logical sections are calculated separately and compared with the score of top 3 ranked and top 5 ranked of the All-Content (i.e. whole content) of research articles. The abstract, methodology, and result sections achieved a better F-measure score than the All-Content (i.e. whole content). Whereas the introduction and literature review section has produced a poor score. Therefore, only abstract, methodology, and result sections have been considered for further experimentations. Then we have taken all possible combinations of three considered sections as demonstrated in Figure 4.25 and calculated the F-Measure score of each combination of the top 3 and top 5 ranked articles.

Firstly, we have calculated the F-measure score of single parameters separately of Valenzuela [1] dataset. The abstract Vs abstract, methodology Vs methodology and result Vs result sections achieved better F-measure scores than the score attained by considering the All-Content parameter of research articles. While introduction Vs introduction and literature review Vs literature review sections scored lower F-measure score than the All-Content parameter of the articles. In a single parameter, abstract Vs abstract is the top scorer parameter with F-measure of 0.70 and 0.69 of the top 3 and top 5 ranked research articles respectively. In the double parameters, two logical sections combined with the help of average and weighted average techniques. The combination of Avg.(abstract, methodology) outperformed other double parameters when combined them with an average approach and scored F-measure of 0.74 and 0.70 of the top 3 and top 5 ranked research articles respectively. The W.Avg.(abstract, methodology) sections combinations have also produced better results than other double parameters when combined with the help of a weighted average technique. This combination achieved an F-measure score of 0.72 and 0.68 of the top 3 and top 5 ranked research articles respectively. In the triple parameters, three logical sections (i.e. abstract, methodology, and result (A, M, R)) combined with the help of average and weighted average techniques. This triple combination is the top scorer parameter when combined with an average technique and scored F-measure of 0.75 and 0.71 of the top 3 and top 5 ranked research articles respectively. When this triple combination took place with the help of the weighted average technique, it achieved an F-measure score of 0.68 and 0.67 of the top 3 and top 5 ranked research articles respectively.

The achieved results of the above experimentations have been compared with the results of the state-of-the-art approaches of Qayyum & Afzal [18], Valenzuela [1] and All-Content based approaches. Valenzuela [1] achieved the accuracy (i.e. precision) of 0.65 by combining all the 12 different features and Qayyum & Afzal [18] achieved the average accuracy (i.e. precision) of 0.72 by combining all the features. The All-Content based approach achieved the F-Measure score 0.63 of the top 3 ranked papers and 0.65 score of top 5 ranked papers. While the best combination of abstract, methodology, and result (A, M, R) sections of our proposed approach achieved the F-Measure score of 0.75. The concluding remarks of this research are that research papers are written with the help of domain-specific terms and knowledge. It is more probable that both papers might use similar vocabulary and terms as they belong or they are closely working on the same topic or extending once work in another work. So, there are more chances that both papers belong to a similar domain and have important citations if citing paper uses domain-specific terms and knowledge in the abstract, methodology, and results section. The logical sections present in a research paper hold diverging importance and more chances that citing and cited paper has used similar vocabulary terms in the same sections of their papers. However, it is necessary to have content in sections that might not always the case. Secondly, there must be a citation relationship between two papers. Furthermore, there is a chance that research named sections distinguishable in their research paper. In the presence of the above cases, the metadata-based approach is recommended.

5.2 Future Work

In this thesis, we have used the freely available annotated dataset that was primarily used in the approach proposed by Valenzuela [1]. There are 465 total annotated paper-citation pairs in this dataset that is a too small amount of data to reach a general conclusion. As this domain contains a very small amount of large annotated datasets, so for future work, the production of large standard annotated datasets are required. The datasets which should cover different authors dispersed geographically and different domains.

Bibliography

[1] M. Valenzuela, V. Ha, and O. Etzioni, "Identifying meaningful citations," in *Workshops at the twenty-ninth AAAI conference on artificial intelligence*, 2015.

[2] J. M. Ziman *et al.*, *Public knowledge: An essay concerning the social dimension of science.* CUP Archive, 1968, vol. 519.

[3] F. Narin *et al.*, *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity.* Computer Horizons Cherry Hill, NJ, 1976.

[4] R. C. Anderson, F. Narin, and P. McAllister, "Publication ratings versus peer ratings of universities," *Journal of the American Society for Information Science*, vol. 29, no. 2, pp. 91–103, 1978.

[5] H. Inhaber and K. Przednowek, "Quality of research and the nobel prizes," *Social Studies of Science*, vol. 6, no. 1, pp. 33–50, 1976.

[6] A. T. Smith and M. Eysenck, "The correlation between rae ratings and citation counts in psychology," University of London, Royal Holloway, Tech. Rep., 2002.

[7] J. Wilsdon, L. Allen, E. Belfiore, P. Campbell, S. Curry, S. Hill, and P. Wouters, "The metric tide: report of the independent review of the role of metrics in research assessment and management. 2015," *Publisher Full Text*, 2015.

[8] R. Benedictus, F. Miedema, and M. W. Ferguson, "Fewer numbers, better science," *Nature News*, vol. 538, no. 7626, p. 453, 2016.

[9] E. Garfield *et al.*, "Can citation indexing be automated," in *Statistical association methods for mechanized documentation, symposium proceedings*, vol. 269. National Bureau of Standards, Miscellaneous Publication 269, Washington, DC, 1965, pp. 189–192.

[10] 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.

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

[12] 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.

[13] T. A. Brooks, "Private acts and public objects: An investigation of citer motivations," *Journal of the American Society for Information Science*, vol. 36, no. 4, pp. 223–229, 1985.

[14] D. O. Case and G. M. Higgins, "How can we investigate citation behavior? a study of reasons for citing literature in communication," *Journal of the American Society for Information Science*, vol. 51, no. 7, pp. 635–645, 2000.

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

[16] 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.

[17] R. Radoulov, "Exploring automatic citation classification," Master's thesis, University of Waterloo, 2008.

[18] 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, 2019.

[19] J. Beel, B. Gipp, S. Langer, and C. Breitinger, "paper recommender systems: a literature survey," *International Journal on Digital Libraries*, vol. 17, no. 4, pp. 305–338, 2016.

[20] C. Nascimento, A. H. Laender, A. S. da Silva, and M. A. Gonçalves, "A source independent framework for research paper recommendation," in *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries.* ACM, 2011, pp. 297–306.

[21] J. Beel, S. Langer, M. Genzmehr, and A. Nürnberger, "Introducing docear's research paper recommender system." in *JCDL*, 2013, pp. 459–460.

[22] F. Ferrara, N. Pudota, and C. Tasso, "A keyphrase-based paper recommender system," in *Italian research conference on digital libraries.* Springer, 2011, pp. 14–25.

[23] S. Pruitikanee, L. Di Jorio, A. Laurent, and M. Sala, "Paper recommendation system: A global and soft approach," in *FUTURE COMPUTING'2012: Fourth International Conference on Future Computational Technologies and Applications*, 2012, p. 7.

[24] I. Nassiri, A. Masoudi-Nejad, M. Jalili, and A. Moeini, "Normalized similarity index: An adjusted index to prioritize article citations," *Journal of Informetrics*, vol. 7, no. 1, pp. 91–98, 2013.

[25] S. Bonzi, "Characteristics of a literature as predictors of relatedness between cited and citing works," *Journal of the American Society for Information Science*, vol. 33, no. 4, pp. 208–216, 1982.

[26] I. Spiegel-Rosing, "Science studies: Bibliometric and content analysis," *Social Studies of Science*, vol. 7, no. 1, pp. 97–113, 1977.

[27] A. Shahid, M. Afzal, and M. Qadir, "Discovering semantic relatedness between scientific articles through citation frequency," *Australian Journal of Basic and Applied Sciences*, vol. 5, no. 6, pp. 1599–1604, 2011.

[28] A. Livne, E. Adar, J. Teevan, and S. Dumais, "Predicting citation counts using text and graph mining," in *Proc. the iConference 2013 Workshop on Computational Scientometrics: Theory and Applications*, 2013.

[29] A. Shahid, M. T. Afzal, and M. A. Qadir, "Lessons learned: The complexity of accurate identification of in-text citations." *Int. Arab J. Inf. Technol.*, vol. 12, no. 5, pp. 481–488, 2015.

[30] 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.

[31] H. Nanba and M. Okumura, "Towards multi-paper summarization using reference information," in *IJCAI*, vol. 99, 1999, pp. 926–931.

[32] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[33] A. Huang, "Similarity measures for text document clustering," in *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, vol. 4, 2008, pp. 9–56.

[34] B. Ribeiro-Neto and R. Baeza-Yates, "Modern information retrieval," *Addison-Wesley*, vol. 4, pp. 107–109, 1999.

[35] T. Korenius, J. Laurikkala, and M. Juhola, "On principal component analysis, cosine and euclidean measures in information retrieval," *Information Sciences*, vol. 177, no. 22, pp. 4893–4905, 2007.

[36] A. Jain, A. Jain, N. Chauhan, V. Singh, and N. Thakur, "Information retrieval using cosine and jaccard similarity measures in vector space model," *Int. J. Comput. Appl*, vol. 164, no. 6, pp. 28–30, 2017.

[37] M. Borg, P. Runeson, J. Johansson, and M. V. Mäntylä, "A replicated study on duplicate detection: Using apache lucene to search among android defects," in *Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement.* ACM, 2014, p. 8.

[38] Y. Zhou, X. Wu, and R. Wang, "A semantic similarity retrieval model based on lucene," in *2014 IEEE 5th International Conference on Software Engineering and Service Science.* IEEE, 2014, pp. 854–858.