CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

A Novel Formula to calculate Similarity between Paper Title and Tweets by employing Dis-similarity and sub-sequences

by

Mariam Nawaz

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

in the

Faculty of Computing

Department of Computer Science

2020

This study is wholeheartedly dedicated to my beloved parents, who have been my source of inspiration and gave me strength when I thought of giving up, who continually provide their moral, spiritual, emotional, and financial support. To my teachers, who shared their words of advice and encouragement to this study. And lastly, I dedicated this thesis to the Allah Almighty, thank you for the guidance, strength, power of mind, protection and skills and for giving me a healthy life.

CERTIFICATE OF APPROVAL

A Novel Formula to calculate Similarity between Paper Title and Tweets by employing Dis-similarity and sub-sequences

by

Mariam Nawaz

(MCS181017)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Mansoor Ahmed	COMSATS, Islamabad
(b)	Internal Examiner	Dr. Nayyer Masood	CUST, Islamabad
(c)	Supervisor	Dr. Muhammad Tanvir Afzal	CUST, Islamabad

Dr. Muhammad Tanvir Afzal
Thesis Supervisor
December, 2020

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

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

Author's Declaration

I, **Mariam Nawaz** hereby state that my MS thesis titled "**A Novel Formula to calculate Similarity between Paper Title and Tweets by employing Dissimilarity and sub-sequences**" 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.

(Mariam Nawaz)

Registration No: MCS181017

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**A Novel Formula to calculate Similarity between Paper Title and Tweets by employing Dis-similarity and sub-sequences**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly 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.

(Mariam Nawaz)
Registration No: MCS181017

Acknowledgements

I would like to express my gratitude and thanks to ALLAH (S.W.T) for providing me abilities to accomplish this research work. Secondly I would like to express my sincerest thanks to my respected supervisor Dr Muhammad Tanvir Afzal his guidance and encouragement. He has taught me, both consciously and unconsciously, how good experimental work is carried out. Sir you will always be remembered as inspirational teacher. Last but not the least; I would like to thank to my mother, father and my husband for their support, encouragement and prayers.

(Mariam Nawaz)

Registration No: MCS181017

Abstract

E-Learning is the emerging research area for scientific knowledge management. E-Learning has been recognized as multibillion industry recently. To identify the relevant resource at the right time is very important factor for E-Learners. There is a lot of literature available about recommendation in E-Learning like recommending books and papers etc. but these recommendations some time not fulfill the individual needs of user. It establishes the need of specialized system in which help is provided to the E-Learners from some growing trends related to their task at hand from the social media.

Twitter is fast growing social media. Millions of tweets are shared on daily basis. Users post, answer questions and share ideas and resources and work to each other.

This thesis has identified that very limited efforts have been made to identify relevant tweets form Twitter for E-Learners. From the recent state-of-the-art, research gap has been identified and worked on in this thesis. It has been identified that the state-of-the-art has identified relevant tweets from Twitter using the similarity scores. However, no one has used the dis-similarity scores and subsequences of title and tweet to identify the relevant tweets. Therefore, this thesis has proposed a novel formula which includes not only the similarity between the paper title and tweet, instead the following parameters have been proposed, implemented, and evaluated as well such as: (1) calculation of dis-similarity scores and (2) calculation of subsequence matching and giving more weights to the higher order n-grams for computation. The proposed formula has been compared with the standard approaches such as: Cosine and Jaccard. Furthermore, the proposed formula has been evaluated using standard evaluation parameters such as: Precision, Recall, and F-Measure. The proposed formula has overall outperformed all approaches and in some of the cases has better precision and comparable recall with the compared approaches.

Contents

List of Figures

List of Tables

Abbreviations

ACM	Association of Computing Machinery
CCS	Computer and Communication Security
IWM	Individual Words Between Paper Title and Tweet
LMS	Learning Management System
SL	Matched sub sequence length between paper and title
TSL	Total Match Subsequence Length
WPT	Number of Total Words in Paper Title
WTT	Number of Total Words in Tweet

Symbols

$\sum$	Summation
$\cup$	Union
$\cap$	Intersection

Chapter 1

Introduction

This chapter gives the overview and motivation of my research area. Furthermore, this chapter elaborates the scope of the work followed by the problem statement. Afterwards research questions are formulated and explained. Subsequently, the applications and implications of my research work and contributions are presented. Important definitions and abbreviations are also discussed in this chapter. Finally, organization of thesis has been outlined at the end of this chapter.

1.1 Overview

The aim of this research is to recommend relevant tweets to learners according to the research papers they are reading. A large number of research papers are published each year [1] and millions of tweets are posted daily on twitter [2]. There is a huge amount of learning material available on digital libraries and citation indexes. When users are reading them, there is large number of tweets shared by the twitter's users related to those research topics [3]. The objective of this thesis is to present the relevant tweets to the user based on user's current context, history and profile. E-learning is referred to as a learning tool that is not limited to a physical classroom environment. Different social networks are used for recommendation in E-Learning like Facebook, twitter, YouTube etc. Twitter has an opportunity to share different learning material in precise and short text. Therefore we use tweets from in our work. It recommends the most relevant tweets to learners.

1.2 Motivation/Thesis Objective

1.2.1 Motivation

E-Learning is a multibillion industry has been estimated through the market value of E-Learning [4]. Current worth of E-learning is $187.77 billion in the year of 2020 and it is expected to be reached $319.167 billion in 2025[1]. So E-Learning market is expecting significant growth opportunities in the next five years. This means that organizations are putting a lot of investments for the E-Learning environment and facilitates the E-Learners and motivation of my research is further emphasized and has increased in the current time in which COVID-19 has spread everywhere and people are restricted in houses due to lockdowns. It is expected to further increase the demand for E-Learning platform therefore, online education has been adapted at least temporally by almost all universities worldwide even schools, colleges and universities [5].

1.2.2 The Scope of the Thesis

The scope of thesis is to evaluate the available techniques using benchmark dataset of tweets using paper's metadata. Furthermore, the innovative formula to compute the similarity will be developed. The developed formula can find out the most relevant tweets according to the focused research paper. We compare proposed technique's results with already available techniques of similarity like Cosine and Jaccard and evaluate which technique is calculating best result.

[1] http://www.reuters.com

We also compare our results with technique of lexical and find out which technique work best.

1.2.3 Problem Statement

The state-of-the-art research has used only the similarity based approaches to gauge the relevance between paper and tweets. Those approaches have missed two important considerations [6].

1. Computation of dissimilarities between metadata of paper and tweet.
2. Calculating the similarity between metadata of paper and tweets using the sub-sequences.

1.2.4 Research Questions

1.2.4.1 Research Question 1

What is the best approach to find the relevance between research paper and tweet from the state-of-the-art approaches?

1.2.4.2 Research Question 2

What will be the effect if we compute relevance between paper title and tweets based on sequence of words by incorporating the dissimilarity score?

1.2.5 Application of Proposed Approach

Individual learning

Specialized system delivers the appropriate content at the right time to support individual learning. Such a system can help learners in their perspective domain learning and can assist them according to their particular needs, context, profiles, history and collaboration.

a. LMS

Our technique can be used in general LMS like Moodle etc. that can make the LMS more efficient.

b. Digital Library

This system can be used in digital Library.

c. Citation Index

This technique can be used in citation index

d. Search Engines

Search engines can use this technique to recommend most relevant material to users

1.2.6 Organization of the Thesis

First chapter describes the overview of system with scope, motivation and problem statement.
Chapter 2 provides the overview of the Literature. It illustrates recommender system can help the E-Learner to find the relevant knowledge from a social network.
Chapter 3 covers the methodology which answers the raised research questions described in first chapter.

Chapter 4 presents extensive experimentation and evaluation of proposed technique.

Chapter 5 concludes the research along with the future work.

Chapter 2

Literature Review

This chapter discusses the research work based on existing research on E-Learning system and different recommendation systems for E-Learners. This chapter focuses on the critical review of the state-of-the-art approaches. We have divided this chapter in following sections. Section 2.1 shows the overview of E-Leaning system. Section 2.2 shows overview of recommendation systems. Section 2.3 demonstrates limitations of traditional E-Learning system. Section 2.4 shows recommendation approaches and 2.5 shows critical analysis of the state-of-the-art approaches.

2.1 Overview of E-Learning Systems

A learning system is based on formal education but with the help of electronic resources like CD/DVD known is E-Learning. E-learning is treated as alternative to classroom learning. E-Learning is not restricted to physical presence in and out of the classroom. It provides the facility to accessing the educational material at anytime and anywhere in the world using computer and internet and any helping device [7].

Many researchers have recommended and praised E-Learning as a very useful tool for distance learning, to minimize number of students or lectures going to their institutions from different homes [8, 9].

Different E-Learning software like blackboard, Moodle and WebCT are available around the world[2]. Thousands of academics center like universities and colleges using these software.200 above universities have been used E-Learning system [10].

The basic features that are common to all of these web-based learning systems are summarized in the section below.

2.1.1 Course Management and Design

It gives the facility to instructor of management of course material and student activities in the course. Instructor can easily schedule their upcoming events and announcements.

2.1.2 Performance Evaluation and Feedback

It provides the facility of performance evaluation of students through conduct of quizzes and online tests. Results show the performance of students. Teacher's comments, remarks and explanation have a great importance for performance evaluation. These are incorporated through the feedback procedure.

2.1.3 Interactive Communication

It provides tools for communication and interaction between students and instructor. Students and instructor can communicate easily even distance is long.

2.1.4 Course Evaluation

Course evaluation is done through students. It helps to improve the course delivery. And course evaluation is a paper or electronic questionnaire, which requires a written or selected response answer to a series of question.

[2] http://E-Learningindutry.com/top-10-E-Learning-staticis-for-2014-you-need-to-know

2.2 Overview of Recommendation System

Recommender system can be considered as a black box. Input is given to the black box in the form of user profile and then matches it against the candidate set to suggest unseen items [11]. These unseen items are considered the most relevant recommendation for the user.

2.2.1 Recommendation in E-Learning

The E-learning recommendation system helps learner to choose alternatives without personal experience and this is especially important in times of information is blowing up. The web-based learning environment is being used extensively in education. Educational resources have been increased through this situation. Services were permanently integrated into the systems and diversified among learners to utilize and access to this educational material. However, in general this educational material is provided to all learners in a way that focuses on the styles used in different ways or the differences between their profiles and individual needs. That reason made the personalization mandatory in the E-learning and not considered as option [12].

Recommendation appropriate links an example of adaptive navigation support technology. Recommender system of E-Learning is used to find relevant items that are according to the already available learner. These Learner's profile and interest used for recommendation. Because these learners are consider good learners according to their score in learning material [13].

Recommendation in E-Learning helps the student to increase their performance. Kl Gauth work shows that recommender system in E-Learning increased the performance of students. E-Learning's Recommender systems are based on good learners. It can help improve student performance. In E-Learning recommender system, aims to address this issue by incorporating good learners ranking strategy to guide students through study materials that are highly recommended by the best students in an effort to enhance their learning performance. Content-based filtering approach applied to ascertain that learning material that is recommended remains within the current learning task or not. Best students are represented by term used “Good leaners”. Describe good leaners as student who scored more than 80% in the conducted research experiment. The result shows that student performance has increased through the recommendation system [14].

Personalized learning occurs when the E-Learning system makes a conscious effort to design educational experiences tailored to the needs, goals, qualification and interests of learners. Author describes a system name “Protus” which is recommendation base module of the programming tutoring system. This system adapts automatically user interest and knowledge level. Recommendation is based on Collaborative filtering approach. This recommendation improves the quality of E-Learning [15].

2.2.2 Recommendation Through Social Media

Social media has gain a lot of success in recent years. Millions of users are visiting different sites like Facebook, Twitter etc. These sites rely primarily on their users for creating content, interpreting other’s people content using Tags and comments; to establish online relationship and join communities that are working online.

There is a lot of data about the emergence of social networks and their users that makes them an important source of personal information about the users of the recommender system. Various social media used in recommendation system like Facebook, Twitter, and YouTube etc.

Recommendations in social media are based on two items peoples and Tags. Information about the relationship people, tags and items is collected from various ways with in the enterprise. These collective relationships are used in the system for recommend items that are related to people and tags according to user. Ido Guy evaluates the recommender system through user study and results shows that Tag-based recommender system is better than people-based [16].

As stated by Berkovici, social media is considered as big source of entertainment. Social media can have a great impact in education therefore recently, after highlighting the scientific community in a traditional perspective; higher learning or education is fetching attention to the use of social media in education [17].

Anderson outlines some of conditions about usage of social media that can help in lead to learning about active collaboration in higher education [18].

It is obvious that students also use social media a lot to learn the Quran and Hadith. Students can share and exchange information with their fellows using these platforms. This work was to explore the effects of a number of factors on collaborative learning and student satisfaction that led to better learner performance [19].

Recommendation system usually aims to predict the classification or relevance of items that no user has seen, and consider items to be the best fit for an active user. Author investigate the effectiveness of existing data from the social media for the process of recommendation specially Facebook. Extract published content from the Facebook about their personal pages favorite items and preferences in the recommendation domain and statistics about other domain preferences that allow cross-domain recommendation. Data is examined that is related to recommendation domain and 44 other domains. Information can then be collaboratively aggregated as input recommender system during the calculation of similarity and prediction phase. This is for the new user when no ratings are available. Data obtained from Facebook can also strengthen board ranking data to improve the performance of recommender system [20].

2.2.3 Why Twitter-based E-Learning is Necessary

Twitter is a US micro blogging and social networking service where user interacts and post tweets. In twitter, per month 320 users are active and 500 million tweets are sent daily. Users send message to each other up to a limit of 140 characters. People post and answer questions, share ideas and resources and work to each other on different issues. Twitter is more interactive social media than Facebook among public specially students and teachers to give fast way to share information [21].

In normal LMS, before making a query has to log in and search a suitable blog and wait to reply. So, the level of interest may decrease during this time. Twitter character demarcation is focused on the questioner so the question and queries are exact and accurate. In the same way, the learner gets the exact answer to his question. Author work to find out the trends in smart learning by using the research papers that investigated by search sites since 2007. Experiments had conducted to find out the effectiveness of twitter in education. Author compared the traditional methods of learning to the twitter base learning methods [22].

Experiments have been conducted to show the student's performance after using the twitter. The results indicate that after using the twitter students got better grades than those who did not use twitter [23].

Author discussed about effectiveness of twitter in learning point of view like learn English in a foreign course. Analyzed tweets of Japanese's students and found that students were continuously active to using twitter for attending course [24].

Junco et al make two different classes and distributed students in both classes. Experiment conducted to check the impact of learning. So in one class twitter use was mandatory and in second, twitter use was optional. Result indicated that twitter usage had a great impact on learning [25].

Kassens did a study to visualize the role of twitter instead of learning of class. Kassens proves that in the exam condition student remember better contents if they receive daily tweet related to their content [26].

An experimental study was conducted to gauge the impact of twitter on the learning. The result of twitter's use shows that it provides a useful tool for collaboration and sharing information with students. Students who were activity tweeting had a better performance grades than who did not utilize twitter [27].

Author analyzed positive association of university students and faculty with twitter. The participation of students in university activities was encouraging [28]. Furthermore, author conducted the user case study and results show that twitter has a positive effect on user learning, teaching environment and experience of students. Results also show that before the start of course, many students had interest toward Facebook about 79% and only 57% were utilizing twitter. This proves that the student is already looking for a helpful tool in their learning environment, such as social network like twitter [29].

2.2.4 Twitter Usage as Recommendation System

There is a huge collection of news reading sources around the world but online reading news has become an interesting and famous way for the people. Users want to read news articles according to their interest but there is a lot of flood of articles and news. Recommender system use to help the users to find out news according to user's interest. Twitter use in research to developing the recommender system based on personalized articles and news. Twitter is use in ranking of news articles according to the public tweets timeline. Furthermore, users create profiles according to their interest and news articles are categorize based on user's profile match. Hybrid (combination of content and collaboration approaches) recommender system has been developed that suggest news to the user that is relevant and interesting according to the user view [30].

User generated a lot of contents on Twitter platform according to their interest or profile. Twitter has no restriction to anyone for posting the contents. Therefore any one can post tweets on the twitter platform and language utilization is creative. As a result, growing number of tweets are awaiting complete analysis on Twitter. In particular, in the event of catastrophe, there is an urgent need for a solution to detect accurate and complete information. Usage supports relevant tweet submission analysis. The sheer volume of contents that is generated by user makes it difficult to find relevant content and information. Therefore, the key is to make it easier to extract psychological, semantic and syntactic features in form of terms. This shows that there are benefits to handling key terms violation separately like disaster investigator are able to effectively detect irrelevant and relevant information [31].

Web mining is used with twitter as knowledge discovery in tweets and clarifying the application of methods using the title of physical activity. Two methods are described structure mining and content mining. In structure mining structure like (meso, micro and macro) are discovered after utilization of analysis of social network. In content mining sentiments and n-gram based analysis used to discover tweet related content. Twitter is used to get the opinion of public. These methods are helpful to get understanding about physical activity and may be utilizing to mine social media to solve the purposes that are related to health [32].

2.2.5 Twitter-based Recommendation in E-Learning

Today learners have the ability to use powerful social networking tools like twitter where learners can freely create and distribute content. Therefore, the "Digital Native" learners get a common LMS structure complex and boring. The research class is active in making the learning experience more effective in relation to the individual needs of the learners. Recommender system based on semantics is used for E-Learners that facilitate the E-Learners with effective E-Learning. Tweets that are relevant to the learners are recommended [6].

2.3 Limitation of Traditional E-Learning System

In traditional E-Learning system, learner's individual needs are not catered. It only considered as course management system to deliver course content according to instructor [33]. As a result, critics have questioned the effectiveness of the E-Learning system and provide alternative to class learning experience [34]. These are disadvantages of E-Learning systems. It requires strong self-motivated and time management skills.

2.4 Recommendation Approaches for E-Learners

The establishment of novel technologies has driven some new and innovative approaches in web-based education. Online courses do not meet the individual needs of the learners. They only work according to the static solution of queries. It is a challenge of management system of E-Learning that to recommend relevant content or information to the E-Learners. Relevant content recommendation can fulfill the individual needs of the learners.

Web-based learning provides the facility to store the learner's patterns of learning in large data set. Data mining techniques can be helpful for creating personalized learning profile [35]. It aims to provide personalized learning task and activities tailored to individual learning needs and consequently to improve the overall learning experience.

Activities and tasks that are related to already complete task by learners or their peers are recommended to the learners. Recommended system gives a best environment for personalized learning to the learners. Web has a great bundle of resources so recommender system recommends interesting and relevant contents to learners. Recommendation may be based on history (resources of learners that were previously viewed and selected) that is utilized by learners or the other learners preference and rating. To avoid the pitfall of one technique, combinations of different techniques are used. Suggestion can be in any form like webpage, task and tutorial.

LMS like WebCT and blackboard are not such an intelligent LMS because do not provide best and intelligent learning environment. These LMS has also deficiency of dynamic and personalized learning. In this case the researchers have begun questions about functionality of traditional E-Learning system [36].

The following portion discussed the pros and cons of approaches and also discussed about challenges task that will be beneficial for researchers as future task. And it will able to work in different research areas of these recommendation approaches to recommend the relevant material.

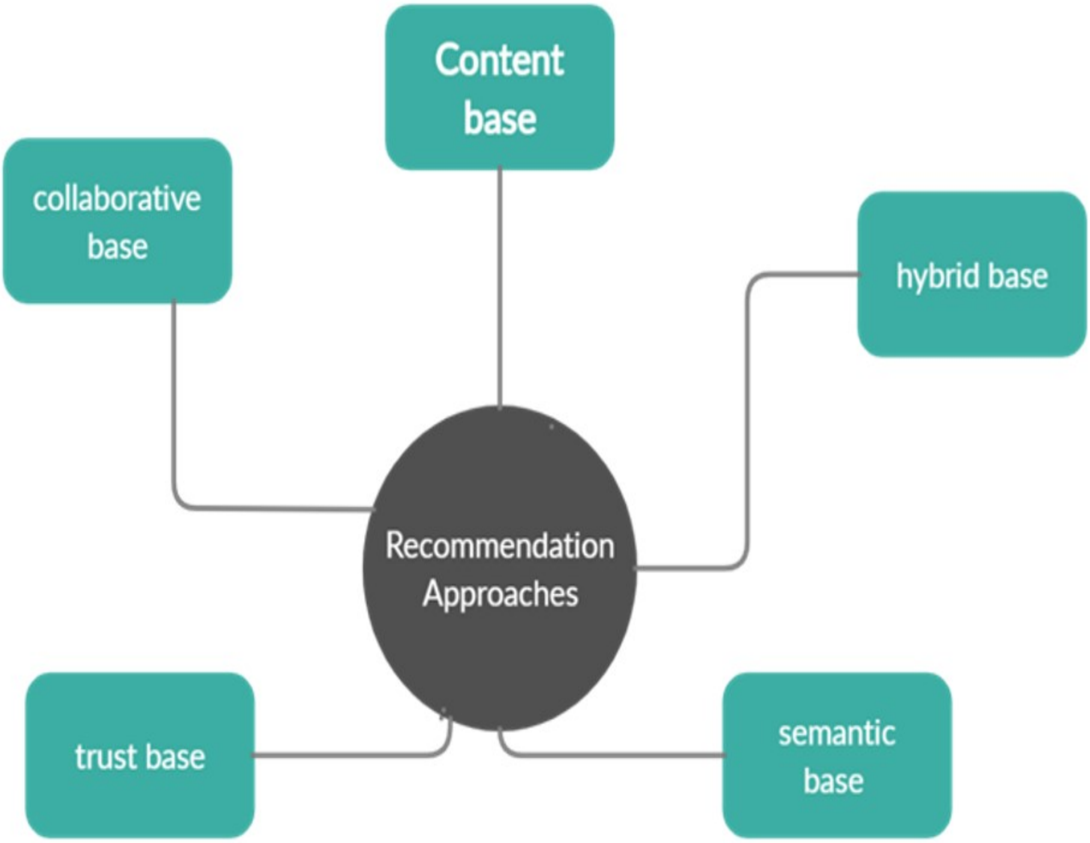


Figure 2.1: Recommendation Approaches

2.4.1 Content-based Filtering Approach

Content base approaches contain the textual information so in recommendation system content base filtering technique is useful to recommend content to the learners for example articles, URLs etc [37]. The learners are recommended new and interesting learning objects so that they can learn the characteristics of the learner's object to meet the object learner's characteristics. Information about user preferences and needs are represented by user profile and it can be obtained through questionnaires and surveys or secular information.

The preferential learner pattern is created by extracting information about the learning object and the learner profile feature.

By applying the TF-IDF weighting method, content similarities are usually estimated using vector space models. Object of learning and learner's profile are represented by weighted term vector. In proposed Content-based systems, unlike statistical data where information has a distinct value, object information is represented in text form.

2.4.2 Collaborative Filtering Approach

David Goldberg et al invented the term Collaborative filtering (CF). In the process of retrieving information, influence of human can play an important role and author was inspired with this knowledge [38]. The technique of collaborative filtering makes recommendation based on analyzing user rating matrix or usage history items.

The basic concept of CF is that if a user F implemented in a similar rating of any set A and B, they should be identical to the classification of other items [39]. To make accurate predictions and recommendation, a vast dataset of user rating datasets is required.

The CF model creates model based on training data and provides data services predicted rate of new items. This procedure is implementing various techniques from data mining and machine learning.

CF techniques are divided in to methods that are memory and model based [40]. The K nearest neighbor (KNN) is the most widely used algorithm in the memory based CF modes [41].

2.4.3 Hybrid-based Model

The hybrid approach combines the two or more techniques to overcome the limitation of a technique. Chen et al work has an example of hybrid model in which collaborative approach combined with content filtering approaches [42].

2.4.4 Trust-based Model

Recommendation system of E-Learning is different from traditional recommender system. Learners that have more experienced can give better recommendation than less experienced learners [43].

A trust base recommendation system assigns a level of trust based on the user's ability to interact with the system [44].

Another model that is trust base which give an association between already available knowledge through similarities and common values necessary to establish trust [45].

Most experienced and reliable learners are two levels that are suggested by Dwivedi [46]. These levels are used for filtering the resources of learning recommendation. Author has conducted experiments and result shows that experience and trust can play an important role in correctness of recommendations. It proves that collaborative filtering work better than traditional collaborative filtering [47].

2.4.5 Semantic Model

A semantic base model can give different benefits in personalized suggestion system. The interest of learners dynamically expressed in specific domain [48]. Semantics could be beneficial in the process of personalization so recommender of next generation should consider it and social media [49]. Semantic web offers better possibilities for improving metadata related to learning content. E-Learning methods that already exist can be expanding through this [50].

Semantic web provides better prospects to improve the metadata associated with learning content. It also offers an excellent opportunity to expand the existing ELearning methods.

The goal is to keep the learning content independent of a particular content provider technology. This will enable new and innovative learning experiences through existing learning objects.

2.5 Critical Analysis

The overview of these techniques are describe below:

Table 2.1: Critical Analysis of existing Recommendation techniques in literature

Author	Dataset	Model/strategies	Limitation
Kl Gauth et al, 2010	Good learners who got score 80% or above in learning material their rating and method treated as benchmark	Content base filtering approach with cosine similarity	Not effective in word sense disambiguation
Nirmal et al,2013	280 RSS news articles 202,224 tweets	Hybrid(content and collaborative filtering) Recommendation model with cosine similarity to recommend news based on user interest rather than presenting whole articles in order of their occurrence	Not effective processing of queries that are written in natural language
B Shapira et al, 2013	Collected explicit user rating on 170 popular movies	Collaborative filtering recommendation technique	Apply simple heuristic procedure for extraction of data

A Magnuson et al, 2015	Collect event organization site and their related tweets	Item base collaborative filtering use to recommend event using tweets	Does not considered the importance of dissimilarity and distance based on sub-sequence
Khalid et al, 2017	100351 research papers	Collaborative filtering approach	Does not calculate distance based of sub-sequence
X Zhao et al, 2018	9,136,976 items	Deep recommender system (DEERS) model with cosine similarity	Does not considered the importance of dissimilarity
S Manoharan et al, 2019	25 users search history	MFIS (Mamdani fuzzy inference system) with content base approach	Use small dataset
Zeinab et al, 2020	2856 Tweets and 567 articles	Collaborative filtering approach	Use only similarity score for calculate relevance
Z xu et al, 2020	1843 users, 3508 tags	Semantic model	Require exact matching between a query term and ontology concept

Different approaches have been used for recommendations in literature. We studied these approaches and analyzed the approaches for the critical review.

After the comprehensive analysis of state-of-the-art approaches, only similarity techniques between the content has been used to compute the relevance for recommendation. None of the contemporary approaches have computed relevance between content based on dissimilarity and have not calculated sub-sequences lengths. These measures may compute best relevance between titles of research papers and tweet. Such important scenarios have been explained below:

2.5.1 Why important to calculate dissimilarity?

Let's have a scenario in which we will demonstrate that calculating the dissimilarity could be also useful for measuring the overall similarity between paper title and tweet title considering the following example. Here we have a paper whose title is written and two tweets which are relevant to that paper are described below. We have highlighted words in paper title and also highlighted words in both tweets that are similar with paper title.

- **Paper Title:** Information and pattern discovery on the World Wide Web
- **Tweet 1:** World Wide Web is big resource of pattern discovery
- **Tweet 2:** Information is important source of knowledge discovery and pattern reorganization in world wide web

In above example, Tweet 1 is more relevant to paper title because it is actually describing about paper tile and tweet 2 is not as much relevant to paper title as Tweet 1 is. However, when we calculated similarity, tweet 2 is ranked higher than tweet 1 because in tweet 2 "6" words are matching with paper title words and in tweet 1 "5" words are matched. But according to dis-similarity, tweet 1 is ranked higher than tweet 2. So it is evident that dis-similarity could be another important individual measure that can be applied along with the calculation of similarity for comprehensively calculating the relevance between paper title and tweet.

2.5.2 Why important to calculate distance base on sub-sequence?

Here we have another scenario in which we will calculate the relevance between content using the sub-sequence measure. This measure could be useful to calculate best relevance between paper title and tweet. Here we also take one paper title and two tweets and calculate relevance which tweet is related to paper title.

- **Paper Title:** Feature weighting in Content based recommendation system using social network analysis
- **Tweet 1:** Attributes used for content based recommendation are assign feature weighting depending on their importance to users.
- **Tweet 2:** Social network analysis has a great impact to get information about any topic and content. Also assign weighting according to comment and interest. Different features used for this purpose.

We can see in above example, the manual inspection reveals that tweet 1 is describing about paper title but if we calculated similarity then tweet 2 will be ranked high than tweet 1 because in tweet 2 "6" words are matched with paper title and in tweet 1, only 5 words are matched. However, if we take sub-sequences, then tweet 1 is ranked higher than tweet 1 .In tweet 1, two sub-sequences are matched with paper title and in tweet 2, one subsequence is matched with paper title. The remaining highlighted words are single matched to paper title. The point we want to make here is that calculating the relevance between paper title and tweet, it's better to compute the similarity based on subsequences. The matching of bigger subsequence might mean more relevance. Therefore, it's important to thoroughly explore the subsequences for calculating the relevance between paper title and tweets.

Both of the above considerations have been exploited in this thesis. This thesis has introduced a novel formula which has incorporated the dissimilarity and subsequences.

Chapter 3

Research Methodology

Research study in the previous chapters shows that researchers have proposed recommender systems that are based on similarity. They do not calculate dissimilarity and sequence of words for recommender systems. Similarity base recommender systems give relevance but some time they do not give precise results. Therefore, these two factors (dissimilarity and sequence of words) should also consider important to calculate relevance. This thesis focuses to calculate relevance based on dissimilarity and sequences of words to retrieve most relevant tweet according to the paper that can be recommended to the learners.

In this chapter, we have discussed the detailed methodology of proposed system. In our proposed approach, similarity as well as dissimilarity and sequence of words have been utilized to find most relevant tweet. Furthermore, the results have been verified using a dataset that contain research paper metadata and tweets. The standard evaluation measures (Precision, recall and F-measure) have been used to check the accuracy of model. The contemporary similarity techniques (Cosine and Jaccard) have been used to compare the results. Furthermore, the proposed formula has been compared with the State-of-the-art approach presented recently by [6]. The detail of the each part of our proposed system is given below.

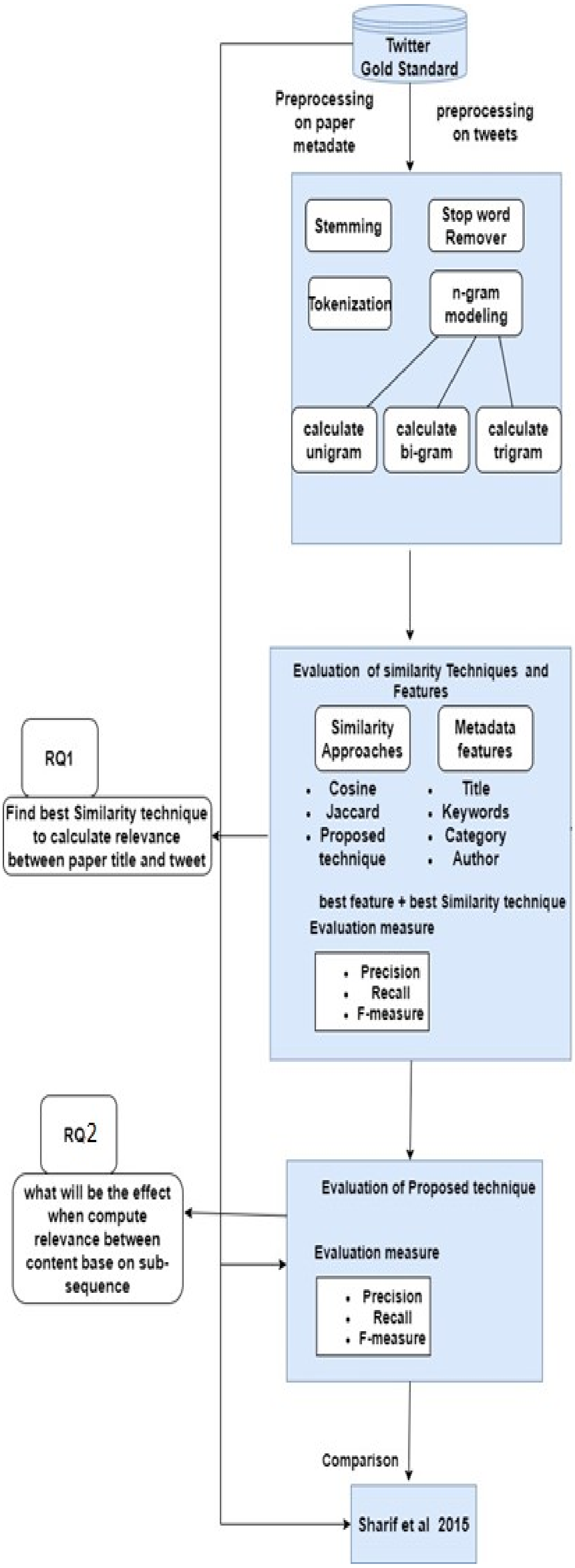
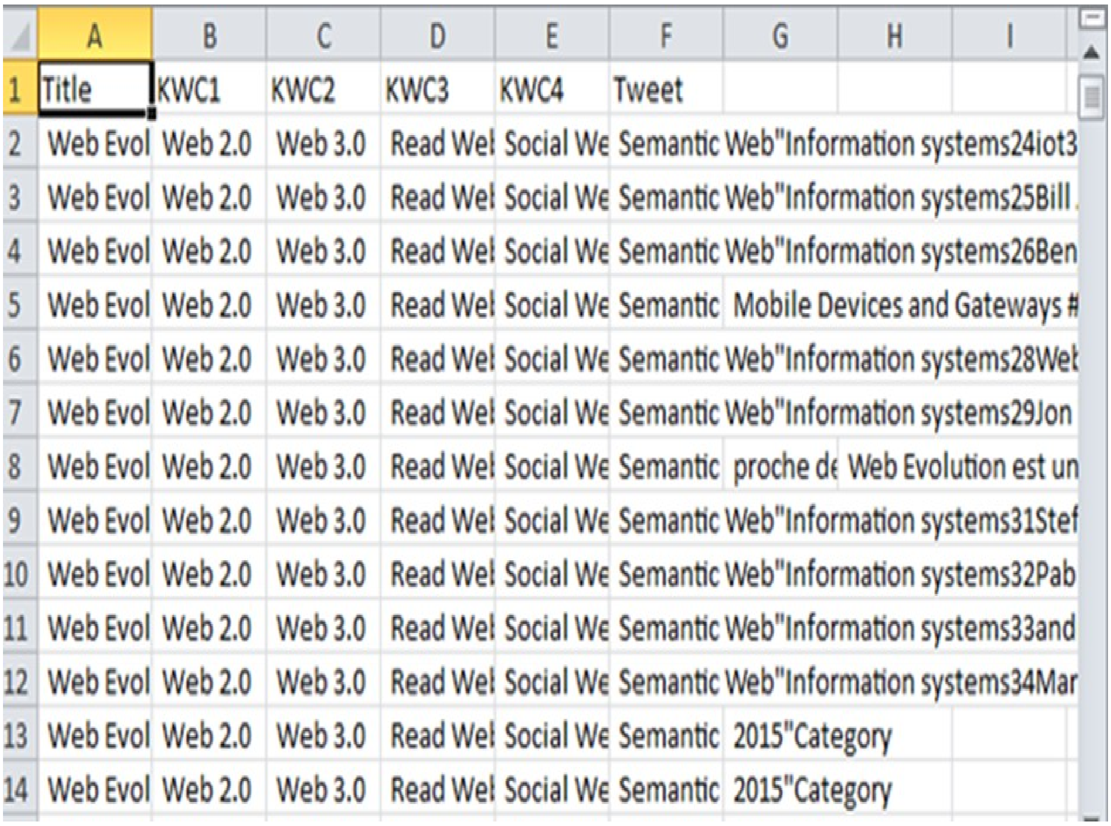


Figure 3.1: Framework Proposed work

3.1 Gold Standard Dataset

We required the dataset which could fulfill the following characteristics: (1) we need research paper that should be belonging to different topics. (2) Tweets should be selected for each paper based on user's profile, history and context of user. Fortunately, we were able to find such the dataset from [6]. This dataset contains 220 research papers selected from ACM diversified topics. Furthermore, for each paper Sharif et al performed a unique user study. In this study, each user was given a paper (which formed the context, history and profile). Participants of the user study was asked to select suitable tweets from twitter. For each research paper users have selected suitable tweets that range from 4 to 22. In this manner they were able to gather 2957 tweets. For each paper selected the relevant tweets formed the Gold standard. For any of the implemented technique (Cosine, Jaccard, Proposed formula and Sharif et al approach) the Gold standard acted as relevant tweets and all other tweets were considered noise.



	A	B	C	D	E	F	G	H	I
1	Title	KWC1	KWC2	KWC3	KWC4	Tweet			
2	Web Evol	Web 2.0	Web 3.0	Read Wel	Social We	Semantic Web"Information systems24iot3			
3	Web Evol	Web 2.0	Web 3.0	Read Wel	Social We	Semantic Web"Information systems25Bill			
4	Web Evol	Web 2.0	Web 3.0	Read Wel	Social We	Semantic Web"Information systems26Ben			
5	Web Evol	Web 2.0	Web 3.0	Read Wel	Social We	Semantic	Mobile Devices and Gateways #		
6	Web Evol	Web 2.0	Web 3.0	Read Wel	Social We	Semantic Web"Information systems28Wet			
7	Web Evol	Web 2.0	Web 3.0	Read Wel	Social We	Semantic Web"Information systems29Jon			
8	Web Evol	Web 2.0	Web 3.0	Read Wel	Social We	Semantic	proche de	Web Evolution est un	
9	Web Evol	Web 2.0	Web 3.0	Read Wel	Social We	Semantic Web"Information systems31Stef			
10	Web Evol	Web 2.0	Web 3.0	Read Wel	Social We	Semantic Web"Information systems32Pab			
11	Web Evol	Web 2.0	Web 3.0	Read Wel	Social We	Semantic Web"Information systems33and			
12	Web Evol	Web 2.0	Web 3.0	Read Wel	Social We	Semantic Web"Information systems34Mar			
13	Web Evol	Web 2.0	Web 3.0	Read Wel	Social We	Semantic	2015"Category		
14	Web Evol	Web 2.0	Web 3.0	Read Wel	Social We	Semantic	2015"Category		

Figure 3.2: Gold Set

3.2 Input for Tweet Ranking

The proposed technique provided two different types of input namely, paper metadata and tweets. Metadata of paper consists of the title of the research paper. In addition to metadata, a complete set of tweets will also be provided.

3.3 Preprocessing

A set of preprocessing steps will be performed to clear extracted data. Let's describe all steps in detail.

3.3.1 Stop Word Removal

In English, there are several words that are called stop words (for example the, is, a, which, at, in etc.) because they have no meaning. These words are common but meaningless. These words are used to combine words that do not participate in the content of the text document. These words are found almost in every text and often found in the title and tweets. Therefore, it is important to remove them to get the unique words. In the stop word removal step, these stop words will be removed from the text data by matching through available list of stop words. We used the dataset on the base of standard list of rainbow statistical text.

3.3.2 Punctuation Removal

Punctuations are characters such as full stop, comma and brackets. These words are used to separate words and clarify the meaning of sentence. Punctuation will be removed to clean the data.

3.3.3 Tokenization

Text is a collection of sequence of word and symbols. Most of the time, before any text processing, the text needs to be sorted into pieces for example number, alphanumeric, words etc. this process is called tokenization.

3.3.4 Stemming

Stemming helps to convert a word to their root word, for example the word "discussion", "discussing" and "discussed" would be converted in to discuss. Stemming is largely based on the assumption that retrieving information in text mining involves creating a query with presenting those incudes all the documents that contain the words presentation and presented. The advantage of stemming is that it can reduce the size of indexing by up to 50% [51]. The preprocessing phase is resulted in the formation of Unigrams.

3.4 Similarity Approaches

In the literature, state-of-the-art approaches available to calculate the similarity between objects. We have used two similarity approaches Cosine and Jaccard. We have applied these approaches on our Gold standard Dataset and calculate result. These results will be compared to our purposed technique.

3.4.1 Cosine Similarity

Cosine similarity measures the similarity between two products or vectors. It is used to determine about similarity of two entries. This similarity score range is between 0 and 1. Cosine similarity is usually used in high-dimension positive spaces e.g. information retrieval and text mining, each term is theoretically assigned a different dimension and a document has a vector property where the value of each dimension is equal the number of times the term appears in the document. Cosine similarity then provides a useful measure of the likelihood that the two documents will be similar in the case of his articles [52]. The Cosine Similarity of two documents (A and B) is:

$$Similarity(A,B) = \frac{A.B}{||A| * |B||} \tag{3.1}$$

Cosine Similarity is very popular matching measure that applies to textual documents such as in multiple information retrieval requests and in clustering [53].

Cosine similarity handles the words rearrangement and other differences in strings [54, 55]. Therefore we use Cosine similarity to compare the tweets with paper's title to compute the relevance. We compute the dot vector of Gold set (Title and tweets). If both have share many term same then dot product is higher otherwise low.

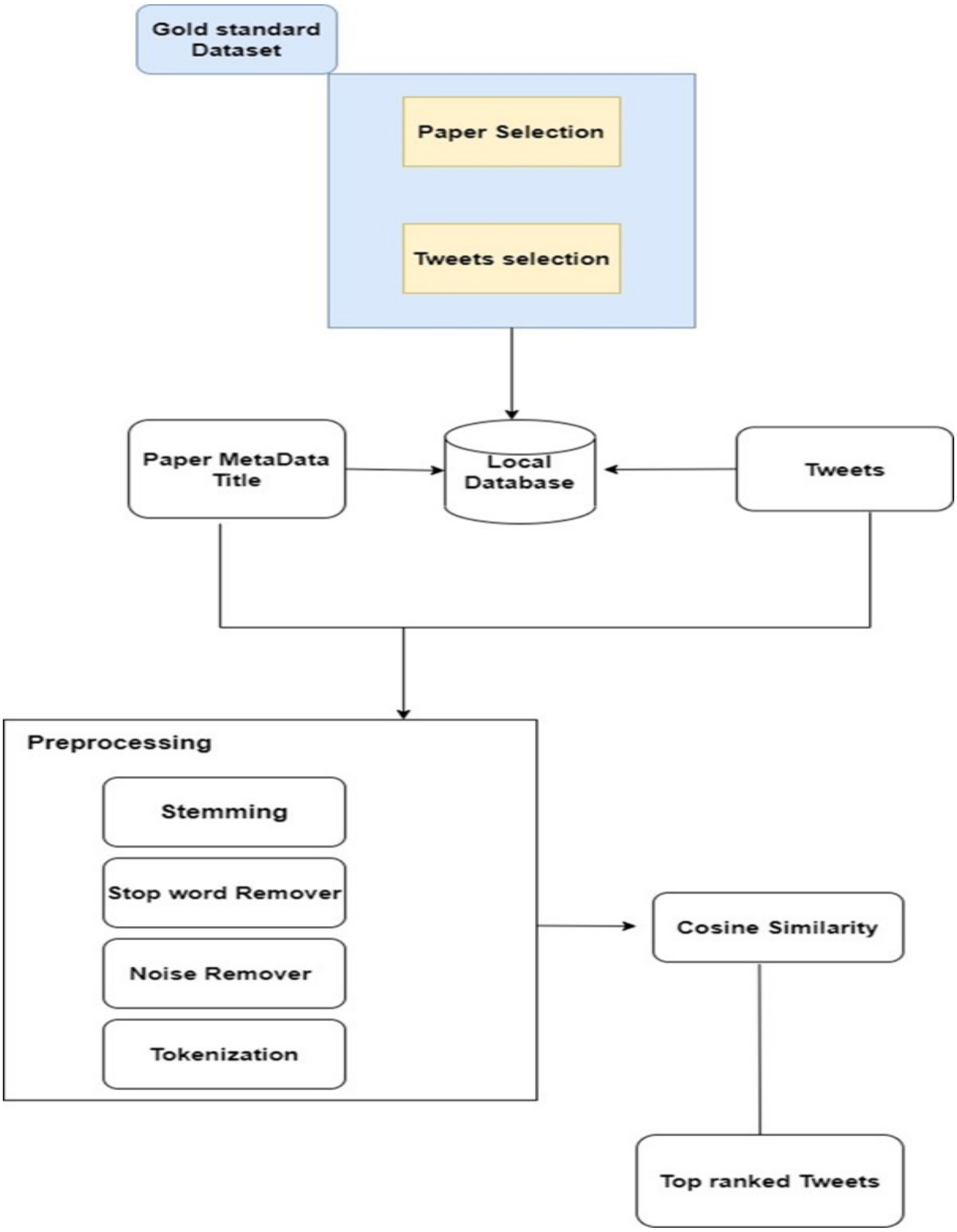


FIGURE 3.3: Dataset with Cosine Similarity

3.4.2 Jaccard Similarity

Jaccard Similarity also called Jaccard Coefficient. It is statistical model that is used to understand similarities between two sets. It compares members form two sets to see that which member is common and which are distinct. Similarity is calculated between finite sample sets.

The Jaccard Similarity between two documents (A and B) is:

$$J(A, b) = \frac{A \cap B}{A \cup B} \tag{3.2}$$

Similarity is measured by calculating the intersection of object that is divided by the union of objects. The range of resulted value is between 0 and 1. It is 1 when both objects are same like A=B and 0 when both objects are disjoint and completely different otherwise value between 0 and 1 [56].

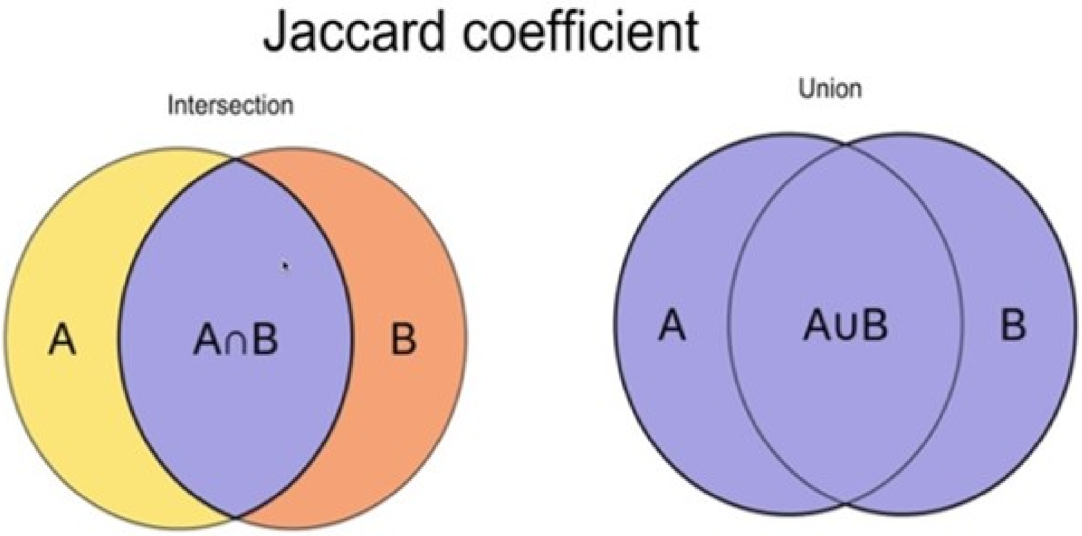


Figure 3.4: Intersection and union of Jaccard Similarity

We applied Jaccard technique on our Gold set. Jaccard Similarity calculates the similarity between paper title and tweet. Thousands of tweets are available against paper's title. Jaccard Similarity technique gives the relevant tweet of that paper.

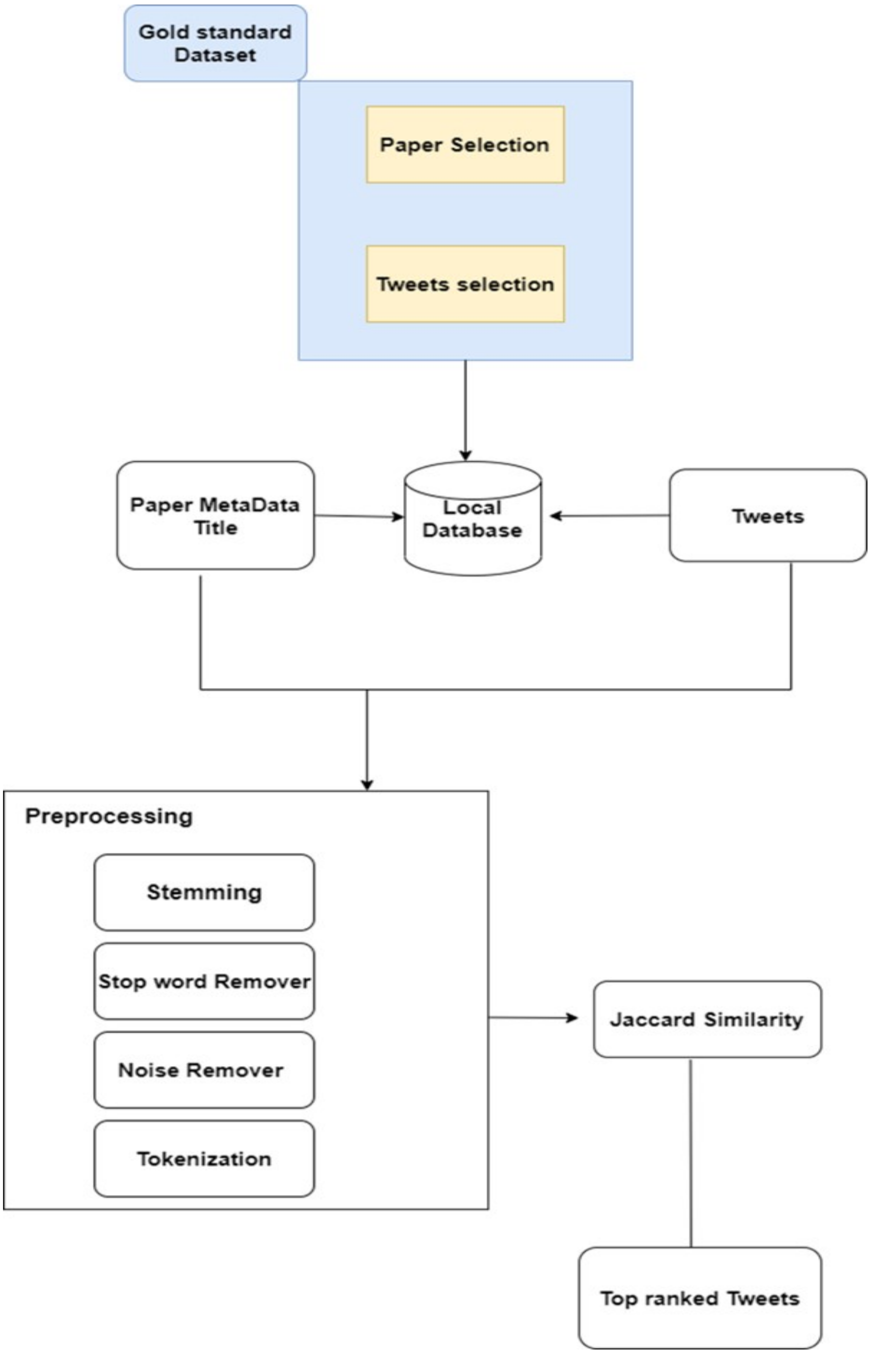
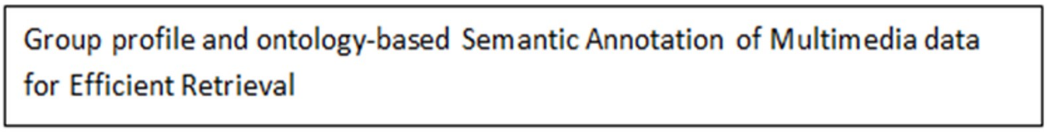


Figure 3.5: Dataset with Jaccard Similarity

3.5 Parameter Extraction

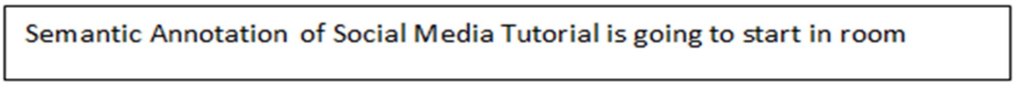
After getting the list of all papers, the next step is to extract information from the paper profiles for further experiments. We extract paper metadata (title, Author name), paper category and keywords. After getting all parameters, one by one we compare all parameters to tweet to get most relevant tweet related to focus paper. At the end we selected title of paper for further work. We analyzed that other parameters like Author name, keywords etc may be same of two or more papers but title of paper always distinct. So to continue experiments titles are extracted. In figure 3.6 title are shown



Group profile and ontology-based Semantic Annotation of Multimedia data for Efficient Retrieval

Figure 3.6: Extracted Title

We get tweets from domain experts that are related to papers. In figure 3.7 tweet is shown



Semantic Annotation of Social Media Tutorial is going to start in room

Figure 3.7: Extracted Tweet

3.6 Proposed Technique

The State of art approaches used similarity technique to identify relevant data. Our work is identifying tweets that are relevant to the research paper. For this purpose, paper title and tweets are compared using our proposed technique in which paper title and tweet is matching base on sequence of words. As much as more they have subsequence, have more similarity. To calculate sequence base similarity use this formula

$$R(P_t, T_t) = \frac{\sum_{i=1}^{m}(SL)^2}{P_w * T_w} + \frac{\sum_{j=1}^{n}(M_{iw})_j}{(P_w - \sum_{i=1}^{m} Wcms_i + T_w - \sum_{i=1}^{m} Wcms_i)^2} \tag{3.3}$$

where

- R = Relevance

- P_t = Paper title
- T_t = Tweet title
- SL = Matched Sequence length
- P_w = Length of paper title
- T_w = Length of tweet title
- M_{iw} = Matched individual words
- $Wcms_i$ = Total matched Sequence length

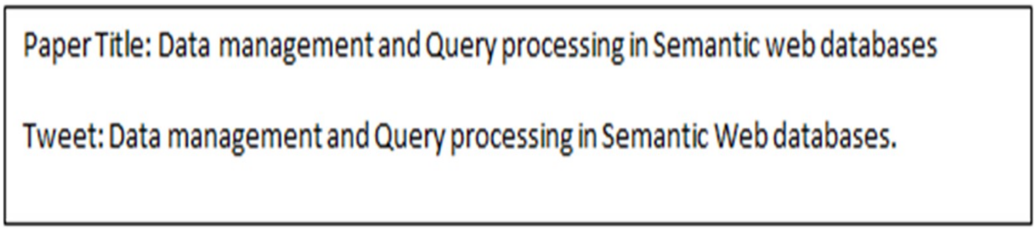
3.6.1 Variations

Here are the some variations for calculating sequence based similarity between paper title and tweet.

- Full match
- Matched as sub sequences
- Matched as individual words
- Matched as both sub sequences and individual words

Now each scenario has been explained with example of each variation

- **Full Match**



Paper Title: Data management and Query processing in Semantic web databases

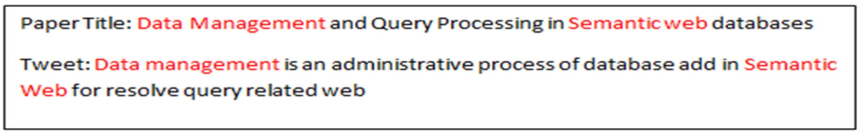
Tweet: Data management and Query processing in Semantic Web databases.

Figure 3.8: Full Match

In figure 3.8, all words of tweet and paper title are matched and we get value 1.

$$= \frac{7^2}{7+7} + 0 = 1$$

- **Matched as Sub Sequences**



Paper Title: Data Management and Query Processing in Semantic web databases

Tweet: Data management is an administrative process of database add in Semantic Web for resolve query related web

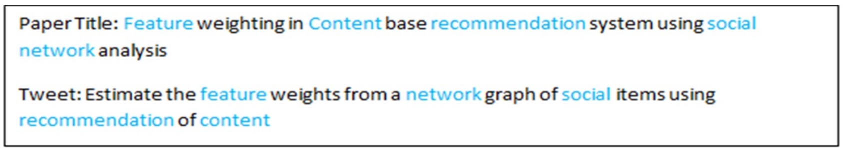
FIGURE 3.9: Subsequence Match

In figure 3.9, paper title and tweets words are matched as sub sequence. Two subsequences are made.

$$= \frac{2^2 + 2^2}{7 * 12} + 0 = 0.04$$

Take square of both sub-sequences and divided by total words of tweet and title. There is no single words matched therefore add 0 and calculate the result.

- **Matched as individual words**



Paper Title: Feature weighting in Content base recommendation system using social network analysis

Tweet: Estimate the feature weights from a network graph of social items using recommendation of content

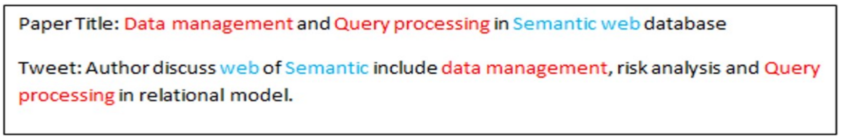
FIGURE 3.10: Individual words match

In figure 3.10, distinct words are matched between paper title and tweet.

$$= 0 + \frac{5}{(9-0) \cup (9-0))} = \frac{5}{(18)^2} = 0.015$$

There is no sub-sequences therefore add 0 and five individual words are matched.

- **Matched with both variation individual and sub sequence**



Paper Title: Data management and Query processing in Semantic web database

Tweet: Author discuss web of Semantic include data management, risk analysis and Query processing in relational model.

FIGURE 3.11: Individual and Sub sequence match

In figure 3.11, both variations are fulfilled. Words are matched as subsequence and individually.

$$\frac{2^2+2^2}{7*10}+\frac{2}{((7-4)+(10-4))^2}=\frac{4}{70}+\frac{2}{(9)^2}=0.024$$

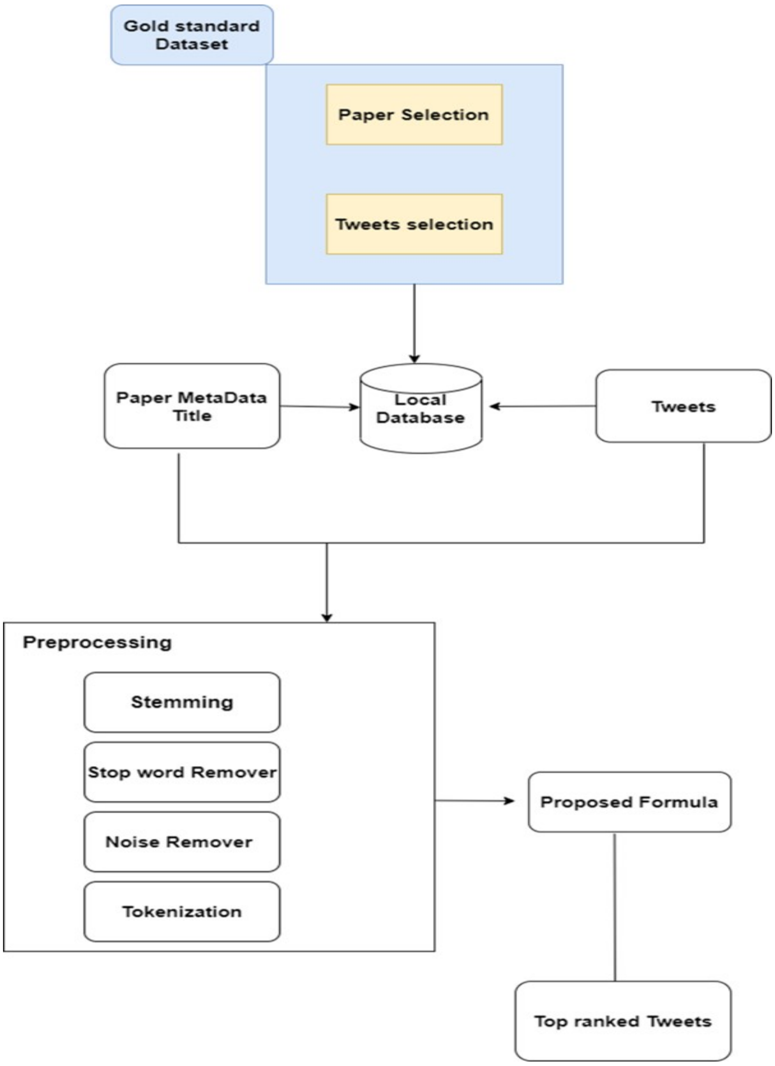


FIGURE 3.12: Dataset with proposed formula

In order to evaluate the proposed technique, evaluation measures are used.

3.7 Evaluation Measures

Precision and recall are the measures of understanding the results of the relevance. Both are important evaluation metrics that are used to evaluate the performance of different systems like classification or information retrieval [57].

3.7.1 Precision

Precision is defined as the number of relevant documents (items that are retrieved are useful for the users and fulfill his query) over the number of total retrieved documents.

$$Precision = \frac{Relevant\ retrieved\ documents}{Total\ retrieved\ documents} \tag{3.4}$$

Precision is the ratio of correct results that are divided by the all returned results.

3.7.2 Recall

Recall is the ratio of relevant retrieved documents over the total relevant documents. Recall measures that a particular query execution system is able to retrieve the related items that the user is interested in seeing.

$$Recall = \frac{Relevant\ retrieveddocuments}{Total\ relevant\ documents} \tag{3.5}$$

In the example of search text in the set of documents recall is the fraction of true results that should be divided by the number of results that should be returned.

3.7.3 F-measure

F-measure combines the recall and precision, is the harmonic mean of recall and precision.

$$F = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.6}$$

F-measure is more useful and best than accuracy, especially if your class divisions are uneven. It is the measure of the test's accuracy [58].

All three measures are applied and comparison have been made on returned results by the proposed approaches Cosine, Jaccard and State-of-the-art approach of Sharif et al [6].

Chapter 4

Experiments and Results

This chapter provides in depth detail of experimental setup and results achieved by applying methodology. Moreover, comparison of similarity techniques with proposed technique is presented in the chapter.

4.1 Dataset Collection

Dataset used in experiments contain tweets and paper's metadata. The dataset has been acquired Sharif et al (Sharif et al, 2015). There was no benchmark available for testing or evaluation of techniques. Gold set was created through ACM classification system by extracting the topics. Different domain experts were available. 60 domain experts were selected who published their papers in relevant research topic. Five research papers were taken form domain experts and requested to provide ten related tweets according to paper metadata. So total Gold set contained 220 research papers and 2957 tweets.

Table 4.1: JUCS Dataset Records

Research Papers	221
Tweets	2957

4.2 Metadata Extraction

The next step was extraction of metadata from the available dataset. Available dataset contains paper metadata (Title, Author, keywords and Category). We use only paper Title for execution of our experiments. This process is done manually. 220 paper's titles have been extracted and save in separate column. Title is distinct item of research papers therefore used for comparison with tweets to get relevant tweet for knowledge discovery.

4.3 Preprocessing

It plays an important role for achieving good results. In this stage all noise is removed. All features needed to be clean. Following are points that are used in preprocessing.

4.3.1 Tokenization

We tokenized the tweets and paper's title. String Tokenizer Class is used to make tokens of given data. Tokenization has converted all strings into tokens (single words). Here is an example of tokenization in which made tokens of one paper title.

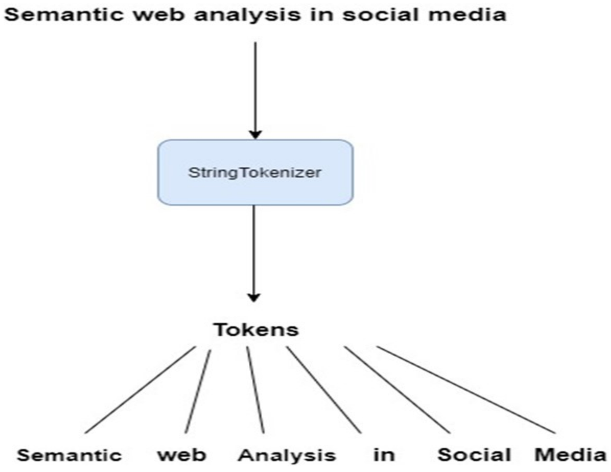


FIGURE 4.1: Paper's Title convert in tokens

Same in the case of tweets, all tweets are converted in tokens. These tokens are added in list and apply stop words and stemming on this list.

4.3.2 Stop Words Remover

Stop words removed from the dataset on the base of standard list of rainbow Statistical text[3]. Before stop words removing, dataset was the large size but afterwards the size of dataset decreased. There are only meaningful token are left. In this process, all the words are initializing in array string then we have sorted them in ascending order for making them ready for efficient comparison.

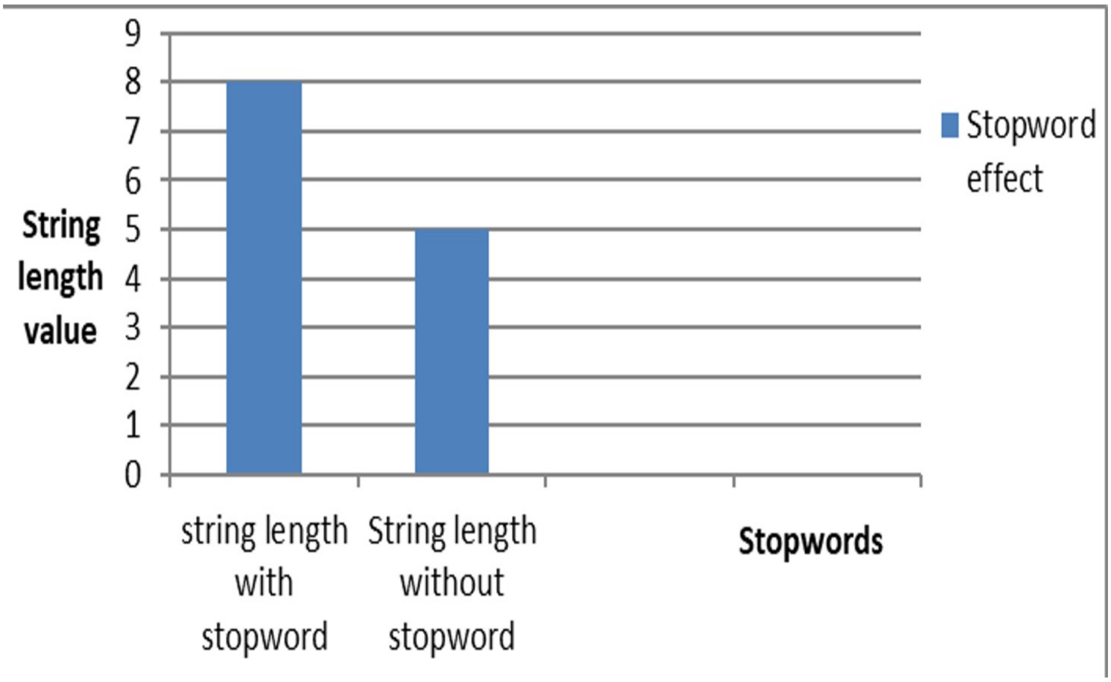


Figure 4.2: Stop words removal

4.3.3 Stemming

In this process, stemming is performed on the dataset which convert the words to their root form. Snowball stemmer is used [51] .Snowball is a language of string processing that is designed for the creating stemming algorithms for information

[3] Gist.github.com

retrieval. It decreases the dictionary size. After performing the preprocessing steps, clean and reduced data is available for further processing.

4.4 Similarities Techniques

Cosine and Jaccard are two similarity techniques that are used to calculate the similarity score between papers' Title and tweets and determine how much tweets are relevant to paper.

4.4.1 Cosine Similarity

Cosine Similarity is working on the vector space model. We made vector of each string (paper's title and tweets). TF.IDF vector is used to calculate the term frequencies and inverse document frequencies. Term frequency is calculated for each token that is available in the string. Afterwards, Cosine measure was used to calculate the similarity. We have 220 research papers and 2957 tweets. Cosine similarity is calculated against on 220x2957=650,540 rows of dataset. Retrieval result assigns a value to tweets that show which tweets are most relevant to paper. These resulted tweets are matched against the Gold Standard Dataset to check the performance of technique.

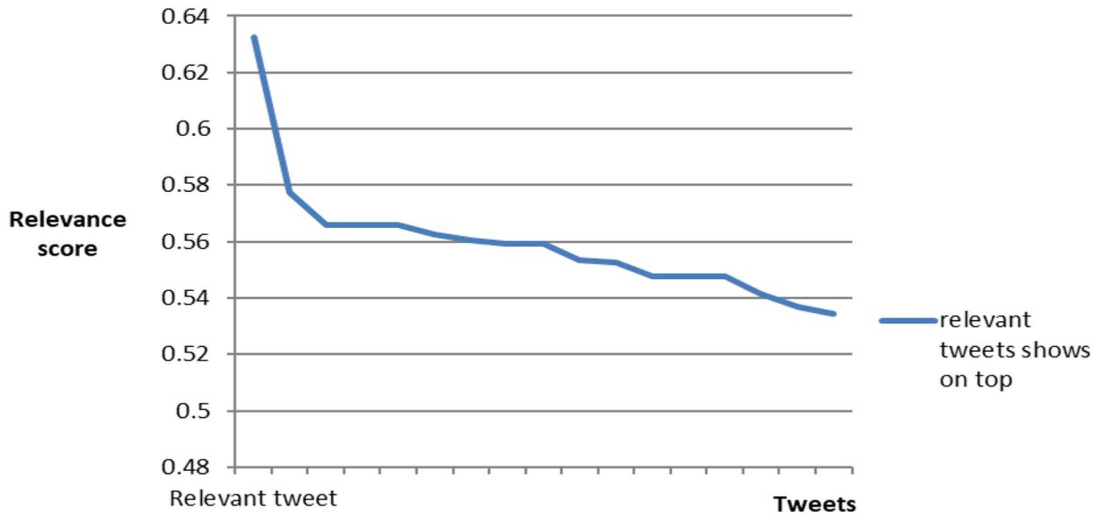


FIGURE 4.3: Relevant tweets shows on the top

This is the example of one paper's title. When applied Cosine similarity on the dataset most relevant tweets comes on top.

We calculated results on Gold standard dataset with benchmark dataset top 3% selected for calculation of precision, recall and f-measure. The details of the evaluation are shown in the figure 4.4. Cosine similarity was able to bring most of the relevant tweets but the retrieved results contained some noise as well therefore precision is not very high but recall is significant. This is what cosine similarity is known for other domains like search engine retrieved results. The recall of cosine is better than the precision is not good.

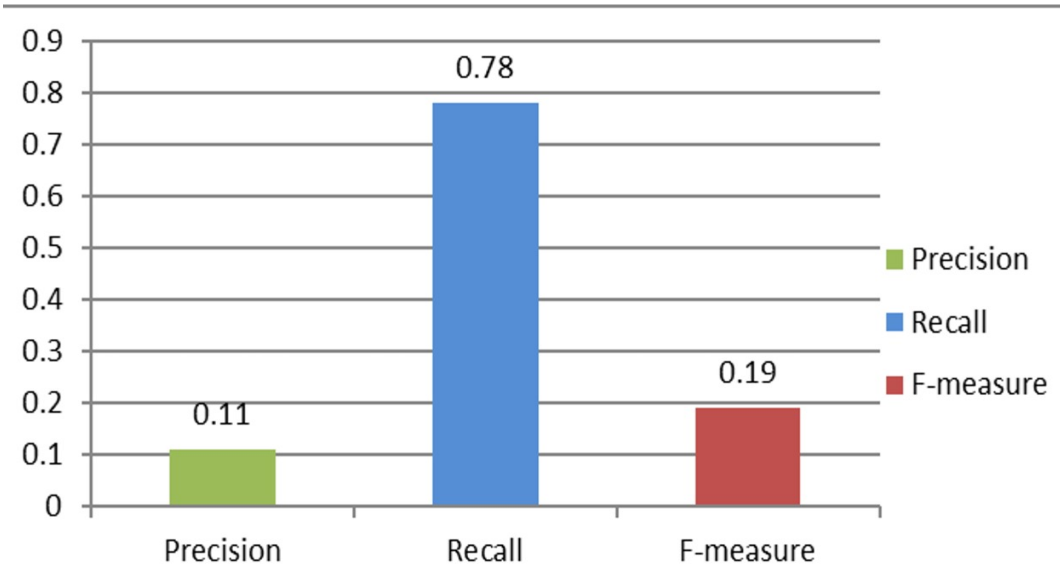


FIGURE 4.4: Cosine Similarity

4.4.2 Jaccard Similarity

Jaccard similarity contains two sets union and intersection. We build two functions for calculating intersection and union of given dataset. Jaccard Similarity calculated the score between paper title and tweets. And ranked tweets according to Similarity. Output file is compared to benchmark dataset. And results are shown in figure 4.5. The recall of Jaccard similarity is high than precision. Therefor it retrieved relevant results but irrelevant results also retrieved. Its mean retrieved results of Jaccard similarity contained some noise.

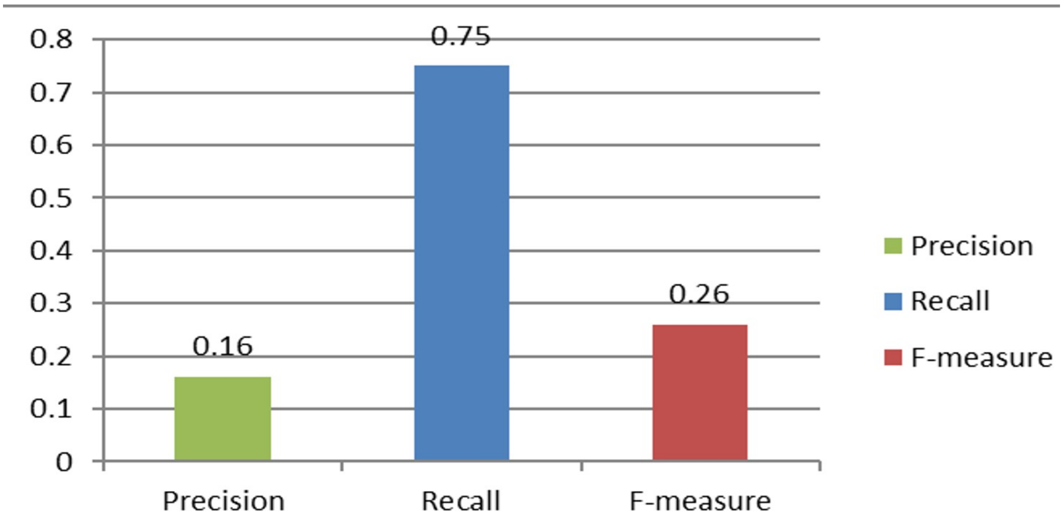


FIGURE 4.5: Jaccard Similarity

4.4.3 Proposed Technique

In our proposed technique, we compare paper title with tweets based on sub sequences. One or many sub-sequences can be present in the title string. If tweet has many sub-sequences according to paper title then those tweets are ranked high and considered most relevant to paper title.

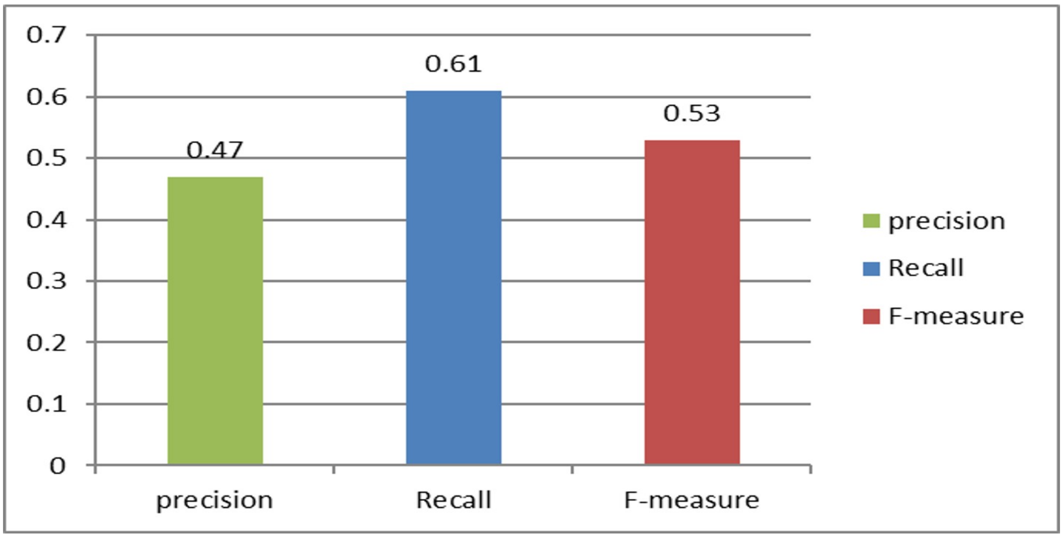


FIGURE 4.6: Proposed formula

4.5 Evaluation

The formula of Precision, Recall and F-measure is used for evaluate the techniques. Cosine, Jaccard and proposed formula based on sub-sequences are using to calculate relevance between paper title and tweets. The reason of using these techniques is because of their high usage in literature to calculate similarity between source and target content. The result of top 3 is reported in this section. Evaluation is done by answering the following question.

RQ1: What is the best approach to find out the relevance between Paper's title and tweet?

In figure 4.7, 4.8 and 4.9 shown the performance of approaches using the evolution measure.

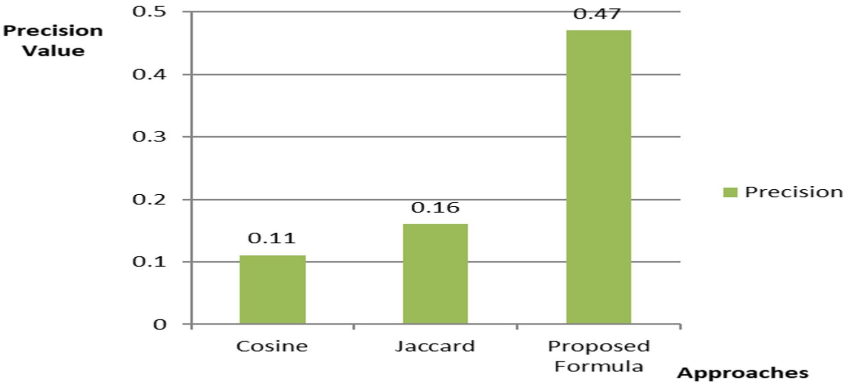


Figure 4.7: Precision of all techniques

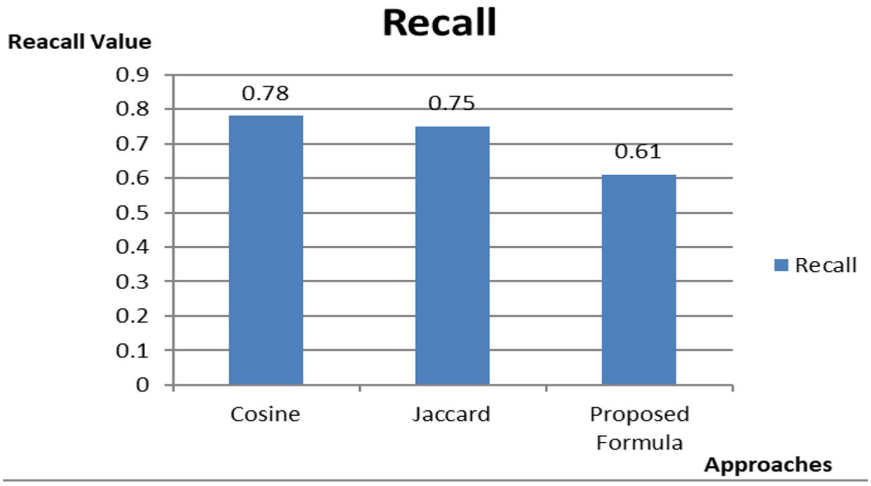


Figure 4.8: Recall of all techniques

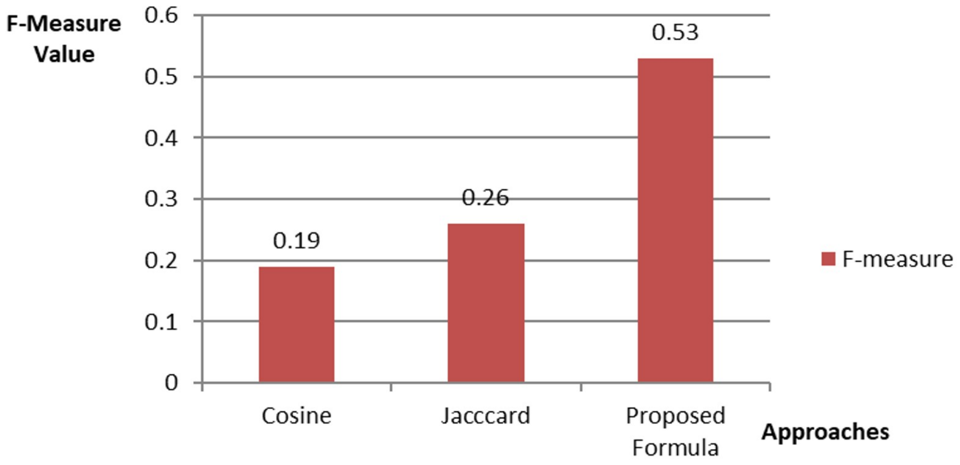


Figure 4.9: F-measure of all techniques

Precision, recall and f-measure is calculated of all techniques. As figure 4.7 shown precision values of all approaches. The value of Cosine similarity is 0.11, Jaccard is 0.16 and proposed approach is 0.47. Our proposed technique has high precision value means better than state-of-the-art-approaches. As the precision of the proposed approach is higher than the state-of-the-art approaches, therefore, this approach is best suited in the situation when one needs more relevant results in the retrieved results.

As figure 4.8 shown the Recall value of all techniques where recall value of the Proposed is 0.61. The recall value of cosine technique and jaccard are 0.78 and 0.75 respectively. The Cosine has high recall value as compared to both Jaccard and the proposed formula. Its mean Cosine similarity returns more relevant results but also introduce more noise as compared to the proposed approach.

As figure 4.9 shown the E-measure values that is harmonic mean of precision and recall. F-measure shows the combined value of precision and recall. F-measure is not biased to any high value from precision or recall. Therefore, the F-measure is considered more reliable to measure the effectiveness of both precision and recall. The F-measure cannot be high if one of the values from precision or recall is high and the second value is low. The F-measure of the proposed approach is better than both cosine and jaccard. This means the proposed approach has improve significant precision and the recall difference with other approaches is negligible.

RQ2: What will be the effect if we compute relevance between paper title and tweets based on sub-sequences by incorporating the dissimilarity score?

As we discussed above if we compute relevance base on sub-sequences and by incorporating the dissimilarity, then it returned more relevant results. The score of precision significantly improved while recall in comparisons with other approaches reduced a little bit. However, overall, the proposed approach has outperformed in terms of F-measure.

4.6 Comparison

In this section comparison are performed with Sharif et al [6]. They have proposed technique in which they performed lexical matching to recommend relevant tweet on current learning topic. They work on the same Gold standard dataset. They analyzed tweets against extended key terms. But they did not use similarity, dissimilarity and sub-sequences to calculate the relevance between paper and tweet. They used different metadata features (keywords, Category and Title) in their work. But title has high precision and recall than other features.

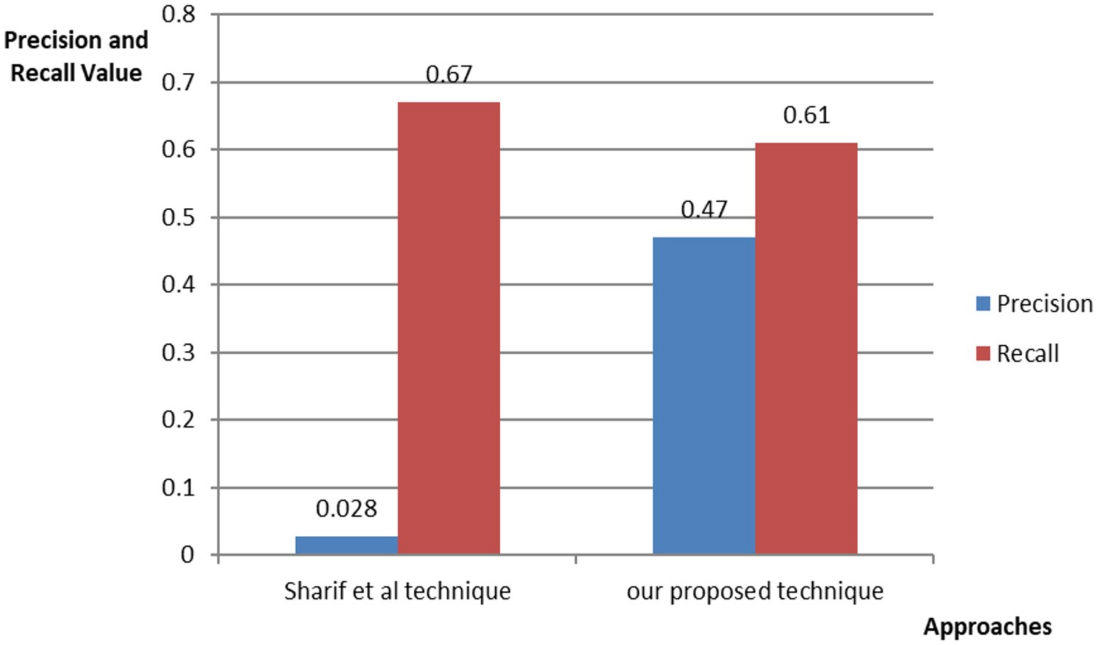


FIGURE 4.10: comparisons between techniques

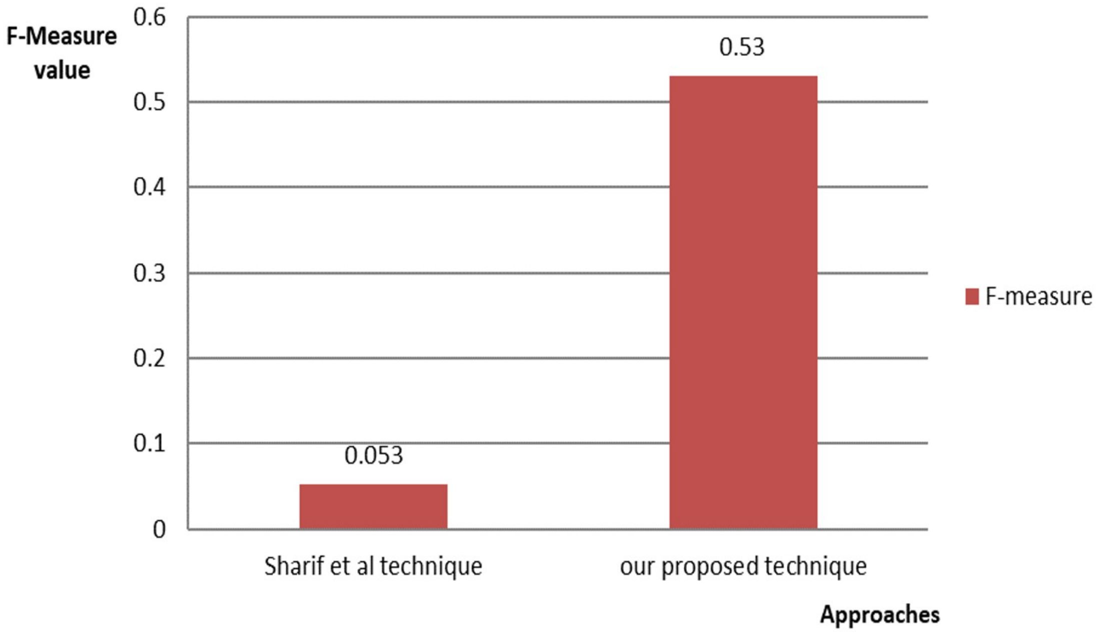


FIGURE 4.11: comparisons between techniques

As shown in figure 4.10, the precision value of Sharif's technique was 0.028 and recall is 0.67.

The precision of our proposed technique remained is 0.47 and recall remained 0.61. Therefore the proposed technique has performed better in term of precision and both techniques are performed approximately equally in term of recall.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Recent and classical papers have been critically reviewed in this thesis about E-Learning and their recommender system that identify the similar content. Recommender systems are helpful to E-Learnes to get relevant resources. In state of the art research, books and research papers are widely used as recommendation content. However social media has been little exploited for relevant recommendation to e learners. Therefore, we use twitter based identification in E-Learning system. Twitter is a micro blogging service. Literature review has shown that twitter has a great impact on education. Literature review concluded that there is a research gap to utilize twitter in E-Learning. To fulfill this gap we have worked on this area and use Sharif et al Gold standard dataset that contain 220 research papers and 2957 tweets. The aim of this research is to retrieve tweets that are relevant to paper titles. We have proposed innovative formula that calculates the Similarity, Dis-similarity and sub-sequence based similarity between paper and tweets.

The thesis highlights the strength and limitation of approaches (Cosine, Jaccard, proposed formula, and Sharif et al). The proposed formula has outperformed by gaining F-measure is 0.53 as compared to F-measure of Cosine is 0.19 and Jaccard is 0.26 respectively furthermore the precision of the proposed approach is higher. This formula can be used where quality recommendations are required. However the recall of the proposed formula is slightly low but this difference is not significant.

5.2 Future Work

In this thesis, we have used the Gold standard dataset from one domain that is computer science. Therefore one future direction could be to use the dataset of diversified domains like Mathematics, Bioinformatics, and Physics etc. We used only one best performing metadata that is title for identifying relevant tweets for research papers. Therefore one possible direction could be to explore the combination of different metadata elements.

Bibliography

[1] S. Bethard and D. Jurafsky, "Who should i cite: learning literature search models from citation behavior," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, 2010, pp. 609–618.

[2] D. Yu, D. Xu, D. Wang, and Z. Ni, "Hierarchical topic modeling of twitter data for online analytical processing," *IEEE Access*, vol. 7, pp. 12 373–12 385, 2019.

[3] C. Nishioka, J. Hauke, and A. Scherp, "Influence of tweets and diversification on serendipitous research paper recommender systems," *PeerJ Computer Science*, vol. 6, p. e273, 2020.

[4] S. Shishehchi, S. Y. Banihashem, and N. A. M. Zin, "A proposed semantic recommendation system for e-learning: A rule and ontology based e-learning recommendation system," in *2010 international symposium on information technology*, vol. 1. IEEE, 2010, pp. 1–5.

[5] S. A. Abdalmenem, S. S. Abu-Naser, M. J. Al Shobaki, and Y. M. Abu Amuna, "Relationship between e-learning strategies and educational performance efficiency in universities from senior management point of view," 2019.

[6] N. Sharif, M. T. Afzal, and A. Muhammad, "Semantic based e-learning recommender system," *European Scientific Journal*, 2015.

[7] J. Wrubel, D. White, and J. Allen, "High-fidelity e-learning: Sei's virtual training environment (vte)," Carnegie-Mellon Univ Pittsburgh PA Software Engineering Inst, Tech. Rep., 2009.

[8] M. N. Yakubu and S. Dasuki, “Assessing elearning systems success in nigeria: An application of the delone and mclean information systems success model,” *Journal of Information Technology Education: Research*, vol. 17, pp. 183–203, 2018.

[9] H. M. Selim, “Critical success factors for e-learning acceptance: Confirmatory factor models,” *Computers & education*, vol. 49, no. 2, pp. 396–413, 2007.

[10] T. M. Hameed, Z. B. H. Hassan, and R. Sulaiman, “Is social network an effective e-learning tool: A survey,” *Middle-East Journal of Scientific Research*, vol. 23, no. 1, pp. 119–126, 2015.

[11] 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.

[12] N. Hoic-Bozic, M. H. Dlab, and V. Mornar, “Recommender system and web 2.0 tools to enhance a blended learning model,” *IEEE Transactions on education*, vol. 59, no. 1, pp. 39–44, 2015.

[13] M. K. Khribi, M. Jemni, and O. Nasraoui, “Automatic personalization in e-learning based on recommendation systems: An overview,” in *Intelligent and Adaptive Learning Systems: Technology Enhanced Support for Learners and Teachers.* IGI Global, 2012, pp. 19–33.

[14] K. I. Ghauth and N. A. Abdullah, “An empirical evaluation of learner performance in e-learning recommender systems and an adaptive hypermedia system,” *Malaysian Journal of Computer Science*, vol. 23, no. 3, pp. 141–152, 2010.

[15] A. Klašnja-Milićević, M. Ivanović, B. Vesin, and Z. Budimac, “Enhancing e-learning systems with personalized recommendation based on collaborative tagging techniques,” *Applied Intelligence*, vol. 48, no. 6, pp. 1519–1535, 2018.

[16] I. Guy, N. Zwerdling, I. Ronen, D. Carmel, and E. Uziel, "Social media recommendation based on people and tags," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 2010, pp. 194–201.

[17] J. Bercovici, "Who coined "social media"? web pioneers compete for credit,"," *Forbes. Disponível*, 2010.

[18] Q. Anderson and L. Rainie, "Millennia's will benefit and suffer due to their hyper connected lives. 2012," *The Pew Research Center's Internet and American Life Project*, 2012.

[19] W. M. Al-Rahmi and A. M. Zeki, "A model of using social media for collaborative learning to enhance learners' performance on learning," *Journal of King Saud University-Computer and Information Sciences*, vol. 29, no. 4, pp. 526–535, 2017.

[20] B. Shapira, L. Rokach, and S. Freilikhman, "Facebook single and cross domain data for recommendation systems," *User Modeling and User-Adapted Interaction*, vol. 23, no. 2-3, pp. 211–247, 2013.

[21] M. Ebner, C. Lienhardt, M. Rohs, and I. Meyer, "Microblogs in higher education–a chance to facilitate informal and process-oriented learning?" *Computers & Education*, vol. 55, no. 1, pp. 92–100, 2010.

[22] J. C. Dunlap and P. R. Lowenthal, "Instructional uses of twitter," *Chapter*, vol. 8, pp. 45–50, 2009.

[23] A. Acar and N. Kimura, "Twitter as a tool for language learning: The case of japanese learners of english," *Special Issue of International Journal of the Computer, the Internet and Management*, vol. 19, no. 1, p. 14, 2012.

[24] I. Ha and C. Kim, "The research trends and the effectiveness of smart learning," *International Journal of Distributed Sensor Networks*, vol. 10, no. 5, p. 537346, 2014.

[25] R. Junco, C. M. Elavsky, and G. Heiberger, "Putting twitter to the test: Assessing outcomes for student collaboration, engagement and success," *British Journal of Educational Technology*, vol. 44, no. 2, pp. 273–287, 2013.

[26] E. Kassens-Noor, "Twitter as a teaching practice to enhance active and informal learning in higher education: The case of sustainable tweets," *Active Learning in Higher Education*, vol. 13, no. 1, pp. 9–21, 2012.

[27] I. Ha and C. Kim, "Understanding user behaviors in social networking service for mobile learning: a case study with twitter," *Malaysian Journal of Computer Science*, vol. 27, no. 2, pp. 112–123, 2014.

[28] C. Evans, "T witter for teaching: Can social media be used to enhance the process of learning?" *british Journal of educational technology*, vol. 45, no. 5, pp. 902–915, 2014.

[29] P. Reed, "Hashtags and retweets: using twitter to aid community, communication and casual (informal) learning," *Research in Learning Technology*, vol. 21, 2013.

[30] N. Jonnalagedda and S. Gauch, "Personalized news recommendation using twitter," in *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, vol. 3. IEEE, 2013, pp. 21–25.

[31] A. Hürriyetoğlu, J. Wagemaker, N. Oostdijk, and A. van den Bosch, "Analysing the role of key term inflections in knowledge discovery on twitter," 2016.

[32] S. Yoon and S. Bakken, "Methods of knowledge discovery in tweets," in *NI 2012: 11th International Congress on Nursing Informatics, June 23-27, 2012, Montreal, Canada.*, vol. 2012. American Medical Informatics Association, 2012.

[33] L. Stojanovic, S. Staab, and R. Studer, "elearning based on the semantic web," in *WebNet2001-World Conference on the WWW and Internet.* Citeseer, 2001, pp. 23–27.

[34] J. Lennon and H. A. Maurer, "Why it is difficult to introduce e-learning into schools and some new solutions," *J. UCS*, vol. 9, no. 10, p. 1244, 2003.

[35] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert systems with applications*, vol. 33, no. 1, pp. 135–146, 2007.

[36] N. Sharif, T. Afzal, and D. Helic, "Learning management systems," *The IPSI BgD Internet Research Society*, p. 44.

[37] M. J. Pazzani and D. Billsus, "Content-based recommendation systems," in *The adaptive web.* Springer, 2007, pp. 325–341.

[38] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Communications of the ACM*, vol. 35, no. 12, pp. 61–70, 1992.

[39] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins, "Eigentaste: A constant time collaborative filtering algorithm," *information retrieval*, vol. 4, no. 2, pp. 133–151, 2001.

[40] Y. Ren, G. Li, and W. Zhou, "A survey of recommendation techniques based on offline data processing," *Concurrency and Computation: Practice and Experience*, vol. 27, no. 15, pp. 3915–3942, 2015.

[41] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen, "Collaborative filtering recommender systems," in *The adaptive web.* Springer, 2007, pp. 291–324.

[42] W. Chen, Z. Niu, X. Zhao, and Y. Li, "A hybrid recommendation algorithm adapted in e-learning environments," *World Wide Web*, vol. 17, no. 2, pp. 271–284, 2014.

[43] D. Helic, "Managing collaborative learning processes in e-learning applications," in *2007 29th International Conference on Information Technology Interfaces.* IEEE, 2007, pp. 345–350.

[44] W. Yuan, D. Guan, Y.-K. Lee, S. Lee, and S. J. Hur, "Improved trust-aware recommender system using small-worldness of trust networks," *Knowledge-Based Systems*, vol. 23, no. 3, pp. 232–238, 2010.

[45] G. Pitsilis and L. F. Marshall, "Modeling trust for recommender systems using similarity metrics," in *IFIP International Conference on Trust Management.* Springer, 2008, pp. 103–118.

[46] P. Dwivedi and K. K. Bharadwaj, "Effective trust-aware e-learning recommender system based on learning styles and knowledge levels," *Journal of Educational Technology & Society*, vol. 16, no. 4, pp. 201–216, 2013.

[47] Dwivedi and K. K. Bharadwaj, "Effective resource recommendations for e-learning: A collaborative filtering framework based on experience and trust," *International Conference on Computational Intelligence and Information Technology*, pp. 166–170, 2011.

[48] S. M. Kumar, K. Anusha, and K. S. Sree, "Semantic web-based recommendation: Experimental results and test cases," *International Journal of Emerging Research in Management & Technology*, vol. 4, no. 6, pp. 215–222, 2015.

[49] L. aDepartament de Llenguatges i Sistemes Informàtics, "Taking advantage of semantics in recommendation systems," in *Artificial Intelligence Research and Development: Proceedings of the 13th International Conference of the Catalan Association for Artificial Intelligence*, vol. 220. IOS Press, 2010, p. 163.

[50] W. Malik, "Visual semantic web. ontology based e-learning management system," 2008.

[51] A. G. Jivani *et al.*, "A comparative study of stemming algorithms," *Int. J. Comp. Tech. Appl*, vol. 2, no. 6, pp. 1930–1938, 2011.

[52] 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.

[53] K. Park, J. S. Hong, and W. Kim, “A methodology combining cosine similarity with classifier for text classification,” *Applied Artificial Intelligence*, vol. 34, no. 5, pp. 396–411, 2020.

[54] L. Gravano, P. G. Ipeirotis, H. V. Jagadish, N. Koudas, S. Muthukrishnan, L. Pietarinen, and D. Srivastava, “Using q-grams in a dbms for approximate string processing,” *IEEE Data Eng. Bull.*, vol. 24, no. 4, pp. 28–34, 2001.

[55] N. Koudas, A. Marathe, and D. Srivastava, “Flexible string matching against large databases in practice,” in *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, 2004, pp. 1078–1086.

[56] M. Besta, R. Kanakagiri, H. Mustafa, M. Karasikov, G. Rätsch, T. Hoefler, and E. Solomonik, “Communication-efficient jaccard similarity for high-performance distributed genome comparisons,” in *2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 2020, pp. 1122–1132.

[57] D. M. Powers, “Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation,” 2011.

[58] J. Guo, A. Nomura, R. Barton, H. Zhang, and S. Matsuoka, “Machine learning predictions for underestimation of job runtime on hpc system,” in *Asian Conference on Supercomputing Frontiers*. Springer, Cham, 2018, pp. 179–198.