

# **Relevant Information Extraction from Twitter during Time-Critical Situations**



**Author**

**Zoha Sheikh**

**MS(IT)-15**

**NUST201463936MSEEC60014F**

**Supervisor**

**Dr. Sharifullah Khan (T.I)**

**Department of Computing**

A thesis submitted in partial fulfillment of the requirements for the degree of  
Masters of Science in Information Technology (MS IT)

**SCHOOL OF ELECTRICAL ENGINEERING & COMPUTER  
SCIENCE (SEECs),  
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY  
(NUST), ISLAMABAD  
(July, 2018)**

# Approval

It is certified that the contents and form of the thesis entitled “Relevant Information Extraction from Twitter during Time-Critical Situations” submitted by Ms. Zoha Sheikh have been found satisfactory for the requirement of the degree.

Advisor: Dr. Sharifullah Khan

Signature: \_\_\_\_\_

Date: \_\_\_\_\_

Committee Member 1: **Dr. Muhammad Ali Tahir**

Signature: \_\_\_\_\_

Date: \_\_\_\_\_

Committee Member 2: **Dr. Asad Shah**

Signature: \_\_\_\_\_

Date: \_\_\_\_\_

Committee Member 3: **Dr. Muhammad Imran**

Signature: \_\_\_\_\_

Date: \_\_\_\_\_

# THESIS ACCEPTANCE CERTIFICATE

Certified that the final copy of MS thesis written by Ms. Zoha Sheikh, (Registration No NUST201463936MSEEC60014F), of MSIT-15 has been vetted by undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: \_\_\_\_\_

Name of Supervisor:

Dr. Sharifullah Khan

Date: \_\_\_\_\_

Signature (HOD): \_\_\_\_\_

Date: \_\_\_\_\_

Signature (Dean/Principal):

\_\_\_\_\_

Date: \_\_\_\_\_

# Abstract

People suffering during disasters and emergencies; look for quick feedback to their queries. People post irrelevant information and there is a sudden rise in this activity during high impact events. Government and relief organizations look for situational awareness information to launch relief operations but due to increase in irrelevant information, they would not be able to take necessary measures on time. Existing studies explored text or user relevancy has using GloVe, pseudo relevance feedback and rule based approaches in Twitter. GloVe approach has shown very low performance. Existing approaches have used unstructured and redundant tweets and no measure has been taken to remove the redundant tweets. Existing systems have focused on text relevancy or user relevancy independently but none of the system has provided the both. The main objective of this research is to increase the content relevancy and finding out sources most relevant to a topic. A novel approach has been proposed to provide access to the relevant information on Twitter. There are two major parts, first is identifying the relevant tweets and second is identifying relevant sources. In the first part, automated technique has been proposed to make the system dynamic and independent enough to prepare its ground truth. To achieve this automation, genism third party domain specific embeddings are used to expand the initial queries and based on the relevance feedback mechanism relevant messages are shown to the user. This continuous relevance feedback would help in generating the ground truth automatically. In the second part text relevancy score taken from the first part, user specific characteristics of sources and tweet specific characteristics have helped in evaluating the source relevancy by classifying them in to different ranks. Initial experiments and user studies have been performed using a real world disaster dataset that shows the significance of the proposed approach. Evaluation of the system is performed using different measures like mean average precision and Normalized Discounted Cumulative Gain (NDCG). The mean average precision of the proposed system is 89% while the NDCG score is 95%.

# Certificate of Originality

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics which has been acknowledged.

Author Name: **Ms. Zoha Sheikh**

Signature: \_\_\_\_\_

# Acknowledgment

I owe my deepest gratitude to Allah Almighty. I also like to acknowledge the efforts of my parents to make such opportunity available for me. I would like to give special thanks to my Supervisor Dr. Sharifullah Khan and GEC members for guiding me throughout this struggle. I also would like to thank Dr. Muhammad Imran and Qatar computing research institute (QCRI) for providing the dataset. I would like to appreciate the efforts made by Dr. Madiha Liaqat for helping in the review of thesis. I want to thank all my lovely friends and specially Ms. Hira Masood for supporting and sharing her valuable advices. I would like to thank everyone who directly or indirectly contributed in this journey.

# Table of Contents

Chapter 1.....	13
Introduction and Motivation .....	13
1.1 Introduction .....	13
1.1.1 Influence of Social Media on Everyday Life .....	14
1.1.2 Relevant Information Extraction.....	15
1.1.3 Relevant Source Extraction.....	17
1.2 Motivation.....	18
1.3 Problem Statement.....	18
1.4 Research Objectives.....	19
1.5 Organization of Thesis.....	19
Chapter 2.....	20
Literature Review.....	20
2.1 Background .....	20
2.1.1 Relevance Based Prediction.....	21
2.1.2 Information Extraction using word2vec.....	21
2.1.3 Cosine Similarity .....	21
2.1.4 Rocchio Algorithm.....	22
2.1.5 Twitter Stream API .....	23
2.1.6 Relevant Source Extraction.....	23
2.1.6 Relevant Tweets Extraction .....	23
2.2 Related Work .....	23
2.6 Critical Analysis .....	28
Chapter 3.....	35
Design and Methodology .....	35
3.1.1 Tweets Pre-processing.....	36
3.1.1.1 Drop Duplicates .....	37
3.1.1.2 Topic / Hashtag .....	38
3.1.1.3 Lowercase .....	38
3.1.1.4 Tweet Language.....	39
3.1.1.5 HTML Encoding.....	39

3.1.1.6 Special Char/ Extra Spaces/ Stop Words Removal.....	39
3.1.1.7 Lemmatization .....	40
3.1.2 Tweets Relevance Prediction.....	40
3.1.3 Tweets Source Features Extraction.....	47
3.1.4 Relevance Ranking .....	48
3.1.4.1 Relevant Source Ranking.....	48
3.1.4.1.1 Source Popularity Measurement .....	48
3.1.4.1.2 Ranking of Sources .....	51
3.1.4.2 Relevant Tweets Ranking.....	53
Chapter 4.....	55
Experiments and Results.....	55
4.1 Dataset Specifications.....	55
4.2 Dataset Annotation.....	55
4.3 Performance Measures.....	58
4.4 Graphical Representation of Precision, Recall and F1 .....	61
4.4.1 PR Curve.....	61
4.4.2 ROC Curve.....	62
4.5 Classifiers Training .....	62
4.5.1 Support Vector Machines.....	62
4.5.2 Random Forest.....	63
4.5.3 Multinomial Naïve Bayes .....	63
4.5.4 10-Fold Cross Validation .....	64
4.6 Experiments and Results.....	64
4.6.1 Results of Jaccard Similarity Measure.....	64
4.6.2 Evaluation of Tweets Relevancy Prediction .....	67
4.6.3 Evaluation of Relevancy Feedback Classifier 60:40 .....	68
4.6.4 Evaluation of Relevancy Feedback Classifier 10-Folds .....	78
4.6.5 Evaluation of Relevant Source Extraction .....	87
4.6.6 Evaluation of Relevant Source Extraction using NDCG .....	90
4.6.7 Evaluation of Relevant Tweets Extraction.....	91
4.6.8 Evaluation of Relevant Tweets Extraction using NDCG.....	92
4.7 Comparative Evaluation.....	93
Chapter 5.....	94



Conclusion .....	94
5.1 Contributions .....	94
5.2 Discussion.....	94
5.3 Limitation and Future Work.....	96
Bibliography .....	97

# Table of Figures

Figure 1 Growth in use of Social Media for News [21].....	15
Figure 2 Cosine Similarity [8] .....	22
Figure 3 Block Diagram of Relevant Tweets and Relevant Source Extraction .....	36
Figure 4 Flow Diagram of Tweet Relevance Prediction System .....	40
Figure 5 Tweet Relevance Prediction System Diagram .....	44
Figure 6 Twitter Relevance Feedback System.....	46
Figure 7 Ground Truth for Evaluation of Source Ranking .....	56
Figure 8 Ground Truth for Evaluation of Tweets Ranking.....	57
Figure 9 PR Curve for MNB using TF .....	69
Figure 10 ROC Curve for MNB using TF .....	70
Figure 11 PR Curve for RF using TF.....	70
Figure 12 ROC Curve for RF using TF .....	71
Figure 13 PR Curve for SVM using TF.....	72
Figure 14 ROC Curve for SVM using TF.....	72
Figure 15 PR Curve for MNB using TF-IDF.....	74
Figure 16 ROC Curve for MNB using TF-IDF .....	74
Figure 17 PR Curve for RF using TF-IDF .....	75
Figure 18 ROC Curve for RF using TF-IDF.....	76
Figure 19 PR Curve for SVM using TF-IDF .....	76
Figure 20 ROC Curve for SVM using TF-IDF.....	77
Figure 21 PR Curve for MNB using TF .....	79
Figure 22 ROC Curve for MNB using TF .....	80
Figure 23 PR Curve for RF using TF.....	81
Figure 24 ROC Curve for RF using TF .....	81
Figure 25 PR Curve for SVM using TF.....	82
Figure 26 ROC Curve for SVM using TF.....	83
Figure 27 PR Curve for MNB using TF-IDF.....	83
Figure 28 ROC Curve for MNB using TF-IDF .....	84
Figure 29 PR Curve for RF using TF-IDF.....	85
Figure 30 ROC Curve for RF using TF-IDF.....	85

Figure 31 PR Curve for SVM using TF-IDF .....	86
Figure 32 ROC Curve for SVM using TF-IDF.....	87
Figure 33 Comparison between Base and Proposed Formula .....	89
Figure 34 Distribution of Sources in Each Bin .....	90

# Table of Tables

Table 1 Critical Analysis of Different Papers .....	29
Table 2 Research Gap Table for preprocessing steps .....	30
Table 3 Research Gap Table for tweet specific features.....	31
Table 4 Research Gap Table for user specific features.....	32
Table 5 Research Gap Table for Classifiers.....	33
Table 6 Research Gap Table for Evaluation Metrics .....	33
Table 7 Parameters of formula for Source Evaluation.....	50
Table 8 Relevant Tweets without using Jaccard .....	66
Table 9 Relevant Tweets using Jaccard .....	67
Table 10 Relevance Feedback System Precision and Recall TF (pre shows precision) .....	69
Table 11 Relevance Feedback System Precision and Recall TF-IDF (pre shows precision) .....	73
Table 12 Relevance Feedback System Precision and Recall TF using 10-Fold Validation (pre shows precision) .....	78
Table 13 Relevance Feedback System Precision and Recall TF-IDF using 10-Fold Validation (pre shows precision) .....	79
Table 14 Classification of Users (pre is precision and rec is recall).....	89
Table 15 Source Evaluation using NDCG .....	91
Table 16 Relevancy Based Classification.....	68
Table 17 Tweets Evaluation.....	92
Table 18 Tweets Evaluation using NDCG.....	92

# Chapter 1

## Introduction and Motivation

### 1.1 Introduction

Social networking sites such as Facebook and Twitter, have become very popular in the recent years [5]. People use these sites for a wide range of activities such as finding updates on an on-going event, looking for new friends, posting personal updates to get in touch with their family and friends, to learn about others' activities and events happening around and worldwide. Among others, the microblogging platform: Twitter has been used by a larger number of people. It provides an easy-to-use microblogging service to users where they can post short messages; which is also called tweets, of 140 characters length [3]. Tweets usually comprise textual content, URLs, mentions of other Twitter users and hashtags. Tweets posted by a user are received by his/her followers and by users who are interested in hashtags/keywords used in the tweets [1]. One can follow updates on a specific topic or an event by following one or more hashtags or keywords related to that event.

The adaptation of microblogging platforms such as Twitter has increased recently during crises, emergencies, and time-critical events [5]. Easy access to social networks provides ways to produce and retrieve information in different forms, such as textual messages, images, and videos. For instance, at the onset of a crisis event, people use social media platforms to fulfill their quick information needs. Access to critical information becomes more important, especially in the first few hours [13], when other information sources such as, traditional media and news channels are

available. Moreover, rapid access to important information can help humanitarian organizations gain situational awareness, early decision-making, and to launch relief efforts accordingly [15].

Recent studies have shown that during emergency situations, 45% of the users have started using social sites and readily click on links shared by their friends and spread the information without investigating whether the information is correct or not [16]. This can actually bring more harm than good if the information is not reliable. For example, through retweet, a message can immediately reach to wider audiences which will then impact larger real-world audiences. For instance, one tweet from a hacked AP account resulted in a dip in the stock market [18].

### **1.1.1 Influence of Social Media on Everyday Life**

Social media has a lot of positive impact on society and individuals, it can shape politics in taking decision based upon the popularity of the politicians, business, world culture, education, careers and many more [8]. Pew Research [21] has claimed that on an average 62% of the people get the latest updates from the social media. People usually rely on social media sites for new updates and innovations. It can be seen that 66% Facebook users get the news from the Facebook site [21]. Nearly 59% Twitter users use Twitter as their source of information [21]. The most used site to get the news updates is Reddit, around seven-in-ten every Reddit user get the news from the Reddit [21]. On Tumblr, the ratio is bit lower, only 31% of its users get the news from Tumblr [21]. The growth of social media with time can be seen in the figure 1. It is evident from the plots that 62% people use social media for getting the news rather than traditional media and remaining 38 use other sources. From 2013 to 2016, it can be seen that the usage of social media platforms have increased a lot.

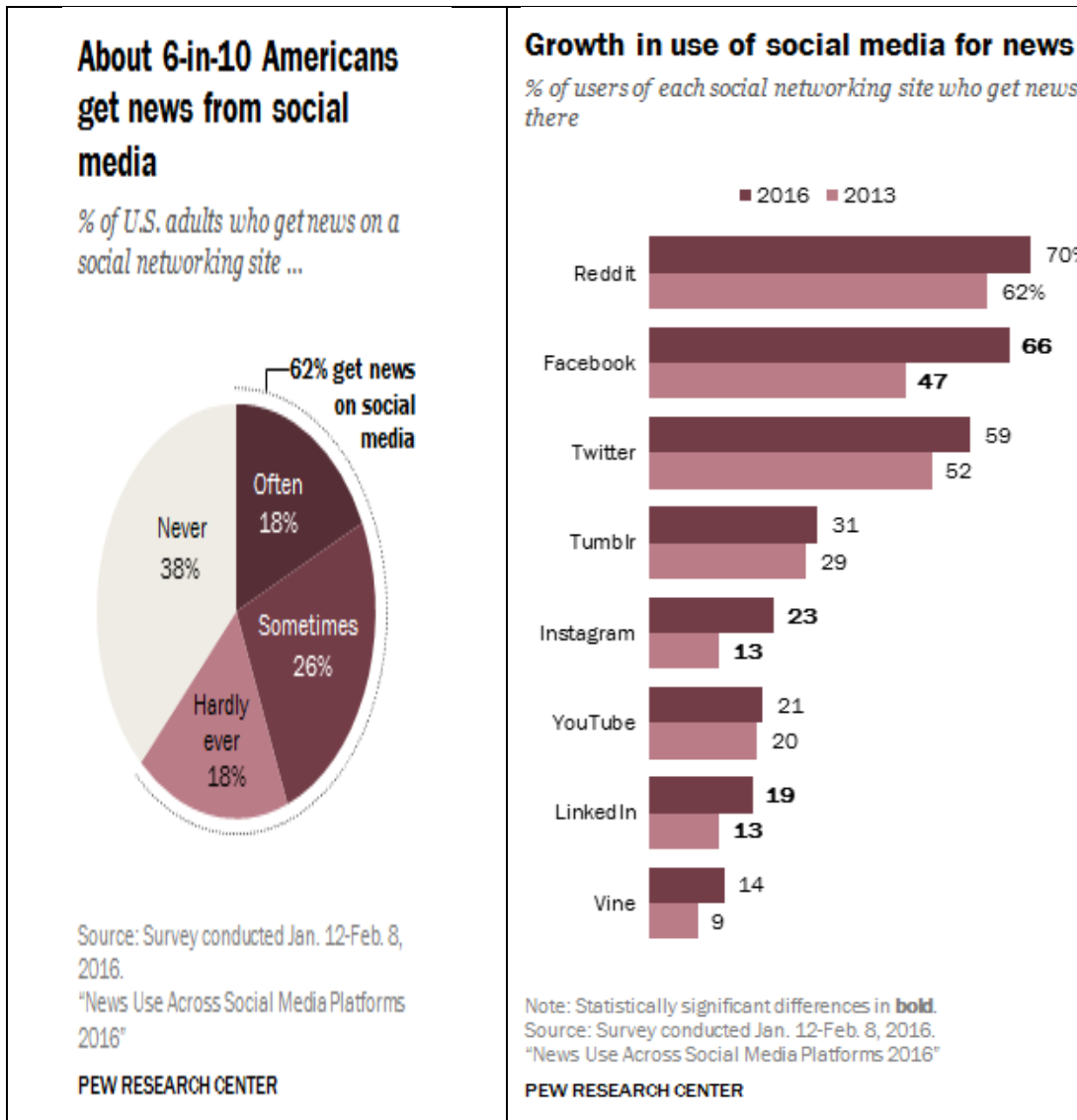


Figure 1 Growth in use of Social Media for News [21]

## 1.1.2 Relevant Information Extraction

Relevancy is the term that describes how much the content fulfills the user's query [12]. Search engine is the biggest example where a user enters the query and search for the most relevant documents. It is obvious if one search engine is not retrieving the relevant results, then a user will

start using an alternative search engine. In case of microblogging service, people need the most relevant tweets according to their search in Twitter. If their experience would be bad, they would move towards other social media sites or news channels.

Twitter has progressed from being a medium of sharing personal updates, opinion and finding new friends into a platform to share information about an ongoing events. During initial hours of an event, people come to Twitter and share an opinion. But not all the content they post is relevant and useful in providing the information about the situation [22]. Finding relevant information and satisfy the user needs is still a challenging task.

It can be seen that during high impact event such as 2010 earthquake in Chile, the hurricane sandy in 2012 [20], the amount of irrelevant data regarding certain topic increases to distract the users and also to create bad image of that particular social networking site [20]. People who retweet that irrelevant information are adding more complexity as the information retweeted more aggressively in the initial hours [19].

Twitter is flooded with information and finding the relevant information from that flood is a non-trivial task. There are many approaches that have been proposed including automatic approach, crowdsourcing approach and a hybrid combination of both. One of the application is Artificial Intelligence for Disaster Response [14], but this application uses the supervised classification, i.e. it needs the labeled data to process that flood of messages. However, training the data on the onset of the emergency event is very cumbersome and requires volunteers and those volunteers would cost much more. This procedure of labeling the data and waiting for the volunteers would obviously delays the process.



In the initial hours of crisis there is no labeled data available and to get the data labeled, is a time consuming task. For example, it can be seen during the 2012 Sandy hurricane, the highest peak observed was around 16k tweets/min [14] posted on Twitter. So, from this pool of unlabeled information, we need to first the labeled tweets to find the relevant information. To achieve end-users' relevant and critical information needs on an onset of an emergency situation it is impossible to train a classifier on an unlabeled data. During emergency situation people can't afford the delays, so, the system should be independent enough to generate the ground truth and train the classifier and provide the relevant information to the users on time so, the government and relief organizations get benefit from it.

### **1.1.3 Relevant Source Extraction**

Twitter has provided platform to common people to post information. People who post the information (tweets) are the *source* of those tweets. These common people are anonymous and from diverse backgrounds. All the activities they do are unmonitored. So, many people took advantage of it and post irrelevant content either to grab attention or their personal interest is associated with it.

Finding the relevant sources based on their profile histories, their expertise and their reputation on the network would help in increasing relevancy of information. Measuring source reputation is an important aspect because that reputation can give a clue about the activity of the source on the network. Finding the relevant sources is also very challenging because there are no discrete measures available to measure the sources. To measure user expertise and reputation, we use several different measures that are considered to be highly relevant to twitter [13]. Sentiments and opinion of the people is one of another way of finding the source relevancy.

The tweets that people post are usually public and anyone can see it and pass it on further. To view the information it is not necessary that you need to be the Twitter user. But there is a restriction in retweeting and using other features for that you need to register. Once the user has been registered he/she can follow people and those people can follow back. In this way the whole network builds. The follow relationship is asymmetrical on Twitter: the user being followed does not necessarily follow his/her follower. Twitter is considered to be a news medium as well as a social networking medium [13]. If people consider it a new medium so, the information should reach the audience on time and the information should be relevant. In order to extract the relevant sources, a novel approach should be proposed which would incorporate all the source specific and relevant tweet specific features.

## **1.2 Motivation**

People suffering during disasters and emergencies; look for quick feedback to their queries. People post a lot of information and there is a sudden spur of posts (tweets) during high impact events. Not all the information posted on Twitter is reliable and relevant in providing information about the event. There is a need to increase the relevancy so that relevant information reaches the needed audience on the right time. Government and relief organizations look for situational awareness information to launch relief operations.

## **1.3 Problem Statement**

In the process of relevant information extraction, redundant tweets are used and there is no mechanism for dealing with duplicate tweets which result in a non-valuable information to user. Relevance feedback or query expansion mechanism are not applied that improves the information relevancy results and covers the bigger set of tweets. Semi-automatic ground truth generation is

not addressed in the previous studies due to which problem of over fitting and under fitting can come while model training. Appropriate source specific feature set is not used for classification of sources. In the best of my knowledge, previous systems have not provided text relevancy and source relevancy under one umbrella.

## **1.4 Research Objectives**

This research has the following main objectives:

- To remove exact similar or 70% similar tweets from that would provide more valuable and relevant information to the users.
- To increase the relevancy of information during high impact events in order to have better coverage of information.
- To make system dynamic and independent enough to generate it's ground truth semi-automatically.
- To select appropriate source specific feature set for better classification of the sources.

## **1.5 Organization of Thesis**

The thesis is structured as follows. Chapter 2 contains literature review which describes the previous work that has been carried out to check the credibility and relevancy of the tweets with the topic. Chapter 3 describes the proposed system. Chapter 4 presents the experimental setup and results. Chapter 5 describes the contributions, conclusion, limitation and future work of the system.

# Chapter 2

## Literature Review

People suffering during disasters and emergencies; look for quick feedback to their queries. People post irrelevant information and there is a sudden rise in this activity during high impact events. Government and relief organizations look for situational awareness information to launch relief operations but due to increase in irrelevant information, they would not be able to take necessary measures on time. Existing studies explored text or source relevancy using GloVe [34], pseudo relevance feedback [22] and rule based approaches [35] in Twitter. GloVe approach has shown very low performance. Existing approaches have used unstructured and redundant tweets [30] and no measure has been taken to remove the redundant tweets. Existing systems have focused on text relevancy [22] or source relevancy [13] independently but none of the system has provided the both.

This chapter is organized as follows. Section 2.1 describes the background which contains the overview of important terms and theories that would be used in next chapters. Section 2.2 describes all the related work that has been done on in this topic so far.

### **2.1 Background**

The background contains the brief overview of important theories, terms, concepts and ideas are mentioned below:

## **2.1.1 Relevance Based Prediction**

During the disaster and emergency situations people are looking for the relevant information which can fulfill their search as Twitter is widely used social media platform so, people hugely rely on it. There is a dire need that the retrieval system should provide relevant and accurate results so that the user experience and trust factor of the users on Twitter does not break or undergo. Relevance based prediction involves dividing the tweets to relevant and irrelevant classes based upon the user feedback. Feedback is generating the ground truth by itself and this can help in training the classifier and then classifier would predict the relevant tweets and people can get the relevant information on time without any delay.

## **2.1.2 Information Extraction using word2vec**

Word2vec is a technology which is self-explanatory as it converts the word to a vector [6]; it's a two layered neural net for text processing. It inputs text and outputs a vector, i.e., feature vectors for words in the corpus. Though it's not a deep neural network, but it turns the text to the numerical form that deep neural networks understands. The main purpose of this approach is that human intervention is not needed, it simply converts the words to the vectors and group the similar vectors together in a vector space. This means it detects the similarities mathematically and all the related words, i.e., either contextually related or by words, are placed in a similar cluster.

## **2.1.3 Cosine Similarity**

Cosine similarity is used to find the distance between vectors, to check how closely the two vectors are. This is metric which is used to normalized dot product of the two attributes. This helps to find

the cosine angle between two vector spaces. The cosine of  $0^\circ$  is 1, and it is less than 1 for any other angle.

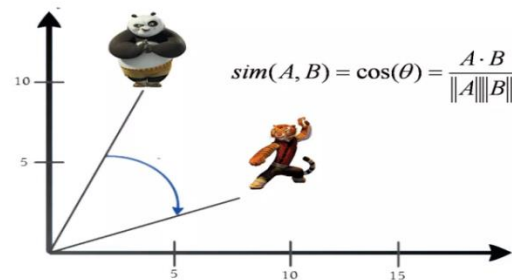


Figure 2 Cosine Similarity [8]

Two vectors with the same direction have a cosine similarity of 1, two vectors at  $90^\circ$  have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude [8]. Figure 2 cosine similarity is showing the formula of measuring the distance. For sparse vectors, it is used to evaluate efficiently the distance between the two vectors. It is used in positive vector space, where the outcome is neatly bounded to 0 to 1.

## 2.1.4 Rocchio Algorithm

A lot of information is available on the internet but relying on all the information available and getting the relevant and reliable information from the whole pool is a dire need [9]. Getting the reliable information manually is a cumbersome process so automated tool is needed to divide the information in to different categories. One of the most widely applied learning algorithms for text categorization is the Rocchio relevance feedback method developed in 1971 [9] and used in information retrieval. The main purpose of this algorithm is to optimize the queries using relevance feedback, and the algorithm can further be used for categorizing the text.

### **2.1.5 Twitter Stream API**

API stands for Application Programming Interface [8]. This tool has endpoints which are used to do interactions with the web services in a more easy way. It is used to access the Twitter data in a programmatic way. You can fetch the source information using Twitter ID or Twitter handles [10].

### **2.1.6 Relevant Source Extraction**

Relevant source extraction is the extraction of relevant sources who post relevant content and it is based upon the profile histories and in an automatic way using analysis of data of social media platform. The major elements used in classifying the sources includes deep analysis of text of posts, links included and the frequent and non-frequent behavior of sources in the classification [13].

### **2.1.6 Relevant Tweets Extraction**

Relevant tweets extraction is the extraction of relevant tweets during high impact events. This would increase the relevancy of content and would decrease the irrelevancy.

## **2.2 Related Work**

It can be seen that during high impact event such as 2010 earthquake in Chile, the hurricane sandy in 2012 [20], the amount of irrelevant data regarding certain topic increases to distract the users and also to create bad image of that particular social networking site. The tweets that are posted in the initial hours of any emergency situation, they are mostly don't carry the relevant content.

People who retweet that irrelevant information, they are adding more complexity as the information retweeted more aggressively in the initial hours [19].

Twitter is flooded with an information and finding the relevant information from that flood is a non-trivial task. There are many approaches have been proposed including automatic approach, crowdsourcing approach and a hybrid combination of both. One of the application is Artificial Intelligence for Disaster Response [14], but this application uses the supervised classification means it needs the labeled data to extract the relevant information. It can be seen during the 2012 Sandy hurricane, the highest peak observed was around 16k tweets/min [14] posted on Twitter. So, this data is unlabeled and labeling it with the help of volunteers would rise the cost and time as well.

Surajit et. al. [34], a very novel technique has been proposed name as GloVe, with the help of this approach tweet text have been converted to vector space word embeddings. But this issue is they have used unstructured and redundant tweets. There were no certain mechanism took place to remove the redundant tweets. Once user inserted the query that query has also converted to query vectors. One both vectors has been created, the distance between two vectors have been found using cosine similarity. The lower the distance between the vectors the higher is the similarity score. This system was created in 2016 but this system have lower accuracy. The reason behind is that, unstructured and redundant tweets have been used for finding the text relevancy.

DL4J et. al. [6], another approach has been proposed to convert the tweet text to vectors using Word2vec approach which makes the highly accurate guesses in putting the word in a same cluster. The model of similar words clusters were created and trained based on the previous appearances. This past appearances were used to establish a word's association with other words like "man" is



associated to “boy” as well and same for “woman” is to “girl”. This Word2vec helps to classify the clusters by topics. Once the word2vec vectors have been created, they can be used to find the relevant aspects and information from the topics and find the most nearest words which are contextually similar.

Andrew et. al. [29], authors have proposed a novel approach to label the unlabeled data and to improve sequence learning which is also semi supervised learning. Two approaches have been proposed so far one was what comes next in a sequence which was a language model in NLP. The second one deals with the word2vec which was a sequence auto encoder, which read the input sequence in to a vector and then predicts the sequence again which was converted to vectors using word2vec and predicts the next sequence by finding the distance between the vectors. The data obtained from pre training step was then used as a starting point for other supervised training models. This also explains that word2vec is semi supervised model which learns and predicts the next occurrences. This helps in finding and predicting the relevant information.

Manjeet et. al. [35], multiple relevant information extraction approaches have been mentioned. It was written in detail the pros and cons of each approach. The two techniques which have been highlighted were the hand-coded and rule based statistical. The first approach involved a programmer to write rules and regular expressions to find the relevant information. The second approach involved the rule base in which certain set of rules are defined if these match, then that information is relevant to the topic. But all these approaches involved manual ground truth generation and they were no mechanism defined for relevance feedback for improving the system automatically.

Fernando et. al. [32], author did an analysis on continuous space word embeddings. According to him, this has achieved a great importance in natural language processing and machine learning communities. It has the great potential of finding similarities and relationships between the texts. The use of term relatedness in the context of query expansion helped a lot in finding the most relevant text according to the query. It can be seen that word embeddings such as word2vec and GloVe, when trained locally, give much better retrieval results as compare to when trained globally.

One of the most widely applied learning algorithms for text categorization is the Rocchio relevance feedback method developed in 1971 and used in information retrieval. The main purpose of this algorithm is to optimize the queries using relevance feedback, and the algorithm can further be used for categorizing the text.

Deepak et. al. [9] authors did analysis on the Rocchio algorithm, where they grouped the documents in to different categories and classes. Rocchio query optimization makes the optimal query vector, it maximizes the similarity to the relevant documents while on the other hand minimizes the similarity to irrelevant documents. Formally, it can be seen in eq. 20:

$$\mathbf{Q}_{opt} = \operatorname{argmax}[sim(Q_{org}, D_r) - sim(Q_{org}, D_{nr})] \quad Eq. 20$$

$Q_{opt}$  represents optimal query vector,  $Q_{org}$  represents original query vector,  $D_r$  represents relevant documents and  $D_{nr}$  represents non-relevant documents. All the relevant terms are added in the query and once users feedback is taken, all those terms which user has marked as irrelevant will be removed from the query. In each iteration, the most optimal query will be found with highest term relevancy.

Using the relevance feedback mechanism, the optimal query optimization runs until it converges (i.e., an optimal query is obtained).

T.Joachims et. al. [28], authors used the same technique of rocchio and explored the use of Support Vector Machines (SVMs) for learning relevant text classifiers from certain set of examples. They also described the performance of SVM as compared to other classifiers for learning text relevancy. Empirical results support the theoretical findings. SVMs performed the best on multiple texts which makes it really robust. They proposed the total automatic process without any need to involve human parameters tuning.

Aditi et. al. [22], a novel approach has been proposed for finding the credible information during emergency situations. The approach used the tweet and source specific features to find the initial ranks of the tweets and then these ranks are again re-ranked based upon the BM25 [22] metric. This metric involves pseudo relevance feedback and based upon that tweets have ranked from highly relevant to the least relevant. And the performance of the ranking has been measured using NDCG which came out to be 83%.

Pal et. al. [3], authors took a different approach to studying trustworthiness of source on Twitter. Their approach was bit unique and effective instead of checking the reliability of a tweet, they proposed checking the reliability of an author and his behavior on social media. This approach gave a lot of good results, they basically compared the age of the user profile with the number of people following the author.

Westermann et al. [5] took a different approach to the problem by examining the effect of system-generated reports of connectedness on reliability. The researchers took few experiments on the ratio of followers and following counts. The results were bit shocking because those profiles which

had too many followers or too less, were showing less trustworthiness. On the other hand, a narrow gap between follows and followers led to higher assessments of trustworthiness.

Majed Alrubaian et al. [13], in his research proposed a novel approach that combines analysis of the user's reputation on a given topic within the social network, as well as users sentiments were given to measure the relevant sources of information against certain topic. The user influence plays an important role to judge the relevancy of a user. Some users post irrelevant information that can create chaos. In his research, much importance was given to the sentiment score that is assigned against each tweet. Following user specific features were given much importance: followers, favorites, tweets number, retweets number and mentions. It has been seen from the history that non-reliable users tend to have more mentions and hashtags than reliable ones because those features are the best way to connect many people in the network and propagate the information. Account age was also calculated and non-reliable users seem to have fewer followers and more friends than reliable users. At the end user had been ranked from highly reliable to the least reliable users against certain topic. To evaluate the performance of the proposed method, two supervised machine learning techniques were used one was logistic regression (LR) and other was naïve Bayes model with a feature ranking algorithm (FR\_NB). Naive Bayes performed really well better than the other. The tests included the kappa statistic, specificity, precision, recall (also known as sensitivity), F-measures, and receiver operating characteristic (ROC) curve [13].

## **2.6 Critical Analysis**

Following papers were critically analyzed and compared with the proposed system.

- P1 -> Word Embedding's for Information Extraction from Tweets [34]
- P2 -> Twitter Based Information Extraction [35]

- P3 → Information Credibility on Twitter [2]
- P4 → Understanding Information Credibility on Twitter [30]
- P5 → Credibility Ranking of Tweets during High Impact Events [22]
- P6 → Reputation-based credibility analysis of Twitter social network users [13]

<b>Paper</b>	<b>Preprocessing</b>	<b>Removing Redundant Tweets</b>	<b>Text Relevancy</b>	<b>Relevance Feedback</b>	<b>Tweet Specific Features</b>	<b>User Relevancy</b>	<b>User Specific Features</b>
Dasgupta, 2016	✓	✓	✓	×	×	×	×
Kumar, 2017	✓	✓	✓	×	×	×	×
Castillo, 2011	✓	✓	✓	×	✓	×	✓
Byungkyu Kang, 2013	✓	✓	✓	×	✓	×	×
Gupta, 2012	✓	✓	✓	×	✓	×	✓
Alrubaian, 2016	✓	✓	×	×	×	✓	✓

*Table 1 Critical Analysis of Different Papers*

Twitter is flooded with information and getting the relevant tweets which would cover the bigger domain is a dire need. Table 1 presents the critical analysis of different papers. From the above table, it can be seen that all the papers have performed preprocessing but duplicate or redundant tweets were not removed which does not add the valuable information to the user's queries. In finding the tweet relevancy, some papers have used text relevancy and tweet specific features but ignored user specific features which can help in improving the accuracy of classification. Classification was dependent on manual ground truths which involve human effort and domain

knowledge. Current systems are not independent enough to generate their ground truth automatically or semi-automatically.

Paper	Pre-Processing												
	Language	Topic	Hashtag	URL	Mention	Case Folding	Lemmatization/stemming	Twitter specific words	Special chars	Stop words	Extra spaces	Duplicates	
												Exact same	Similar
Dasgupta, 2016	✓	✓	✓	✓	✓	×	×	×	×	✓	✓	×	×
Kumar, 2017	✓	✓	×	×	×	×	×	×	×	×	✓	×	×
Castillo, 2011	✓	✓	×	×	×	×	×	×	×	×	✓	×	×
Byungkyu Kang, 2013	✓	✓	×	×	×	×	×	×	×	×	✓	×	×
Gupta, 2012	✓	✓	×	×	×	×	×	×	×	×	✓	×	×
Alrubaian, 2016	✓	✓	×	×	×	×	×	×	×	×	✓	×	×

Table 2 Research Gap Table for preprocessing steps

Table 2 shows the preprocessing steps that have been used in existing systems. The preprocessing steps are: restricting the tweet language, topics and removing extra spaces. In P1, hashtag, URL, mentions and stop words have also been removed. Preprocessing is the first and the most important

step in information retrieval for removing noise from the data and bringing the data to consistent form.

Paper	Tweet Specific Features						
	Sentiment	Similarity	Mentions	Hashtags	Retweet	URLs	Length
Dasgupta, 2016	x	x	x	x	x	x	x
Kumar, 2017	x	x	x	x	x	x	x
Castillo, 2011	✓	x	✓	✓	✓	✓	✓
Byungkyu Kang, 2013	✓	x	✓	✓	✓	✓	✓
Gupta, 2012	✓	x	✓	✓	✓	✓	✓
Alrubaian, 2016	✓	x	✓	x	✓	x	x

Table 3 Research Gap Table for tweet specific features

Table 3 shows tweet specific features that have been used to evaluate the ranking of the sources and tweets. P3, P4, P5 and P6 used sentiment scores while in proposed system similarity score has given much importance because it calculates the relevancy of content posted by the users. Mentions, hashtags, retweets, URL's and length are used in P3, P4 and P5 while P6 just used mentions and retweets for calculating the rank of the user. In the proposed system mentions, retweets and URL counts were used.

Paper	User Specific Features			
	No. of tweets	Followers	Followings	Favorites
Dasgupta, 2016	x	x	x	x
Kumar, 2017	x	x	x	x

Castillo, 2011	✓	✓	✓	✗
Byungkyu Kang, 2013	✓	✓	✓	✓
Gupta, 2012	✓	✓	✓	✗
Alrubaian, 2016	✓	✓	✓	✗

*Table 4 Research Gap Table for user specific features*

Table 4 depicts user specific features that have been used to calculate the ranking of the users. In P3, P4 and P5 number of tweets, followers, followings were used while in P4 favorites counts were also used to calculate the rank. In P6, two features, i.e., number of tweets and followers count, were used. In proposed system, number of tweets, followers and favorites counts were used.

Paper	Classifiers						
	SVM	Random Forest	Decision Tree	Naïve Bayes	J48	Logistic Regression	Proposed
Dasgupta, 2016	✗	✗	✗	✗	✗	✗	✗
Kumar, 2017	✗	✗	✗	✗	✗	✗	✗
Castillo, 2011	✓	✗	✓	✓	✓	✗	✗
Byungkyu Kang, 2013	✗	✗	✗	✗	✗	✗	✓
Gupta, 2012	✓	✗	✗	✗	✗	✗	✗
Alrubaian, 2016	✗	✗	✗	✓	✗	✓	✓



Table 5 Research Gap Table for Classifiers

Table 5 shows different classifiers used in previous studies and in our proposed technique. P3 used lots of classifiers SVM, Decision Tree, Naïve Bayes and J48.

Paper	Dataset for evaluation	Measure Parameter							System Efficiency
		Precision	Recall	F1	Accuracy	NDCG	PR Curve	RO Curve	Best Scores
Dasgupta, 2016	Manually Labeled	✓	✓	×	×	×	×	×	8%
Castillo, 2011	Manually Labeled	✓	✓	✓	×	×	×	×	88%
Byungkyu Kang, 2013	Manually Labeled	×	×	×	✓	×	×	✓	96%
Gupta, 2012	Manually Labeled	×	×	×	×	✓	×	×	0.73
Alrubaian, 2016	Manually Labeled	✓	✓	✓	×	×	×	✓	94%

Table 6 Research Gap Table for Evaluation Metrics

Table 6 shows the evaluation measures that have been used to evaluate the accuracy and efficiency of the algorithm. In P2, P3 and P4, dataset or ground truth was generated manually and precision, recall and F1 measure was used only in P2. In P5, ground truth was generated using semi-automated techniques and precision, recall, F1 and ROC curve was used. While in proposed

technique, ground truth was generated manually and precision, recall, F1, NDCG, PRC and ROC curve evaluation measures were used.

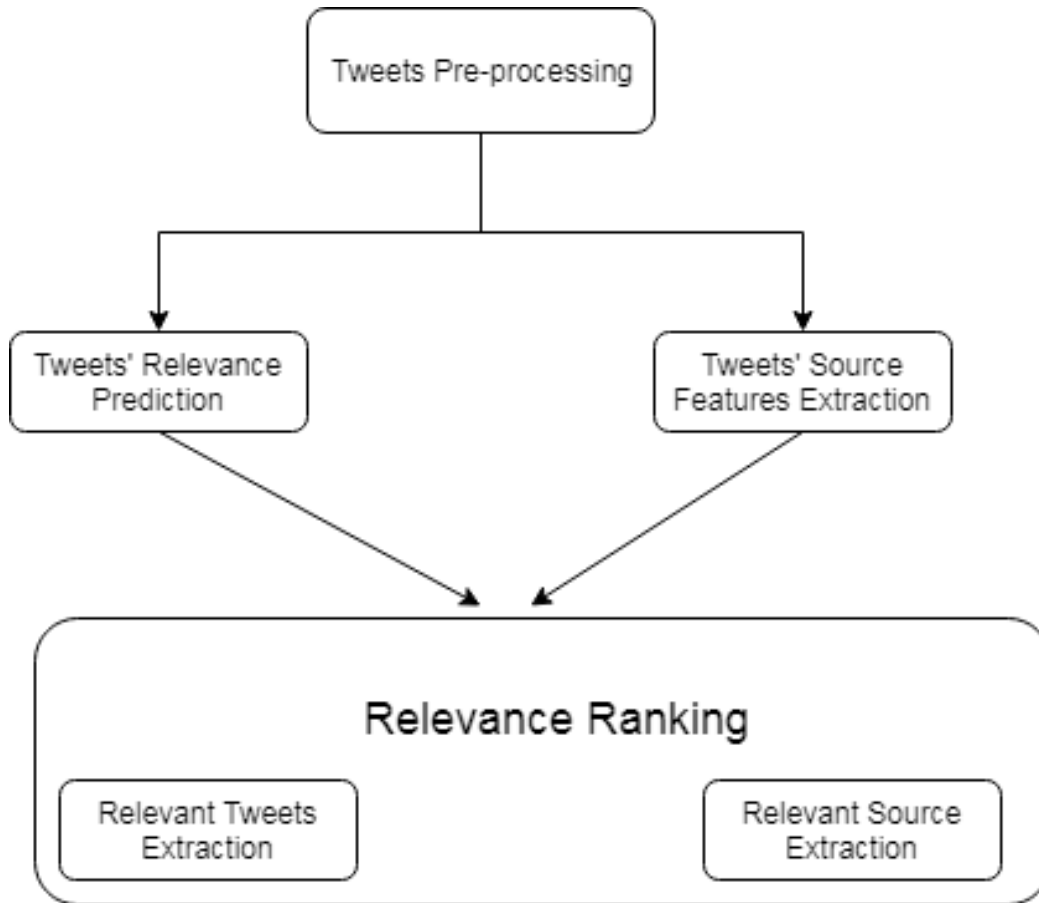
# Chapter 3

## Design and Methodology

This chapter is structured as follows. Section 3.1 gives an overview of the proposed system. Section 3.2 describes the data gathering and preprocessing techniques. Section 3.3 describes the relevance based classification and 3.4 describes the source feature extraction module. Section 3.5 explains the relevance based ranking module which is further categorized to relevant tweets extraction and relevant source extraction.

### **3.1 Proposed System**

The block diagram of proposed system has been shown in figure 3. The input to the proposed system is the tweets that are initially preprocessed for cleaning and fine tuning. These processed tweets are then used for the prediction of relevant tweets and for extracting the tweets source features from the Twitter using Twitter Stream API. For the prediction of relevant tweets, relevance feedback has been used that helps in getting the relevant information. Once relevant tweets have been predicted and source features have been extracted, next step is to rank the tweets and sources from the highly relevant to the least relevant.



*Figure 3 Block Diagram of Relevant Tweets and Relevant Source Extraction*

The next section gives a brief detail of the steps involved in extracting the relevant tweets and sources.

### **3.1.1 Tweets Pre-processing**

The experiment is carried out on Nepal Earthquake dataset. The dataset was provided by QCRI (Qatar Computing Research Institute). The data was initially in the form of Comma-Separated Values (CSV) files that was then merged into a single file. To make the process faster and avoid the usage of filing, all the data has been populated in to database. MySQL DB has been used in this research.

Tweets are so small of 140 characters and considered as written in a way that a machine could hardly process and get valuable information from them so for thy need to be cleaned [31]. Real world data is generally incomplete, noisy, contains errors and outliers, spelling mistakes and duplicates.

The major purpose of data preprocessing is to clean the noisy data, makes it smooth and remove the outliers. Few techniques that have been used for making the data smooth have been listed here.

### **3.1.1.1 Drop Duplicates**

Twitter is the real-time information engine [44]. Anything that is posted don't pass from a verification step. When the same tweet has been retweeted many times it comes under the category of duplicates. In order to extract the relevant information, showing them multiple times would not bring any valuable information to the people and at the end it would ruin the user experience as well. So, in order to make the user experience better and showing the relevant content on time it is necessary to remove duplicate tweets [34]. In Python there are pre-built methods to drop duplicates. In proposed methodology, it was very important to remove the duplicates because unique set of tweets were required to train the classifier. All the tweets that appear exactly same would be simply removed by this method. Keeping these duplicate tweets was totally useless, so to make the processing of data faster all the duplicates have already been dropped.

The drop duplicates will only remove those tweets that appear exactly the same but as seen and studied there were many tweets which are exactly the same in meaning but somehow different in words, those tweets were also required to be removed. So, for this Jaccard Similarity [11] has been applied.

In this research, each tweet is checked with other to find the Jaccard score of both tweets, if the similarity score is more than certain threshold then that tweet is considered to be a duplicate. In this case threshold was set to 0.7 or 70% [43]. Jaccard is calculated by using the eq. 16 where doc1 and doc2 are the 2 tweets among which the similarity score is measured.

$$Jaccard\ Similarity = \frac{len(Intersection)}{len(Union)} \quad Eq. 16$$

There were many other similarity measures like Cosine Similarity, Euclidean distance, Manhattan Distance, Minkowski Distance and Cheby Shev [11] that has been used but Jaccard has given the best results.

### **3.1.1.2 Topic / Hashtag**

A word preceded by a hash sign commonly used on social media especially Twitter to identify the specific topic. It is really common to use a lot of hashtags in your tweets as it helps in increasing the trends and also in searching [44]. These trends express which topic is hotter in twitter these days. This also expresses how many people are talking about this similar topic. Hashtags have been removed because this data was specifically on one topic. If there are multiple diverse datasets then this step could be skip. To remove hashtags, the algorithm searches for the hash sign and where it finds that word would be removed.

### **3.1.1.3 Lowercase**

In information retrieval, bringing text to presentable form is one of the initial and important step [34]. The tweets are written in a very informal way so some are written in capital letter; some in a

lower case and some are totally out of the way [44]. So, just to keep the consistency maintained it's better to make all the tweets lowercase. To bring the text to lowercase python lower() function is used.

### **3.1.1.4 Tweet Language**

There are 33 languages on twitter [45]. In order to process languages other than Twitter their dictionaries would require. So, to avoid this tweets language is restricted to English. The scope of this research was limited to English Tweets. To restrict the tweets, regex has been applied to search for those tweets having lang = eng.

### **3.1.1.5 HTML Encoding**

In order to fetch the data from Twitter Stream API which is the REST API, the dataset of Tweets contains a lot of html encoded text like text=&lt;script&gt;. This HTML encoded text do not give any relevant information so, removing them is necessary [34]. To remove the HTML encodings, a special regex has been designed to search for all the HTML encoded characters in the text.

### **3.1.1.6 Special Char/ Extra Spaces/ Stop Words Removal**

Stop words are the most common words that occur frequently in the text like “is”, “am” and “the” [34]. In order to train word2vec model [6] only on the important uni-grams and bi-grams these special character, extra spaces and stop words have been removed. To remove special characters and extra spaces regex has been used. To remove stop words python NLTK corpus library / stop words removal package has been used.

### 3.1.1.7 Lemmatization

Lemmatization usually refers to doing things in a more proper way with the use of an English vocabulary and analysis of words, normally it helps to remove the ending of the words and bringing the word to the base form of a dictionary word which is also called as lemma [46]. To train the word2vec model [6] on the standard set of words it was necessary to apply lemmatization. Tweets lemmatization has been done using NLTK Word Net lemmatization package.

### 3.1.2 Tweets Relevance Prediction

It is commonly known that many same words can have different meanings i.e. polysemy. Two different words can have the same meaning i.e. synonymy.

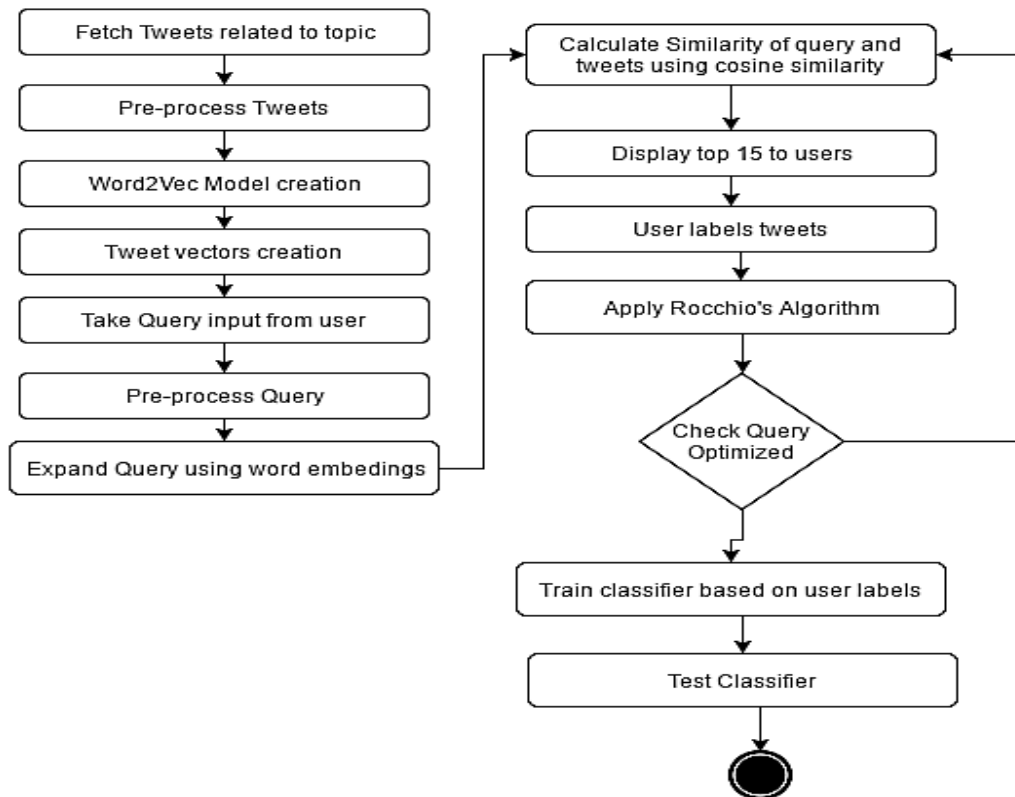


Figure 4 Flow Diagram of Tweet Relevance Prediction System



.While searching, if doing exact match then we'll surely going to miss most of the tweets because vocabulary of searcher might not match to that of the documents or tweets. Consider the query = {earthquake damages} as this is relatively unambiguous with respect to meaning and context of each word, exact matching will miss many tweets containing destructions, injury, or harm. Relevance feedback and query expansion aim to overcome the problem of synonymy or context related issues [12]. Figure 4 shows the flow diagram of tweets relevance prediction system.

From the set of corpus only fetch those tweets that are related to certain topic. In this case, topic would be Nepal Earthquake. Once the topic related tweets are fetched, those tweets need to be preprocessed. All the duplicates, HTML encoding, special characters and spaces would be removed from the tweets. Word2Vec model creation would be the next step but that would be on the full corpus not on the preprocessed as it'll reduce the vocabulary of the model. Word2vec *Gensim* model takes the parameters which are number of features and workers. Number of features tells the dimensionality of the vector while number of workers tells how many parallel threads need to run to do the process faster. In this case *number of features = 300* and *workers = 4*. Once the model is ready, tweets which were preprocessed are then tokenized and pass to the word2vec model, this model will create the vectors based on the training of the model.

Once the vectors are created, average tweet needs to be created. Next, the user is requested to input the query that is tokenized and expanded using word2vec embeddings. Once the query has been expanded, its vectors and average query vector would be created. In order to find the relevant tweets, the spatial cosine distance needs to be measured between the query vector and tweet vector. Lesser the distance the more relevant will be the results, tweets are sorted in descending order based upon similarity score and top 15 tweets will be shown to the user and user will mark those tweets as relevant or not. Based upon the user feedback, query will be optimized using Rocchio

relevance feedback. Rocchio aims to find the most optimal query which can give you the most similar documents / tweets related to the query. Eq. 17 shows that all the relevant terms are added in the query and all the irrelevant terms are discarded from the query.

$$Q_{opt} = \operatorname{argmax}[sim(Q_{org}, D_r) - sim(Q_{org}, D_{nr})] \quad Eq. 17$$

$Q_{opt}$  represents optimal query vector,  $Q_{org}$  represents original query vector,  $D_r$  represents relevant documents and  $D_{nr}$  represents non-relevant documents.

Based upon the user feedback all the positive terms will be added in the query and all the negative terms will be removed from the query in the next iteration and with each iteration you'll get the most optimal query [12].

$$q_m = \alpha q_0 + \beta \frac{1}{|D_r|} \sum_{d_j \in D_r} d_j - \gamma \frac{1}{|D_{nr}|} \sum_{d_j \in D_{nr}} d_j \quad Eq. 18$$

In the above eq. 18, it can be seen that we can assign weights to certain documents to give them higher priorities on others. The weights are given as follows  $\alpha=1$ ,  $\beta=0.75$ ,  $\gamma=0.15$ .  $D_r$  = set of known relevant doc vectors;  $D_{nr}$  = set of known irrelevant doc vectors;  $q_m$  = modified query vector;  $q_0$  = original query vector;  $\alpha$ ,  $\beta$  and  $\gamma$  are weights (hand-chosen or set empirically). The new and the optimized query moves toward relevant documents and away from irrelevant documents. Tradeoff  $\alpha$  vs.  $\beta/\gamma$ : If a lot of judged documents are available,  $\beta/\gamma$  is most probably be higher than some weights in query vector can go negative. Negative term weights are ignored set to 0 [12]. On the basis of new query, its vectors will be again created and new set of tweets will be shown to the user unless we get the most optimized version of the query [14].

To check if the query has been optimized or not, the new query vector will be compared with the previous one. If both are same then this will be the most optimized version of the query. Once we get the most optimized query and the relevant set of tweets, those tweets needs to be sent to the classifier for training and prediction of the tweets. Random forest using 100 trees, SVM and 10-fold cross validation has been used.

The retrieval systems rely on the query words (in the simple setting). It is not possible to give the accurate and all the results of a query in a single iteration. Systems need chances to get the results better and for that system needs the feedback from the users. As the query gets better the results and the gap also starts filling. With every iteration, more query terms are added and query expands and with this expansion the end results will be more accurate and will cover a much bigger circle of tweets which are contextually similar to the query. This whole framework is called relevance feedback. User issues a query usually short and simple query. The system returns some results. The user marks some results as relevant or non-relevant. The system computes a better representation of the information need based on feedback. Relevance feedback can go through one or more iterations.

Different classifiers such as Multinomial Naïve Bayes, Random Forest and SVM, are used. All these classifiers work best with text or document classification having binary classes [41]. These classifiers are trained on the labeled data given by the user's feedback. Data is divided in 60:40 proportions. Classifier was trained on 60% dataset while the prediction and testing was done on the remaining 40%. The model was trained using uni-gram and bi-gram based features relevant and irrelevant dataset. Unigrams and bigrams are converted to TF and TF-IDF. The model evaluation was performed using 10-fold cross-validation technique. Evaluation results from

various experiments are presented in the next section. Once a model is trained, automatic categorization of subsequent tweets from Twitter live stream starts.

Count vectorizer that is the part of sklearn (*sklearn.feature\_extraction.text*) library has been used and it converts the collection of text document to a matrix of token counts. Once all the counts has been created. We need TF-IDF transformer and in Python TF-IDF transformer feature is used (*sklearn.feature\_extraction.text (use\_idf=False/True)*). When passing *use\_idf* as true, it now computes all the counts as an IDF.

Figure 5 shows the complete view of tweets relevance prediction system. Figure has been explained with the help of an example.

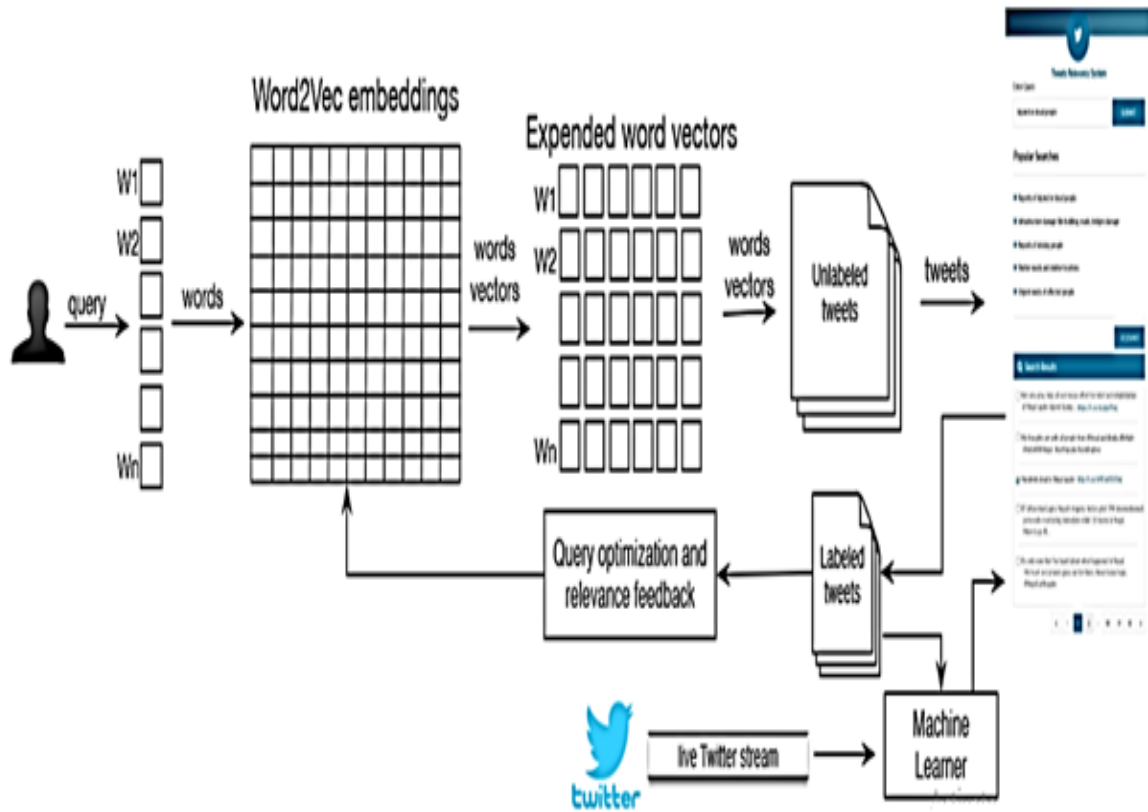


Figure 5 Tweet Relevance Prediction System Diagram

For example, consider a query “*Reports of dead and injured people*”. This query will be preprocessed same as of tweet. After preprocessing, the query is expanded using relevancy word2vec model. It takes the number of features or dimensions of the vector. The worker parameter helps you divide the process amongst multiple threads to make the process faster. In this implementation Gensim word2vec library is used shown in eq. 19.


$$\text{gensim.models.Word2Vec}(\text{min}_{count} = 1, \text{workers} = 4, \text{iter} = 10, \text{size} = \text{num}_{features}) \quad \text{Eq. 19}$$

Gensim vocabulary is built on bigram transformer, i.e., *gensim.models.phrases*. Word2vec model is trained on bigram vocabulary. Trained model can be displayed using *sklearn.decomposition* and *matplotlib pyplot*. Vector dimensions are 300, tweet is first tokenized and each word passes to word2vec which returns the vector of each word. Average vector is created by adding all the word vectors and then dividing it with the word vector count.

For tokenizing the tweets, NLTK tweet tokenizer is used which takes two arguments one is strip handles which removes all the handles from the tweets and also reduce the length of the tweet.

T1 = '@remy: This is waaaaayyy too much for you!!!!!!' → [':', 'This', 'is', 'waaayyy', 'too', 'much', 'for', 'you', '!', '!', '!']

Figure 6 shows the prototype of the proposed system of relevance feedback that was built using Photoshop CS5.1.



## Tweets Relevancy System

Enter Query

SUBMIT

### Popular Searches

---

- Reports of injured or dead people
- Infrastructure damage like building, roads, bridges damage
- Reports of missing people
- Shelter needs and shelter locations
- Urgent needs of affected people

---

RELEVANT

**Search Results**

Not only pray, help all out rescue effort for relief and rehabilitation of Nepal quake injured &... <https://t.co/xLqtpdYzuj>

---

My thoughts are with all people from #Nepal and #India #BeSafe #IndiaWithNepal #earthquake #sendinglove

---

Hundreds dead in Nepal quake <http://t.co/vMCw0OJ0wj>

---

RT @KanchanGupta: Nepal's tragedy, India's grief. PM @narendramodi personally monitoring immediate relief. 10 teams in Nepal. More to go #I...

---

It's only now that I've heard about what happened in Nepal. My heart and prayers goes out for them. Never loose hope. #NepalEarthquake

<
1
2
3
...
50
51
52
>

*Figure 6 Twitter Relevance Feedback System*

### 3.1.3 Tweets Source Features Extraction

Tweets sources are the ones who posted the tweets during different time intervals on a certain topic. In order to evaluate the source relevancy, a lot of source features I.e. followers count, favorites count, retweet count, followings count and much more needs to evaluate. More than 140 million active users post tweets of 140 characters every day [2]. These tweets are available to researchers and practitioners at no cost. Twitter has made the data available to the users and the several attributes using the Twitter streaming API [8]. The API can only be accessed using authenticated requests. You can only make certain number of requests at a time window called rate limit. You can access the API using a server side language. Server side scripting language make requests to API and result returned is in a JSON format. OAuth authentication is required to access the API. Using an API one can easily fetch 3200 most recent tweets of any user using their retweets as well. API returns the JavaScript Object Notation (JSON) formatted tweet objects. Following attributes are retrieved when Twitter Streaming API returns the result against each user:

- Friends count
- Followers count
- User tweet count
- Favorite count
- Retweet count
- Hashtag count
- URL count

Data of each source who have posted tweets on the topic of Nepal Earthquake and their labelled tweets, i.e., relevant or irrelevant, from the relevance feedback module are fetched from API.

### **3.1.4 Relevance Ranking**

In this section, the relevance ranking of the tweets and sources have been explained in detail.

#### **3.1.4.1 Relevant Source Ranking**

Ranking or categorizing source based upon the relevant information they have posted against certain topic needs certain attributes of the user to be deeply examined. The attributes includes the reputation of the source, i.e., the followers and the following count. This tells the network credibility of the source and how much this information can be further propagated in the form of the retweet [13].

The other attributes like listed count, follow request count, source tweet count, favorite count, hashtag and URL count also tells us the reputation and the popularity of the user. By analyzing all this we would be able to find out the credible sources of information.

##### **3.1.4.1.1 Source Popularity Measurement**

To measure source expertise and reputation, certain features or attributes were considered. The most highly relevant are retweet of the source on a certain topic and most important is that how many times a relevant tweet has been further retweeted [13].

The tweet that has been marked relevant by the relevance feedback method, there is a need to calculate how many times that tweet has been retweeted further. This retweet factor has the highest weightage among all other features or attributes. Table 6 is showing all the list of parameters used in the eq. 10.



The similarity score of the tweet helps us in identifying the relevant tweets. Suppose that there are set of users  $U$  who have more than one tweet on a given topic  $p \in P$ . Given a set of tweets  $T$ , we calculate the tweets of each user  $t_{ui}$  over time  $T$  mention in eq. 1.

$$I^2(u_i) = \left\{ \sum_{u \in U, p \in P} \frac{t_{ui}^p}{|T|}, \text{if } t_{ui} \in |T|0; \text{otherwise} \right\} \quad \text{Eq. 1}$$

It has been seen that relevant retweets, favorites, and similarity score / relevancy score are the best indicators of user popularity from a quantitative perspective, based on the assumption that a tweet that is relevant and plus it has been retweeted many times is considered to be credible source of information to many sources. All the attributes mentioned previously were quantitative ones means the tweet count of the source. Both the quantitative and qualitative attributes are of equal importance. The followers count of the source and the ones who are reading the tweet and further retweeting it. There must be some relationship between the two, that relationship factor comes under the qualitative attribute. These attributes gives the brief definition about the source expertise. Table 7 contains all the parameters that are used further to evaluate the rank of the user:

<b>Parameter</b>	<b>Explanation</b>
No. Flw ( $u_i$ )	Count of followers of a particular source
No. Ufav ( $u_i$ )	Count of favorites of a particular source
No. Twt ( $u_i$ )	Count of total tweets of a source
No. RT ( $u_i$ )	Count of retweets of a particular source
No. men	Count of mentions of a particular source
No. url	Count of links of a particular source
$\Delta_u$	Similarity score of the tweet that has been marked as relevant or not.

$\Delta_s$	Sentiment score of a tweet.
$R^P(u_i)$	Reputation rank of the source
$I^{p \in P}(u_i)$	Activity of source $u$ on a certain topic $p$
$EE^{p \in P}(u_i)$	Event engagement of source $u$ on topic $p$
$\phi^{p \in P}(u_i)$	Event engagement of source $u \in U$ on topic $p \in P$ using the no. of favorites $\phi$
$v^{p \in P}(u_i)$	Event engagement of source $u \in U$ on topic $p \in P$ using the no. of retweets $v$
$m^{p \in P}(u_i)$	Event engagement of source $u \in U$ on topic $p \in P$ using the no. of mentions $m$
$ul^{p \in P}(u_i)$	Event engagement of source $u \in U$ on topic $p \in P$ using the number of links $ul$
$w^{p \in P}(u_i)$	User influence of source $u$ on a certain topic $p$
$\vartheta^{p \in P}(u_i)$	Social popularity of $u \in U$ on a given topic $p \in P$

Table 7 Parameters of formula for Source Evaluation

Social popularity of the user  $\vartheta$  of  $u \in U$  on a given topic  $p \in P$  can be calculated by the following eq. 2:

$$\vartheta^{p \in P}(u_i) = \log(\text{No. Flw}(u_i)) / \max(\log(\text{No. Flw}(U, p))) \quad \text{Eq. 2}$$

We then calculate the number of favorite's  $\phi$  and the number of retweets  $v$  of source  $u$ 's posts on topic  $p$  using eq. 3 and eq. 4, respectively.

$$\phi^p(u_i) = \log(\text{No. UFav}(u_i)) / \max(\log(\text{No. UFav}(U, p))) \quad \text{Eq. 3}$$

$$v^p(u_i) = \log(\text{No. RT}(u_i)) / \max(\log(\text{No. RT}(U, p))) \quad \text{Eq. 4}$$

$$m^p(u_i) = \log(\text{No. men}(u_i)) / \max(\log(\text{No. men}(U, p))) \quad \text{Eq. 5}$$

$$ul^p(u_i) = \log(\text{No. url}(u_i)) / \max(\log(\text{No. url}(U, p))) \quad \text{Eq. 6}$$

From these components, the event engagement of user  $u \in U$  on topic  $p \in P$  is determined using eq. 7.

$$EE^{p \in P}(u_i) = \varphi^p(u_i) + v^p(u_i) + m^p(u_i) + ul^p(u_i) \quad Eq. 7$$

Subsequently, for the given topic  $p \in P$ , the user's influence  $\omega(u)$  can be computed as follows in eq. 8:

$$\omega^{p \in P}(u_i) = \vartheta^{p \in P}(u_i) + \frac{EE^{p \in P}(u_i)}{\text{Log}(N)} \quad Eq. 8$$

In eq 8,  $N$  denotes the number of users considered with respect to topic  $p$ .

### 3.1.4.1.2 Ranking of Sources

The final step is to rank the source according to their reputations, which can be calculated as follows:

$$R^p(u_i) = \Delta_s \times I^p(u_i) + (1 - \Delta_s) \times \omega^p(u_i) \quad Eq. 9$$

$$R^p(u_i) = \Delta_{ui} \times \omega^p(u_i) + (1 - \Delta_{ui}) \times EE^{p \in P}(u_i) \quad Eq. 10$$

In eq. 9,  $\Delta_s$  was used which was the sentiment score but the results computed from it was not up to the mark and this formula has been mentioned in the paper [13]. So, in this implementation sentiment score has been replaced with the similarity score  $\Delta_{ui}$  as it can be seen in eq. 10.

Sources are then need to be classified to the bins according to the value of  $R^p(u_i)$ . The source with the highest  $R^p(u_i)$  values are considered to be the most trusted source on a given topic, and the source with the lowest priority values are considered to be the least trusted.

The bins are created by using Pandas cut functions of Python. This function divides data in to 4 equal bins. Bin-1 shows the highly irrelevant source to the given topic, Bin-2 shows the probable irrelevant source, Bin-3 shows probable relevant source to a certain topic and Bin-4 shows the highly relevant source to a certain topic. Values have been inserted in all the formulas just to get the understanding, how source ranks have been evaluated. Below is the walk through example:

$\vartheta^{p \in P}(u_i)$  is the social popularity of the user which includes the followers count of the source and that follower count is divided by the maximum followers count.

$$\vartheta^{p \in P}(u_i) = \log(25)/\max(\log(10097)) = \frac{4.64}{13.30} = \mathbf{0.34}$$

There is a need to find out number of favorite's  $\varphi$  and the number of retweets  $v$  of source  $u$ 's posts on topic  $p$  using the below equations respectively:

$$\varphi^p(u_i) = \log(90)/\max(\log(5932)) = \frac{6.49}{12.53} = \mathbf{0.51}$$

$$v^p(u_i) = \log(120)/\max(\log(8563)) = \frac{6.90}{13.06} = \mathbf{0.52}$$

$$m^p(u_i) = \log(56)/\max(\log(2480)) = \frac{5.80}{11.27} = \mathbf{0.51}$$

$$ul^p(u_i) = \log(23)/\max(\log(1875)) = \frac{4.52}{10.87} = \mathbf{0.41}$$

From the above components, the event engagement of the source is calculated:

$$EE^{p \in P}(u_i) = \varphi^p(u_i) + v^p(u_i) + m^p(u_i) + ul^p(u_i) = 0.51 + 0.52 + 0.51 + 0.41 = \mathbf{1.95}$$

The source influence source influence score is calculated as follows:

$$\omega^{p \in P}(u_i) = \vartheta^{p \in P}(u_i) + \frac{EE^P(u_i)}{\text{Log}(5000)} = 0.34 + \frac{1.95}{10.28} = \mathbf{0.52}$$

The rank of the source is computed using the below formula:

$$R^P(u_i) = 0.75 \times 0.52 + (1 - 0.75) \times 1.95 = 0.39 + 0.4875 = \mathbf{0.8875}$$

The value lies in 0.75 - 1.0 scale which considers it as the highly relevant source so, this source would be placed in bin 4.

### 3.1.4.2 Relevant Tweets Ranking

Relevant tweets extraction works in the same way as that of relevant source extraction except in this instead of taking average of relevancy score of tweets, each tweet relevancy score would be used against different queries. So, tweets has been ranked against particular query.

$$R^P(u_i) = \Delta_{ui} \times \omega^P(u_i) + (1 - \Delta_{ui}) \times EE^{p \in P}(u_i) \quad \text{Eq. 12}$$

$\Delta_{ui}$  is the relevancy score of each tweet. The following part of the equation  $\omega^P(u_i) + (1 - \Delta_{ui}) \times EE^{p \in P}(u_i)$  calculates the source relevancy score. User influence score and event engagement score of the source has been taken from the previous step. After inserting the values in eq. 12

$$R^P(u_i) = 0.81 \times 0.52 + (1 - 0.81) \times 1.95 = 0.42 + 0.37 = \mathbf{0.79}$$

The value lies in 0.75 - 1.0 scale which considers it as the highly relevant tweet so, this tweet would be placed in bin 4. After tweets have been ranked, the tweets would pass from the same Pandas cut functions of Python. This function divides data in to 4 equal bins. Bin-1 shows the

highly irrelevant tweets, Bin-2 shows the probable irrelevant tweets, Bin-3 shows probable relevant tweets and Bin-4 shows the highly relevant tweets to a certain topic.

# Chapter 4

## Experiments and Results

This chapter is organized as follows. Section 4.1 explains the dataset used for evaluation. Section 4.2 describes the annotation scheme for ground truth generation. Section 4.3 describes the evaluation metrics used. Section 4.4 explains the graphical representation of precision, recall and F1. Section 4.5 explains which classifiers have been used to train the model. Section 4.6 describes experiments carried out in this research and their results. Comparative evaluation has been shown in section 4.7.

### **4.1 Dataset Specifications**

To evaluate the performance of proposed algorithm, Nepal Earthquake dataset has been used. The dataset consists of more than 1 million tweets that are not publicly available. This dataset was collected by QCRI Qatar Computing Research Institute for research purposes. The dataset contains wide range of tweets from diverse users during emergency situations.

### **4.2 Dataset Annotation**

This section describes the annotation of set of tweets and users in order to obtain the ground truth. Annotations were performed by 20 different professionals, in the age group (25 to 34 years), and their designation was Software Engineer and Quality Assurance. Each participant was provided 500 tweets and 350 sources to rank. In the relevance feedback, 3,000 tweets were manually labeled

as relevant and non-relevant and 2,000 classifiers had predicted the results. So, all in all we had 5,000 labeled dataset of tweets as relevant and non-relevant.

In order to generate the ground truth of sources, human annotators were provided with the attributes (user id, tweet author, similarity score, user influence and user activity, etc.) about the users. Human annotators need to rank the sources and place them in the different bins. Firstly, they were given a brief session regarding the attributes of the sources, their importance and how they can assign rank and classify them in to different bins from highly relevant to highly irrelevant. To rank the sources and divide them in to different bins, annotators were provided with the Excel sheets attached below for generating the ground truth as shown in figure 7.

	A	B	C	D	F	G	H	I	J	K	L
1	<b>Rank the sources based upon the session deliver to you</b>										
2	<b>Note:</b> You need to rank the sources and assign bins to them based upon the scores. Details of each scale is mentioned for a										
3	reference										
4											
5											
6	<b>Source Rank Scale:</b>										
7	<input type="radio"/> 0-25 (Highly irrelevant) <input type="radio"/> 26-50 (Probably irrelevant) <input type="radio"/> 51-75 (Probably releaveant) <input checked="" type="radio"/> 76-100 (Highly relevant)										
8											
9	<b>Source Bins:</b>										
10	<input type="radio"/> Bin1 (Highly irrelevant) <input type="radio"/> Bin2 (Probably irrelevant) <input type="radio"/> Bin3 (Probably releaveant) <input checked="" type="radio"/> Bin4 (Highly relevant)										
11											
12											
13											
14	user_id	fav_engagement_score	relevancy_score	retweet_engaj	author	event_engager	scoial_popularity_score	user_influence_score	user_activity_score	source_rank	soource_bin
15	1002516470	0.37855571	0.817794871	0.127950148	Akash94ks	0.506505858	0.343524478	1.092371613	0.0000011	<input type="radio"/> 0-25 <input type="radio"/> 26-50 <input checked="" type="radio"/> 51-75 <input type="radio"/> 76-100	<input type="radio"/> Bin1 <input type="radio"/> Bin2 <input checked="" type="radio"/> Bin3 <input type="radio"/> Bin4
16	101553559	0.166291264	0.759260514	0.178063424	upadhyaaisir	0.344354688	0.310692054	0.841798739	0.0000046	<input type="radio"/> 0-25 <input checked="" type="radio"/> 26-50 <input type="radio"/> 51-75 <input type="radio"/> 76-101	<input type="radio"/> Bin1 <input checked="" type="radio"/> Bin2 <input type="radio"/> Bin3 <input type="radio"/> Bin4
17	101988381	0.518924183	0.802040144	0.428169193	baudelaireMe	0.947093376	0.427724419	1.766774511	0.000007	<input type="radio"/> 0-25 <input type="radio"/> 26-50 <input checked="" type="radio"/> 51-75 <input type="radio"/> 76-102	<input type="radio"/> Bin1 <input type="radio"/> Bin2 <input checked="" type="radio"/> Bin3 <input type="radio"/> Bin4
18	1020058453	0.537067875	0.706808802	0.362260844	tuscanyuae	0.899328719	0.787568167	2.167826481	0.000014	<input type="radio"/> 0-25 <input type="radio"/> 26-50 <input checked="" type="radio"/> 51-75 <input type="radio"/> 76-103	<input type="radio"/> Bin1 <input type="radio"/> Bin2 <input checked="" type="radio"/> Bin3 <input type="radio"/> Bin4
19	1030581656	0.408472212	0.773633546	0.768925683	wostey	1.177397895	0.320404045	1.924821093	0.0000011	<input type="radio"/> 0-25 <input type="radio"/> 26-50 <input checked="" type="radio"/> 51-75 <input type="radio"/> 76-104	<input type="radio"/> Bin1 <input type="radio"/> Bin2 <input checked="" type="radio"/> Bin3 <input type="radio"/> Bin4
20	103245739	0.488149213	0.680710483	0.615505925	Jaskirat5B	1.103655138	0.679798565	2.291911376	0.000036	<input type="radio"/> 0-25 <input type="radio"/> 26-50 <input checked="" type="radio"/> 51-75 <input type="radio"/> 76-105	<input type="radio"/> Bin1 <input type="radio"/> Bin2 <input checked="" type="radio"/> Bin3 <input type="radio"/> Bin4
21	10331222	0.462938845	0.829315757	0.561606429	Analia_Fiesta	1.024545273	0.468597157	1.91883317	0.000084	<input type="radio"/> 0-25 <input type="radio"/> 26-50 <input type="radio"/> 51-75 <input checked="" type="radio"/> 76-106	<input type="radio"/> Bin1 <input type="radio"/> Bin2 <input type="radio"/> Bin3 <input checked="" type="radio"/> Bin4
22	104061513	0.666425061	0.760980533	0.432516544	Algheryni	1.098941605	0.398140609	1.923896175	0.000035	<input type="radio"/> 0-25 <input type="radio"/> 26-50 <input checked="" type="radio"/> 51-75 <input type="radio"/> 76-107	<input type="radio"/> Bin1 <input type="radio"/> Bin2 <input checked="" type="radio"/> Bin3 <input type="radio"/> Bin4
23	10409342	0.50569135	0.615444418	0.605578023	elidiolatorre	1.111269373	0.39624966	1.937308502	0.000066	<input type="radio"/> 0-25 <input type="radio"/> 26-50 <input checked="" type="radio"/> 51-75 <input type="radio"/> 76-108	<input type="radio"/> Bin1 <input type="radio"/> Bin2 <input checked="" type="radio"/> Bin3 <input type="radio"/> Bin4
24	104193112	0.6502583	0.720118537	0.282112162	KarimBarolia	0.932370463	0.42818892	1.748451065	0.000004	<input type="radio"/> 0-25 <input type="radio"/> 26-50 <input checked="" type="radio"/> 51-75 <input type="radio"/> 76-109	<input type="radio"/> Bin1 <input type="radio"/> Bin2 <input checked="" type="radio"/> Bin3 <input type="radio"/> Bin4

Figure 7 Ground Truth for Evaluation of Source Ranking



At the top of each sheet they were provided with the scale which would help them to rank the sources. All this procedure has been repeated for ranking the tweets as well. Professionals were provided with the excel sheets of same format having all the attributes listed. This time they need to rank and classify the tweets instead of the sources. At the end we had 5000 ranked tweets and around 3500 ranked sources. The ground truth excel sheet of tweets ranking is attached below in figure 8.

	A	B	C	D	F	G	H	I	J	K	L
1	<b>Rank the tweets based upon the session deliver to you</b>										
2	<b>Note:</b> You need to rank the tweets and assign bins to them based upon the scores. Details of each scale is mentioned for a										
3	reference										
4											
5	<b>Tweets Rank Scale:</b>										
6	<input type="radio"/> 0-25 (Highly irrelevant) <input type="radio"/> 26-50 (Probably irrelevant) <input checked="" type="radio"/> 51-75 (Probably relevant) <input type="radio"/> 76-100 (Highly relevant)										
7											
8	<b>Tweets Bins:</b>										
9	<input type="radio"/> Bin1 (Highly irrelevant) <input type="radio"/> Bin2 (Probably irrelevant) <input checked="" type="radio"/> Bin3 (Probably relevant) <input type="radio"/> Bin4 (Highly relevant)										
10											
11											
12											
13											
14	tweet_id	fav_engagement_score	relevancy_score	retweet_engage	author	event_engager	social_popularity_score	user_influence_score	user_activity_score	source_rank	source_bin
15	5.91904E+17	0.37855571	0.657261435	0.127950148	Akash94Ks	0.506505858	0.343524478	1.092371613	0.0000011	<input type="radio"/> 0-25 <input type="radio"/> 26-50 <input checked="" type="radio"/> 51-75 <input type="radio"/> 76-100	<input type="radio"/> Bin1 <input type="radio"/> Bin2 <input checked="" type="radio"/> Bin3 <input type="radio"/> Bin4
16	5.92184E+17	0.166291264	0.539251468	0.178063424	upadhyaysir	0.344354688	0.310692054	0.841798739	0.0000046	<input type="radio"/> 0-25 <input checked="" type="radio"/> 26-50 <input type="radio"/> 51-75 <input type="radio"/> 76-101	<input checked="" type="radio"/> Bin1 <input type="radio"/> Bin2 <input type="radio"/> Bin3 <input type="radio"/> Bin4
17	5.91958E+17	0.518924183	0.459838028	0.428169193	baudelaireMe	0.947093376	0.427724419	1.766774511	0.0000007	<input type="radio"/> 0-25 <input type="radio"/> 26-50 <input checked="" type="radio"/> 51-75 <input type="radio"/> 76-102	<input type="radio"/> Bin1 <input checked="" type="radio"/> Bin2 <input type="radio"/> Bin3 <input type="radio"/> Bin4
18	5.9224E+17	0.537067875	0.646215855	0.362260844	tuscanyusue	0.899328719	0.787568167	2.167826481	0.000014	<input type="radio"/> 0-25 <input type="radio"/> 26-50 <input checked="" type="radio"/> 51-75 <input type="radio"/> 76-103	<input type="radio"/> Bin1 <input checked="" type="radio"/> Bin2 <input type="radio"/> Bin3 <input type="radio"/> Bin4
19	5.92227E+17	0.408472212	0.64403049	0.768925683	wostey	1.177397895	0.320404045	1.924821093	0.0000011	<input type="radio"/> 0-25 <input type="radio"/> 26-50 <input checked="" type="radio"/> 51-75 <input type="radio"/> 76-104	<input type="radio"/> Bin1 <input checked="" type="radio"/> Bin2 <input type="radio"/> Bin3 <input type="radio"/> Bin4
20	5.92263E+17	0.488149213	0.698850151	0.615505925	Jaskirat58	1.103655138	0.679798565	2.291911376	0.000036	<input type="radio"/> 0-25 <input type="radio"/> 26-50 <input checked="" type="radio"/> 51-75 <input type="radio"/> 76-105	<input type="radio"/> Bin1 <input type="radio"/> Bin2 <input checked="" type="radio"/> Bin3 <input type="radio"/> Bin4
21	5.91989E+17	0.462938845	0.554879685	0.561606429	Analia_Fiesta	1.024545273	0.4688597157	1.91883317	0.000084	<input type="radio"/> 0-25 <input type="radio"/> 26-50 <input type="radio"/> 51-75 <input checked="" type="radio"/> 76-106	<input type="radio"/> Bin1 <input type="radio"/> Bin2 <input checked="" type="radio"/> Bin3 <input type="radio"/> Bin4
22	5.91907E+17	0.666425061	0.577774637	0.432516544	Algherny	1.098941605	0.398140609	1.923896175	0.0000035	<input type="radio"/> 0-25 <input type="radio"/> 26-50 <input checked="" type="radio"/> 51-75 <input type="radio"/> 76-107	<input type="radio"/> Bin1 <input type="radio"/> Bin2 <input checked="" type="radio"/> Bin3 <input type="radio"/> Bin4
23	5.92355E+17	0.50569135	0.593581351	0.605578023	elidliatorre	1.111269373	0.39624966	1.937308502	0.000066	<input type="radio"/> 0-25 <input type="radio"/> 26-50 <input checked="" type="radio"/> 51-75 <input type="radio"/> 76-108	<input type="radio"/> Bin1 <input type="radio"/> Bin2 <input checked="" type="radio"/> Bin3 <input type="radio"/> Bin4
24	5.91961E+17	0.6502583	0.463487744	0.282112162	KarimBarolia	0.932370463	0.42818892	1.748451065	0.000004	<input type="radio"/> 0-25 <input type="radio"/> 26-50 <input checked="" type="radio"/> 51-75 <input type="radio"/> 76-109	<input type="radio"/> Bin1 <input type="radio"/> Bin2 <input checked="" type="radio"/> Bin3 <input type="radio"/> Bin4

Figure 8 Ground Truth for Evaluation of Tweets Ranking

To check the reliability of results of source ranking obtained via annotation, each source has been classified and ranked by two annotators. Both agreeing on the same option then that source had been added to the annotated dataset. All those sources had been discarded on which two annotators gave different rank.

In order to check the reliability of results of tweets ranking obtained via annotation, each tweet has been classified and ranked by two annotators. Both agreeing on the same option then that tweet had been added to the annotated dataset. All those tweets had been discarded on which two annotators gave different rank.

In order to train the relevance feedback classifier dataset has been divided into training and test set. Before dividing the data into training and test set, it was completely shuffled in order to avoid under fitting or over fitting. Training, validation and test data has been divided in ratio of 60:20:20. Cross validation [15] has also been applied in order to avoid over fitting on the data.

### 4.3 Performance Measures

Multiple experiments have been performed in order to evaluate the performance of the proposed algorithm. There are standard measures that are available in order to check the performance of the proposed algorithm. These measures include Precision, Recall, F1 and Mean Average Precision (MAP) for evaluating the performance of any system. Stanford NLP measure called Normalized Discounted Cumulative Gain (NDCG) has also been used as this system is based on ranking or recommendation.

Precision of any system tells how accurate your model is, i.e., how much a system can accurately predict the positive [37]. Adding all the precisions and taking its mean by taking average is mean average precision (MAP). Precision is defined as follows:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad Eq. 25$$

Recall tells how many actual positives our model capture through labeling it as positive (true positive) [37]. Recall is defined as follows:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad Eq. 26$$

F1 measure involves both precision and recall measures in order to evaluate the accuracy of methodology [37]. F1 is defined as follows:

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad Eq. 27$$

Apart from the above mentioned measures, macro and micro averaging methods have also been used. Macro is considered as global and this method is very simple. In this technique, the average of the precision and recall of the system on different sets is calculated [40] and is defined as:

$$P_M = \frac{1}{|L|} \times \sum_{l \in L} \frac{|TP_l|}{|TP_l| + |FP_l|} \quad Eq. 30$$

$$R_M = \frac{1}{|L|} \times \sum_{l \in L} \frac{|TP_l|}{|TP_l| + |FN_l|} \quad Eq. 31$$

$$F1_M = 2 \times \frac{P_M \times R_M}{P_M + R_M} \quad Eq. 32$$

Eq. 30, eq. 31 and eq. 32 are showing the formula of macro precision, macro recall and macro F1.

Micro [40] is totally opposite from Macro. In Micro-average method, you add all the individual true positives, false positives, and false negatives of the system for different sets and at the end take an average. Micro is much better for multiclass and is defined as:

$$P_\mu = \frac{\sum_{l \in L} |TP_l|}{\sum_{l \in L} (|TP_l| + |FP_l|)} \quad Eq. 33$$

$$R_{\mu} = \frac{\sum_{l \in L} |TP_l|}{\sum_{l \in L} (|TP_l| + |FN_l|)} \quad Eq. 34$$

$$F1_{\mu} = 2 \times \frac{P_{\mu} \times R_{\mu}}{P_{\mu} + R_{\mu}} \quad Eq. 35$$

Eq.33, eq. 34 and eq. 35 are showing the formula of micro precision, micro recall and micro F1.

Discounted cumulative gain (DCG) [36] is a measure of ranking documents. In information retrieval discounted and normalized cumulative gain is used in search engines to rank the result set. NDCG measures the relevancy and gain of the document and also its position in the result set and is defined as mention in eq. 24.

$$Cumulative\ Gain\ at\ p = CG_p = \sum_{i=1}^p rating(i) \quad Eq. 21$$

$$Discounted\ Cumulative\ Gain\ at\ p = DCG_p = \sum_{i=1}^p \frac{rating(i)}{\log_2(i+1)} \quad Eq. 22$$

$$Ideal\ Discounted\ Cumulative\ Gain\ at\ p = IDC G_p = \sum_{i=1}^{|REL|} \frac{rating(i)}{\log_2(i+1)} \quad Eq. 23$$

Where |REL| is the number of best ratings up to position p (Note: |REL| <= p)

$$Normalized\ DCG_p = \frac{DCG_p}{IDCG_p} \quad Eq. 24$$

In order to check the accuracy or success of ranking based system, there is a need to calculate the NDCG score. As these scores varies from 0 to 1.0. So, more it is towards 1.0 the more accurate the system is. As it is a clear fact that during high impact events, people are more interested in

highly relevant tweets than the ones that are not that much relevant. In order to cater this issue, this NDCG will help us in ranking the system. NDCG always give the least importance to the lowest rank documents or tweets in our case.

The lower the ranked position of relevant tweets, the less useful to show it to the audiences, since it is less likely to be user interested in. This uses graded relevance as a measure of usefulness, or gain, from examining a document or a tweet. Gain is higher at the top and as it goes down it also starts decreasing.

The proposed approach has been compared with Majeed Alrubaian et al. [13] so in order to make a fair comparison of both techniques, we used the same dataset. In this paper their main focus was on sentiment score of a tweet and the user activity but our main focus was on similarity score and user influence which includes the relevant retweet importance.

## **4.4 Graphical Representation of Precision, Recall and F1**

In this section, precision, recall and F1 have been shown using graphs, i.e., PR curve and ROC curve.

### **4.4.1 PR Curve**

The precision-recall curve [39] shows the curve between the precision and recall. A higher area under the curve shows high recall and high precision, whereas high precision shows having a low false positive rate, and high recall shows having a low false negative rate.

## 4.4.2 ROC Curve

ROC curve [39] is between true positive rate and false positive rate. Accuracy of a model is measured by area under the ROC curve. The following statistics shows the range and which system is considered good, fair or excellent.

- .90-1 = excellent (A)
- .80-.90 = good (B)
- .70-.80 = fair (C)
- .60-.70 = poor (D)
- .50-.60 = fail (F)

## 4.5 Classifiers Training

Multiple classifiers have been trained including support vector machines, random forest and multinomial naïve bayes. To train the classifier on a bigger set of data 10-folds cross validation has been used to achieve maximum efficiency and accuracy.

### 4.5.1 Support Vector Machines

Support Vector Machines [41] is the supervised machine learning that is based on the concept of decision planes that define decision boundaries. A decision plane is also called the hyperplane that differentiates these two decision boundaries. Hyperplane in the linear SVM can be made using a linear kernel. Eq. 36 is showing a linear kernel.

$$f(x) = B(0) + \text{sum}(a_i * (x, x_i)) \quad \text{Eq. 36}$$

The polynomial kernel can be written as:

$$K(x, x_i) = \exp(-\gamma \sum (x - x_i)^2) \quad \text{Eq. 37}$$

Eq. 37 is showing the polynomial kernel. Linear kernel is used where when the problem is linearly separable but when the problem is not linearly separable there is a need to use polynomial kernel to deal with complex data. This classifier works best for text classification.

### 4.5.2 Random Forest

Random forest classifier [41] as its name says it has a lot of decision trees I.e. in proposed methodology 100 decision trees are used and these decision trees all vote for the new test object based upon the learnt from the training set. At the end, votes are aggregated from different decision trees to decide the final class of the test object.

$$P(c|f) = \sum_1^n P_n(c|f) \quad \text{Eq. 38}$$

Eq. 38 is showing the formula how votes of different decision trees are combined to predict the class of new test object. This classifier works best for document classification.

### 4.5.3 Multinomial Naïve Bayes

The Naive Bayes Classifier [41] has its roots from the so-called Bayesian theorem and it is preferable for those inputs that have high dimensionality. Despite its simplicity, it can often outclass more refined classification methods. Naïve Bayes works on the probabilities of assigning particular class to a particular word. It finds the probability by estimating the conditional probability of a particular word given a class as the frequency. Multinomial works on the term

frequency by counting the words. Eq. 39 shows how probabilities have been calculated whenever a new test object comes.

$$p(\vec{F} | C_i) = \frac{p(F^* | C_i)p(C_i)}{\sum p(F^* | C_j)p(C_j)} \quad Eq. 39$$

#### **4.5.4 10-Fold Cross Validation**

Cross-validation [15] is a technique to evaluate predictive models by dividing the original sample to predictive, training and testing set. In case of 10 folds cross validation only one part is left for testing and model is trained on nine sub samples.

### **4.6 Experiments and Results**

Multiple experiments have been performed in order to evaluate the performance of proposed technique. In all these experiments, different preprocessing pipelines have been used and evaluations have been performed.

Performance and accuracy of each experiment was evaluated using different performance measures. These measures have already been described in detail in the sections before.

#### **4.6.1 Results of Jaccard Similarity Measure**

First experiment has been performed for the evaluation of relevance feedback technique. A comparison using Jaccard and without Jaccard similarity measure has been carried out. By using Jaccard all the duplicate tweets has been removed. And it has been seen from the experiment that it is useless to show a user a duplicate tweet as a user need to mark them as relevant or irrelevant multiple times which will be hectic for the ones who is giving relevance feedback.



So, removing the duplicate tweets using Jaccard similarity is really an important and necessary step. Table 8 and table 9 are showing the top tweets against Q1 and Q2.

**Q1** → Reports of injured and dead people

**Expanded Q1** → 'dead', 'morning\_there', 'fatalities', 'reports', 'so\_far', 'casualties', 'lives\_lost', 'injured', 'killed', 'people', 'deaths', 'reported', 'injuries', 'houses'

**Q2** → Shelter needs and shelter locations

**Expanded Q2** → 'sanitation', 'locations', 'food\_shelter', 'medical\_care', 'immediate\_needs', 'clean\_water', 'food\_water', 'shelters', 'shelter', 'water\_food', 'medicine'.

Query	Top Tweets
Q1	<ol style="list-style-type: none"> <li>1. RT @Arab_News: #Nepal #earthquake: - reports of avalanches in #Everest - temples flattened - unknown numbers killed &amp; injured - #bbcbreakin</li> <li>2. Nepal earthquake: Hundreds die, many feared trapped: 970 dead, over 1,700 injured; 539 killed in Kathmanda valley.. <a href="http://t.co/YDqFOeMpBn">http://t.co/YDqFOeMpBn</a></li> <li>3. RT @JaskiratSB: 1457 dead. And counting. <a href="http://t.co/G3d3Ku4n3n">http://t.co/G3d3Ku4n3n</a> <a href="http://t.co/6fjB6P1Ak8">http://t.co/6fjB6P1Ak8</a></li> <li>4. <b>Nepal #earthquake latest: - reports of avalanches in Everest - temples flattened - unknown numbers killed &amp; injured</b> <a href="http://t.co/Qg97ZRymbV\x94">http://t.co/Qg97ZRymbV\x94</a></li> <li>5. <b>#Nepal #earthquake latest: - reports of avalanches in Everest - temples flattened - unknown numbers killed &amp; injured</b> <a href="http://t.co/LS2tcRVaxT">http://t.co/LS2tcRVaxT</a></li> </ol>

<b>Q2</b>	<ol style="list-style-type: none"> <li>1. RT @ArtofLivingABC: #NepalQuake #ArtOfLiving centre in Nepal providing food relief and shelter. Organising Blood Donation. <a href="http://t.co/Dz5o">http://t.co/Dz5o</a></li> <li>2. RT @NeelakshiGswm: Rajasthan CM announces 10,000 relief kits ( food packets, plastic sheets, medicines, water bottles &amp; blankets) for #Nepa</li> <li>3. <b>RT @TomvanderLee: Oxfam experts preparing to fly from UK with supplies to provide clean water, sanitation &amp; emergency food supplies #Nepal</b></li> <li>4. <b>@oxfamgb experts preparing to fly from UK with supplies to provide clean water, sanitation &amp; emergency food supplies. #NepalEarthquake</b></li> <li>5. Nepal Govt needs: Search&amp; Rescue capacity; Medical teams &amp; supplies, tenting x hospitals; Heavy equipment for rubble removal; Helicopters</li> </ol>
-----------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

*Table 8 Relevant Tweets without using Jaccard*

Query	Top Tweets
<b>Q1</b>	<ol style="list-style-type: none"> <li>1. RT @PzFeed: QUAKE DEVASTATION -7.8 quake near Nepal's capital - Nepal: 1,457 dead -India: 34 dead -China: 12 dead -Hundreds missing, trapped</li> <li>2. Earth-quake Slams Nepal 2400 dead and 6200 injured.</li> <li>3. #Nepalearthquake: Death toll crosses 550, several injured <a href="http://t.co/4OjwaeACTW">http://t.co/4OjwaeACTW</a></li> <li>4. Over 400 people dead, 1000 people injured. #PrayForNepal</li> <li>5. At least 26 dead, 113 injured in Bihar; 10 dead, 34 injured in Uttar Pradesh; 3 dead, 30 injured W. Bengal. <a href="http://t.co/mGI76ImUgM">http://t.co/mGI76ImUgM</a> #NepalQuake</li> </ol>

<b>Q2</b>	<ol style="list-style-type: none"> <li>1. Bangladeshi Government send medicine with Doctors, food packets, drinking water and blankets to Nepal. <a href="http://t.co/7YDtJNP8sP">http://t.co/7YDtJNP8sP</a></li> <li>2. Death toll climbs to 2263 in Nepal; shortage of Medical supplies water, food critical; survivors staying outdoors #NepalEarthquake</li> <li>3. ... toilets, food, shelter and emergency aid to people. #NepalEarthquake</li> <li>4. #NepalEarthquake #RedCross @JChapagain @federation topping up relief stocks, plus basic health units, water and sanitation to fill rural gap</li> <li>5. #Fpitch @Oxfam: Clean water &amp; emergency food. You can help #NepalEarthquake <a href="http://t.co/QBT0re8Ed2">http://t.co/QBT0re8Ed2</a> <a href="http://t.co/Esu3ZBQDhc">http://t.co/Esu3ZBQDhc</a></li> </ol>
-----------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

*Table 9 Relevant Tweets using Jaccard*

## 4.6.2 Evaluation of Tweets Relevancy Prediction

In this experiment, evaluation of relevancy based classification has been performed. It can be seen that how query is optimizing with each iteration. The comparison has been done between the system relevant tweets marked by the system and what feedback user has given on them. Table 16 is showing the relevancy based classification results of each iteration.

- **Q1** -> Reports of injured or dead people
- **Q2** -> Infrastructure damage like building, roads, bridges damage
- **Q3** -> Reports of missing people
- **Q4** -> Shelter needs and shelter locations
- **Q5** -> Urgent needs of affected people

Method	Iter1			Iter2			Iter3		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1

<b>Q1</b>	0.91	0.87	0.89	0.93	0.85	0.89	0.97	0.90	0.93
<b>Q2</b>	0.77	0.68	0.72	0.88	0.79	0.83	0.91	0.85	0.88
<b>Q3</b>	0.81	0.85	0.83	0.90	0.93	0.91	0.93	0.95	0.94
<b>Q4</b>	0.89	0.92	0.90	0.92	0.95	0.93	0.96	0.89	0.92
<b>Q5</b>	0.85	0.81	0.83	0.89	0.76	0.82	0.72	0.68	0.70
<b>MAP</b>	<b>0.84</b>			<b>0.90</b>			<b>0.89</b>		

Table 10 Relevancy Based Classification

### 4.6.3 Evaluation of Relevancy Feedback Classifier 60:40

In the second experiment, evaluation of relevance feedback technique has been performed using means average precision measure. This measure has been used by many researchers to calculate the accuracy of the system using precision and recall. There are three classifiers that were used Naïve Bayes, Random Forest and SVM (Support Vector Machine). These classifiers were trained on both TF (term frequency) and TF-IDF (term frequency- inverse document frequency). It has been shown that in one class TF has performed better while in another class TF-IDF has performed better but overall SVM has given much better results than the other two. Table 10 shows the precision and recall of TF while Table 10 shows the precision and recall of TF-IDF.

**Tweet converted to TF**

Method	Irrelevant			Relevant			Average/ Total		
	Pre	Recall	F1	Pre	Recall	F1	Pre	Recall	F1
MNB	0.68	1.00	0.81	0.99	0.38	0.55	0.81	0.73	0.69
RF	0.80	0.03	0.05	0.34	0.99	0.51	0.64	0.35	0.21
SVM	0.92	1.00	0.96	0.99	0.83	0.90	0.94	0.94	0.94

Table 11 Relevance Feedback System Precision and Recall TF (pre shows precision)

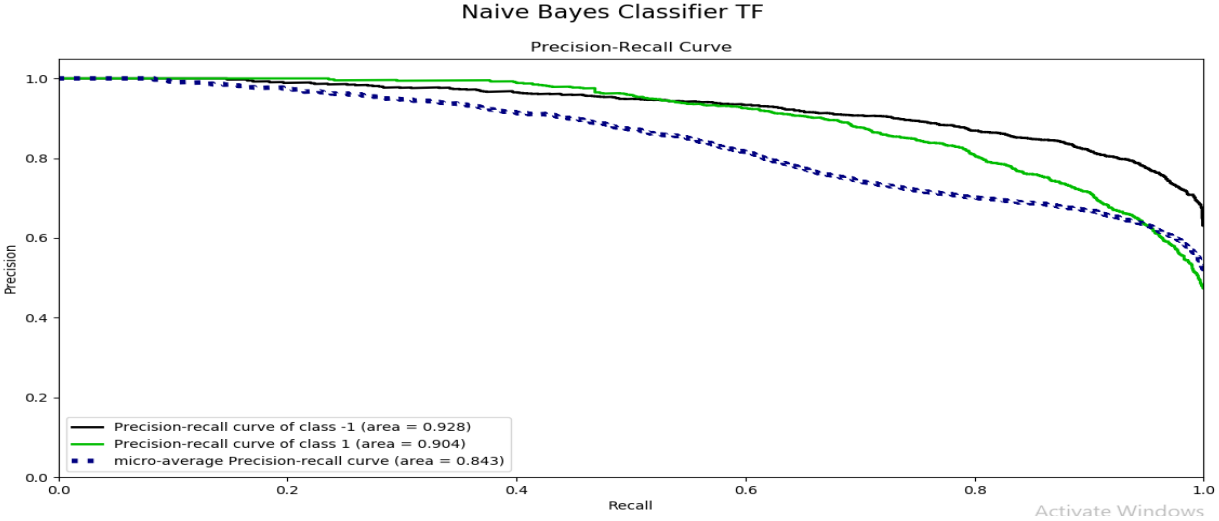


Figure 9 PR Curve for MNB using TF

Figure 9 is showing the PR curve of Naïve Bayes using term frequency; the classifier has shown really good results. It is predicting irrelevant class correctly 92% of the time while the relevant classes are predicting 90% of the time correctly. It is covering a lot of area under the curve which is a sign of good precision and recall.

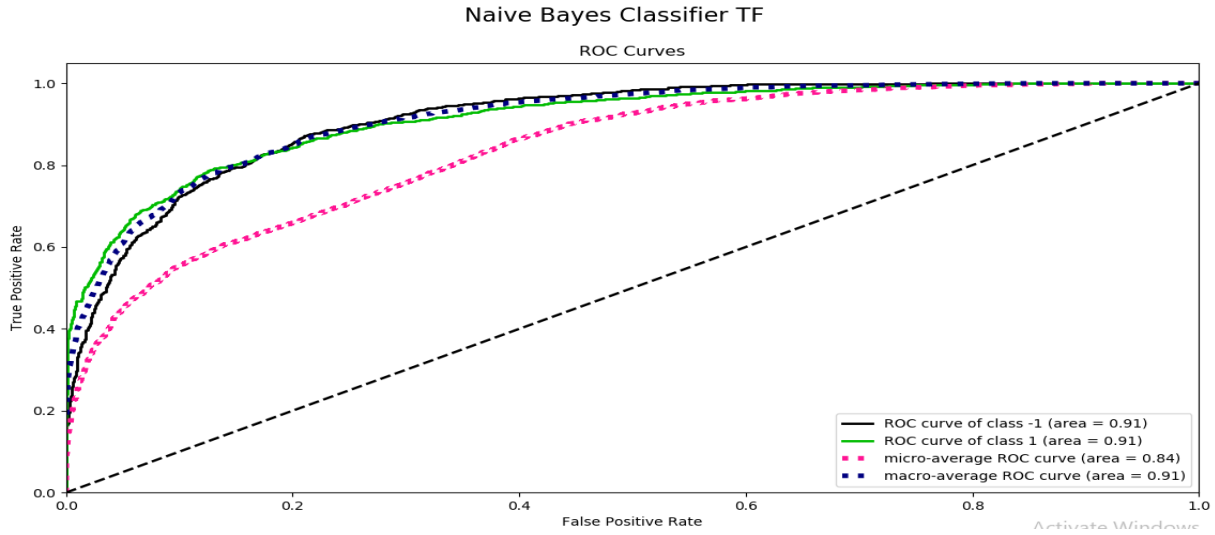


Figure 10 ROC Curve for MNB using TF

Figure 10 is showing the ROC curve of Naïve Bayes using term frequency; the classifier has shown really good results. It is predicting irrelevant class correctly 91% of the time while the relevant classes are predicting 91% of the time correctly. It is covering a lot of area under the curve which is a sign of good precision and recall.

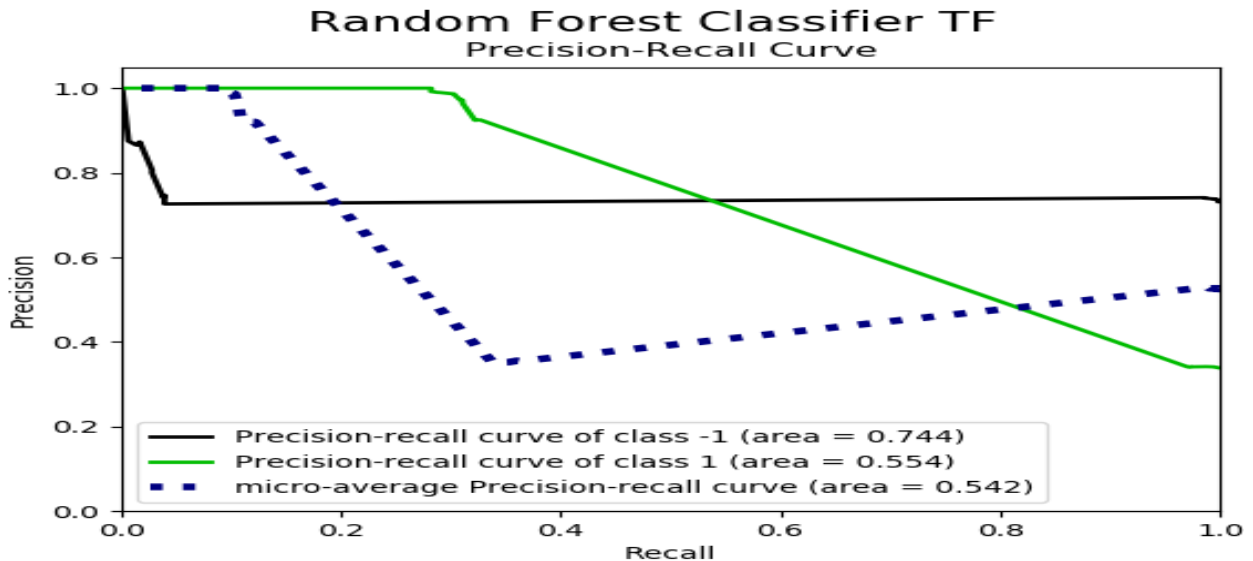


Figure 11 PR Curve for RF using TF

Figure 11 is showing the PR curve of Random Forest using term frequency; the classifier has shown average results. It is predicting irrelevant class correctly 74% of the time while the relevant classes are predicting 54% of the time correctly. It is covering less area under the curve which is a sign of average precision and recall.

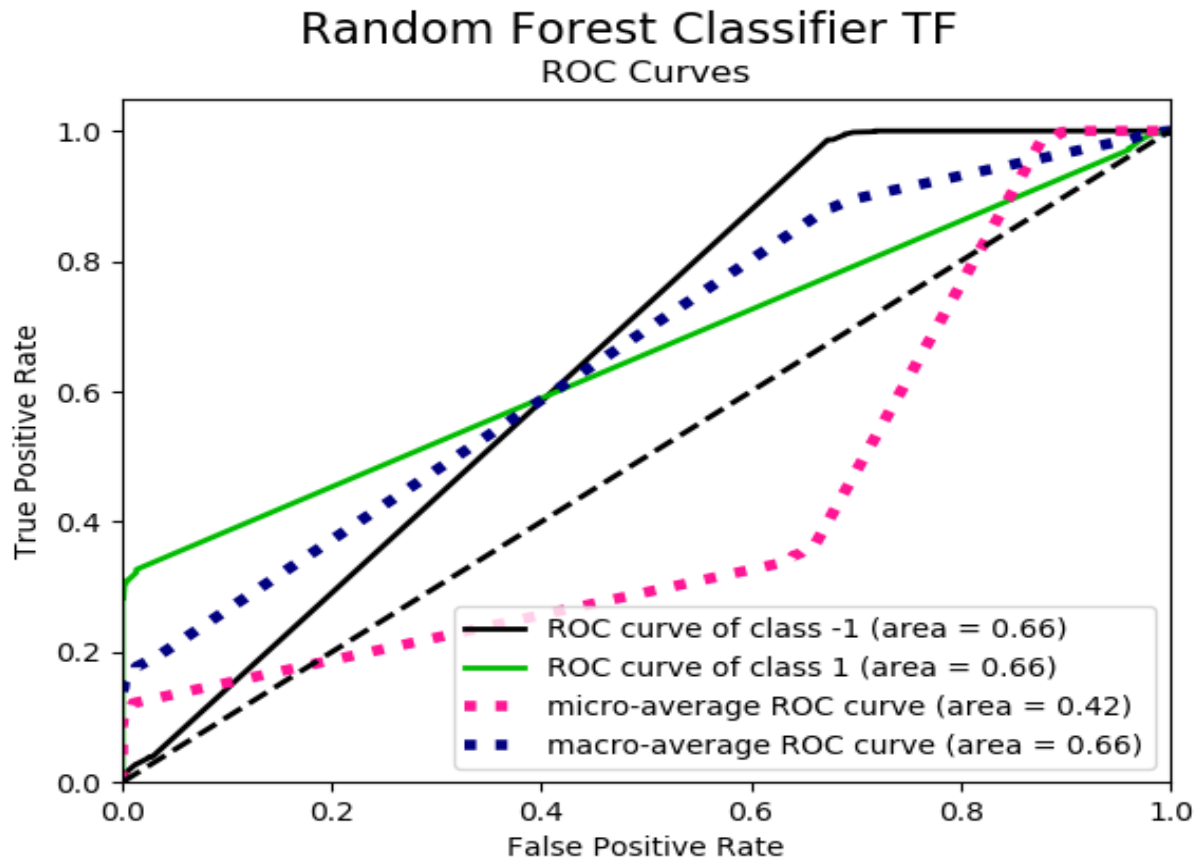


Figure 12 ROC Curve for RF using TF

Figure 12 is showing the ROC curve of Random Forest using term frequency; the classifier has shown average results. It is predicting irrelevant class correctly 66% of the time while the relevant classes are predicting 66% of the time correctly. It is covering very less area under the curve which is a sign of average precision and recall.

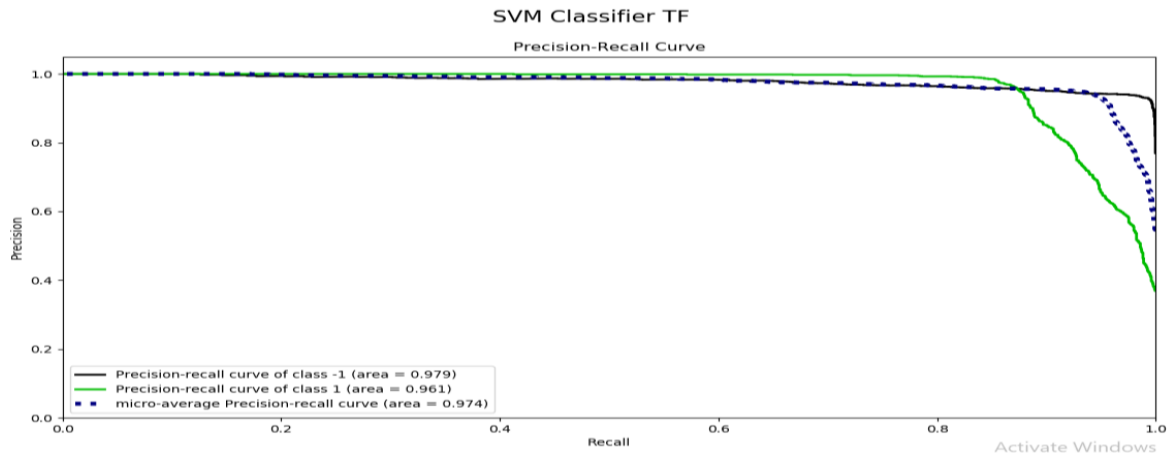


Figure 13 PR Curve for SVM using TF

Figure 13 is showing the PR curve of Support Vector Machine (SVM) using term frequency; the classifier has shown really good results. It is predicting irrelevant class correctly 97% of the time while the relevant classes are predicting 96% of the time correctly. It is covering a lot of area under the curve which is a sign of good precision and recall.

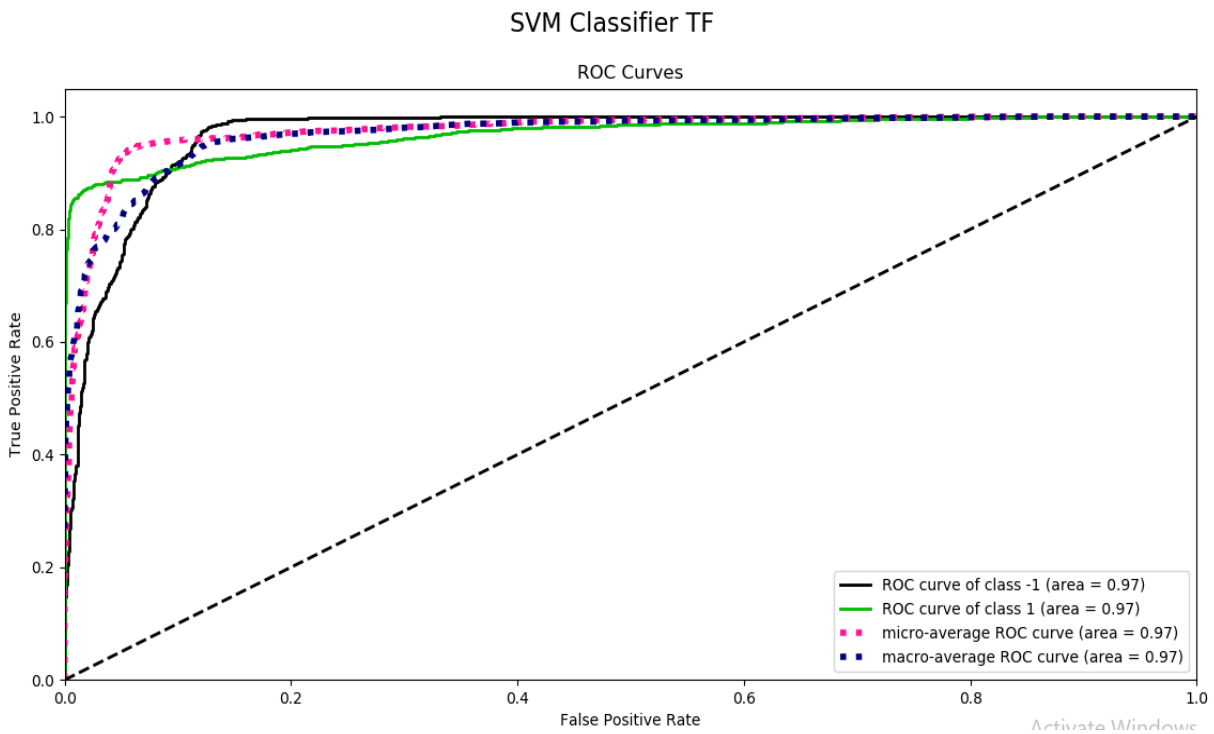


Figure 14 ROC Curve for SVM using TF



Figure 14 is showing the ROC curve of Support Vector Machine (SVM) using term frequency; the classifier has shown really good results. It is predicting irrelevant class correctly 97% of the time while the relevant classes are predicting 97% of the time correctly. It is covering a lot of area under the curve which is a sign of good precision and recall.

<b>Tweet converted to TF-IDF</b>									
<b>Method</b>	<b>Irrelevant</b>			<b>Relevant</b>			<b>Average/ Total</b>		
	<b>Pre</b>	<b>Recall</b>	<b>F1</b>	<b>Pre</b>	<b>Recall</b>	<b>F1</b>	<b>Pre</b>	<b>Recall</b>	<b>F1</b>
<b>MNB</b>	0.67	0.99	0.80	0.98	0.36	0.53	0.80	0.72	0.68
<b>RF</b>	0.82	0.98	0.89	0.92	0.57	0.71	0.85	0.84	0.83
<b>SVM</b>	0.89	0.99	0.94	0.96	0.77	0.85	0.92	0.91	0.91

*Table 12 Relevance Feedback System Precision and Recall TF-IDF (pre shows precision)*

Figure 15 is showing the PR curve of Naïve Bayes using term frequency-inverse document frequency; the classifier has shown really good results. It is predicting irrelevant class correctly 91% of the time while the relevant classes are predicting 88% of the time correctly. It is covering a lot of area under the curve which is a sign of good precision and recall.

### Naive Bayes Classifier TF-IDF

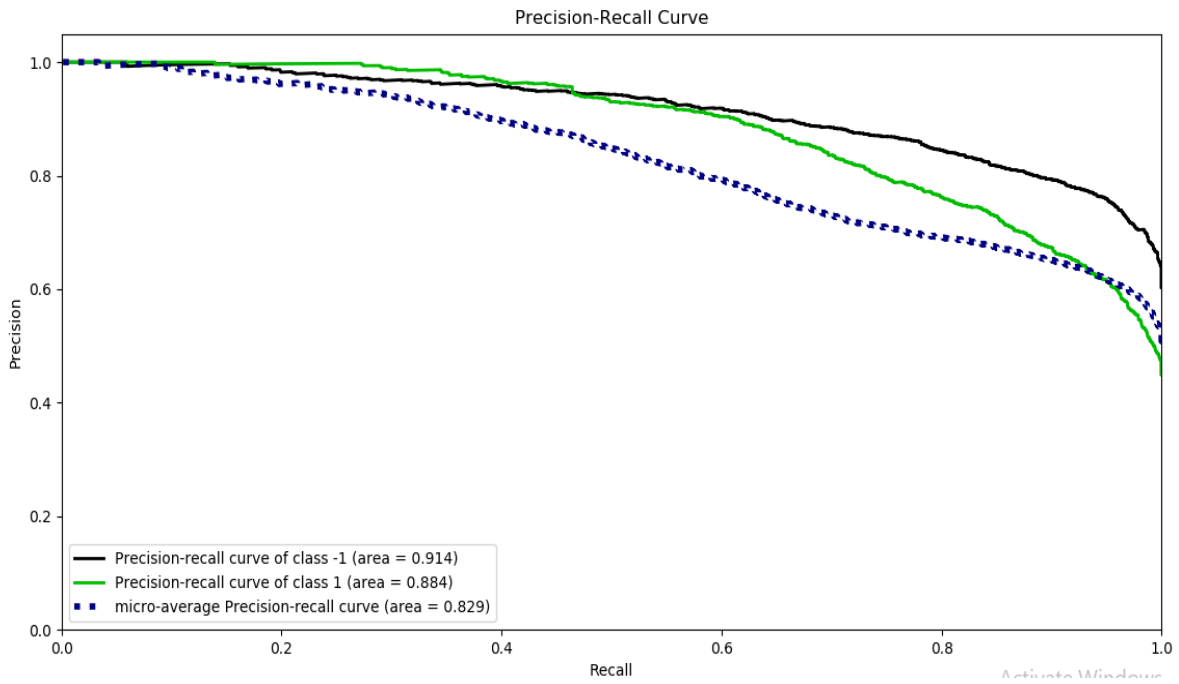


Figure 15 PR Curve for MNB using TF-IDF

### Naive Bayes Classifier TF-IDF

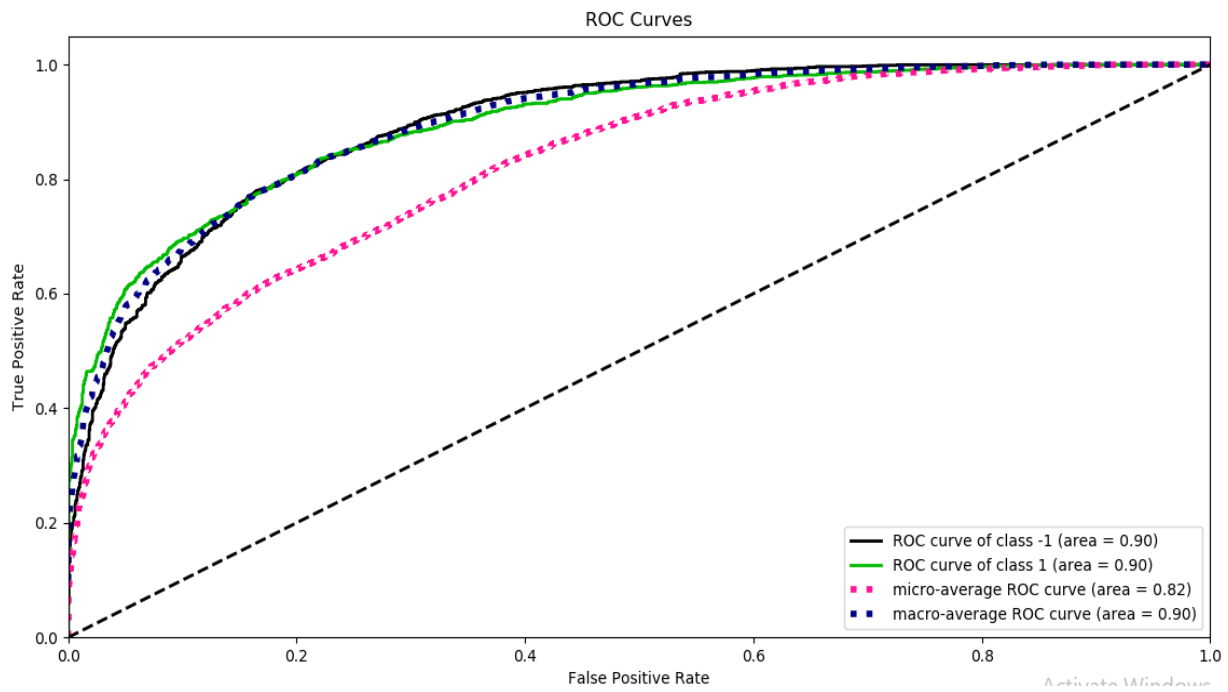


Figure 16 ROC Curve for MNB using TF-IDF

Figure 16 is showing the ROC curve of Naïve Bayes using term frequency-inverse document frequency; the classifier has shown really good results. It is predicting irrelevant class correctly 90% of the time while the relevant classes are predicting 90% of the time correctly. It is covering a lot of area under the curve which is a sign of good precision and recall.

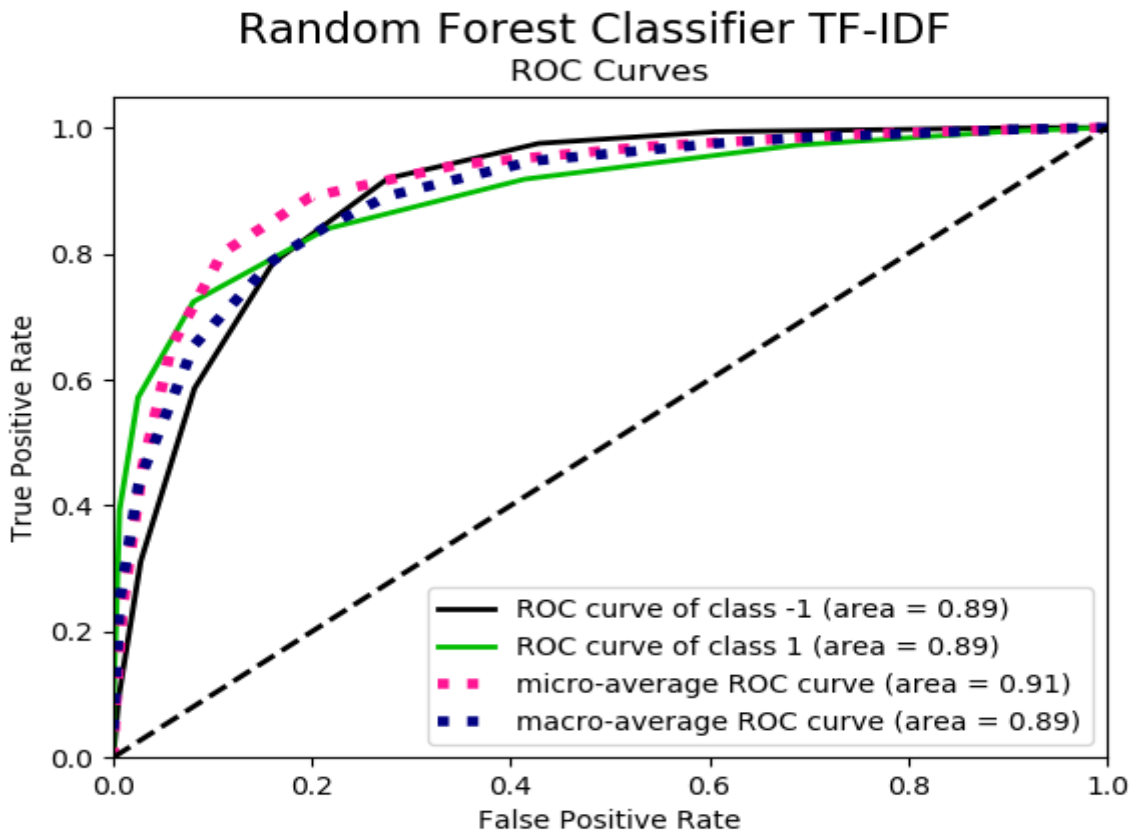


Figure 17 PR Curve for RF using TF-IDF

Figure 17 is showing the PR curve of Random Forest using term frequency-inverse document frequency; the classifier has shown good results. It is predicting irrelevant class correctly 89% of the time while the relevant classes are predicting 89% of the time correctly. It is covering good area under the curve which is a sign of better precision and recall.

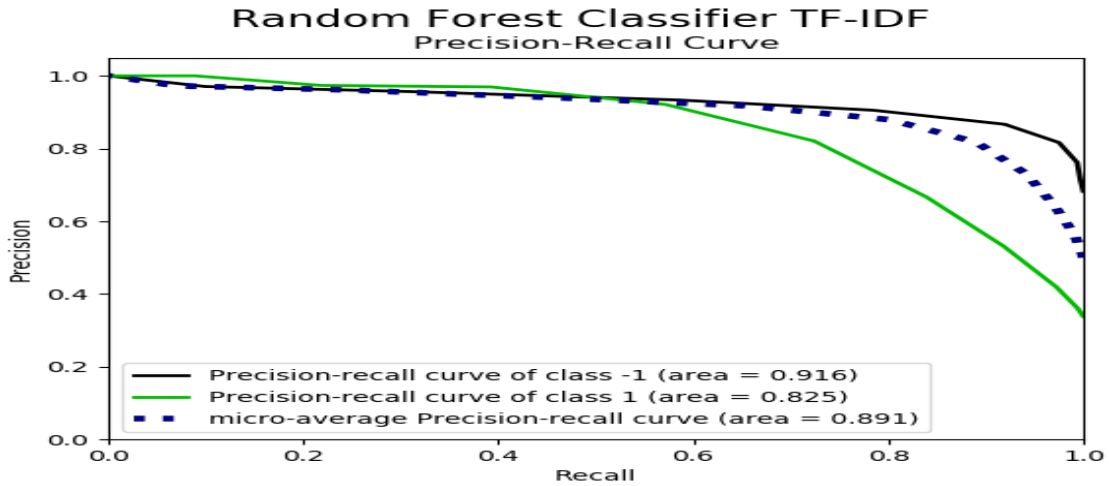


Figure 18 ROC Curve for RF using TF-IDF

Figure 18 is showing the ROC curve of Random Forest using term frequency-inverse document frequency; the classifier has shown good results. It is predicting irrelevant class correctly 91% of the time while the relevant classes are predicting 82% of the time correctly. It is covering a lot of area under the curve which is a sign of good precision and recall.

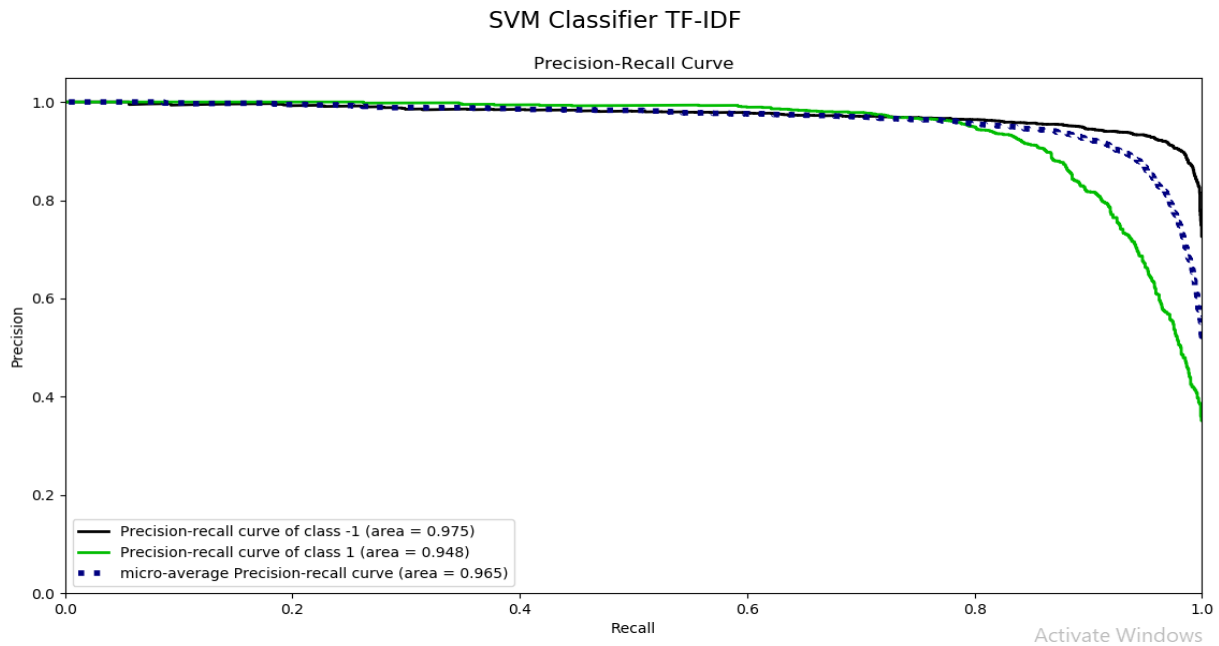


Figure 19 PR Curve for SVM using TF-IDF

Figure 19 is showing the PR curve of Support Vector Machine (SVM) using term frequency-inverse document frequency; the classifier has shown really good results. It is predicting irrelevant class correctly 97% of the time while the relevant classes are predicting 94% of the time correctly. It is covering a lot of area under the curve which is a sign of good precision and recall.

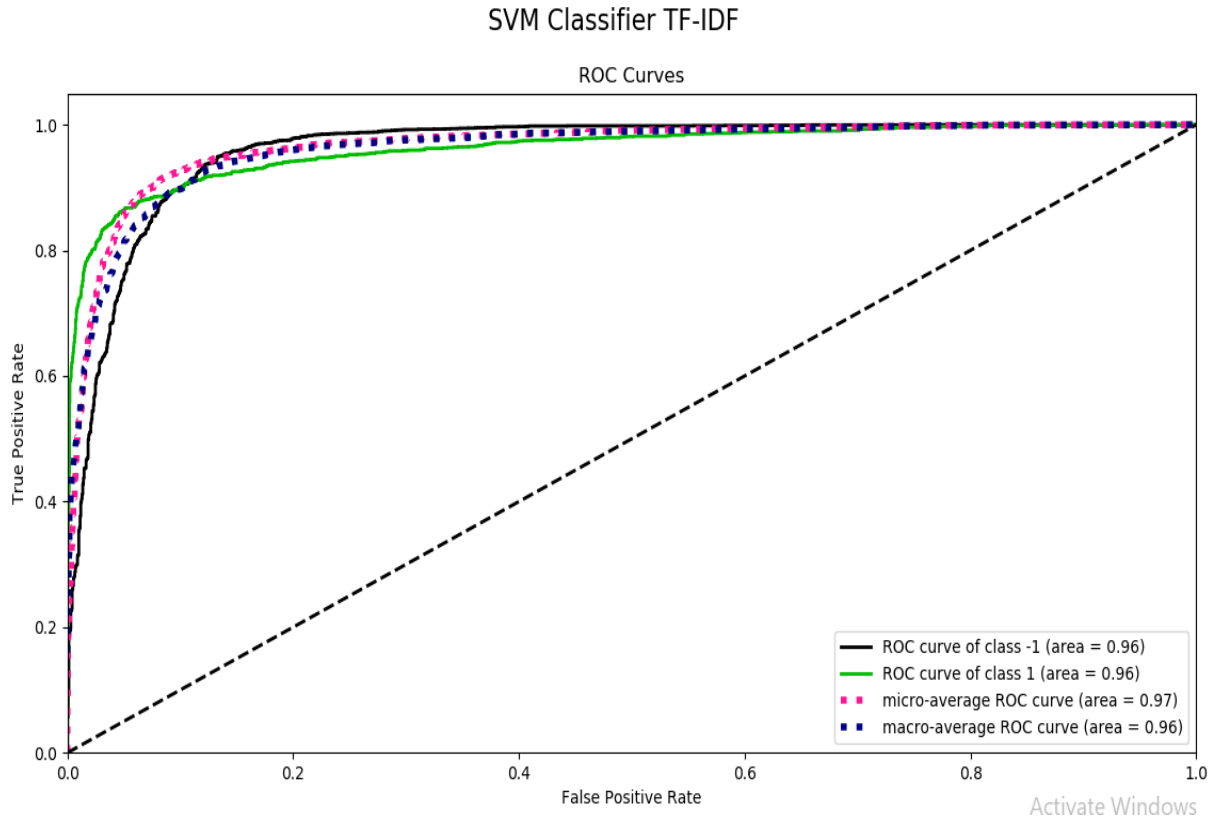


Figure 20 ROC Curve for SVM using TF-IDF

Figure 20 is showing the ROC curve of Support Vector Machine (SVM) using term frequency-inverse document frequency; the classifier has shown really good results. It is predicting irrelevant class correctly 96% of the time while the relevant classes are predicting 96% of the time correctly. It is covering a lot of area under the curve which is a sign of good precision and recall. From all the classifiers, it can be seen that SVM is performing the best amongst all.

## 4.6.4 Evaluation of Relevancy Feedback Classifier 10-Folds

In another experiment, evaluation of relevance feedback module has been performed using 10-Fold cross validation. During 10-fold cross validation, the total dataset has been divided in 10 subsamples. From these subsamples, 9 subsamples are used for training while only one subsample is use for testing and predicting. This shows that dataset is trained or larger dataset. There are three classifiers that are used Naïve Bayes, Random Forest and SVM (Support Vector Machine). These classifiers are trained on both TF (term frequency) and TF-IDF (term frequency- inverse document frequency). It has been shown from the experiment that Naïve Bayes and SVM are performing well better than Random Forest. If we go little bit deeper then we can see Random Forest is predicting relevant classes much better than the irrelevant ones. Table 12 shows the precision and recall of TF using 10-fold validation while Table 13 shows the precision and recall of TF-IDF using 10-fold validation.

Method	Irrelevant			Relevant			Average/ Total		
	Pre	Recall	F1	Pre	Recall	F1	Pre	Recall	F1
<b>MNB</b>	0.83	0.95	0.89	0.92	0.74	0.82	0.87	0.86	0.86
<b>RF</b>	0.57	1.00	0.73	0.97	0.02	0.03	0.74	0.57	0.42
<b>SVM</b>	0.94	0.97	0.95	0.95	0.92	0.94	0.95	0.95	0.95

*Table 13 Relevance Feedback System Precision and Recall TF using 10-Fold Validation (pre shows precision)*

Method	Irrelevant			Relevant			Average/ Total		
	Pre	Recall	F1	Pre	Recall	F1	Pre	Recall	F1
<b>MNB</b>	0.82	0.94	0.88	0.90	0.73	0.81	0.86	0.85	0.85
<b>RF</b>	0.79	0.96	0.87	0.93	0.66	0.77	0.85	0.83	0.83
<b>SVM</b>	0.95	0.95	0.95	0.93	0.93	0.93	0.94	0.94	0.94

Table 14 Relevance Feedback System Precision and Recall TF-IDF using 10-Fold Validation (pre shows precision)

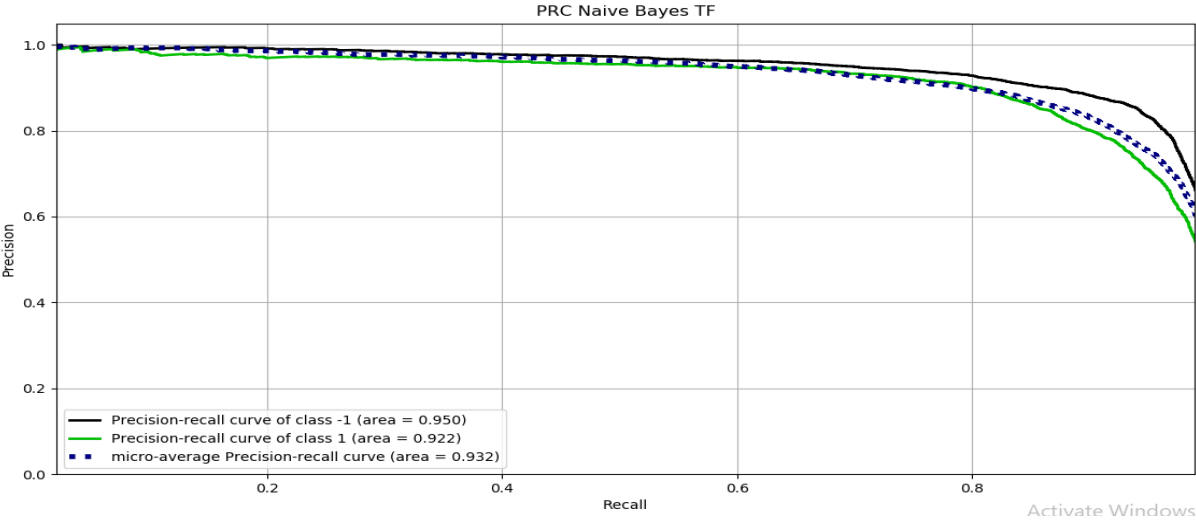


Figure 21 PR Curve for MNB using TF

Figure 21 is showing the PR curve of Naïve Bayes using term frequency; the classifier has shown really good results. It is predicting irrelevant class correctly 95% of the time while the relevant classes are predicting 92% of the time correctly. It is covering a lot of area under the curve which is a sign of good precision and recall.

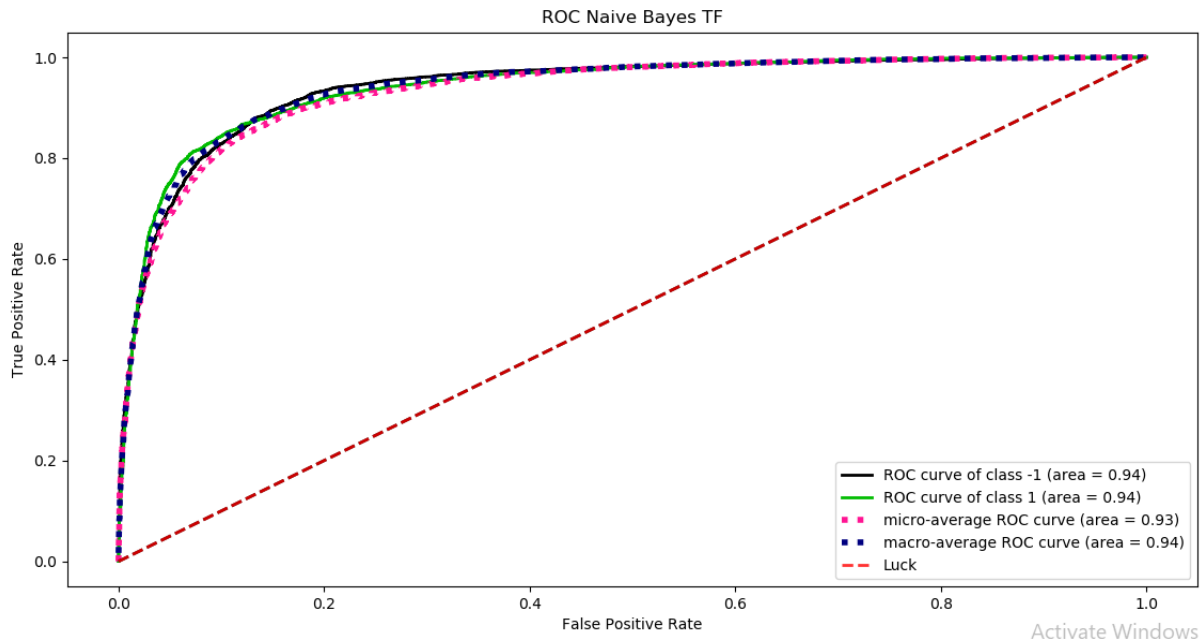


Figure 22 ROC Curve for MNB using TF

Figure 22 is showing the ROC curve of Naïve Bayes using term frequency; the classifier has shown really good results. It is predicting irrelevant class correctly 94% of the time while the relevant classes are predicting 94% of the time correctly. It is covering a lot of area under the curve which is a sign of good precision and recall.

Figure 23 is showing the PR curve of Random Forest using term frequency; the classifier has shown average results. It is predicting irrelevant class correctly 65% of the time while the relevant classes are predicting 58% of the time correctly. It is covering very less area under the curve which is a sign of not good precision and recall.



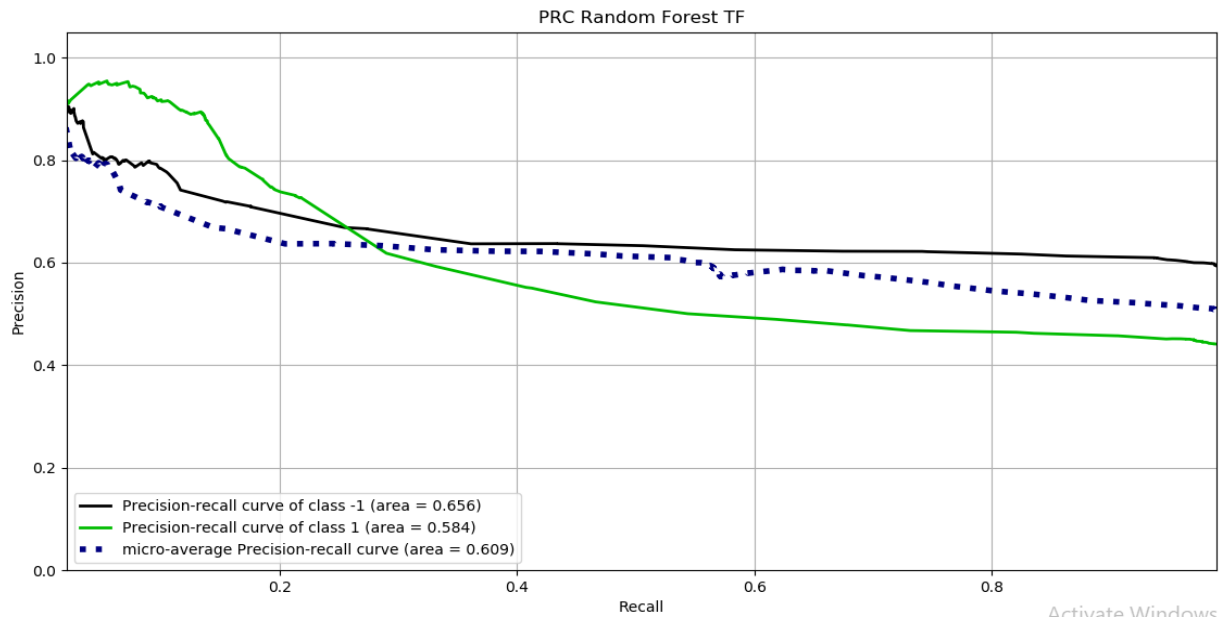


Figure 23 PR Curve for RF using TF

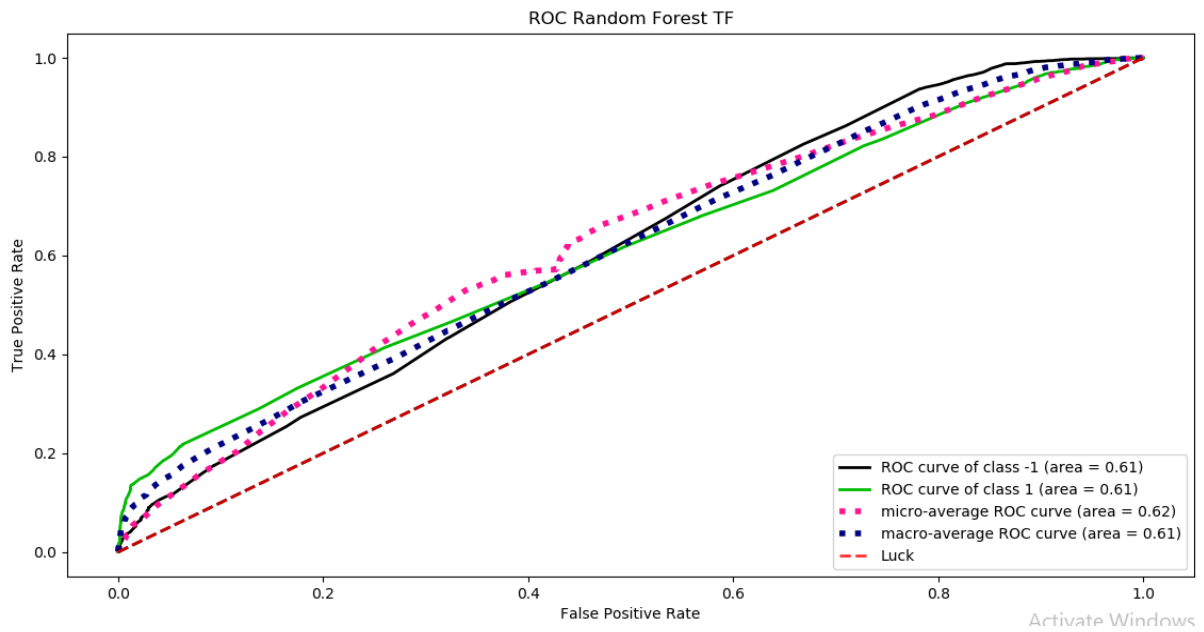
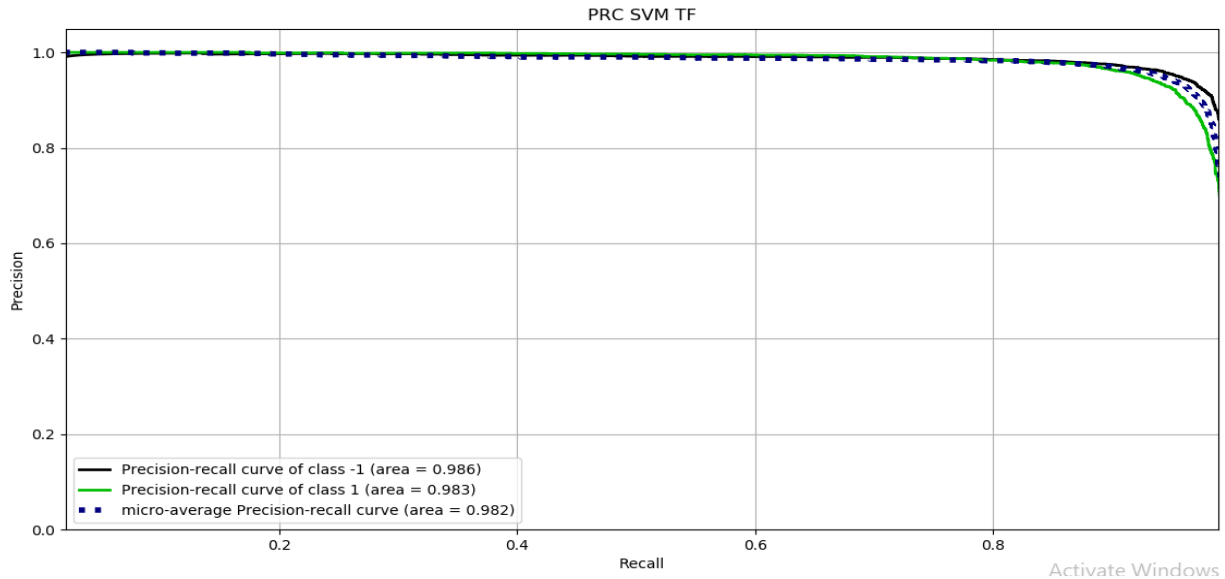


Figure 24 ROC Curve for RF using TF

Figure 24 is showing the ROC curve of Random Forest using term frequency; the classifier has shown average results. It is predicting irrelevant class correctly 61% of the time while the relevant classes are predicting 61% of the time correctly. It is not covering a lot of area under the curve which is a sign of not good precision and recall.



*Figure 25 PR Curve for SVM using TF*

Figure 25 is showing the PR curve of Support Vector Machine using term frequency; the classifier has shown really good results. It is predicting irrelevant class correctly 98% of the time while the relevant classes are predicting 98% of the time correctly. It is covering a lot of area under the curve which is a sign of good precision and recall.

Figure 26 is showing the ROC curve of Support Vector Machine using term frequency; the classifier has shown really good results. It is predicting irrelevant class correctly 98% of the time while the relevant classes are predicting 98% of the time correctly. It is covering a lot of area under the curve which is a sign of good precision and recall.

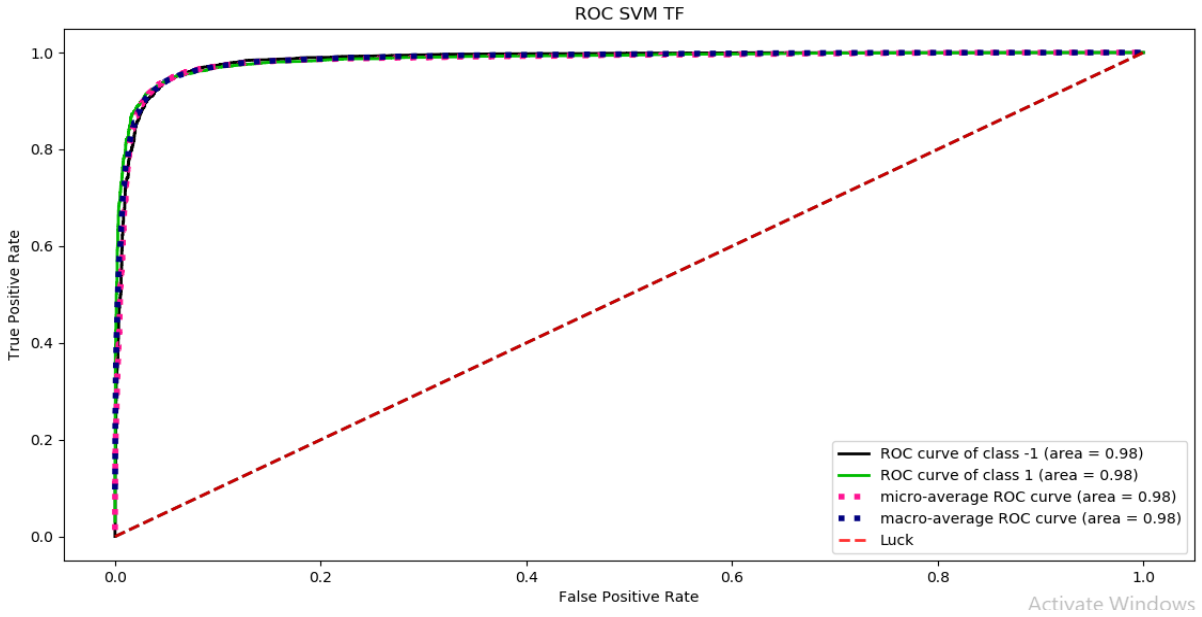


Figure 26 ROC Curve for SVM using TF

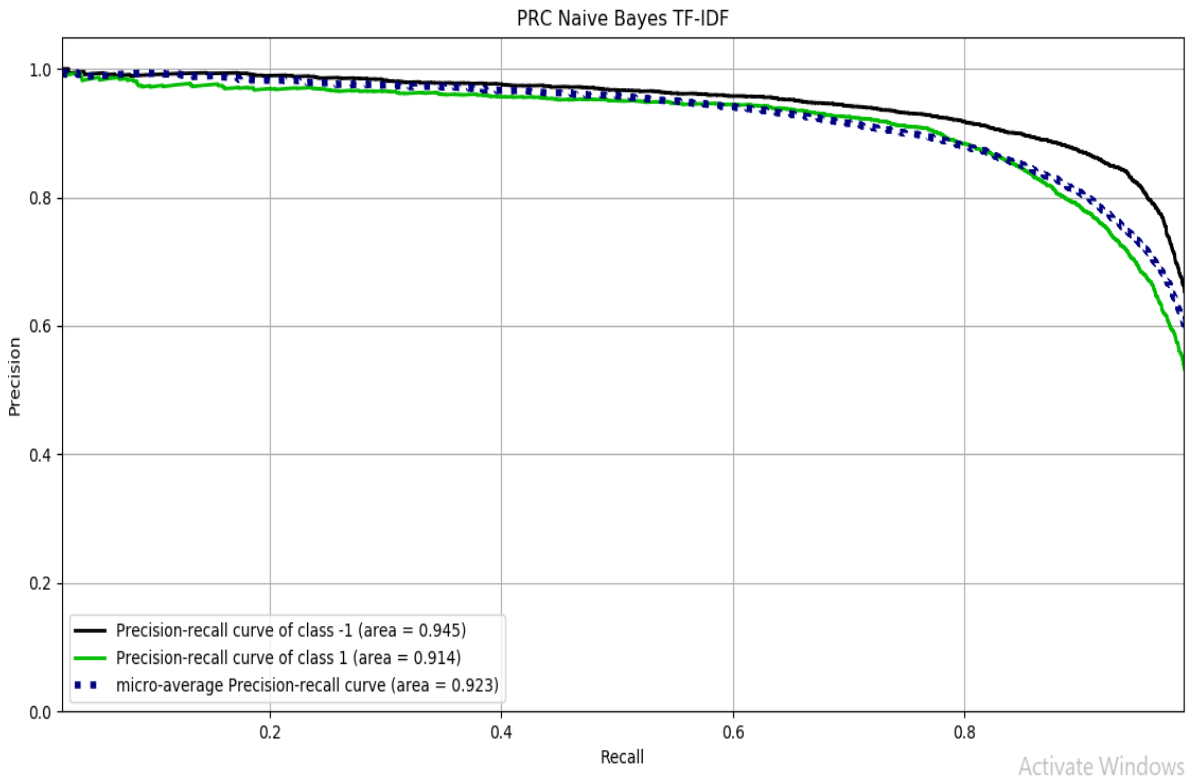


Figure 27 PR Curve for MNB using TF-IDF

Figure 27 is showing the PR curve of Naïve Bayes using term frequency-inverse document frequency; the classifier has shown really good results. It is predicting irrelevant class correctly 94% of the time while the relevant classes are predicting 91% of the time correctly. It is covering a lot of area under the curve which is a sign of good precision and recall.

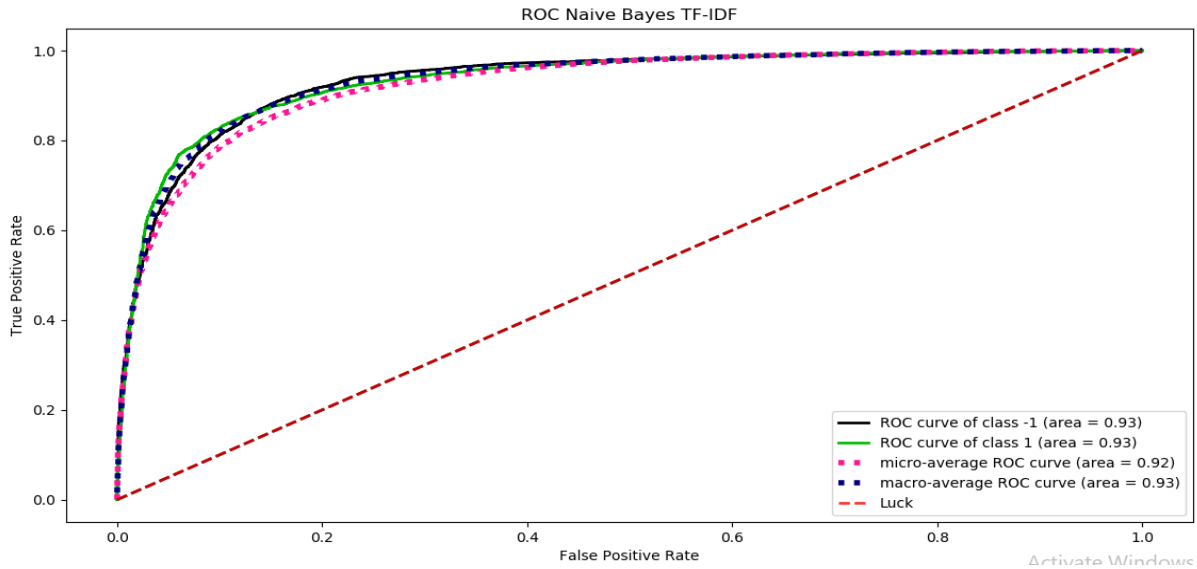


Figure 28 ROC Curve for MNB using TF-IDF

Figure 28 is showing the ROC curve of Naïve Bayes using term frequency-inverse document frequency; the classifier has shown really good results. It is predicting irrelevant class correctly 93% of the time while the relevant classes are predicting 93% of the time correctly. It is covering a lot of area under the curve which is a sign of good precision and recall.

Figure 29 is showing the PR curve of Random Forest using term frequency-inverse document frequency; the classifier has shown really good results. It is predicting irrelevant class correctly 94% of the time while the relevant classes are predicting 90% of the time correctly. It is covering a lot of area under the curve which is a sign of good precision and recall.

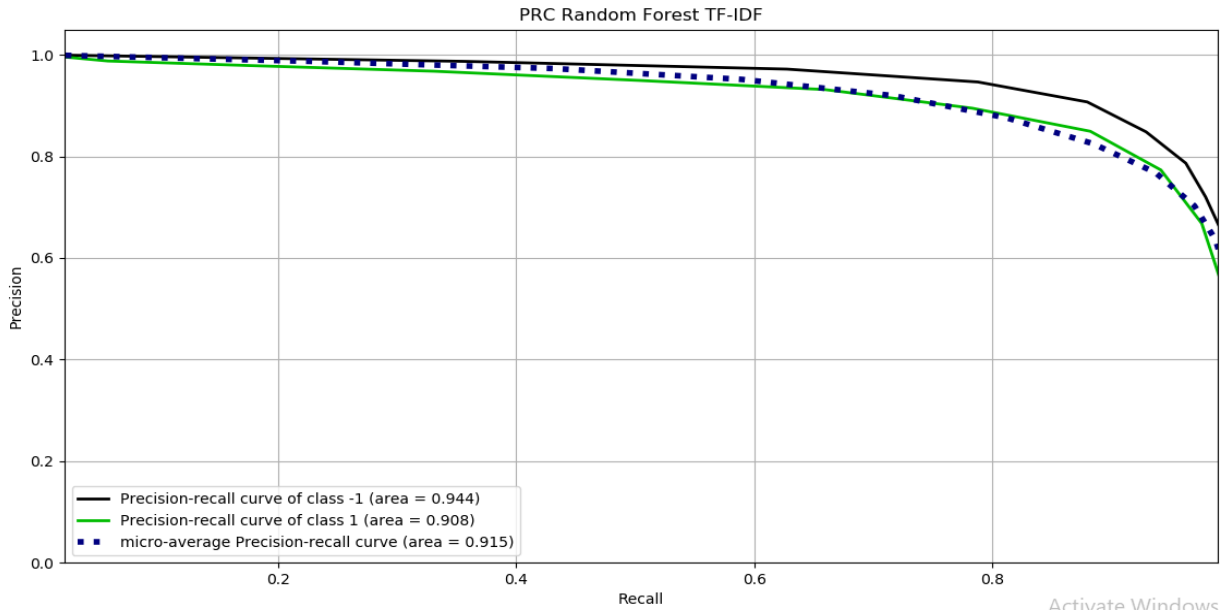


Figure 29 PR Curve for RF using TF-IDF

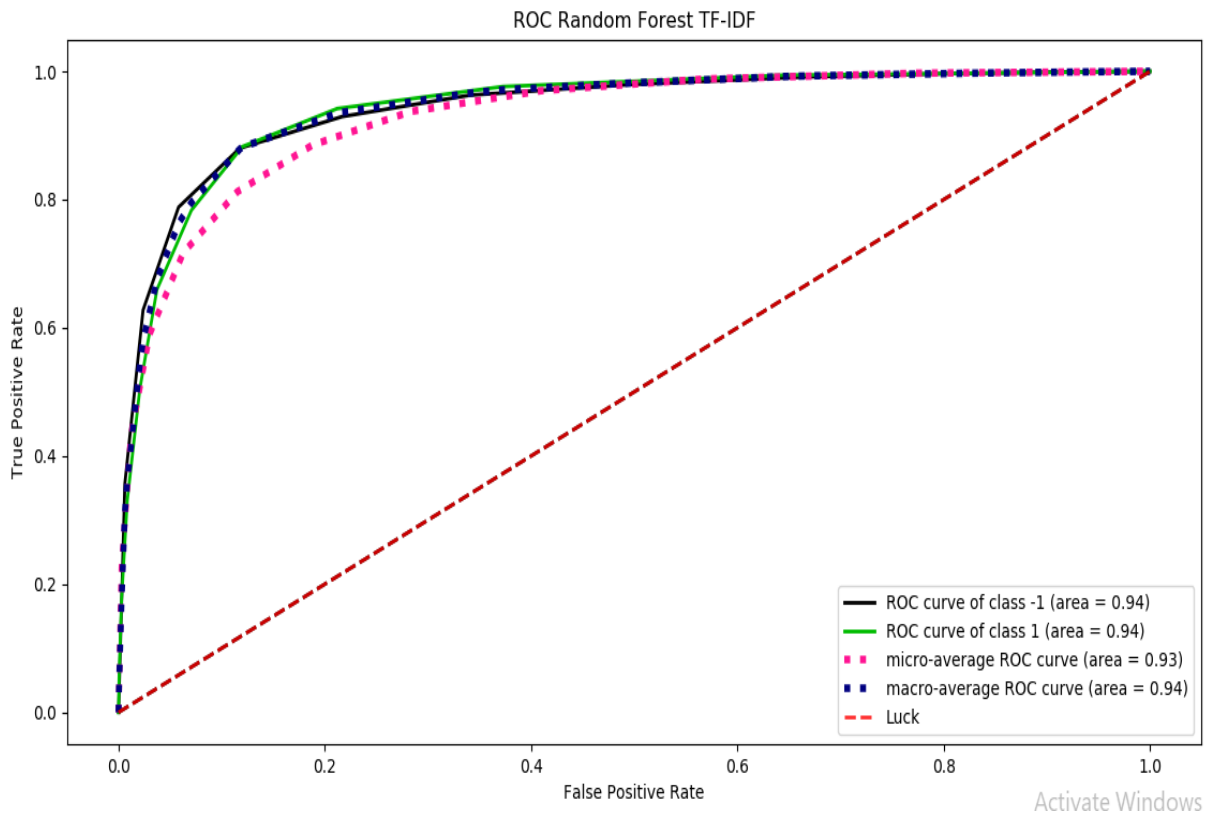
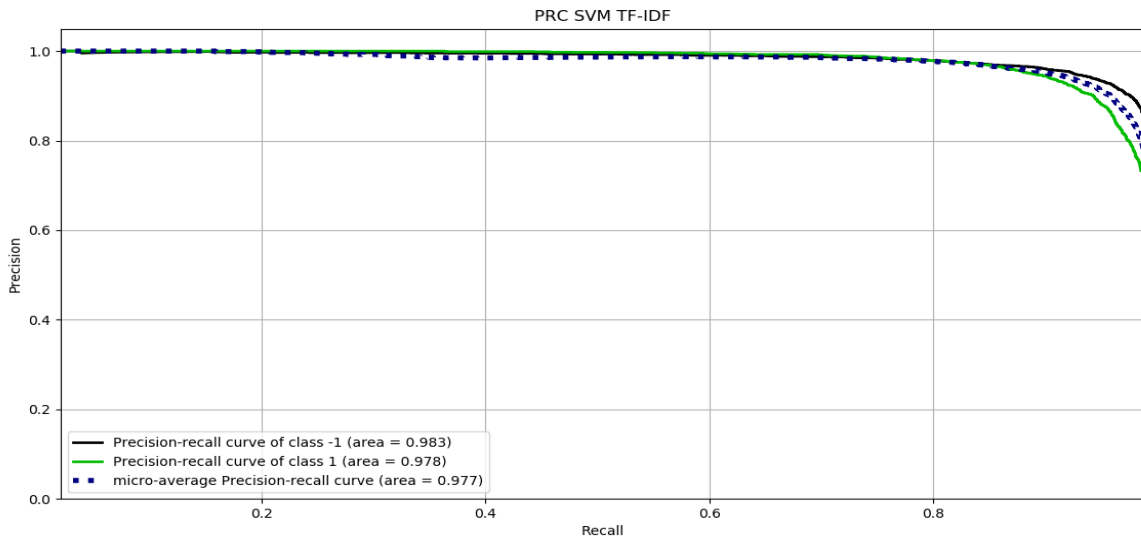


Figure 30 ROC Curve for RF using TF-IDF

Figure 30 is showing the ROC curve of Random Forest using term frequency-inverse document frequency; the classifier has shown really good results. It is predicting irrelevant class correctly 94% of the time while the relevant classes are predicting 94% of the time correctly. It is covering a lot of area under the curve which is a sign of good precision and recall.



*Figure 31 PR Curve for SVM using TF-IDF*

Figure 31 is showing the PR curve of Support Vector Machine (SVM) using term frequency-inverse document frequency; the classifier has shown really good results. It is predicting irrelevant class correctly 98% of the time while the relevant classes are predicting 97% of the time correctly. It is covering a lot of area under the curve which is a sign of good precision and recall.

Figure 32 is showing the ROC curve of Support Vector Machine (SVM) using term frequency-inverse document frequency; the classifier has shown really good results. It is predicting irrelevant class correctly 98% of the time while the relevant classes are predicting 98% of the time correctly. It is covering a lot of area under the curve which is a sign of good precision and recall. From all the classifiers, it can be seen that SVM is performing the best amongst all.

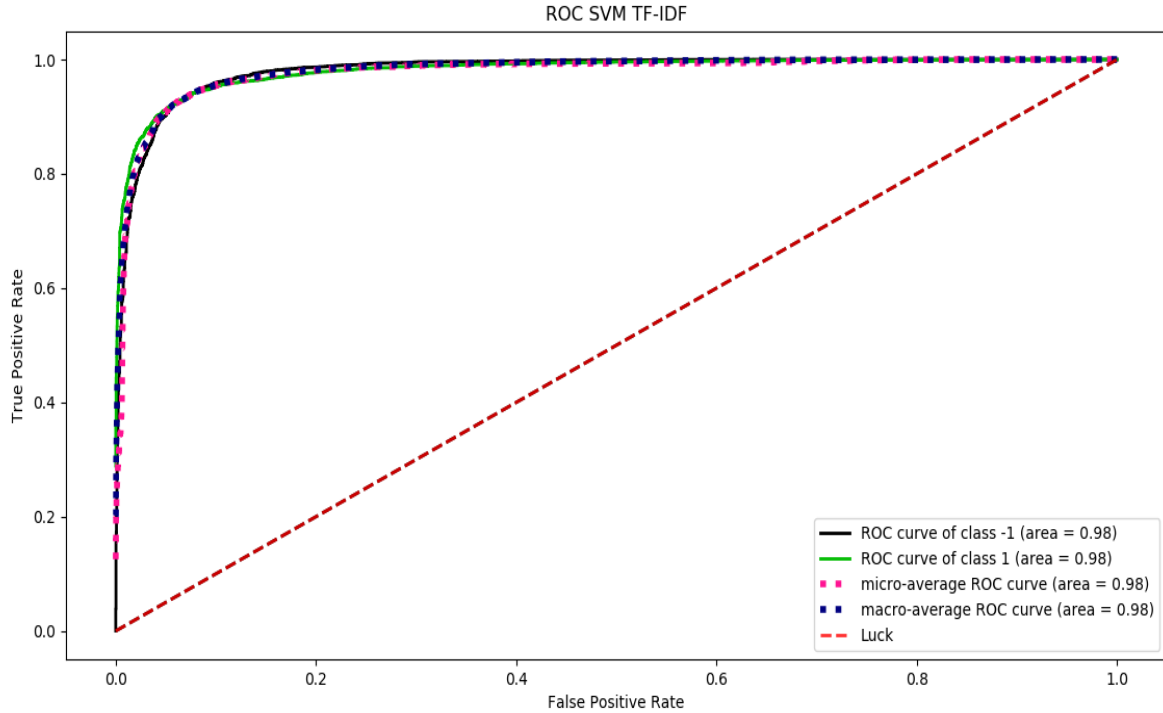


Figure 32 ROC Curve for SVM using TF-IDF

### 4.6.5 Evaluation of Relevant Source Extraction

In this experiment, evaluation of binning or classifying module has been performed. In this module the source has been classified to from highly relevant to highly irrelevant to certain topic. The sources have been classified from bin1 to bin4. Bin1 is showing the sources who are highly irrelevant to a certain topic while in bin4 contains those sources who are highly relevant to certain topic.

It can be seen that the system is verified and tested on two formulas in original formula which has taken from the paper [13] as a base system of classifying sources, that original formula is giving really bad results on our dataset while when the formula has been changed which is our proposed formula or technique then it can be seen that model is predicting all the bins really well.

It can also be seen, if we take individual feature then that information is too less to classify a bins or users and results were also not up to the mark. Only the results of updated formula or proposed technique are good.

Table 14 is showing different precision and recall scores of different bins against different features.

Method	Bin 1			Bin 2			Bin 3			Bin 4		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
<b>Similarity Score</b>	0.15	0.5	0.2	0.12	0.5	0.20	0.28	0.30	0.29	0.99	0.28	0.44
<b>User Activity</b>	0.99	0.12	0.22	0.07	0.29	0.01	0.01	0.60	0.26	0.09	0.33	0.19
<b>User Influence</b>	0.59	0.91	0.72	0.31	0.70	0.43	0.32	0.40	0.36	0.94	0.31	0.47
<b>User Event Engagement</b>	0.77	0.66	0.71	0.32	0.66	0.43	0.37	0.46	0.41	0.80	0.32	0.46
<b>Favorite Engagement</b>	0.69	0.60	0.64	0.30	0.57	0.40	0.46	0.44	0.45	0.57	0.31	0.40



<b>RT Engagement</b>	0.87	0.38	0.53	0.20	0.39	0.26	0.31	0.43	0.36	0.47	0.29	0.36
<b>Social Popularity</b>	0.41	0.56	0.47	0.62	0.42	0.50	0.39	0.43	0.41	0.11	0.34	0.17
<b>Original / Base Formula</b>	0.41	0.33	0.37	0.73	0.32	0.44	0.35	0.55	0.43	0.09	0.03	0.04
<b>Updated / Proposed Formula</b>	0.89	0.78	0.83	0.79	0.84	0.81	0.73	0.88	0.80	0.81	0.78	0.79

Table 15 Classification of Users (pre is precision and rec is recall)

### Comparison Base and Proposed Formula

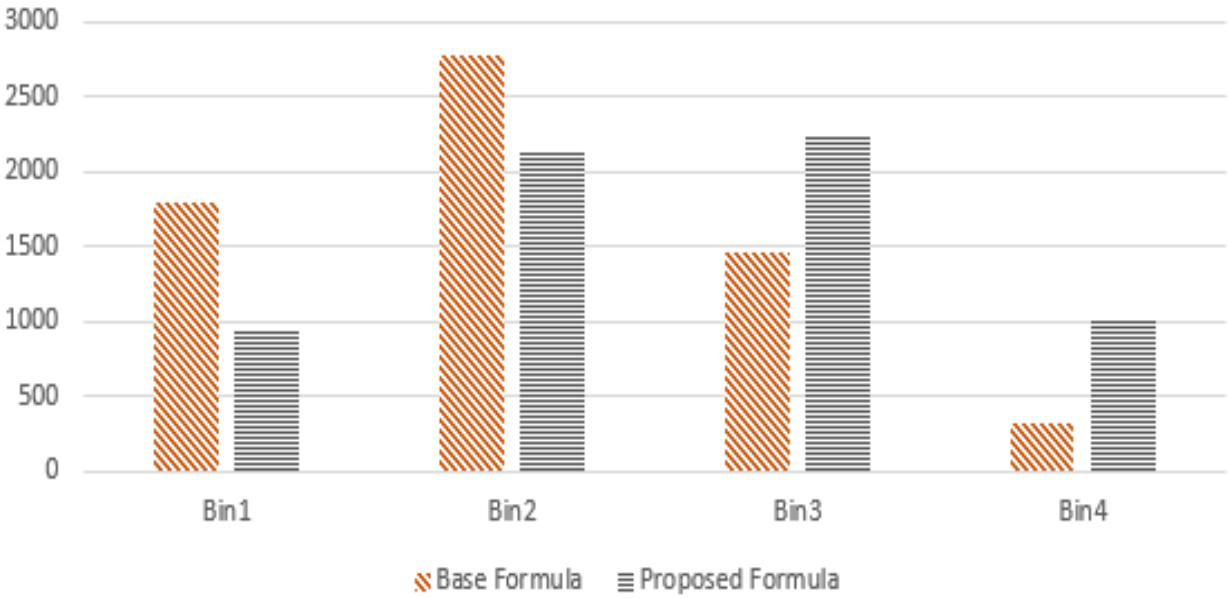


Figure 33 Comparison between Base and Proposed Formula

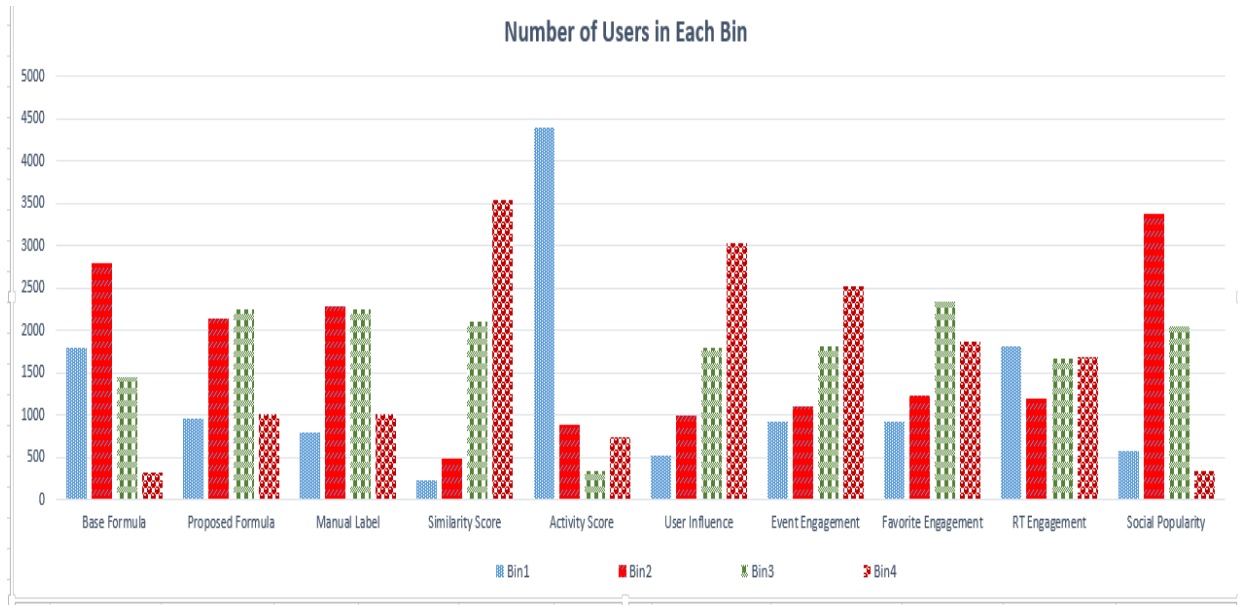


Figure 34 Distribution of Sources in Each Bin

Figure 33 shows comparison between proposed formula and base system and figure 34 is showing the plots of different measures and distribution of sources in each bin. It can be seen clearly that proposed formula has balanced distribution of users. And it can be seen that mostly people are laying in bin2 and bin3 because these are the points where it is difficult to differentiate much. But the extremes like bin1 and bin2 is having average number of sources.

So, it can be seen from the above mentioned results that the proposed technique as compared with all the measures used have given much better results, can be seen in table 13 in proposed formula section.

#### 4.6.6 Evaluation of Relevant Source Extraction using NDCG

It's really common to check the accuracy of the model using precision and recall measure but NDCG (normalized discounted cumulative gain) is the standard Stanford NLP measure to check the accuracy of the model. The binning classification will be checked using this measure and we

can see from the results that the system is performing really well. Table 15 is showing users evaluation using NDCG measure.

Name	Bin1	Bin2	Bin3	Bin4
<b>Ground Truth DCG</b>	15.23	124.88	119.45	31.7 3
<b>Normalized Base Formula Rank DCG</b>	10.05	88.31	58.39	8.27
<b>Normalized Proposed Formula Rank DCG</b>	14.50	111.80	104.43	29.7 5
<b>Max DCG</b>	15.23	124.88	119.45	31.7 3
<b>Normalized Base Formula Rank NDCG</b>	0.61	0.70	0.483	0.26
<b>Normalized Proposed Formula Rank NDCG</b>	0.95	0.89	0.87	0.93

*Table 16 Source Evaluation using NDCG*

The MAP is 80% while the NDCG is 91% which shows that NDCG has performed much better.

#### 4.6.7 Evaluation of Relevant Tweets Extraction

In this experiment, evaluation of raking tweets and extracting the relevant tweets have been performed. In this experiment the tweets have been classified to from highly relevant to highly irrelevant to certain topic. The tweets have been classified from bin1 to bin4. Bin1 is showing the tweets who are highly irrelevant to a certain topic while in bin4 contains those tweets who are highly relevant to certain topic. Evaluation is performed on multiple queries as mentioned above. Table 17 is showing different precision and recall scores of different bins against different queries.

Method	Bin 1	Bin 2	Bin 3	Bin 4
--------	-------	-------	-------	-------

	<b>Pre</b>	<b>Rec</b>	<b>F1</b>	<b>Pre</b>	<b>Rec</b>	<b>F1</b>	<b>Pre</b>	<b>Rec</b>	<b>F1</b>	<b>Pre</b>	<b>Rec</b>	<b>F1</b>
<b>Q1</b>	0.95	0.87	0.91	0.92	0.85	0.88	0.89	0.83	0.86	0.91	0.88	0.89
<b>Q2</b>	0.82	0.79	0.80	0.77	0.69	0.73	0.71	0.79	0.75	0.87	0.83	0.85
<b>Q3</b>	0.87	0.83	0.85	0.79	0.87	0.83	0.68	0.75	0.71	0.94	0.87	0.90
<b>Q4</b>	0.92	0.86	0.89	0.82	0.76	0.79	0.87	0.81	0.84	0.80	0.72	0.76
<b>Q5</b>	0.89	0.80	0.84	0.74	0.81	0.77	0.78	0.84	0.81	0.97	0.91	0.94
<b>MAP</b>	0.86			0.80			0.78			0.89		

Table 17 Tweets Evaluation

#### 4.6.8 Evaluation of Relevant Tweets Extraction using NDCG

It's really common to check the accuracy of the model using precision and recall measure but NDCG (normalized discounted cumulative gain) is the standard Stanford NLP measure to check the accuracy of the model. The binning classification will be checked using this measure and we can see from the results that the system is performing really well. Table 18 is showing users evaluation using NDCG measure.

<b>Name</b>	<b>Bin1</b>	<b>Bin2</b>	<b>Bin3</b>	<b>Bin4</b>
<b>Ground Truth DCG</b>	88.50	196.23	121.82	60.75
<b>Normalized Proposed Formula Rank DCG</b>	80.49	167.28	101.95	56.77
<b>Max DCG</b>	88.50	196.23	121.82	60.75
<b>Normalized Proposed Formula Rank NDCG</b>	0.90	0.85	0.83	0.93

Table 18 Tweets Evaluation using NDCG

The MAP is 83% while the NDCG is 87% which shows that NDCG has performed much better.

## **4.7 Comparative Evaluation**

There are many experiments performed and each experiment is showing really up to the mark results. The system has been evaluated against two base systems, one proposed in [13] and the other one is evaluated against the ground truth labeled by professionals from different fields. Professionals were provided with the sheet having all the features listed and their scores and they need to rank the relevant sources and relevant tweets based upon their thinking. As it involves human annotation so chances of error are high. Classifying the sources and place them in the bins has been evaluated against the base system mentioned in [13]. Same data set has been passed and provided the same environment. The end result was the ranks calculated by the system to classify the sources among most relevant to the least relevant. We can see from figure 16, where there is a comparison of old formula proposed in [13] and a proposed formula, proposed formula is classifying sources in to bins with much more accuracy than the one proposed in [13]. The average precision of the base system using our dataset is 0.34 while with the proposed technique it is 0.80 so, overall the system is performing really well. The NDCG calculated for the source ranking is 0.91 and this NDCG has been compared against the Ground truth generated by the professionals. The mean average precision of tweets ranking is 0.83 while the NDCG is 0.87.

# Chapter 5

## Conclusion

This chapter summarizes the whole research. Section 5.1 presents the contributions followed by conclusion in section 5.2. Limitations and future work are described in section 5.3.

### 5.1 Contributions

This research has following main contributions:

- A new model has been proposed which can provide relevant tweets and relevant sources and classifies them with higher accuracy.
- The proposed system is independent and semi-automated to generate the ground truth automatically using the relevance feedback mechanism.
- The proposed system removes the redundant tweets and increases the relevancy of the content during high impact events and provide the information which is valuable to them.
- The source of the tweets have been categorized in to different ranks from highly relevant to the least relevant in the proposed system. This approach not only depends on the source specific attributes such followers count or relevant retweet count, but also on the tweet specific attributes such as hashtag count, URL's count, mentions count and relevancy score of the tweet, which is the more reliable and appropriate way of classifying and ranking.

### 5.2 Discussion

In this research, a new system has been proposed that provides the relevant tweets and relevant sources. The system has used state of the art technique of removing the redundant tweets using Jaccard measure. This would provide much valuable information to the user in which they are interested. The proposed technique have incorporated the relevance feedback, that helps in covering the bigger set of tweets by expanding the query and adding more words to the query. System is independent and dynamic enough to generate the ground truth semi-automatically. This continuous relevance feedback mechanism would help in training the system in a way that it improves by itself based upon the feedback of the user. The relevancy score and the relevant tweets has been incorporated to rank the relevant tweets and relevant sources along with other source specific attributes. Appropriate feature set selection including relevancy score, source specific features like followers count, mentions count, relevant tweets count and URLs count have helped in achieving better mean average precision score and higher Normalized Discounted Cumulative Gain (NDCG). The system provides the relevant sources and relevant tweets with much higher accuracy. The system would not only provide you the relevant sources and tweets but it also provides the ranks, i.e., relevant, probably irrelevant or highly irrelevant, of the user.. This diversity of information would help the Government and humanitarian organizations to do appropriate measures and launch relied operations on time to deal with the emergency situations. The proposed system for classifying sources performs better as compared to the baseline system [13]. The proposed technique has been evaluated using multiple measures such as mean average precision, 10-folds cross validation and Normalized Discounted Cumulative Gain (NDCG). There are multiple classifiers that have been trained like Naïve Bayes, Random Forest and SVM (Support Vector Machine). All the classifiers have predicted the results really well but SVM has performed up to the mark. The base system [13] for finding the relevant sources have an average NDCG of

0.5 while the proposed system has an average NDCG of 0.91. This shows that the proposed system is highly effective in ranking the sources. The NDCG of finding the relevant tweets is 0.87.

### **5.3 Limitation and Future Work**

The evaluation shows that in terms of performance, accuracy, and usability, it is possible to increase the relevancy of the content and the sources on Twitter by decreasing the irrelevant content. At the same time, we can see that there are many challenges such as Twitter API call limit. API can fetch only 3200 most recent tweets of the particular source, we cannot fetch the older tweets of any source to ensure that their behavior is suspicious since the start or they have now started posting irrelevant content. Word2vec model is also vocabulary dependent, more the data; better the model would be trained. The relevancy of the content module is dependent upon the user feedback, the more serious and accurate feedback is, the more accurate classifier would be trained and more accurate would be the results. All the mentioned challenges cannot be addressed in the future, i.e., it is not possible for now to fetch the tweets older than 3200, you can only fetch most recent 3200 tweets of the source. This is the limitation of Twitter Stream API which is a third party API. The future work includes, testing and evaluating the proposed system using different social media datasets, i.e., Reddit and Facebook and Word2vec model can be trained on a larger dataset (containing more disaster tweets) to increase the vocabulary.



# Bibliography

- [1] M. AlRubaian, M. Al-Qurishi, M. Al-Rakhami, S. M. M. Rahman, and A. Alamri, “A *Multi-stage Credibility Analysis Model for Microblogs*” presented at the Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, Paris, France, 201
- [2] C. Castillo, M. Mendoza, and B. Poblete, “*Information credibility on twitter*” presented at the Proceedings of the 20th international conference on World wide web, Hyderabad, India, 2011
- [3] Pal, A. and Counts, S. What’s in a @name? How Name Value Biases Judgment of Microblog Authors. in Proc. ICWSM, AAAI (2011)
- [4] M. Mendoza, B. Poblete, and C. Castillo, “*Twitter Under Crisis: Can we trust what we RT?*” in Proceedings of the first workshop on social media analytics, 2010, pp. 71-79
- [5] Westerman, D., Spence, P.R., and Van Der Heide, B.: “*A social network as information: The effect of system generated reports of connectedness on credibility on Twitter*”, Computers in Human Behavior, 2012, 28, (1), pp. 199-206
- [6]DL4J Word2Vec. [Online]. Available: <https://deeplearning4j.org/word2vec.html>
- [7] Cornell publications. [Online]. Available: [https://www.cs.cornell.edu/people/tj/publications/joachims\\_97a.pdf](https://www.cs.cornell.edu/people/tj/publications/joachims_97a.pdf)
- [8]Twitter Streaming API. [Online]. Available: <https://www.slideshare.net/kaleemmalick/twitter-api-59270361>
- [9] Deepak P. and Sutanu Chakraborti “*Finding Relevant Tweets*” Springer-Verlag Berlin Heidelberg 2012.
- [10] Playing with Twitter Stream API. [Online]. Available: <https://medium.com/@ssola/playing-with-twitter-streaming-api-b1f8912e50b0>
- [11] Five most important similarity measures. [Online]. Available: <http://dataaspirant.com/2015/04/11/five-most-popular-similarity-measures-implementation-in-python/>
- [12] Relevancy lecture. [Online]. Available: <https://www.cs.ucy.ac.cy/courses/EPL660/lectures/lecture8-rel.pdf>

- [13] Majed Alrubaian, Muhammad Al-Qurishi, Mabrook Al-Rakhami, Mohammad Mehedi Hassan\*, † and Atif Alamri “*Reputation-based credibility analysis of Twitter social network users*” Published online 21 May 2016 in Wiley Online Library (wileyonlinelibrary.com).
- [14] ZohaSheikh, Hira Masood, Sharifullah Khan, Muhammad Imran “*User-Assisted Information Extraction from Twitter During Emergencies*” conference ISCRAM 2017 France.
- [15]Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection." *Ijcai*. Vol. 14. No. 2. 1995.
- [16]L. Bilge et al, “*All your contacts are belong to us: automated identify theft attacks on social networks*”, Proceedings of ACM World Wide Web Conference, 2009.
- [17]M. Mowbray, “*The Twittering Machine*”, Proceedings of the 6th International Conference on Web Information and Technologies, April 2010.
- [18]The power of one wrong tweet by Heather Kelly, CNN. [Online]. Available: <http://edition.cnn.com/2013/04/23/tech/social-media/tweet-ripple-effect/>
- [19] Visualized: Incorrect information travels farther, faster on Twitter than corrections by Craig Silverman. [Online]. Available: <http://www.poynter.org/2012/visualized-incorrect-information-travels-farther-faster-ontwitter-than-corrections/165654/>
- [20] Faking Sandy: Characterizing and Identifying Fake Images on Twitter during Hurricane Sandy. [Online]. Available: [http://precog.iitd.edu.in/Publications\\_files/9psosm4-gupta.pdf](http://precog.iitd.edu.in/Publications_files/9psosm4-gupta.pdf)
- [21]News Use Across Social Media Platforms 2016. [Online]. Available: <http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/>
- [22]Aditi Gupta, Ponnurangam Kumaraguru “*Credibility Ranking of Tweets during High Impact Events*”, Proceedings of the 1st Workshop on Privacy and Security in Online Social Media.
- [23] Chao Yang “*A Taste of Tweets: Reverse Engineering Twitter Spammers*”, Published 2014 in ACSAC.
- [24] Muhammad Imran<sup>1</sup>,Prasenjit Mitra<sup>1</sup>,Carlos Castillo<sup>2</sup> “*Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages*”, In Proceedings of the 10th Language Resources and Evaluation Conference (LREC), pp. 1638-1643. May 2016, Portorož, Slovenia.
- [25] Aditi Gupta “*A survey on Analyzing and Measuring Trustworthiness of User-Generated Content on Twitter during High-Impact Events*”.

- [26]Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier “*TweetCred: Real-Time Credibility Assessment of Content on Twitter*”, arXiv:1405.5490v2 [cs.CR] 30 Jan 2015.
- [27] He, B.; Macdonald, C.; He, J.; and Ounis, I. 2008. “*An effective statistical approach to blog post retrieval*”. In CIKM Proceedings of the 17th ACM conference on Information and knowledge management.
- [28]T. Joachims. “*Text categorization with support vector machines: Learning with many relevant features*”, Technical Report 23, Universität Dortmund, LS VIII, 1997.
- [29]Andrew M. Dai, Quoc V. Le “*Semi-supervised Sequence Learning*”, Advances in Neural Information Processing Systems 28 (NIPS 2015).
- [30]Sujoy Sikdar, Byungkyu Kang, John O'Donovan, Tobias Höllerer, Sibel Adah “*Understanding Information Credibility on Twitter*”, SOCIALCOM '13 Proceedings of the 2013 International Conference on Social Computing.
- [31]Axel J. Soto, Cynthia Ryan, Fernando Peña Silva, Tobias Höllerer, Sibel Adah “*Data Quality Challenges in Twitter Content Analysis for Informing Policy Making in Health Care*”, Proceedings of the 51st Hawaii International Conference on System Sciences | 2018.
- [32]Fernando Diaz, Bhaskar Mitra, Nick Craswell “*Query Expansion with Locally-Trained Word Embeddings*”, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).
- [33]GloVe vs word2vec revisited. [Online]. Available: <http://dsnotes.com/post/glove-enwiki/>
- [34]Surajit Dasgupta\*, Abhash Kumar\*, Dipankar Das “*Word Embeddings for Information Extraction from Tweets*”, Published 2016 in FIRE.
- [35] Manjeet Kumar, Abhishek Garg, Anuj Munjal, AkanshaTanwar “*Twitter Based Information Extraction*”, International Journal of New Technology and Research (IJNTR) March 2017.
- [36]Discounted Cumulative Gain. [Online]. Available: [https://en.wikipedia.org/wiki/Discounted\\_cumulative\\_gain](https://en.wikipedia.org/wiki/Discounted_cumulative_gain)
- [37] Accuracy, Precision, Recall or F1? [Online]. Available: <https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9>
- [38]TF-IDF [Online]. Available: <http://www.tfidf.com/>

- [39] The Relationship between Precision-Recall and ROC Curves. [Online]. Available: <https://www.biostat.wisc.edu/~page/rocpr.pdf>
- [40] Micro- and Macro-average of Precision, Recall and F-Score. [Online]. Available: <http://rushdishams.blogspot.com/2011/08/micro-and-macro-average-of-precision.html>
- [41] Types of classification algorithms in Machine Learning [Online]. Available: <https://medium.com/@sifium/machine-learning-types-of-classification-9497bd4f2e14>
- [42] Nihalahmad R. Shikalgar, Arati M. Dixit “JIBCA: Jaccard Index based Clustering Algorithm for Mining Online Review”, International Journal of Computer Applications (0975 – 8887) Volume 105 – No. 15, November 2014.
- [43] The Ecological Status of European Rivers: [Online]. Available: <https://books.google.com.pk/books?id=JbLlv82wZeoC&pg=PA466&lpg=PA466&dq=jaccard+similarity+70%>
- [44] How to Clean Up and Fine Tune Your Twitter Feed: [Online]. Available: <https://lifehacker.com/how-to-clean-up-and-fine-tune-your-twitter-feed-1514738479>
- [45] Tweets Language: [Online]. Available: <https://gigaom.com/2013/07/28/twitter-is-available-in-33-languages-but-its-universal-language-is-the-emoji/>
- [46] Stemming and Lemmatization: [Online]. Available: <https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html>