Rule Based Sentiment Analysis of Punjabi Tweets Using Vector Evaluation Method

Dissertation submitted in partial fulfilment of the requirements for the award of degree of

Master of Engineering

in

Information Security

Submitted By:

Mehak

(Roll No. 801533012)

Under the supervision of:

Mr. Varinderpal Singh — System Analyst

Dr. Rajiv Kumar — Assistant Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA - 147004

JULY 2017

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "Rule based Sentiment Analysis of Punjabi Tweets using Vector Evaluation Method", in partial fulfilment of the requirements for the award of degree of Master of Engineering in Information Security submitted in Computer Science and Engineering of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of Dr. Rajiv Kumar & Mr. Varinderpal Singh and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any degree of this or any other University.

Mehak

Mehak

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

(Dr. Rajiv Kumar)
Assistant Professor

(Mr. Varinderpal Singh)
System Analyst

ACKNOWLEDGEMENT

I am highly grateful to the authorities of Thapar University, Patiala for providing the opportunity to carry out this thesis work. I would like to express my deeps thanks and gratitude to my guide Mr. Varinderpal Singh and Mr. Rajiv Kumar, Assistant Professor, Computer Science and Engineering Department, Thapar University, Patiala for their sincere and invaluable guidance. Their encouragement and valuable advice during the entire period has made it possible for me to complete my work. This thesis work was enabled and sustained by their vision and ideas.

I am thankful to Dr. Maninder Singh, Head of Computer Science Engineering Department, Thapar University for setting high standards for his students and encouraging them time to time so that they can achieve them as well. I would also like to thank entire faculty and staff of Computer Science and Engineering Department and my friends who devoted their valuable time in completion of this work. I would also like to express my gratitude to all the library staff for their services.

Above all, I owe it all to Almighty God for granting me the wisdom, health and strength to undertake this research work and enabling me to its completion. I would also like to thank my parents for their years of unyielding love and encouragement. They have wanted the best for me and I admire their sacrifice and determination.

Mehak

(801533012)

ABSTRACT

Due to the exponential enhancement in the Internet usage and replacement of public opinions, Sentiment Analysis becomes an important process in today's life. Sentiment Analysis is a process of extracting information from opinions generated by the users. Twitter is a micro-blogging platform which provides a tremendous amount of data which can be used for various applications of Sentiment Analysis like predictions, reviews, elections, marketing, etc. The thoughts or opinions of other people provide information that helps in decision making process. But, Sentiment Analysis is a challenging task because it is very difficult to find the exact sentiment from text as there are so many challenges like entity identification, subjectivity detection in performing sentiment analysis.

The project work carried out in the dissertation is focused on Rule based sentiment analysis. Python is simple yet powerful, high-level, interpreted and dynamic programming language, which is well known for its functionality of processing natural language data by using NLTK (Natural Language Toolkit). The goal of this dissertation is to classify twitter data into sentiments (News, Personal, Opinions and Entertainment) by calculating weight of each word and the words with highest weights are selected as features which will be then compared to the rest of the words in order to find the most suitable category for Punjabi Tweets.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

User-generated content is a vital source of information to determine the sentiment or opinion of people on services. Better connectivity, high computational and user-friendly technology has led to wide spread use of online journals, forums, electronic news and the social networking platforms like Facebook and Twitter. Currently, opinions have been made on all fields at different blog forums and other social networks. Thus data mining and identifying user emotions, wishes, likes and dislikes is a vital task that has attracted research community from last decade. Sentiment Analysis (SA) or Opinion Mining (OM) is estimation study of user's opinions, behaviors and sentiments concerning an entity [1]. SA is a classification task which classifies input data into different classes or categories. There are two types of classification:

- Binary classification: classifies input objects into one of the two classes, like whether the movie review is positive or negative.
- Multi-class classification: classifies input objects into one of the multiple classes. For example, classifying news into different classes such as sports news, agriculture news, political news, business news.

As compared to binary classification, which requires piercing between the two classes, the multi-class classification is a more challenging task [2].

SA is an important Natural Language Processing (NLP) task to determine the sentiment of the text as is the best way to understand the text used. Research on SA has been investigated from different perspectives. The most popular aspect is to categorize these studies into three different levels which are document level [41], sentence level [45], and entity and aspect level, also known as phrase level [46].

Different techniques applied for SA are broadly classified as supervised learning, unsupervised learning and Rule-Based learning.

Supervised learning is defined as a task of deriving a function from labelled training data, which can be used for mapping new data [62]. It is fast and accurate as compared to unsupervised learning which derives a function using unlabelled data to find the hidden structure. But building of proper training and testing set is a crucial task for supervised learning. Unsupervised learning can be used to group together the input data into classes depending on their statistical properties. SA performed by using unsupervised approaches is done by comparison.

In this dissertation, supervised rule-based learning for SA of Punjabi tweets is implemented. In this technique, the data space is formed with a bunch of rules here [51]. The left side depicts a condition applied on the feature set while the right side shows the class category. The conditions are applicable on the term presence. Term absence is hardly used as it is least descriptive. There're a variety of criteria available for generating rules. The training phase builds all the rules depending on these criteria. The most common criteria are support and confidence [48]. The *Support* is defined as the number of items in the training data set which are relevant to the rule. The *Confidence* points to the conditional probability that the right side of the rule is fulfilled if the left side is fulfilled.

Confidence criteria of Rule-based learning for SA is used in this dissertation. Vector Evaluation Method for SA has been used to carry out this work. In this method, the weight of each word of each data is calculated. Weight is calculated by multiplying the Term Frequency (TF) and Inverse Term Frequency (IDF) of each word. The words having highest weights are selected as features. Combination of rule-based learning and Vector Evaluation method has been used for SA.

Since the available text is in the unstructured form, there is a need to convert this unstructured data into structured data. It can be done by finding word frequency of the labelled document, weighing each word using TD-IDF scheme or by using Bag-Of-Word model.

- **TF-IDF model:** This technique is used to weigh every feature of the text document [3]. *'TF'* refers to the Term Frequency of a word, i.e. the total sum of the number of occurrences of a particular word in a document. But TF by itself has some short comings. For example, if the documents are about "Google search algorithm", the term *'Google'* is very likely to occur multiple times. The emphasis of the document is not about the company Google but the search algorithm they employ. Hence, to reduce the effect of the word Google, IDF (Inverse Document Frequency) comes in action. Document frequency (DF) refers to a number of documents in the collection that contain a specific word. Higher the value of DF, lower the importance of the feature. Equation 1.1 depicts IDF for a feature is calculated as follows::

$$IDF = log\ (N/DF) \tag{1.1}$$

  Here *'N'* refers to the total sum of documents in the corpus. TF-IDF score for a feature is calculated as shown in the following equation 1.2:

$$TD - IDF = TF * IDF \tag{1.2}$$

- **Bag-Of-Word model:** It involves breaking down the text into words. Each broken word represents a feature. This process is also referred to as *'Tokenization'*. A group of features extracted forms a feature vector for the document. There are several ways to prune this vector because this vector becomes too large. Techniques like stop word removal and stemming are applied most commonly. Stop word removal includes removing words which have insignificant value to the document. Stemming is the process for reducing derived words to their stem, base or root form – generally a written word form [5]. But BOW model does not return accurate results as it may contain some features which are least. Also it has dimensionality issue as the total dimension is the size of vocabulary and it doesn't consider the semantic relation between words.

- **Word Frequency:** The frequency or number of occurrences of each word in the collected corpus is calculated. The words with highest frequencies are selected as *features*.

1.1 Need for Sentiment Analysis

1.1.1 **Business intelligence [6]:** SA helps in business as it automatically evaluates online products and review services. The information in customer reviews is of great interest to both the companies and consumers. Companies spend a huge amount of money to find customers' opinions and sentiments, as this information is useful to exploit their marketing-mix in order to affect consumer satisfaction. Moreover, online opinions clearly influence the reputation of the company. Thus, SA is becoming an important component of Customer Relationship Management (CRM) solutions.

1.1.2 **To perform different NLP tasks [7]:** SA would help in following NLP tasks:

1.1.2.1 **Question Answering:** Interests in sharing questions/answers through the Community Question Answering (CQA) systems, such as Stock Exchange, Yahoo! Answers have been increased, so predicting the best answer in such systems can be a challenging task. One of the vital challenges in selecting the best answer is to extract informative features which represent the problem definition good enough.

1.1.2.2 **Summarization [9]:** Summarization is a task of reducing the text content of a document without the loss of its meaning. SA is already being used in various domains for analysis of large scale text data interpretation and opinion mining. SA is often applied to text reviews, news, twitter data or other voluminous text data. Its primary task is to assign a positive, negative and sometimes neutral sentiment to each word of a sentence. In sentiment based summarization there is no requirement of deep semantic analysis [8].

1.1.2.3 **Information Extraction [10]:** Automatic document sense/information extraction is considered to be among crucial problems of NLP. It is proved to be competitive with other previously developed summarization methodologies. Information Extraction is supposed to recognize the sentiment expressions of the specific subject such as a

person, a product or a company and then to calculate the sentiment and the validation of them.

1.2 Sentiment Analysis Applications

Researches indicate that people believe on the opinions of other people that exist outside their social network account, for example, online reviews. Here SA comes into play. Some of the applications of sentiment analysis are:

1.2.1 **Disaster relief:** Many natural disasters all over the world have killed people and affected lot of human lives worldwide. Hence there is a critical need to use a mechanism to best allocate investment for prevention, preparedness, response, and recovery to enhance safety and reduce the cost and social effects of emergencies and disasters. Social media helps to keep informed, locate loved ones, express support or notify authorities during emergencies and disasters. Due to huge popularity and diverse areas of discussion of social media, it has been used recently as a tool for first responders those who provide first-hand aids for disaster relief and crisis management. [59]

1.2.2 **Public opinion polls:** Data on public opinion can be used to determine public awareness, to predict outcomes of various events and to infer characteristics of human behaviors. Indeed, readily available public opinion data is valuable to researchers, policymakers, marketers, and many other groups, but is difficult to generate. Twitter has proved to be a boon for many research enterprises. For example, polling results of various political parties such as BJP, Congress and AAP can be depicted [54].

1.2.3 **Emotions in novels and fairy tales:** Book texts, such as novels, fairy tales, fables and epics have long been channels to convey emotions, both explicitly and implicitly. SA can be used in tandem with effective visualizations in order to quantify and keep a track emotions in both individual books and across very large collections. Large word emotion association lexicon can be used to create a simple emotion analyzer [56].

1.2.4 **Financial Markets:** There are numerous news items, articles, blogs, and tweets about each public company. An SA system can use these various

sources to discover articles that discuss the companies and aggregate the sentiment about them as a single score. It can then be used by an automated trading system. One such system is The Stock Sonar *http://www.thestocksonar.com*. This system (developed by Digital Trowel) graphically shows the daily positive and negative sentiment about each stock alongside the graph of the stock price [12].

1.3 Challenges in Sentiment Analysis

Some of the challenges while performing the SA are as follows [13]:

1.3.1 **Disorganized Data:** The data on the Internet is very unstructured, there are different forms of the data talking about the same entities, persons, places, things and events. The web contains data from different sources for example from books, journals, web documents, health records, companies logs, internal files of an organization and even data from multimedia platforms comprising of texts, images, audios, videos etc. The diverse sources of the data makes the analysis more complex as the information is coming in different formats.

1.3.2 **Noise (slangs, abbreviations):** The content on the web is very noisy. In today's era of 140 characters texting, for their ease people use various abbreviations, slangs, emoticons in normal text which makes the analysis more complex and difficult.

For example, *'mvie ws awsummm :D'*

The web content reports a large number of spelling variations for the same word. Example, a word *'awesome'* can be found in various forms such as *'awsum, awssuummm, awsome'*. The repetition of the characters can be in any combination.

1.3.3 **Sarcasm Detection:** *'Sarcasm'* is defined as a bitter, hurtful or cutting expression or remark; a bitter jibe or taunt is usually conveyed through irony or understatement [14]. It's a hard for human beings to interpret sarcasm, and more difficult task is to make a machine able to understand same. If there is not a presence of any sentiment bearing words, sentence may have an implicit sentiment [15]. For example:

'One should question the strength of mind of the writer who had written this book.'

The above sentence do not have any word that has negative sentiment but it's a negative sentence.

1.3.4 **Contextual Information:** Identifying the type of the text is an important challenge to emphasis on. Consider the following examples:

1. *'The journey was long.'*
2. *'Presentation was long.'*
3. '*Battery life of Apple i-phone is long.'*

In all the examples, meaning of long is same which is indicating the duration or section of time. In (1.) and (2.) *'long'* express the broadness, hence it is a Negative sentiment; on the other hand in (3.), '*long'* depicts the efficiency thus it's a Positive sentiment.

It's clear that a word with same meaning can possess multiple usages by depending upon the context. So, it is important to detect the context in order to find the subjective information in a text.

1.3.5 **Word Sense Disambiguation(WSD):** A word can have multiple meanings and the polarity of a word also changes depending on the sense of its usage. For example, word *'good'* has 21 adjectives, 4 nouns and 2 adverbs whose polarity changes with respect to the sense it is used. It is mentioned in following figure 1.1 :

Sense	Part of Speech	Polarity Scores	Meaning
good-1	Adjective	Pos-1 Obj-0 Neg-0	morally admirable
good-2	Adjective	Pos-0.375 Obj-0.5 Neg-.125	not left to spoil; "the meat is still good"
good-3	Noun	Pos-0 Obj-1 Neg-0	commodity/articles of commerce

Figure 1.1: Multiple senses of word "good"

1.4 Sentiment Analysis(SA) of Short Text

SA involves in analysing the emotions, feelings and the behaviour of a writer from a given text. Sentiments are considered as the expression of our emotions and thoughts which can be subjective and objective in nature. Subjectivity means that the text contains opinionated content while Objectivity depicts that the text is without any opinions. Some examples of subjective and objective level are:

- **Subjective:** *'This movie by Aamir Khan and Kajol is superb.'*
  This sentence has an opinion, it talks about the movie and the user's feelings about the same, "superb" and hence it's subjective.
- **Objective:** *'This movie stars Aamir Khan and Kajol.'*
  This factual sentence or a normal information instead of an opinion or a view of an individual and therefore its objective.

It's mandatory to distinguish between the subjective opinions and objective facts.

SA helps to determine the entities and subject from text towards which sentiment is pointed. Everyone share their information and thoughts on social networking sites, blogs and web-forums. Sometimes it becomes hard to perform SA because of the primary reason of data sparseness. Some of the short text messages or tweets might not contain any sentiment or emotion which later becomes quite difficult for selection of features during feature selection process. In this work, short text used are from Twitter known as *'tweets'*. The size of tweets is maximum 140 characters.

1.5 Overview of Twitter

"*Everyone will be tuned into everything that's happening all the time! No one will be left out"*. These were the lines published by Robert Dennis Crumb, an American artist and an illustrator in the 1960's [16]. Although this was published in a cartoon, this is no longer a vision of the future. Thanks to Jack Dorsey, Twitter originated in 2006 making these cartoon lines a reality.

1.5.1 About Twitter[21]

Twitter [17] is a social networking site which allows people to micro-blog about a variety of topics. Microblogging is defined as *'a form of blogging*

that lets you write brief text updates (usually less than 200 characters) about your life on the go and send them to friends and interested observers via text messaging, instant messaging (IM), email or the web'. [18]. Twitter helps users to connect with other Twitter users around the globe. The messages exchanged via Twitter are referred to microblogs because there is a limit of 140 character foist by Twitter on every tweet. This allows the users to present any information with only a few words that may or may not be followed with a link which gives more detailed information. Therefore, messages of Twitter, called as *'tweets'* are usually concentrated upon. Twitter is identical to Short Message Service (SMS) messages which are exchanged through mobile phones and other hand held devices. The limit of 140 character has also spurred the usage of shortening services of URL such as bit.ly and content hosting services to accommodate content involving multimedia and text greater than 140 characters [19]. Other social platforms like Facebook [20] and Orkut [24] introduced the idea of *'Status'* before Twitter came into existence. But Twitter took a step forward and made these *'statuses'* sharable among people since its creation.

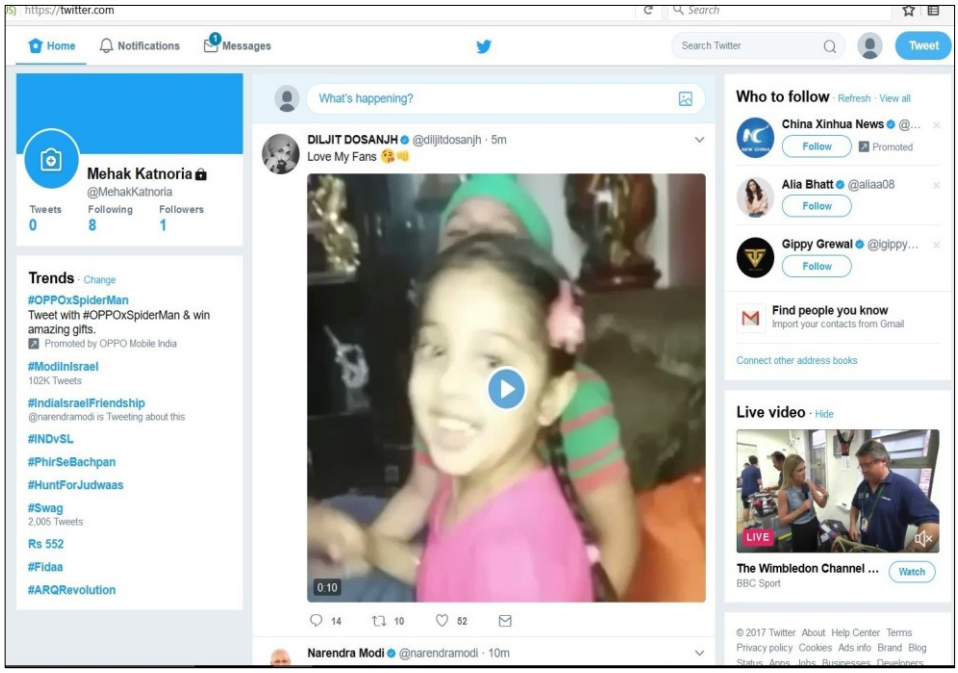


Figure 1.2: Home Page of a Twitter user

Although, Twitter is termed as a social networking website, it has the flavour of a personal log instead of a platform to communicate with other

people. With the rise in popularity of Twitter, it changed its focus to uncover what is happening around a user or where are they, anywhere in the world. The need to use Twitter has changed from recording one's own thoughts to a wider domain. Like users now have become news reporters and can re-tweet someone's else tweets.

1.5.2 Concepts in Twitter

1.5.2.1 **Users:** A Twitter user is a person or a system who shares/writes, messages called tweets. These tweets by default are public to any user until and unless the author specifically sets it to be private. All users have a unique user name and user id. User A can start writing tweets but other users will not be able to read them because the user does not have any followers yet. When A identifies another user (say 'B') to follow, his tweets are visible to B. Consequently, A becomes a follower of B. Thus, B becomes a friend of A. However, friendship need not be two-way. It is possible that A is not a friend of B if B does not follow the tweets of A. So an asymmetrical relationship exists among different users of Twitter. Twitter has also imposed a limit of 2000 friends for every user but there are no limits on the number of followers.

1.5.2.2 **Tweet:** A tweet is a Twitter message. It is short message since it is restricted to be within 140 characters by Twitter. This enforces the users to be concise in what they have to say. This is the reason why users use word shortenings (E.g.: "fr"-for, "cud"–could) and abbreviations. Interestingly enough, there is a rich and well understood set of abbreviations which is surprisingly consistent across user groups, and even across other electronic mediums such as SMS and chat rooms [22]. Twitter is also a challenging medium since users want to convey all they have to say within 140 characters, they could also make spelling mistakes and tweets can be prone to syntactic errors. Most of the times, users usually provide links to external resources when they cannot convey the complete information within 140 characters. These URL links to text, audio or video files are referred to as *'Artifacts'*.

1.5.3 Special Characters of Tweets

1.5.3.1 **Referring to another user:** In order to refer to another user within a tweet, ``@" symbol is used followed by the intended user name. When a user refers to another user at the beginning of a tweet, the tweet becomes a ``Direct Message (DM)". DM are those tweets which are public yet designed as a correspondence among two users of the system. Twitter provides the provision to view only direct messages intended to the user. This ensures that these messages which usually have a higher priority to the intended user do not get lost by the overwhelming stream of other tweets in the user space. For example:

Bob: @Alice, How was the Biology test? (Direct Message)

Trudy: I really had loads of fun at the party. @Bob made it extra special with his cookies.

1.5.3.2 **Re-tweets (RT):** If a tweet attracts other users or is very interesting, users may republish that tweets, commonly known as re-tweeting. A RT is similar to forwarding an e-mail. When a user re-tweets some content, the user is effectively endorsing and sharing the content with their followers [23].

Twitter earlier lacked a specific structure for re-tweets by merely providing a convention on how to re-tweet. Several forms of re-tweets were used, some of the most common being *'retweet @username'*, *'RT @username'* or via *'@username'*, before or after the re-tweeted message. But the current version has an option called *'Retweet'* right next to each tweet.

1.5.3.3 **Hash Tags:** Users can tag their tweets with the help of *'Hash tags'*. Hash tags are of the form *'#<tag-name>'*.

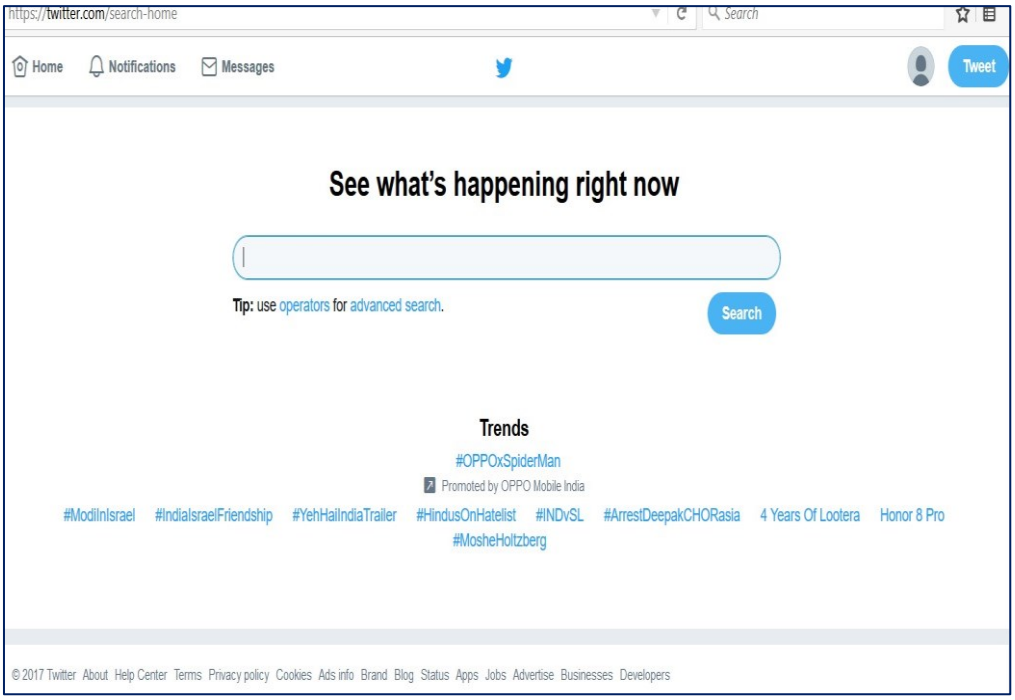


Figure 1.3: Twitter search engine

Users can convey about their tweets with the help of keywords that represent the insights of the tweet. Tweets which are hash tagged, helps Twitter to group similar tweets together with same hash tags. This helps searching easier and faster on Twitter. Most of the Twitter search tools [27] use hash tags to enhance search quality as shown in above figure 1.3.

Figure 1.4: Hash tags analytics for teachers

It is noted that the hash tag itself adds to the character count of the tweet. Shown in figure 1.4 are hash tags associated with the teachers which is related to education [25].

In this work, the data has been collected from twitter and the feature sets are created using TF-IDF scheme, calculating word frequency and BOW model. Initially, the Twitter messages collected are divided into pre-defined classes, namely News, Opinion, Personal and Entertainment. Supervised rule-based SA is performed on the Twitter data. Experiment results show that the proposed algorithm outperforms the traditional BOW technique.

Chapter 2

Literature Survey

2.1 Approaches for Sentiment Analysis

Machine Learning approaches for SA are discussed below:

2.1.1 Supervised Techniques

It is implemented by building a classifier. This classifier is trained by data which is manually labeled based on frequent terms in the documents or can be obtained from user-generated/user-labelled online sources. The supervised learning methods depend on the training documents which are labelled. Supervised techniques perform better than unsupervised techniques [57]. Commonly used supervised techniques for SA are Naive Bayes, SVM and Decision Tree which are discussed as follows:

- **Naive Bayes (NB) Classifier:**

  It is based on Bayesian theorem and convenient when the range of the inputs is high. In spite of its simplicity, NB Classifier performs better than other classification methods [58]. As shown in the figure 2.1, objects can be classified as either RED or GREEN. Main task is the identification of class for the new/testing objects. This decision can be taken on the basis of existing objects. Since, there are double the numbers of GREEN objects than RED as shown figure 2.1. So, it can be thought that new objects will more likely belong to GREEN class. This faith is known as the prior probability in the Bayesian analysis.

Pang, L. Lee, S. Vaithyanathan et al [29] were the first who worked on SA. Their main aim was to classify text by overall sentiment and not just by topic. They used Naive Bayes, SVM and Maximum entropy algorithms.

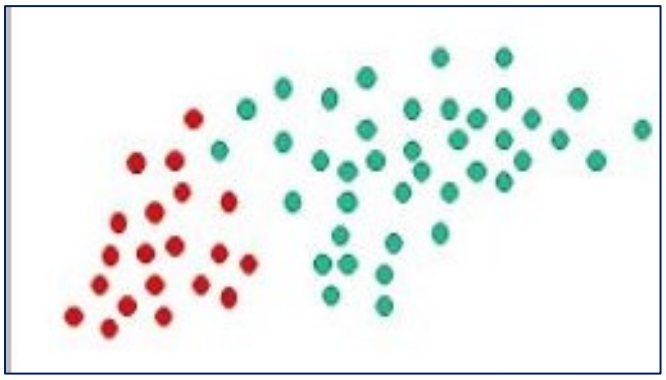


Figure 2.1: Demonstration of Naive Bayes Classifier

O. Almatrafi, S. Parack, B. Chavan et al [30] also work on SA using NB classifier. They worked on Indian general elections of 2014. They performed mining on 6,00,000 tweets which were collected over a period of 7 days for two political parties. They used NB algorithm as a classifier which can classify the tweets in two categories, either positive or negative. They diagnosed the thoughts and opinions of users towards these two political parties in different locations and they plot their finding on India map by using a Python library. An example of their results on tweets of BJP in 2014 is shown in Figure 2.2, which shows BJP got positive reviews in different locations of India.

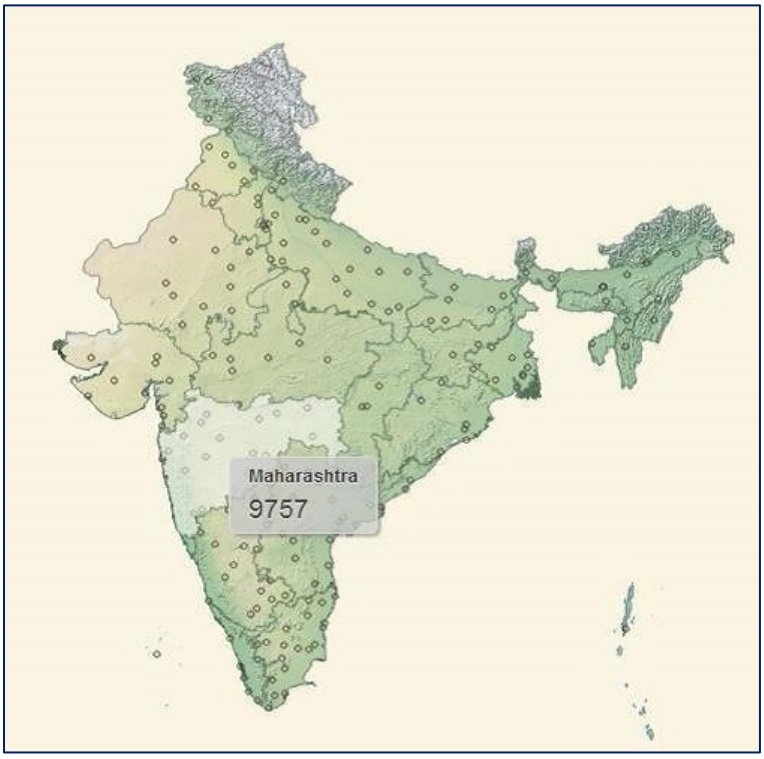


Figure 2.2: Positive tweets of BJP for different states in 2014

- **Support Vector Machines:**

  It works on the idea of decision planes that specify decision boundaries. A set of objects belonging to different class memberships are separated by decision planes [58]. An example to describe the concept of linear SVMs is shown in Figure 2.3. The objects either belong to GREEN class or RED class. The dividing line specifies the decision boundary. On the right hand side of the boundary, all objects are GREEN all objects are RED on the other side. A new object (white circle) will be classified as GREEN if it falls to the right side of the boundary or classified as RED if it falls to the left side of the boundary. An example of hyperplane classifier is shown in Figure 2.4.

P. Pang, L. Lee, S. Vaithyanathan et al [26] used SVM for the SA. They also concluded that classification of sentiment is very challenging which depends over various factors. They depicted that supervised machine learning algorithms are the base for sentiment analysis.

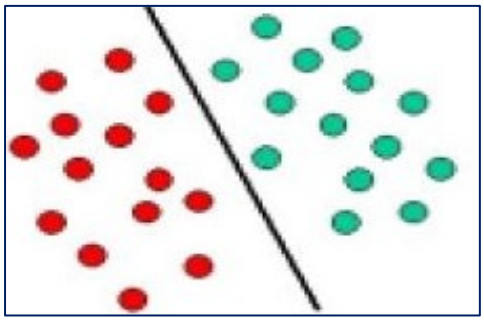


Figure 2.3: Example of linear SVM

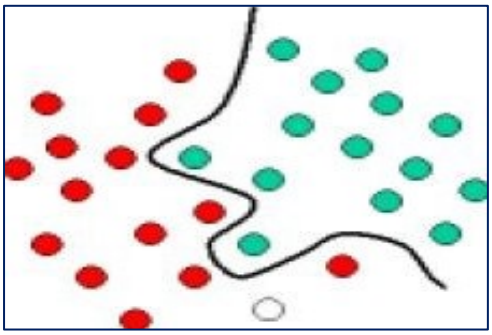


Figure 2.4: Example of Hyperplane Classifier

- **Decision Trees:**

  It provides a hierarchical decomposition of the training data in which a condition on the attribute value is used to divide the data [47]. The goal is to create a model that predicts the value of a target variable based on several input variables. Each node corresponds to one of the input variables. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

C.Chen, F.I.SanJuan, E.SanJuan, C.Weaver [53] used decision trees, SVM and NBC for sentiment based classification. An accuracy of 84.59% was achieved for reviews of book, The Da Vinci Code, collected from Amazon.com.

2.1.2 Unsupervised Techniques

Here sentiment classification is done by comparison. The features of a given text are compared against word lexicons whose sentiment values are decided prior to their use. Hierarchical clustering and partial clustering are mostly utilized algorithms of unsupervised techniques.

- **Hierarchical clustering:** It partitions the objects into tree like structure where each node depicts a cluster. There are 0 or more than 0 child nodes in cluster of the tree. Hence, tree grows in downward direction by its nature [47].
- **Partial clustering:** Here objects are partitioned. Objects can change the clusters on the basis of dissimilarity. K-means clustering algorithm is mostly used algorithm for this purpose [47].

P.D.Turney et al [55] used unsupervised method for sentiment detection. He used "poor" and "excellent" seed words for computing the semantic orientation of phrases. An accuracy of 66% was achieved for movie review domain and 84% for automobile reviews. Li et al. [55] developed an approach by using k-means clustering algorithm to cluster documents into positive group and negative group and achieved an accuracy of 70%. While the machine learning based approaches provided better classification accuracy, but required a lot of training time and pre-

classified training corpus. Moreover semantic orientation based approaches did not gave good performance, but returned results quickly [61].

2.2 SA on the basis of data

Three levels of data on which SA is performed are document level, sentence level and phrase level as:

2.2.1 Document level

It is the most common filtering technique which uses the actual text of message to determine the polarity of the document, for example whether a message is spam or not. The content is very dynamic and it is very challenging to represent all information in a mathematical model of classification. For example, in content-based Spam filtering, the characteristics used by the filter to identify Spam message are constantly changing over time. T.B.Shahi, A.Yadav et al [41] worked on this concept where they classified the Nepali mobile SMSs into two categories, Spam and Ham.

2.2.2 Sentence Level

A.Pak, P.Paroubek et al [42] focused on using Twitter, the most popular microblogging platform for SA. The authors performed linguistic inspection of collected corpus and build a sentiment classifier to determine positive, negative and neutral sentiments of twitter document. They also proposed a system for emoticons available in tweets. They create a corpus for emoticons so that emoticons can be replaced with their respective meaning so that features can be extracted from these emoticons. An example of emoticons and their corresponding meaning is shown is Table 2.1 as follows:

EMOTIONS	MEANING
:D, =), =D, ;-)	Happy
;-(, ;(, =(, =[,)-:	Sad
:P, =P	Joking

Table 2.1: Emoticons and their meaning

2.2.3 Entity/Phrase level:

T.Wilson, J.Wiebe, P.Houffman et al [46] presented a new approach to phrase-level SA that first determines whether an expression is neutral or polar and then disambiguates the polarity of the polar expressions. With this approach, the system is able to automatically identify the contextual polarity for a large subset of sentiment expressions, achieving results that are significantly better than baseline.

2.3 Indian Language Sentiment Analysis

Lot of work has been done on English language. As text is not confined to one language, SA has been performed on various Indian Languages too. India is a land of vastness and continuity with 22 languages. The Constitution of India designates official language of India as Hindi written in Devanagari script, as well as English. Other languages are Kashmiri, Nepali, Punjabi, Gujarati, Marathi, Tamil, Telegu, Bengali, Konkani, Malayalam, Kannada, Odiya, Bhili, Mizo, Manipuri, Nissi, Khasi and Assamese. SA applications are language-specific systems, mostly designed for English but very little amount of work has been done on Indian Languages. SA for Indian Languages is s challenging task as there is a scarcity of resources like automatic tools for tokenization, feature selection and stemming.

Following Figure 2.2 represents SA performed on various Indian Languages. It tells about techniques, dataset, features extracted and tool used on multiple Indian Languages.

AUTHOR	LANGUAGE	TECHNIQUES	DATASET	FEATURES EXTRACTED	TOOL USED (*if used*)
U.Jain, A.Sandhu[33]	Punjabi	*Supervised :* SVM Maximum Entropy Algorithm	Punjabi websites, news paper and various Punjabi blogs	*Linguistic* : Joy, Sadness, Fear, Surprise, Disgust and Anger	MATLAB
M.Patil, P.Game[34]	Marathi	*Supervised:* NB, Centroid, K nearest Classifier and Modified K nearest Classifier.	4000 Marathi text documents prepared using various sources	*Semantic :* economy,geography,history ,literature, botany	
J.J.Patil, N.Bogiri[35]	Marathi	*Supervised :* SVM	Manual dataset of 200 documents	*Semantic :* Sports, Politics, Crime, Economics, Education, Entertainment, Social, Health	LINGO Clusttering Algorithm :
J.Kaur, S.Bhagla[36]	Punjabi	*Supervised:* Naïve Bayes Classifier	Manually collected Punjabi online news	*Semantic:* politics, business, entertainment and health.	Visual basic 2010 using C# language
A.Sharma[37]	Punjabi	*Supervised:* NB	data from different Punjabi newspapers, blogs	*Linguistic:* positive/negative	
N.Desai, A.D.Dave[38]	Hindi	*Supervised :* SVM	Polarity labelled corpora of Hindi sentences generated by Aditya Joshi et al[42].	Non-Sarcastic, Mild Positive Sarcastic, Extreme Positive Sarcastic, Mild Negative Sarcastic and Extreme Negative Sarcastic	
A.K.Mandal, R.Sen[39]	Bangla	*Supervised :* NB KNN SVM DT(C4.5)	Various Bangla websitesEprothom-alo.com,online_dhaka.com,bdnews24.com,dailykalerkantha.com,bbc.co.uk/Bengali, ittefaque.com	*Syntactic*: 5 categories business,Sports,health,techn-ology,education	
S.Amarappa, S.V.Sathyanar-ayana[40]	Kannada	*Supervised :* MNB	Part of EMILLE Corpus ,some part from web article,some self-typed corpus from kannada books	*Semantic:* names of people,companies, government ,Location names, date,time	Sklearn toolkit, python 2.7
T.B.Shahi, A.Yadav[41]	Nepalese	*Supervised:* NB SVM	Nepali SMS dataset	*Linguistic:* Legitimate/SPAM	SVM Light tool for SVM classification. Naive Bayes implemented in java

Figure 2.5: Sentiment Analysis on Indian Languages

Chapter 3

Problem Statement

SA is a process of extracting information from user's opinions which they post on any social networking sites. The decisions of the people get affected by the opinions of other people. If anyone wants to buy something or watch a movie, the person will first look out for the reviews and opinions about the movie on let's say, social media. So there is a need of system that can automatically generate SA from this huge amount of data. Thus, to make this abundant data useful SA is performed, i.e. extracting feature from this data and classify them. Text is not confined to one language. Large amount of data is available in other languages too. Punjabi text is available on number of online forums, blogs, Twitter and other social networking sites. The data collected for dissertation is in Punjabi language. Punjabi is 10th widely spoken language in world [4]. The motive behind working on Punjabi language is that it is difficult for some people to read in English, thus Punjabi is used for SA for better understanding.

3.1 Objectives

Following are the defined objectives for SA on Punjabi data:

1. The main objective of this thesis work is to perform the SA on small texts, called tweets in Punjabi language which are extracted from Twitter.
2. Study the existing approaches and techniques for SA.
3. Extract the twitter posts in Punjabi language by user defined parameters using twitter APIs.
4. Perform pre-processing on the data collected.

5. Calculate each word's weight of each in the Punjabi tweet.
6. Classify tweets into 4 classes namely *News, Entertainment, Personal* and *Opinion*.

3.2 Methodology

Following methodology has been used to achieve all the objectives discussed in above section:

1. Literature survey has been carried out by study of existing approaches like knowledge based, concept based approaches etc. and techniques like supervised, unsupervised and rule-based to perform the SA on data collected from Twitter, blogs and other social networking sites.
2. Study of different methods and techniques applied on Indian languages in order to perform SA.
3. Twitter API is used to get the Punjabi tweets which is saved as google document.
4. Pre-processing of data collected from Twitter is performed using Python language. In this task stop words, unwanted spaces, special characters and punctuation marks are removed.
5. Calculating the each word's weight in the collected document with the help of weighting scheme function which is the product of the term occurrence frequency (tf) and the inverse document frequency (idf).
6. Choosing those keywords that have the largest weight in training dataset.
7. Finding the most suitable category. It is done by comparing the selected keywords as features with key words of each category of tweet.
8. Returning the name of category that matches the most.

Chapter 4

Implementation

In this chapter, the complete SA has been implemented for Twitter data of Punjabi language. For data collection, Twitter API has been used. The data collected from Twitter is divided into training and test datasets. The Twitter data was collected in the month of May. Whole implementation has been done in Python 2.7.12. The process of data collection is described in following section.

4.1 Collection of Data

4.1.1 Twitter API (Application Programming Interface)

Users can collect tweets from Twitter with the help of Twitter API. Two kinds of APIs are provided by Twitter: REST API and Streaming API. The differences between both of them is that with REST APIs, a limited data can be collected at a time as it support connections for short time period. On the other hand, Streaming API provides long time and real time connection. Streaming API has been used for the analysis of Punjabi Tweets. For collecting large amount of tweets, long-lived connections are needed with no limit data rate.

When the Twitter API for collection of data, it is linked with the Gmail account. One can select the language in which they want the tweets to be saved. Similarly the user can input the hashtags if they want tweets regarding their particular area of interest or any topic. The tweets are saved as google docx in an excel sheet. One can later download that sheet our systems for easy use. The saved excel sheet gives every possible information regarding every single tweet, like screen name and full name of twitter account holder, Tweet text, tweet ID, followers of the particular user, Retweets, number of people the user follows, Location, Bio and Profile image.

As the objective of this dissertation is to analyze the sentiment of Tweets for Punjabi language, only tweets in Punjabi Language will be collected. It is depicted in the following figure 4.1 :

2	Date	Screen Name	Full Name	Tweet Text	Tweet ID
3	5-22-2017 4:01:10	@Sunita167777	MSG ki beti Parwati	RT @Gurmeetramrahim: #LionHeartCrazeContinues ਪੰਜਾਬੀਓ! ਜ਼ੋਰ ਲਾ ਕੇ ਭੀ ਆਇਆ ਦੂਜਾ ਸਥਾਨ ! ਲਾਓ ਜ਼ੋਰ ਪੂਰਾ, ਫਿਰ ਵਧੇ ਦੁਹਰਾਡੀ ਸ਼ਾਨ !! https://t.co/S5n...	866609876352602113
4	5-22-2017 4:01:11	@shillinsa	shilpa kaushal	RT @JagbaniOnline: #HalfGirlfriend ਅਤੇ ਹਿੰਦੀ ਮੀਡੀਅਮ 'ਤੇ ਭਾਰੀ ਪਈ #RamRahim ਦੀ #JattuEngineer , ਕੀਤੀ ਸ਼ਾਨਦਾਰ ਓਪਨਿੰਗ https://t.co/cyxHwq9bYH...	866609881259712512
5	5-22-2017 4:01:12	@Jyoti13608540	Jyoti	RT @JagbaniOnline: #HalfGirlfriend ਅਤੇ ਹਿੰਦੀ ਮੀਡੀਅਮ 'ਤੇ ਭਾਰੀ ਪਈ #RamRahim ਦੀ #JattuEngineer , ਕੀਤੀ ਸ਼ਾਨਦਾਰ ਓਪਨਿੰਗ https://t.co/cyxHwq9bYH...	866609885412155392
6	5-22-2017 4:01:25	@mha_pagal	mha pgl	@MRandhawa72 @sukhi_zaildar @virsa22 @rai_pataka @kaurbrar067 Hain?? #ਝੋਟਾਂ ਮੇਨੂੰ ਤਾ ਬੈਡਾਂ ਲਗਦਾ 😂😂😂😂	866609940248612864
7	5-22-2017 4:01:35	@Kavita2650	Kavita	RT @ImRajchugh: ਆਉਣ ਹਨੇਰੀਆਂ ਜਾਣ ਹਨੇਰੀਆਂ ਸਾਰੇ ਸਿਨਮੇਆਂ ਵਿੱਚ ਧੂੰਆ ਪੈਣ ਤੇਰੀਆਂ ਸਕੈਂਟ ਸਿੰਘ ਸਿੱਧੂ ਦੀ ਬੱਲੇ ਬੱਲੇ ਕਾਲ ਪਰੇਮੀਆਂ ਦੀ ਜੁੱਤੀ ਬੱਲੇ ਬੁੱਕ ਲਾ...	866609983508623361
8	5-22-2017 4:01:37	@4seemasetia	4seemasetia	RT @JagbaniOnline: #HalfGirlfriend ਅਤੇ ਹਿੰਦੀ ਮੀਡੀਅਮ 'ਤੇ ਭਾਰੀ ਪਈ #RamRahim ਦੀ #JattuEngineer , ਕੀਤੀ ਸ਼ਾਨਦਾਰ ਓਪਨਿੰਗ https://t.co/cyxHwq9bYH...	866609999005530628
9	5-22-2017 4:02:05	@BrarHaRjeet	H🅰RJEET 🅱R🅰R	ਇੱਕ ਪਾਸੜ ਪਿਆਰ ਚੋਂ ਕੀ ਮਿਲਣਾ ਇੱਕ ਜੇਡ ਯਾਦ ਦੀ ਬਲਦੀ ਦੇ 🔥 ਇੱਕ ਕਮੀਂ ਜੇ ਪੂਰੀ ਨਹੀਂ ਹੋਣੀ ਉੱਡ ਦੁਨੀਆਦਾਰੀ ਚੱਲਦੀ ਦੇ 🌍	866610109358657536
10	5-22-2017 4:02:28	@Bittu78037646	Bittu	RT @JagbaniOnline: #HalfGirlfriend ਅਤੇ ਹਿੰਦੀ ਮੀਡੀਅਮ 'ਤੇ ਭਾਰੀ ਪਈ #RamRahim ਦੀ #JattuEngineer , ਕੀਤੀ ਸ਼ਾਨਦਾਰ ਓਪਨਿੰਗ https://t.co/cyxHwq9bYH...	866610206242983937
11	5-22-2017 4:02:34	@5bishu1	5bishu	RT @JagbaniOnline: #HalfGirlfriend ਅਤੇ ਹਿੰਦੀ ਮੀਡੀਅਮ 'ਤੇ ਭਾਰੀ ਪਈ #RamRahim ਦੀ #JattuEngineer , ਕੀਤੀ ਸ਼ਾਨਦਾਰ ਓਪਨਿੰਗ https://t.co/cyxHwq9bYH...	866610232557879296
12	5-22-2017 4:02:53	@OBittu	Bittu Insan O+	RT @JagbaniOnline: #HalfGirlfriend ਅਤੇ ਹਿੰਦੀ ਮੀਡੀਅਮ 'ਤੇ ਭਾਰੀ ਪਈ #RamRahim ਦੀ #JattuEngineer , ਕੀਤੀ ਸ਼ਾਨਦਾਰ ਓਪਨਿੰਗ https://t.co/cyxHwq9bYH...	866610310324637696
13	5-22-2017 4:03:18	@BikarInsan	Bikar Insan	RT @JagbaniOnline: #HalfGirlfriend ਅਤੇ ਹਿੰਦੀ ਮੀਡੀਅਮ 'ਤੇ ਭਾਰੀ ਪਈ #RamRahim ਦੀ #JattuEngineer , ਕੀਤੀ ਸ਼ਾਨਦਾਰ ਓਪਨਿੰਗ https://t.co/cyxHwq9bYH...	866610413881901056
14	5-22-2017 4:03:19	@dailyajitnews	Daily Ajit	ਬਟਾਲਾ: ਪਿੰਡ ਬਿਸਲੀ ਵਾਲ 'ਚ ਕੁੱਟਮਾਰ ਕਰਨ 'ਤੇ ਇਕ ਵਿਅਕਤੀ ਦੀ ਮੌਤ https://t.co/O1FEdu0lUC	866610418067922944
15	5-22-2017 4:03:36	@nathoinsan	Nathoinsan@gmail.com	RT @JagbaniOnline: #HalfGirlfriend ਅਤੇ ਹਿੰਦੀ ਮੀਡੀਅਮ 'ਤੇ ਭਾਰੀ ਪਈ #RamRahim ਦੀ #JattuEngineer , ਕੀਤੀ ਸ਼ਾਨਦਾਰ ਓਪਨਿੰਗ https://t.co/cyxHwq9bYH...	866610492785262592
16	5-22-2017 4:03:38	@6himanshu1	6himanshu	RT @JagbaniOnline: #HalfGirlfriend ਅਤੇ ਹਿੰਦੀ ਮੀਡੀਅਮ 'ਤੇ ਭਾਰੀ ਪਈ #RamRahim ਦੀ #JattuEngineer , ਕੀਤੀ ਸ਼ਾਨਦਾਰ ਓਪਨਿੰਗ https://t.co/cyxHwq9bYH...	866610497440759808
17	5-22-2017 4:04:13	@gurtej_2	Gurtej Singh 2	RT @sakshiinsan1: @Gurmeetramrahim ਮੈਂ ਤਾਂ ਉੱਡਦੀਗਾ <<<<< #JE17CrOnDay1	866610645512400897

Figure 4.1: Tweets collected using Twitter API

Raw tweets were segregated into 4 different folders to store tweets of different categories/classes as shown in figure 4.2. They were stored in the hard drive from where these can be easily imported to the snippet and further proceed for analysis.

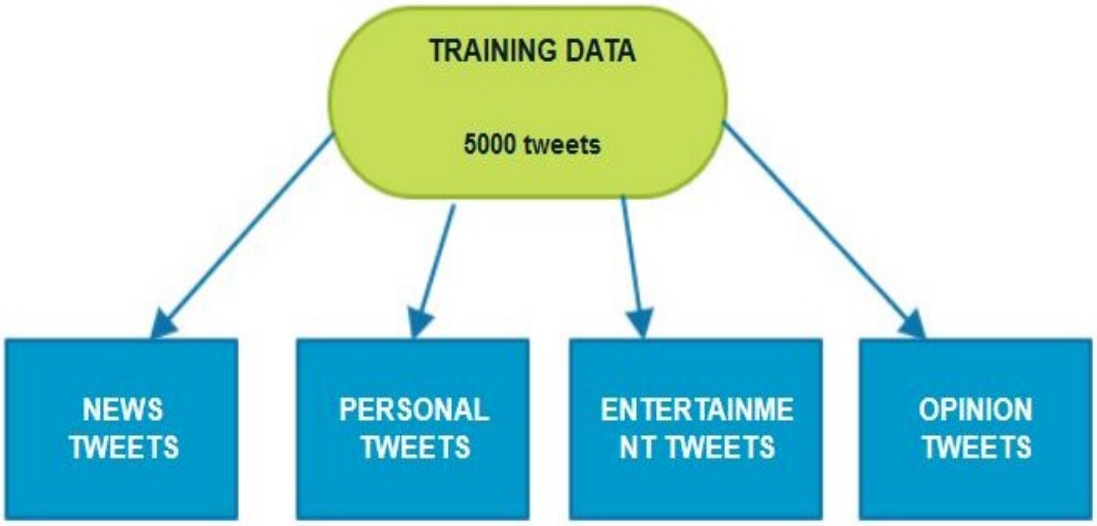


Figure 4.2: Training data segregation

Table 4.1: Number of files in each class

CLASS	NUMBER OF FILES
NEWS	670
ENTERTAINMENT	1,777
PERSONAL	2,075
OPINION	478
TOTAL FILES	**5000**

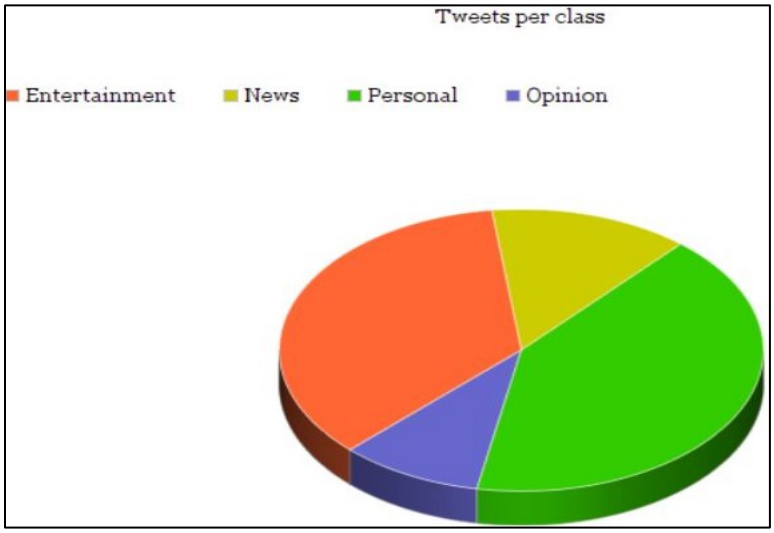


Figure 4.3: Distribution of tweets per class

Table 4.1 depicts the number of files contained in each class. The data collected is segregated into the respective class folders of News, Entertainment, Personal and Opinion as illustrated in figure 4.3. Training data is one on the basis of which an architecture/algorithm is built. Testing data is one on which the work is tested to measure the performance of an architecture. Training data is 5000 tweets and 400 tweets for testing purpose.

4.1.2 Description of pre-determined classes

- **News:** News tweets are understandable by a larger group of audiences and thus it is generic. Generally, they are neutral in nature, i.e. they are not highly opinionated on a particular topic but present the facts. They usually originate from corporate tweeters (E.g.- JagbaniOnline, Ajit, ZeePunjab) although there have been cases where they originated from personal tweeters [32]. They usually

convey the summary information and provide a link to an external detailed resource. They tend to be very structured and short. For example :

"ਪੰਜਾਬ ਸਰਕਾਰ ਵੱਲੋਂ ਬੁਢਾਪਾ ਪੈਨਸ਼ਨ ਤੇ ਦੂਜੀਆਂ ਸਕੀਮਾਂ ਦੇ ਲਾਭਪਾਤਰੀਆਂ ਨੂੰ ਵਿੱਤੀ ਸਹਾਇਤਾ ਦੇਣ ਲਈ 256.74 ਕਰੋੜ ਰੁਪਏ ਦਾ ਬਜਟ ਉਪਬੰਧ ਕੀਤਾ ਗਿਆ "

- **Personal:** Personal tweets are those that generally highlight an individual user's thoughts or his state of mind. It is generally significant to a smaller group of users and does not have global importance. Like news, they tend to be non-opinionated. They usually originate from personal tweeters. They are to the point and convey information via an artefact. They words to convey the thoughts within the 140 character limit. For Example:

  " ਹਾਹਾ ਹਾਹਾ... ਔਕੇ ਜੀ "

- **Opinions:** Opinion tweets convey the opinions of the author of the tweet. They usually originate from twitter users who share their thoughts on diverse entities. They either convey a positive or a negative sentiment through the tweet. They may contain shortening of words and may additionally use emphasis on words to convey stronger opinions. For Example :

  "ਬੰਦਾ ਮਿਹਨਤਾ ਨਾਲ ਕਰਦਾ ਤਰੱਕੀਆ ਪੰਡਤਾ ਦੇ ਨਗ ਨੀ ਕਿਸੇ ਨੂੰ ਤਾਰਦੇ..."

- **Entertainment:** Entertainment tweets are those which gives information about the entertainment industry, like movies reviews, movie collections or about songs release. It can be originated from twitter user or by any news agency. For Example:

  "#HalfGirlfriend ਅਤੇ 'ਹਿੰਦੀ ਮੀਡੀਅਮ' 'ਤੇ ਭਾਰੀ ਪਈ #RamRahim ਦੀ #JattuEngineer , ਕੀਤੀ ਸ਼ਾਨਦਾਰ ਓਪਨਿੰਗ https://t.co/cyxHwq9bYH... "

4.2 SA task workflow

The goal is to achieve SA for data extracted from Twitter using vector evaluation method. This method uses Punjabi tweets and then the weights of each of the words are

calculated to determine the keywords (or features). Then they will be compared with each word of the testing corpus which we have to categorize in order to determine the best suitable category. The workflow is shown in figure 4.4.

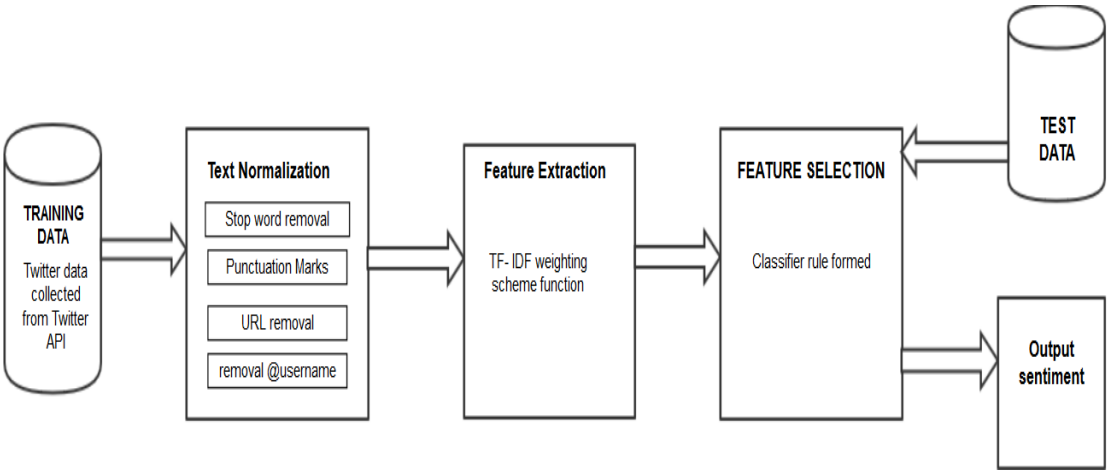


Figure 4.4: SA workflow for implementation

The training data is divided into 4 classes of sentiments which are News, Personal, Entertainment, Opinion.

4.2.1 Operations for SA of Punjabi tweets

The procedure for SA on Punjabi tweets are as follows:

4.2.1.1 Data collection from Twitter:

5000 Punjabi tweets are collected from Twitter by using Twitter API. Most of the tweets include message along with usernames, URLs, punctuation marks, emoticons, special characters, stop words, abbreviations and hash tags. In order to make this data fit for SA, pre-processing of this raw data is mandatory.

4.2.1.2 Text Normalization/Pre-processing:

Code for pre-processing the collected data is created using Python. Cleaning of Twitter data is very important because tweets contain number of syntactic features that are not useful for analysis and may hinder the results. The pre-processing includes the following operations:

- **Stop word removal :** In this step, eliminate the stop words from the tweet using the Punjabi stop words list. It contains 146 words. Figure 4.5 represents the list of stops words for Punjabi.

ਉਹ	ਆਪਣਿਆਂ	ਇਲਾਵਾ	ਹਾਂ।	ਕਾਸਦਾ	ਜੀਹਨੂੰ	ਤੁਹਾਡੀਆਂ	ਵੀ	ਮੇਰੀਆਂ
ਉਹਨੇ	ਆਪੇ-ਆਪਦ	ਸੋ	ਹੈ	ਕਾਹਦਾ	ਜਿਸ	ਤੁਹਾਨੂੰ	ਵਿਚ	ਲਈ
ਉਹਨੂੰ	ਆਪੇ-ਆਪਦੇ	ਸੀ	ਹਨ	ਕਿਸਦੀਆਂ	ਜਿਸਦਾ	ਦਾ	ਵਿੱਚ	
ਉਹਨਾਂ	ਆਪਣਾ-ਆ	ਸਨ	ਹਾਂ	ਕੀਹਦੀਆਂ	ਜਿਸਦੇ	ਦੀ	ਵਿੱਚੋਂ	
ਉਸ	ਆਪਣੇ-ਆਪ	ਸਗੋਂ	ਕਿ	ਕਿਸਦਿਆਂ	ਜਿਸਦੀਆਂ	ਦੇ	ਵਜੋਂ	
ਉਸਨੇ	ਇਨ੍ਹਾਂ	ਸਾਡਾ	ਕੌਣ	ਕੀਹਦਿਆਂ	ਤੂੰ	ਦੀਆਂ	'ਚ	
ਉਸਨੂੰ	ਇਹ	ਸਾਡੇ	ਕੀ	ਕੋਈ	ਤੋਂ	ਨੂੰ	ਆਪਣੀ	
ਉਵ	ਇਹੀ	ਸਾਡੀ	ਕਿਹੜਾ	ਕਈ	ਤੇ	ਨੇ	ਆਪਣੀਆਂ	
ਉਨ੍ਹਾਂ	ਇਸ	ਸਾਡੀਆਂ	ਕਿਹੜੇ	ਕਈਆਂ	'ਤੇ	ਨਾ	ਇੱਕ	
ਉਹਦਾ	ਇਹਨਾਂ	ਸਾਰੇ	ਕਿਹੜੀ	ਚਾਰੇ	ਤੇਥੋਂ	ਨਾਲ	ਇੱਕੋ	
ਉਹਦੇ	ਇਹਨੇ	ਸਾਰਾ	ਕਿਹੜੀਆਂ	ਚੌਂ	ਤੈਨੂੰ	ਪਰ	ਹੈ।	
ਉਹਦੀ	ਇਹਨੂੰ	ਸਾਰਿਆਂ	ਕਿਹੜਿਆਂ	ਜਾਂ	ਤੁਸੀਂ	ਫਿਰ	ਹਨ।	
ਉਹਦੀਆਂ	ਇਸਨੇ	ਹੋਰਨਾਂ	ਕਿਸ	ਜੇ	ਤੁਸਾਂ	ਮੈਂ	ਕਿਸਦੇ	
ਅਤੇ	ਇਸਨੂੰ	ਹੋਰ	ਕਿਸੇ	ਜੋ	ਤੇਰਾ	ਮੈਥੋਂ	ਕੀਹਦੇ	
ਅਸੀਂ	ਇਹਦਾ	ਹੋ	ਕਿਨ੍ਹਾਂ	ਜਿਵੇਂ	ਤੇਰੇ	ਮੈਨੂੰ	ਜਿਹੜਿਆਂ	
ਅਸਾਂ	ਇਹਦੇ	ਹੋ	ਕਿਨ	ਜਦੋਂ	ਤੇਰੀ	ਮੇਰਾ	ਜਿਹਨੇ	
ਆਪਣਾ	ਇਹਦੀ	ਹਨ	ਕਿਸਦਾ	ਜਿਹੜਾ	ਤੇਰੀਆਂ	ਮੇਰੇ	ਤੁਹਾਡੇ	
ਆਪਣੇ	ਇਹਦੀਆਂ	ਹਾਂ	ਕੀਹਦਾ	ਜਿਹੜੇ	ਤੁਹਾਡਾ	ਮੇਰੀ	ਤੁਹਾਡੀ	

Figure 4.5: Punjabi Stop Words

A sample processed tweet is shown in figure 4.6 and 4.57:

" ਫੀਕੋ ਵੱਲੋਂ ਸਨਤਕਾਰਾਂ ਨੂੰ ਪਾਣੀ ਦੀ ਵਰਤੋਂ ਕਰਨ ਅਤੇ ਐਨ ਓ ਸੀ ਲੈਣ ਸਬੰਧੀ ਲਗਾਇਆ ਗਿਆ ਕੈਂਪ https://t.co/A52SeoBJgW"

Figure 4.6: Original Tweet

ਫੀਕੋ ਵੱਲੋਂ ਸਨਤਕਾਰਾਂ ਪਾਣੀ ਵਰਤੋਂ ਕਰਨ ਐਨ ਓ ਲੈਣ ਸਬੰਧੀ ਲਗਾਇਆ ਗਿਆ ਕੈਂਪ URL

Figure 4.7: Processed Tweets

- **URL (Uniform resource locator) removal :** URLs are removed or substituted with *'URL'* word and also include it to stop words list.
- **Username removal :** The @ symbol is removed that comes before the username in a tweet. It also substitutes the @ and the username with a word 'AT_USER' and also include it in the stop words list.
- **Punctuation marks removal :** All punctuation marks, such as !, ?, (,), - , _ , :, %, etc. are deleted as they have no significance in the data**.**
- **Remove Emoticons :** remove emoticons like 😊, 💞, ❤ , 🔌 , etc. from tweets.
- **Remove duplicates :** remove all repeating words from text so that there will be no duplicates.

Table 4.2 shows the various types of contents that are included in tweets and also the actions performed on these contents.

Table 4.2: Removed and Modified Content

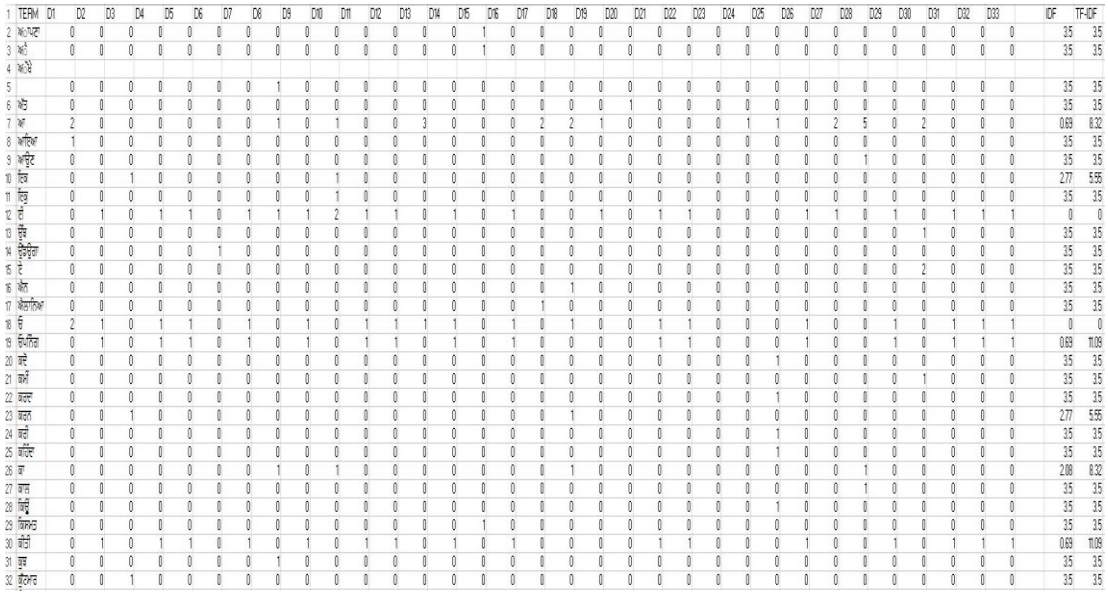
CONTENT	ACTION
Punctuation (! ? , . " : ;)	Deleted
#	Deleted #
@any_user	Delete @any_user or replaced with "AT_USER" and then added in stop words.
URLs and web links	Delete URLs or replaced with "URL" and then added in stop words
Number	Removed
Stop words	Removed
Emoticons	Removed
White spaces	Removed

4.2.1.3 Feature Extraction:

Weight of each word in the collected document is calculated using weighting scheme function by multiplying term occurrence frequency (TF) and the inverse document frequency (IDF). The IDF of the ith term is defined as log (n/DF). Here *'n'* depicts the number of documents in the corpus and *'DF'* represents the number of documents in which the term appears and the weight function is shown below in Equation 4.1:

$$Wij = TFij * log\,(n/DFj) \tag{4.1}$$

Following figure 4.8 represents the sample of calculated TF-IDF of each word in processed tweet:



TERM	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15	D16	D17	D18	D19	D20	D21	D22	D23	D24	D25	D26	D27	D28	D29	D30	D31	D32	D33	IDF	TF-IDF
ਅੰਮ੍ਰਿਤਸਰ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3.5	3.5
ਅਹੀ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3.5	3.5
ਅੰਦੋ																																			
	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3.5	3.5
ਅੰਤ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	3.5	3.5
ਆਂ	2	0	0	0	0	0	0	1	0	1	0	0	3	0	0	0	2	2	1	0	0	0	0	1	1	0	2	5	0	2	0	0	0	0.69	8.32
ਆਇਆ	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3.5	3.5
ਆਉਂਦਾ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	3.5	3.5
ਇਕ	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.77	5.55
ਇਕੱਠ	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3.5	3.5
ਇਜ	0	1	0	1	1	0	1	1	1	2	1	1	0	1	0	1	0	0	1	0	1	1	0	0	0	1	1	0	1	0	1	1	1	0	0
ਉਸ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	3.5	3.5
ਉਹਨਾਂ	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3.5	3.5
ਏ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	3.5	3.5
ਐਲਾਨਿਆ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3.5	3.5
ਐਨ	2	1	0	1	1	0	1	0	1	0	1	1	1	1	0	1	0	1	0	0	1	1	0	0	0	1	0	0	1	0	1	1	1	0	0
ਓਪੀਨੀਓ	0	1	0	1	1	0	1	0	1	0	1	1	0	1	0	1	0	0	0	0	1	1	0	0	0	1	0	0	1	0	1	1	1	0.69	11.09
ਕਸੇ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	3.5	3.5
ਕਸੌਂ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	3.5	3.5
ਕਰਦਾ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	3.5	3.5
ਕਰਨ	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.77	5.55
ਕਰੀ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	3.5	3.5
ਕਰੇਹ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	3.5	3.5
ਕਾ	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2.08	8.32
ਕਾਸ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	3.5	3.5
ਕਿਉਂ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	3.5	3.5
ਕਿਸਾਨ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3.5	3.5
ਕੀਤੀ	0	1	0	1	1	0	1	0	1	0	1	1	0	1	0	1	0	0	0	0	1	1	0	0	0	1	0	0	1	0	1	1	1	0.69	11.09
ਕੁਝ	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3.5	3.5
ਕੁੜਮਨ	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3.5	3.5

Figure 4.8: TF-IDF of each word from processed tweet

4.2.1.4 Feature Selection:

After calculating the weight of each word, keywords called '*features'* with largest weights are chosen from the data. For example:

"RT @JagbaniOnline: #JattuEngineer ਫਿਲਮ ਲੱਗਣ 'ਤੇ ਡੇਰਾ ਪ੍ਰੇਮੀਆਂ ਨੇ ਮਨਾਈ ਖੁਸ਼ੀ https://t.co/aUMZ0GU1wJ @MSGTheFilm @Gurmeetramrahim"

The word 'ਫਿਲਮ' has the highest weight among all the words in the tweet which is 86.12. Thus this word is selected as a keyword or feature.

For the SA of Punjabi tweets, 5400 Punjabi tweets were collected. First 5000 tweets were selected as training set and remaining 400 as testing sets. One of the ways that data can be represented is feature-based. By features, it is meant that some attributes that are thought to capture the pattern of the data are first selected and the entire dataset must be represented in terms of it. Different features such as syntactic features, or semantic features can be used. For example, one can use the keyword lexicons as features as used in this work. Then the dataset can be represented by these features using either their presence or frequency or weight and in this work, the dataset is represented by features using their weights. Attribute selection is the process of extracting features by which the data will be represented before any training algorithm takes place.

Attribute selection is the first task and once the attributes are selected, the data will be represented using the selected attributes. Although, the entire data set is used in this work for the selection of attributes, the representation of the data must be done on a per instance (Twitter post) basis.

Feature vector plays a very important role in classification and helps to determine the working architecture. Feature vector also help in predicting the unknown data sample. Each tweet words are added to generate the feature vectors. In this work, weight of each word is calculated in order to extract the feature. Once these features are extracted from the input data set, they are fed to algorithm for measuring the system performance. The selection of features may seem to be ad hoc at first glance but the features were chosen such that it is tuned for the four classes.

4.2.1.5 Finding the most suitable category:

Once the features are extracted, assign the suitable category by comparing the features with words of each tweet.

Chapter 5

Results and Discussions

In this section, the proposed algorithm has been tested on 400 tweets, known as test data. Tweets are pre-processed and normalized. TF-IDF weighing scheme is used for feature extraction since minimum feature set is required for SA. This method is compared with BOW which uses a large feature set and Word frequency model which calculates frequency of each word. The results for SA of Punjabi tweets using three different methods of feature extraction are compared in this section.

5.1 Test Tweets

Test tweets are collected with the help of Twitter API as shown in figure 5.1:

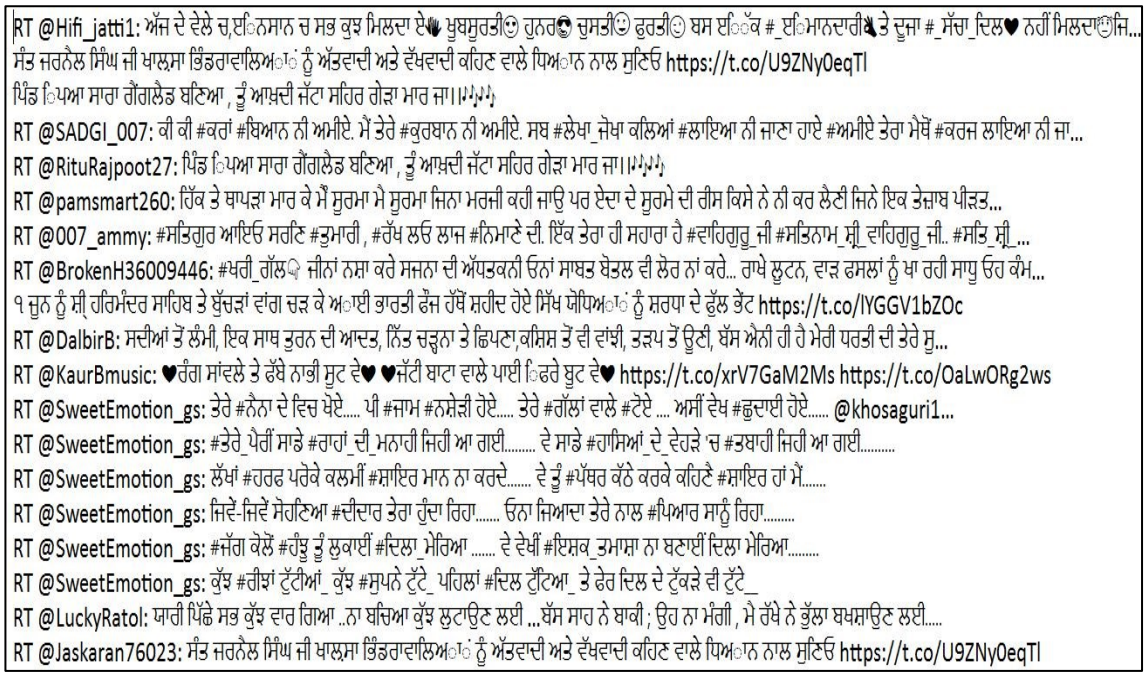


Figure 5.1: Sample of testing tweets

Figure 5.2 shows the number of test tweets in each class.

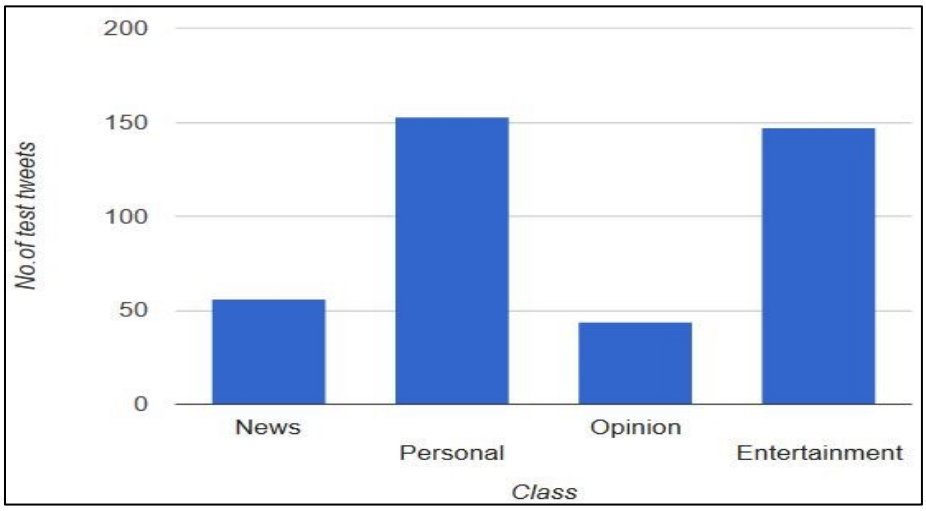


Figure 5.2: Test tweets per class

5.2 Feature Extraction and Selection

The features are extracted using 3 different methods as described earlier, i.e. using TF-IDF, word frequency and Bag-Of-Words. Following figure 5.3 illustrates the features extracted for SA of Punjabi tweets using the above mentioned 3 different approaches.

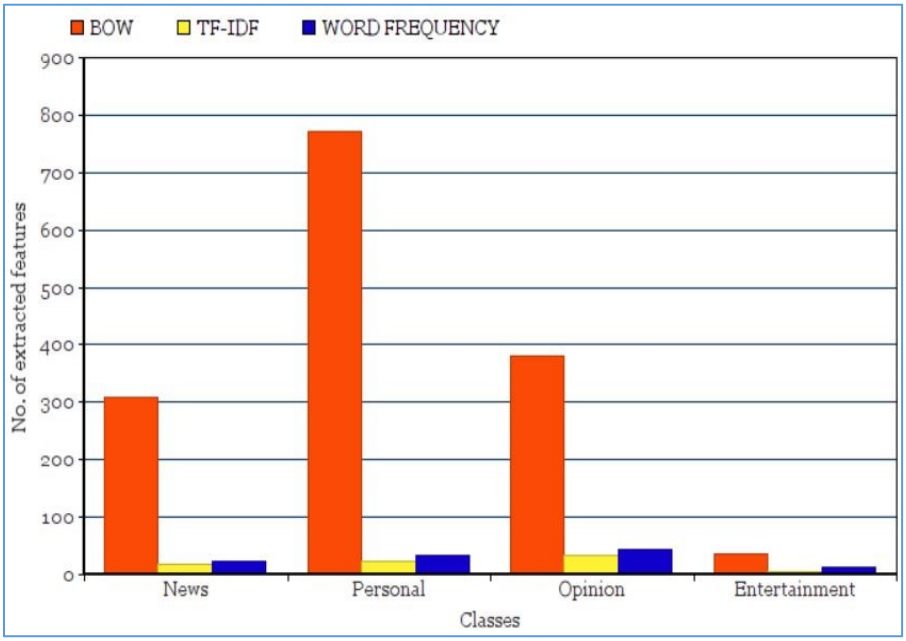


Figure 5.3: Features per class

Clearly BOW model generates maximum number of features, some of which may be of least significance (such as ਜ਼ੋਰ and ਪੁਰਾ) for each class of Punjabi tweets.

The aim of this work is to choose a small feature set for SA of Punjabi tweets and the TF-IDF weighing scheme provides a small feature set.

5.3 Experimental results of SA of Punjabi tweets

The results for proposed framework have been compared on the basis of three properties, namely accuracy, precision and recall.

- **Precision** is defined as a measure of relevance of the results.
- **Recall** tells about how many truly relevant results are returned.
- **Accuracy** means how many documents match properly with total number of documents.

5.3.1 Accuracy

Accuracy means how many tweets are correctly identified among total number of tweets. Following figure 5.4 depicts the overall accuracy of the models for testing of twitter data. As the figure shows, TD-IDF shows more accurate results as compared to other two models.

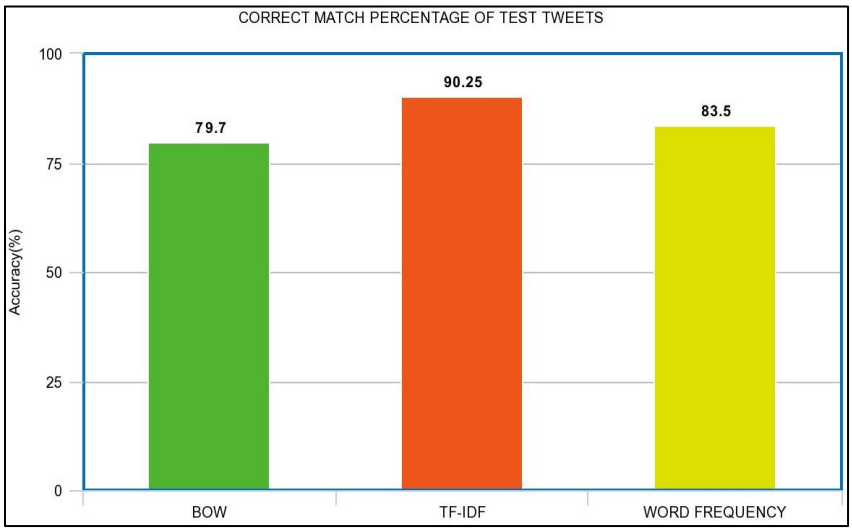


Figure 5.4: Overall accuracy using 3 models

Table 5.1 shows the tweets correctly that have correctly matched with the respective

classes as:

Table 5.1: Correct tweets detected per class

METHOD / CLASS	BOW	TF-IDF	WORD FREQUENCY
News	41/56 = 73.2%	48/56 = 85.5%	44/56 = 78.5%
Opinion	31/44 = 70.4%	36/44 = 81.6%	35/44 = 79.5%
Personal	126/153 = 82.3%	138/153 = 90.1%	131/153 = 85.6%
Entertainment	116/147 = 78.9	128/147 = 87%	123/147 = 83.6%

Figure 5.5 shows the individual accuracies per class of each model.

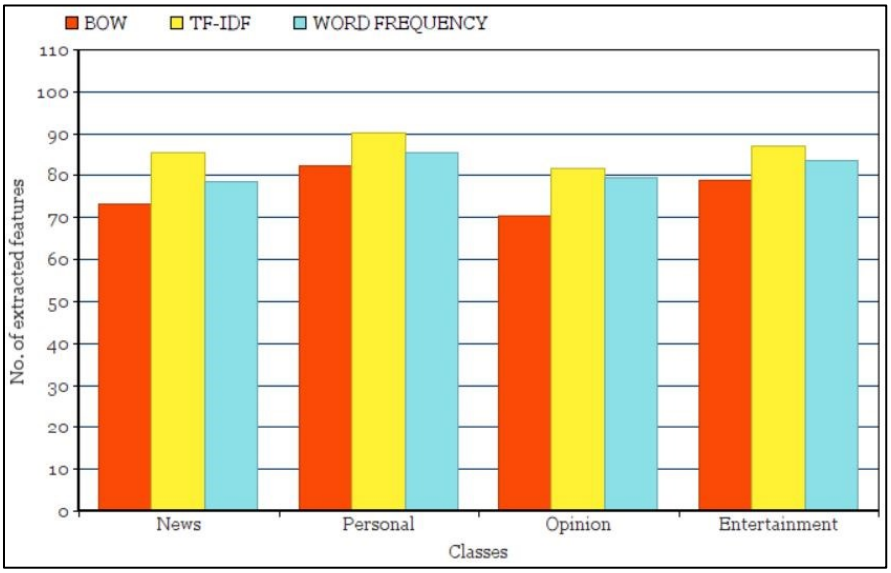


Figure 5.5: Accuracy per class

For a classification system, time for building the model is a critical factor. The time tends to get higher with higher dimensionality. This causes a problem for text classification because standard techniques such as BOW results in a very high dimensionality. Thus, TF-IDF model suits the best.

5.3.2 Precision

Precision means total number of correctly identified tweets divided by number of tweets

that are supposed to be correctly identified. It is calculated as:

$$Precision = T_P / (T_P + F_P) \quad (5.1)$$

Here T_P is True Positive and F_P is False Positive.

Table 5.2 shows the precision per class for each model using Equation (5.1):

Table 5.2: Precision per class

METHOD / CLASS	BOW	TF-IDF	WORD FREQUENCY
News	41/(41+6)= 87.2%	48/(48+4)= 92.4%	44/(44+5)= 89.7%
Opinion	31/(31+4)= 88.5%	36/(36+3)= 92.3%	35/(35+4)= 89.7%
Personal	125/(126+15)= 89.3%	138/(138+10)= 93.2%	131/(131+13)= 90.9%
Entertainment	116/(116+16)= 87.8%	128/(128+12)= 91.4%	123/(123+14)= 89.7%

Following figure 5.6 demonstrate the precision calculated for each class with respect to the methods of feature extraction applied.

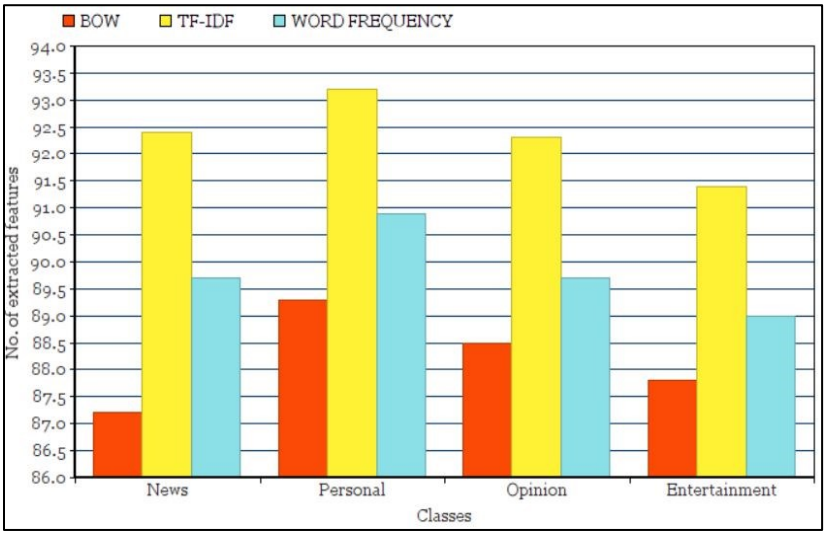


Figure 5.6: Precision per class

5.3.3 Recall

Recall is number of correct matches divided by total number of actual matches. It is calculated as:

$$Recall = T_P / (T_P + F_N) \quad (5.2)$$

F_N is False Negative. Following Table 5.3 shows the Recall per class as:

Table 5.3: Recall per class

METHOD CLASS	BOW	TF-IDF	WORD FREQUENCY
News	41/(41+7)= 85.4%	50/(50+3) = 94.3%	44/(44+4) = 91.6%
Opinion	31/(31+5)= 86.1%	36/(36+3) = 92.3%	35/(35+3) = 92%
Personal	126/(126+11)= 91.9%	141/(141+3)= 96%	131/(131+7) = 94.5%
Entertainment	116/(116+13) = 89.9%	128/(128+6) = 95%	123/(123+9)= 93.1%

Figure 5.7 shows recall of individual class as:

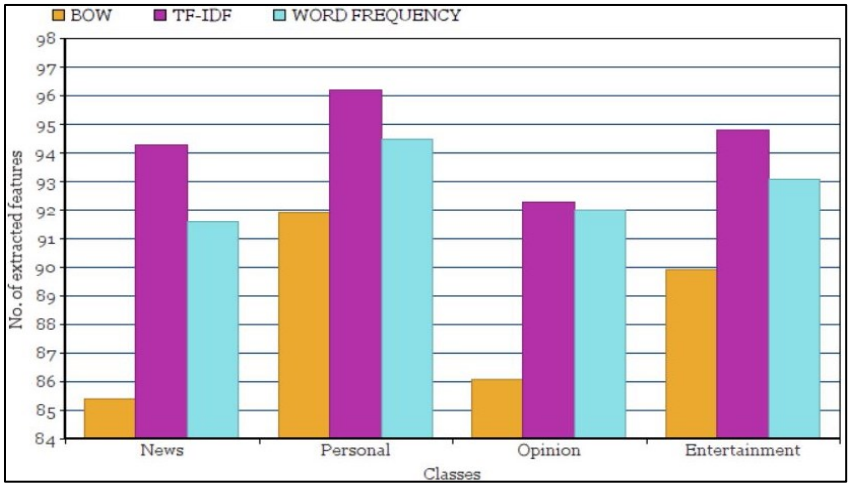


Figure 5.7: Recall per class

The number of words used in BOW after stop words removal was 1,496. When the number of tweets increases, number of words, called *features* also increases which takes more time to build a classifier. Word frequency model finds the frequency of each word after stop words removal. But some least significant words may acquire a high frequency which makes no relevance for feature selection process. While TF-IDF have fixed dimensions since it makes use of small feature sets. So TF-IDF performs better than BOW and Word Frequency model.

In BOW model, misclassified tweets are basically between personal and opinion tweets. In case of TF-IDF and Word Frequency, misclassified tweets are between news and opinion.

Chapter 6

Conclusion and Future scope

6.1 Conclusion

The framework described in this dissertation is a step towards efficient classification of Punjabi short text messages. These are difficult to classify than larger text because they have few word occurrences and hence becomes difficult to gather the semantics of these messages. The efficient ideas are required to increase the accuracy of classification by using minimum feature sets to represent these messages. A framework is proposed to classify the Punjabi Twitter messages which are perfect example of short text messages because of their 140 character limit. In this framework, data is collected from Twitter using Twitter API. Collection of 5000 tweets is used as a training data. This training data is then manually labelled into four classes, namely– news, opinion, personal and entertainment.

Feature selection can done using three methods- TF-IDF weighing scheme, BOW scheme and Word frequency model. In TF-IDF weighing scheme, weight of each word is calculated from the collected document and words with highest weights are selected as features. In BOW, tweet is broken into words. Stop words are removed and every word represents a feature. In case of word frequency, frequency of every word is calculated and words with highest frequencies are selected as features. It is analysed that TF-IDF scheme shows better results than BOW and word frequency method.

In this dissertation, rule based supervised approach using vector evaluation method is implemented for SA of Punjabi tweets.

6.2 Future Scope

Some of future scopes that can be included in our research work are:

1. New machine learning applications and technologies such as deep Learning can be implemented.
2. A web-based application or a tool can be made for classification of data into different categories.
3. Data corpus can be enhanced by adding more tweets.
4. There can be increase in the classification categories for better analysis of tweets.
5. There can be improvement in the system that can deal with sentences of multiple classes.

References

[1] W.Medhat, A.Hassan, H.Korashy, "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal, vol. 5, pp. 1093-1113, April 2014.

[2] S.Har-Peled, D.Roth, D.Zimak, "Constraint Classification for Multiclass Classification and Ranking," in Advances in Neural Information Processing Systems (NIPS),Vancouver, BC, January,2002.

[3] C.D.Manning, P.Raghavan, H.Schutz, Introduction to Information Retrieval, Cambridge University Press, 2008.

[4] "Punjabi Language", [Online] Available: https://en.wikipedia.org/wiki/Punjabi_language

[5] W.Cohen, "William W.Cohen," [Online]. Available: http://www.cs.cmu.edu/ ./wcohen/.

[6] "Stemming", [Online]. Available: http://en.wikipedia.org/wiki/Stemming

[7] L.Plaza, J.C.Albornoz, "Sentiment Analysis in Business Intelligence: A survey", IGI-Global, pp. 231 252, 2011.

[8] F.Eskandari, H.Shayestehmanesh, S.Hashemi, "Predicting best answer using sentiment analysis in community question answering system", Signal Processing and Intelligent Systems Conference (SPIS),pp. 53-57, 2015.

[9] M.Guerini, L.Gatti, M.Turchi, "Sentiment analysis: How to derive prior polarities from SentiWordNet", arXiv preprint arXiv:1309.5843, 2013.

[10] N.Yadav, N.Chatterjee, "Text Summarization Using Sentiment Analysis for DUC Data, International Conference on Information Technology(ICIT), pp. 229-234, July 2017.

[11] J.Liu, J.Yao, G.Wu, "Sentiment Classification Using Information Extraction Technique", Advances in Intelligent Data Analysis VI, Lecture Notes in Computer Science, vol. 3646, Springer, Berlin, Heidelberg.

[12] "Sentiment Analysis", [Online]. Available: https://en.wikipedia.org/wiki/Sentiment_analysis.

[13] S.Vohra, J.Teraiya, "Applications and Challenges for Sentiment Analysis: A Survey", International Journal of Engineering Research & Technology (IJERT), vol. 2, no. 2, pp. 1-5, 2013.

[14] M.Tsytsarau, T.Palpanas, "Survey on mining subjective data on the web", Data Mining and Knowledge Discovery,vol.24, no. 3, pp. 478-514, 2012.

[15] R. Plutchik, "A general psychoevolutionary theory of emotion," in Emotion: Theory, Research and Experience, University of Illinois Press, 1980, pp. 3-33.

[16] "Techniques and Applications for Sentiment Analysis", [Online]. Available: http://www.ceine.cl/en/techniques-and-applications-forsentiment-analysis/

[17] D.M.E.D.M.Hussein, "A survey on sentiment analysis challenges", Elsevier, pp. 1-9, 18 April, 2016

[18] "Sarcasm", [Online]. Available: https://en.wikipedia.org/wiki/Sarcasm

[19] S.Mukherjee, P.Bhattacharyya, "Sentiment Analysis: A Literature Survey", IIT Bombay, Mumbai, 2013.

[20] "TIME," [Online]. Available: http://www.time.com/time/magazine/article/ 0,9171,1044658,00.html

[21] http://www.twitter.com

[22] "Microblogging," [Online]. Available: http://en.wikipedia.org/wiki/Microblogging

[23] "Twitter," [Online]. Available: http://en.wikipedia.org/wiki/Twitter.

[24] http://www.orkut.com.

[25] J. Sankaranarayanan, H. Samet, B. E. Teitler, M.D. Lieberman, J. Sperling, "TwitterStand: News in Tweets",in ACM GIS, Seattle, Washington, pp. 42-51, November 2009.

[26] J.Kurtz, "What s a Retweet?," SocialMediaToday, 27 May 2009.

[Online]. Available: http://www.socialmediatoday.com/SMC/96585.

[27] "Twitter Search", [Online]. Available: http://search.twitter.com/.

[28] "HashTags," [Online]. Available: http://hashtags.org.

[29] P.Pang, L.Lee, S.Vaithyanathan, "Thumbs up? sentiment Classification using machine learning techniques", Proc. ACL-02 conference on Empirical methods in natural language processing, vol.10, pp. 79-86, 2002

[30] O.Almatrafi , S.Parack, B.Chavan, "Application of location-based sentiment analysis using Twitter for identifying trends towards Indian general elections 2014". Proc. The 9th International Conference on Ubiquitous Information Management and Communication, 2015.

[31] L.Jiang, M.Yu, M.Zhou, X.Liu, T.Zhao, "Target-dependent twitter sentiment Classification", Proc. The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 151-160, 2011

[32] P.Pang, L.Lee, "Opinion Mining and Sentiment Analysis. Foundation and Trends in Information Retrieval", vol. 2(1-2), pp.1-135, 2008.

[33] E. U. Jain, A. Sandhu, "Emotion Detection from Punjabi Text using Hybrid Support Vector Machine and Maximum Entropy Algorithm," International Journal of Advanced Research in Computer and Communication Engineering(IJARCCE), vol. 4, no. 11, pp. 89-93, November 2015.

[34] M. Patil, P. Game, "Comparison of Marathi Text Classifiers, " ACEEE Int. J. on Information Technology, vol. 4, no. 1, pp. 11-22, March 2014.

[35] J. J. Patil, N. Bowiri, "Automatic Text Categorization Marathi Documents", International Journal of Advance Research in Computer Science and Management Studies, pp. 2321-7782, March-2015.

[36] J. Kaur, S. Bhagla, "News Classification Using Naˆflve Baye's Classier," International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), vol. 6, no. 4, pp. 698-702, April 2016.

[37] A. Sharma, "Sentiment Analyzer using Punjabi Language," International Journal of Innovative Research in Computer and Communication Engineering(IJIRCCE), vol. 2, no. 9, pp. 5904-5909, September 2014.

[38] N.Desai,A.D.Dave, "Sarcasm Detection in Hindi sentences using Support Vector machine," International Journal of Advance Research in Computer Science and Management Studies(IJARCSMS), vol. 4, no. 7, pp. 8-15, July 2016.

[39] A. K. Mandal, R. Sen, "Supervised Learning Methods For Bangla Web Document Categorization," International Journal of Arti cial Intelligence and Applications(IJAIA), vol. 5, no. 5, pp. 93-105, September 2014.

[40] S. Amarappa, Dr. S. V. Sathyanarayana, "Kannada Named Entity Recognition And Classification (NERC) based on Multinomila Naive Bayes(MNB) Classi er," International Journal on Natural Language Computing (IJNLC), vol. 4, no. 4, pp. 39-52, August 2015.

[41] T. B. Shahi, A. Yadav, "Mobile SMS Spam Filtering for Nepali Text Using Naˆflve Bayesian and Support Vector Machine," International Journal of Intelligence Science(IJIS), vol. 4, pp. 24-28, 2014.

[42] Pak, Alexander, P.Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", LREC, vol. 10, 2010.

[43] N.Cohen, "Twitter on the barricades: Six lessons learned", [Online]. Available: http://www.nytimes.com/2009/06/21/weekinreview/21cohenweb.html, 20 June, 2009.

[44] "Expert System", [Online]. Available: http://www.expert system.com/natural-language-processing-sentimentanalysis/.

[45] M.Hu, B.Liu, "Mining and Summarizing Customer Reviews", KDD, August 2004.

[46] T.Wilson, J.Wiebe, P.Ho man, "Recognizing contextual polarity in phrase level sentiment analysis", ACL, 2005.

[47] J.R.Quinlan, "Induction of Decision Trees", Machine Learning, vol. 1, pp. 81-106, 1986..

[48] "Rule-based KS", [Online]. Available: http://www.cs.uu.nl/docs/vakken/mdks/lectures/lecture4.pdf

[49] Y.Mejova, I.Weber, M.W.Macy, "Twitter: A Digital Socioscope", Cambridge University Press, 2015. Same as 62

[50] E.M.Cody, A.J.Reagan, P.S.Dodds, C.M.Danforth, "Public Opinion Polling with Twitter", arXiv:1608.02024v1, pp. 1-15, August 2016.

[51] S.Mohammad, "From Once Upon a Time to Happily Ever After:Tracking Emotions in Novels and Fairy Tales", 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pp. 105-114, June 2011.

[52] P.Lewicki, Hill T 2006 Statistics: methods and applications, Tulsa, OK. Statsoft.

[53] C.Chen, F.I.SanJuan, E.SanJuan, C.Weaver, "Visual analysis of conicting opinions", IEEE Symposium on Visual Analytics Science and Technology, Baltimore, Maryland: United States, 59-66, 2006.

[54] P.D.Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised Classification of reviews", Association for Computational Linguistics(ACL), Philadelphia, pp. 417-424, 2002.

[55] G. Li, F. Liu, "A clustering-based approach on sentiment analysis", Intelligent Systems and Knowledge Engineering (ISKE), pp. 331-337, 2010.

[56] Y.Mejova, I.Weber, M.W.Macy, "Twitter: A Digital Socioscope", Cambridge University Press, 2015.

[57] M.Mohri, A.Rostamizadeh, A.Talwalkar, “Foundations of Machine Learning”, The MIT Press ISBN 9780262018258, 2012.

[58] “Supervised Learning”, [Online]. Available: https://en.wikipedia.org/wiki/Supervised_learning.

[59] G.Beigi, X.Hu, R.Maciejewski, H.Liu, "An Overview of Sentiment Analysis in Social Media and Its Applications in Disaster Relief," in *Sentiment Analysis and Ontology Engineering* , Springer, March 2016, pp. 313-340.

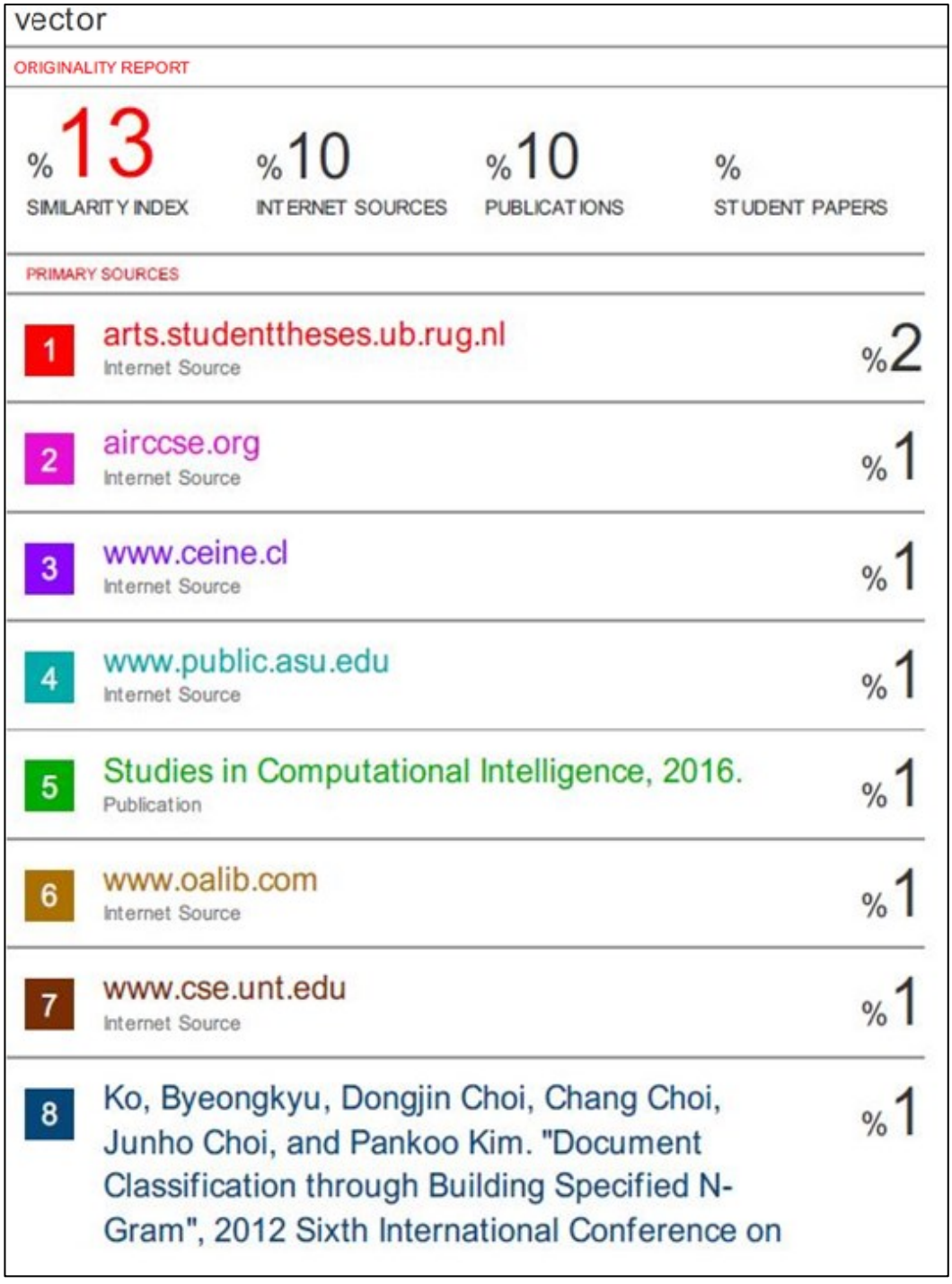
List of Publication

Mehak, Varinderpal Singh and Rajiv Kumar, "Punjabi Document Classification using Vector Evaluation Method", in *International Conference on Computing Methodologies and Communication(ICCMC 2017)*, IEEE[Accepted and Published].

Video Link

https://youtu.be/PmXvu3c6saU

Plagiarism Report



vector

ORIGINALITY REPORT

%13	%10	%10	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	arts.studenttheses.ub.rug.nl Internet Source	%2
2	airccse.org Internet Source	%1
3	www.ceine.cl Internet Source	%1
4	www.public.asu.edu Internet Source	%1
5	Studies in Computational Intelligence, 2016. Publication	%1
6	www.oalib.com Internet Source	%1
7	www.cse.unt.edu Internet Source	%1
8	Ko, Byeongkyu, Dongjin Choi, Chang Choi, Junho Choi, and Pankoo Kim. "Document Classification through Building Specified N-Gram", 2012 Sixth International Conference on	%1