Sentiment analysis for Urdu news Tweets using Decision tree

By

Raheela Bibi

NUST201261292MCEME 117435

MS-16(CSE)

Submitted to Department of Computer Engineering

In fulfillment of the requirements for the degree of

Masters of Science

In

Computer Software Engineering

Thesis Supervisor

Dr. Usman Qamar

College of Electrical and Mechanical Engineering

National University of Science and Technology

In the name of Allah most beneficent most merciful

وَلَا يُحِيطُونَ بِشَيْءٍ مِّنْ عِلْمِهِۦٓ إِلَّا بِمَا شَآءَ ۚ

And they can’tencompass anything from His knowledge, but to extend He wills [2:255]

Sentiment analysis for Urdu news Tweets using Decision tree

Author
RAHEELA BIBI
00000117435

A thesis submitted in partial fulfillment of the requirements for the degree of
MS Computer Software Engineering

Thesis Supervisor:
DR. USMAN QAMAR
Thesis Supervisor's Signature:____

DEPARTMENT OF COMPUTER ENGINEERING
COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY,
ISLAMABAD
JUNE 2018

Language Correctness Certificate

This thesis has been read by an English expert and is free of typing, syntax, semantic, grammatical and spelling mistakes. Thesis is also according to the format given by the university.

Signature of Student

Raheela Bibi
Registration Number
00000117435

Signature of Supervisor

Dr. Usman Qamar

Copyright Statement

Copyright in text of this thesis rests with the student author. Copies (by any process) either in full, or of extracts, may be made only in accordance with instructions given by the author and lodged in the Library of NUST College of E&ME. Details may be obtained by the Librarian. This page must form part of any such copies made. Further copies (by any process) may not be made without the permission (in writing) of the author.

The ownership of any intellectual property rights which may be described in this thesis is vested in NUST College of E&ME, subject to any prior agreement to the contrary, and may not be made available for use by third parties without the written permission of the College of E&ME, which will prescribe the terms and conditions of any such agreement.

Further information on the conditions under which disclosures and exploitation may take place is available from the Library of NUST College of E&ME, Rawalpindi.

ACKNOWLEDGEMENTS

"This is by the Grace of my Lord to test me whether I am grateful or ungrateful! And whoever is grateful, truly, his gratitude is for (the good of) his own self, and whoever is ungrateful, (he is ungrateful only for the loss of his own self). Certainly! My Lord is Rich (Free of all wants), Bountiful" [An-Naml: 40]

I am indebted to NUST College of Electrical and Mechanical Engineering, for providing me an opportunity for Masters Research. First and foremost I offer my sincerest gratitude to my thesis supervisor, Dr Usman Qamar, who has supported me throughout my thesis, with his patience and knowledge whilst allowing me the room to work in my own way. I attribute the level of my Masters degree to his encouragement and effort and without him this thesis, too, would not have been completed or written. One simply could not wish for a better or friendlier supervisor. I would also like to thank my family for the support they provided me through my entire life.

Dedicated
to

I dedicate my thesis research work to my family, A special feeling of gratitude to my loving parents, whose words of encouragement and push for tenacity ring in my ears. My brothers Abrar and Umer have never left my side and are very special.

I also give special thanks to my best friend Munazza Ansar, Asma Shaheen and Qurat-ul-Ain for being there for me throughout the reaearch phase. I will always appreciate all they have done .All of you have been my best cheerleaders.

I also dedicate this to Sir Syed Tahir Hussain Bukhari who has supported and encouraged me all the way.

ABSTRACT

Purpose: *In the last few years, with rapid growth in use of networking sites such as twitter and Face book has been increased greatly. This also attracted the researcher to use social networks data for sentiment analysis. Sentiment analysis is also known as opinion mining is the process of finding out the emotion such as positive, negative and neutral from the series of words. In present, on internet huge amount of data has been generated and to extract useful information from data is also become interest for the researchers. Sentiment analysis has been done mostly in English and Chinese languages. In this paper, sentiment classification is done on Urdu news tweets. The proposed methodology consists upon two steps. In first step data preprocessing is done such as removal of hash tag and removal of stop words is done. In second step feature vector is designed. The feature vector is formulated by through the identification of number of positive words, negative words, and presence of negation and use of POS tags. After formulation of feature vector the decision tree is used as classification algorithm. The decision tree classifies the tweet as positive, negative and neutral. The experimental result of the proposed methodology shows significant success in terms of accuracy and sentiment analysis.*

Methods*: This paper proposed sentence level sentiment analysis. This section presented the methodology of sentiment analysis of Urdu news tweets data. First the detail of dataset collection is providing. Next the annotation of data set is done with help of human annotators. Then data preprocessing carry out. Then feature vector is calculated by considering the more relevant features. Finally the tweets are classified into positive, negative and neutral using decision tree classifier*

Results: *In this research endeavor, we presented a summary of existing state of art for classification of Urdu news tweets. Sentiment classifications on Urdu tweets have been attempted in this research work. The impacts of feature vector and decision tree were analyzed to classify the tweets as positive, negative and neutral.*

The preprocessed form of training data along with feature vector was employed to the algorithm C45 which is used for decision tree.

Keywords: *Urdu sentiment, emotion, decision tree*

TABLE OF CONTENTS

List of Figures

List of Tables

CHAPTER 1
INTRODUCTION

Will, you must forgive me, but I have not the slightest sympathy with what the world calls Sentiment – not the slightest. —

Mark Twain, Letter to Will Bowen

Sentiment analysis is the method of determining the emotion at the back of the words. Opinion analysis is used to get the information about sentiment, opinions and understanding of the attitudes in an online medium. To monitor the social media, sentiment analysis allow to gains the opinions of the user on different topics. Sentiment analysis has wider and powerful applications. To extract the insight information from social media networking is a practice adopted by many organizations all around the world [1].

Sentiment analysis has been correlated with stock market. Sentiment analysis help to understand the customer reviews upon different items. On the other hand it is also true that sentiment analysis is not a perfect science because it is difficult to understand the Human language through machine. For a machine to analyze the grammatical sense, cultural difficulties, and to deal misspelling in online medium is a difficult task. Train the machine to understand the tone of writing material is also a difficult task. Let’s take an example by considering the following sentence:

“My flight‘s been delayed. Brilliant!”

For human it would be easy task to understand the meaning of the sentence. The sentence shows that the person was being ironic. Having a delayed flight for most of the people is not a good experience. This sentence would be negative by considering the contextual understanding. But without considering the contextual, the machine would interpret as positive sentence because of the use of word “Brilliant” in the sentence [2].

That’s completely a different task for the linguistic expert to teach the machine about the basic concept of sentiment analysis. Machine use dictionary for the job of sentiment analysis and dictionary evolves because the language evolves. The language will evolve faster with the use of social media networking. Twitter restrict tweets as 140 characters, this limit need the use of memes which alter the way of talking style. This change brings many challenges for the expert to handle the task of sentiment. Sentiment analysis is not used as 100 % accurately to determine to positive and negative content. It needs so many effort and check and balance by human to look upon the sentiment result produce by machine [50].

Sentiment extraction is performed in two steps:

1) At first level that is also known as lowest level, adjectives and adverb are use in finding the sentiment of sentences.
2) In second level contextual polarity is taken under consideration.
3) In next level the selected words are typically attach to the subject. This level is also known as high level.

Take an example of the sentence “The gouda was abysmal‖” in this sentence the word goudal is an entity. In this sentence negative sentiment is expressed about this entity.

1.1 Structured Text

Structured text name is taken form one of the database languages is known as “Structured Query language”. Structure text means in sequence form. Example of structure data is Excel. Textual files commonly used columns and rows. This type of data is commonly processed. Generally this type of data is easily accessed [47].

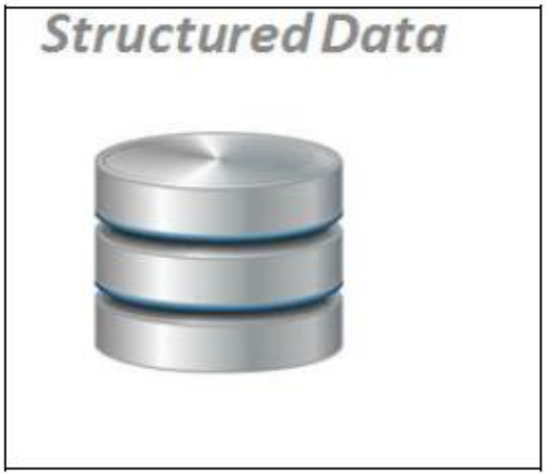


Figure 1.1: Structure Text [47]

1.2 Unstructured Text

Unstructured text has been found on World Wide Web. To retrieve the information from the web some techniques are used for this purpose. World Wide Web represents 80% of the data includes text and multimedia contents like videos, photos, and email etc. unstructured text handling is difficult as they not neatly adjust in database. Machine and human generate the unstructured text.

Figure 1.2: Unstructured Text [47]

It's time to create bridges for the generation of smooth integration of unstructured dataset into structure one.

Human generated unstructured data:

- ✓ Data generated from mobile include text as well as location
- ✓ Twitter, Face book, Linked In all are social media data
- ✓ Internal text of the company such as survey, logos, email and internal Documents etc

Machine generated Text includes:

- ✓ Information about weather through Satellite
- ✓ Oceanic information through Radar
- ✓ Collection of Scientific data that includes atmospheric image

From above discussion it is determine that main aim of opinion mining is to find out the attitude of the speaker or a writer about specific topic. The attitude of the speaker or writer may be positive or it may be negative depends upon his her judgment. Sentiment analysis is also a way of evaluation of specific product by collecting comments from the customers. Everyday a large amount data is created through social media network.

1.3 Online Media Networking

Social Media is providing Platforms to the people to spend more of time on social networks sites. We say that social media become Fundamental tool for the people all around the world to share thoughts and information. If it is used in positive way then it can be very efficient tool. It can be used in many different ways depends upon the user like companies used it to contacts with the customers, track the talents of the employee, measure the competition and find out the employees for the company. Some of the social sites are Twitter, Face book, MySpace etc these entire site are free of charges. Social media provide the facility to stay in touch with friends and family. Number of relationships progressively increases because of the large numbers of the connection of the people. Features of social media are music group, sharing photos and videos blogs and much more. It also facilitates to maintain the business through contacts with the customer and employees. LinkedIn is the best place to share the information about the business, share the professional experiences, meeting with the clients and professionals. Social interactions between people change by the first communication technology internet.   Adoption of internet has increase rapidly since 1990s. In 1990s the expert expected the internet to be trash heap to the history because the user increases rapidly. In 2006 internet is used 63% by American [48].

Internet becomes the fundamental part of our lives. It provides platforms for so many sites such as social networking for the people to stay connected. Through social sites people share their views about specific topics and also share photos and pictures. According to the survey of Amanda Lenhart 555 of the teenagers have personal profile and using the socials sites such as Facebook and Myspace.

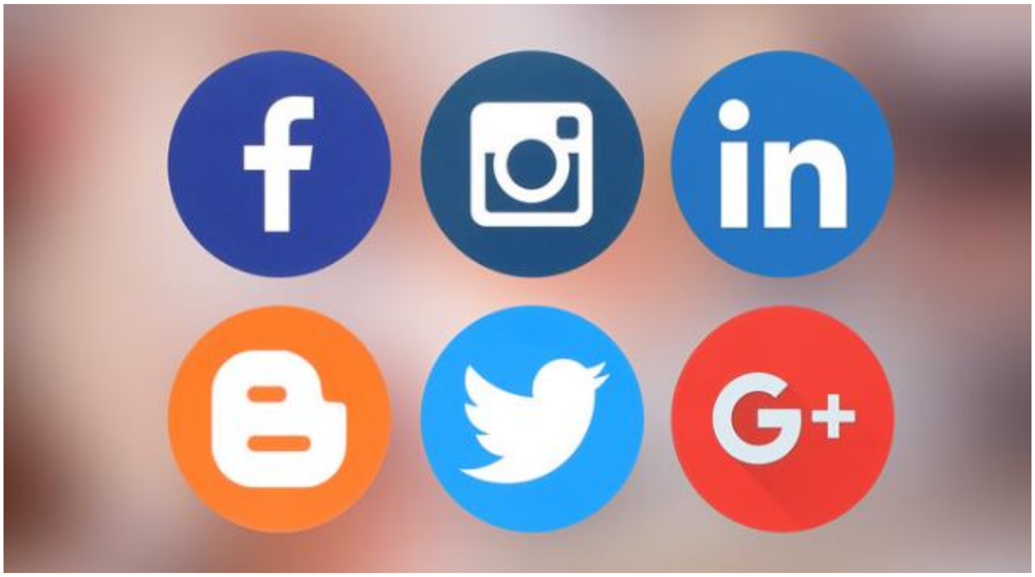


Figure 1.3: On line Social media [48]

Sixdegree.com was the first social site launched in 1997. The basic purpose of this site was to enable users to create the profile and make the online dating smoother. Other social sites were also launched after sixdegree.com but failed after some time. In 2002 Ryze.com and match.com were launched and used by large number of users. Too much use of micro blogging had a negative impact on the user by creating shortage of attention and separating the user from the real world. So use of micro blogging has advantages as well as disadvantages depends upon the how to use them. It also depends upon the user knowledge and interest of when and which site to use [48].

Some of the advantages of social networking sites are like it allows the people share their thoughts with their friends and whoever they want and also meet them. Some people used social media sites to connect with friends and family, even establish relationship to marry. Use of social media sites kept the user updated by giving information what is happening around the world. It makes the life easier and faster.

Educational institutes like Colleges and universities are becoming fond of social media networking sites. It facilitates not only students but faculty members to get connected, make group and share the lectures freely and easily. Recruiting of employees done with helps of social networking by technical sectors to go through the profiles of the employees. On the other hand social media networking also breaks the cultural values around some of the part of the world. Social media done modification in communication by enable different people to connect with their family and loved ones. It brings the world together. But the main disadvantages of social media sites are theft of personal identity.

User’s personal detail is requiring by the social networking to sign up and in order to access the site. Some of the social sites misuse the personal information of the user. Some time users’ privacy evade by advertisers. To find new victim’s criminals and sex offenders often visit the site. For the sake of revenge the young’s one post hates and misuses the photos that will negative impact on the user future. They types of criminals are called cyber criminals and sometimes even this lead to the death of young ones. The peoples addicted to the social networking sites. The developers made this site for the purpose of communication. Social sites delay the ability of the young people to face and meet the people. Social networking site obsolete the socializing.

1.4 Online Micro blogging

Online micro blogging is similar to the blogging. It is a broadcast medium. Micro blogging content is normally smaller in size as compare to the blogging. The major reason of the popularity of the micro blogging is that it allows the user to share very limited content such as short messages, images of individuals and links of the videos. This short limited content on micro blogging is known as micro posts. Micro blogging post is varying from the simple post. The post of micro blogging follow theme such as “most watched movie”. To promote the products and services the micro blogging is used for commercialization purpose.

Privacy setting is the popular feature offer by many platforms of micro blogging. The privacy setting feature facilitates the user to manage the post by showing the micro blog to the specific chosen readers. The post of micro blogging may include the messaging, text, video and Email. Slowly and gradually Micro blogging moves into the mainstream. In United States the election campaign for presidential election was started on twitter by Barrack Obama. Media organizations such as New York Times and BBC send headlines on Twitter. Traditional organization move towards the use of micro blogging [49].

Micro blogging advantages over traditional blogging:

Here we listed some of the reason to use micro blogging over traditional blogging.

- ✓ **Less time taken by developing the contents**

  The traditional blogs take time to complete the content as they are quiet lengthy. On the other hand micro blogging allows user to post short content which take less time to complete it. It also allow user to post news or recently happen incident in less time.

- ✓ **Increase frequently posts**

  Micro blogging is becoming popular among the user as it is a source of sharing information and also source of interaction between people. Most of the people use micro blogging on the mobile devices. Micro blogging proves to be the best choice to share the news in short as compare to the lengthy news that takes time. So short news posting increase the chance of frequent posts.

- ✓ **Share information in an easier way**

  Large numbers of platforms are available for micro blogging to share the post easily. Through these platforms user can easily share the information such as text, video and short messages easily to everyone. User can also share what is happening in their life or any news related to the event.

- ✓ **Direct communication between the followers**

  Micro blogging platforms facilitate the user for better communication with celebrity by liking the post, commenting on the post, tweeting and many more. Celebrity directly makes connection to the user by video call also.

- ✓ **Easy and convenient using with mobile phone**

The use of micro blogging increases day by day. The main reason of gaining attention is the increasing trend of mobile browsing among the users. It is difficult and time consuming to interact through long and lengthy posts where on the other hand micro blogging provide facility to play with smaller and faster posts [49].

1.5 Twitter: The Network

One of the online micro blogging services is twitter. Twitter allow user to read and write the post. The post must be short sentence containing 140 characters. The short sentence on twitter is known as Tweets. To use the service of twitter the user registers them on Twitter. Register users can post on the Twitter where unregistered users allow reading the post only. Twitter was introduced in 2006 by Noah Class and his team members. Twitter application can also be accessible through mobile applications. Monthly active user of Twitter was 310 millions in number. Most of the user of Twitter is employee from all around the world. About 80% of the twitter accounts are from outside the America. Twitter consists upon 3500 employees; it supports more than 40 languages, 35 offices around the world and more than 40% of the employees having technical background [51].

Twitter experienced rapid growth in tweets. About 4, 00, 000 tweets was posted in the year 2007. In 2008 10 millions tweets were posted. In 2010 about 60 millions tweets were posted on twitter.

Figure 1.4: Twitter App [51]

According to Compete.com Twitter has becomes third-largest online micro blogging service. Twitter gains popularity since January 2009. In June 2010, about 65 millions tweets were posted.

The main purpose to use the micro blogging and twitter is as follows:

- ✓ Huge numbers of messages were generated by twitter. The generated messages were increased day by day. Hence the extracted data was also in huge number.
- ✓ The scope of micro blogging increasing day by day. People find it easy to use and they can easily share the information and also give opinions on the concern topic.
- ✓ Twitter contains data from different languages as it is use by users from all over the world.
- ✓ The user of twitter has a large number of varieties such as the user can be film star, celebrity, sportsman and even politician and leaders. Many leaders across the world including Prime minister of Pakistan also use the twitter. So twitter contains posts of different languages, sex and religion also.
- ✓ Twitter provides high density of sentiment
- ✓ To extract tweets for twitter it provide user friendly API
- ✓ Twitter networking is popular as open access network

Figure 1.5: Networking on Twitter [51]

1.6 Sentiment Mining

The method of finding attitude of writer from text such as negative, positive and neutral is known as sentiment mining. For monitoring social media, sentiment analysis is helpful because it allows getting public opinion on wider topics. Its applications are dominant and wide. To extract the opinion from social data is adopted as a practice by many organizations.

A great number of online user interests in social media give motivation to the researcher to examine the emotion after study the text article. Emotional detection has been taken on the words, sentences or articles. Different technologies have been used to determine the emotions that help in many ways such as categorization of text document and recommendation of contextual music.

Large number of data on internet has been produces which help to find information from data present on internet. To finds a positive or negative opinion about specific subject then we used opinion mining. Negation also effects polarity of sentence and hence needs to be special concentration in sentiment analysis [1].sentiment analysis job is to find polarity of emotion expressed in a text [2]. Main aim of sentiment analysis is to find out the attitude of a speaker or writer according to the topic [2]. People use information technology to search out the opinion.

There are also some problems involved with sentiment analysis such as some words in a sentence are show positive sense and some show negative sense. Let's take an example of opinion such as “the mobile size is small” in this sentence the word small give a positive sentiment. Consider the second sentence “the battery time of this mobile is small” now here the word small show negative sentiment.

The text classification and processing depend upon the concept about a minor change in combine two sentences doesn’t create change in meaning [8]. Sentiment analyses differentiate the two sentences such as “the movie was good” is consider opposite as “the movie was not good”. In the above two sentences it is easy for the people to predict sentiment as positive or negative. However in social media medium like twitter, Face book requirement of context make difficulty for the people to understand the comment of someone as positive or negative. To deal different languages such as Chinese, Arabic, English in accordance with their morphological structure is one of the complex tasks in opinion mining [8].

To determine the people point of view on the social media posts then sentiment mining is done. Social media networks and social review sites are challenges for opinion mining as they are becoming popular [4]. The technique of extraction opinion from languages such as Chinese, Hindi, and Arabic is different from one another.

Mostly opinion mining is done in Chinese and English language. Presently a small effort has started towards the Urdu, which is the national language of Pakistan [3]. As an estimated 200 million people use to spoke Urdu language as their national language. It is a combination of many languages like Arabic, Persian, Turkish and Hindi also.

Researcher works on the extraction of sentiment from Chinese and English text. These techniques are not applicable on Asian based languages like Persian and Urdu. All of these Asian based languages have different script and morphological rules [6]. Morphology of Urdu is quiet such that its context is difficult to process. The alphabets (haroof) have many shapes and there is no indication for boundaries of words. Consider the two words (خوب صورت) that have single space as well as consider words (دستگیر) that have no single space. The segmentation of word is split up into two areas that are space insertion and space deletion [7].

1.7 Sentiment analysis Applications in Real world

Here are the few examples of sentiment analysis in real world applications:

1. **Reviews of product and services**

Many websites provided automated feedback for example Google product services.

1. Monitoring of reputation:

Reputation of Twitter and Face book is monitored to measure the reputation.

2. Prediction of Result:

The outcome of the event is predicted by analyzing the sentiment collected from different sources.

1.8 Background and motivation

Natural language processing (NLP) is one of the emerging fields for the processing of structure text and structure text. In late 1940 research on natural language processing was started and in 1954 language conversion is first time done that is Russian conversion into English. Computational linguistics, statistics analysis and language rules are applied for conversion of data. . Now days NLP is one of the area in which many research is carry out with help of different supervised and non-supervised techniques.

Sentiment mining is one of the important areas of research and many techniques have been proposed. Sentiment analysis aims to examine what people think and perceive. To determine the features of specific product or service, people usually showed interest opinion. On internet more than 80% data is UN structured. As social media data is also unstructured so there is need of natural language processing [9].Research has been done to find out ways and system in the field of sentiment analysis. Many tools and techniques have been proposed for sentiment analysis [9]. Descriptive data is mostly focus by researcher for sentiment analysis. Sentiment analysis is one of the popular research area in natural language processing.

Sentiment analysis has different level. The divisions are:

- Document level sentiment analysis
- Sentence level sentiment analysis
- Aspect based sentiment analysis
- Lexicon based sentiment analysis

Document level sentiment analysis is the simplest level of sentiment in which it is assumed the document contains one opinion on specific topic. For document sentiment analysis two types of techniques are used that is supervised and non-supervised techniques. In supervised learning the training data is trained using learning algorithm such as Naive bayes, KNN or SVM. Then new document is tested. Research showed high accuracy is achieved by new advancement such as POS tagging. In unsupervised learning semantic orientation of the phrase is determine in the specific document. [9] In sentence level it is assumed that single sentence may contain more than one opinion. Sentences are analyzed either they are objective or subjective. For classification supervised techniques are used [11].

Different strategies are used to handle different types of sentences such as sarcastic sentences, conditional and question sentences [12]. Aspect level sentiment deals with different opinion of people about entities. Different aspects of reviews are dedicated to the product categories. Aspect based sentiment is consider to focus on the determining the emotion within the document.

The classic approach noun phrase is used to determine the classification [14]. Tagged corpus technique is also used to determine the aspect based sentiment analysis and solved the problem of feature extraction [18] [13]. Sometime user does not give comparable opinion about a product. In this situation sentiment analysis is used to determine comparative opinions sentences. The goal is to extract the entities from each sentence. Relative small numbers of words are used to determine the comparative sentiment analysis such as superlative adjectives, comparative objectives and additional phrases. The result showed high recall but low precision [15].

1.9 Objective and Contribution

Efforts are being made to propose methodologies and techniques to detect and correct the sentiment from the online available data. Efforts are also being made to propose technologies to increase accuracy of detection especially for English language. So much work has been done in English language and many sentiment analysis tools are available.

In this thesis, we focused on proposing a methodology for detection of sentiment in Urdu news tweets. Our proposed methodology consists of three parts, collection of data set, preprocessing of tweets, and detection of positive and negative tweets. In preprocessing part, part of speech tagging is used to separate different parts of speech from text document. Back-end word list is used for detection of positive and negative words. Regular expression is used to match the words.

1.10 Outline

Chapter 2 discusses the work or research done on our research topic. Sentiment analysis approaches, method and architecture are discussed. Chapter 3 discusses the proposed methodology is detail. Techniques used during identification process of sentiment also described in detail. Chapter 4 discusses the results achieved by applying proposed methodology. In chapter 5 conclusion and future work is discussed.

CHAPTER 2

LITERATURE REVIEW

Recently there has been seen dramatic flow of interest in sentiment mining. Twitter data is mostly analyzed by the researchers for opinion mining. On the other hand as concern with languages, mostly the opinion mining is done on Arabic, Chinese and English languages. Urdu is mostly ignore by research community and very little amount of work has been done on Urdu.

2.1 Why Sentiment Analysis?

Sentiment analysis is also described as opinion mining. Opining mining is one of text oriented technique which addresses the problem of extracting, detecting and analyzing the text, determining the positive and negative opinions, and also determining that how an entity such as (people, product, organization etc.) is considered as positive or negative [1]. Sentiment analysis is an interesting research field as more and more people attracted towards the social media networks. This field provides potential for real world applications such as opining mining which help organizations, company or people to take better decision regarding to the product or services. The three main component of sentiment analysis are:

1) The holder of opinion
2) The target of opinion
3) The opinion itself [16]

2.1.1 Different level of Analysis

Remarkable research is done on word or phrase level of sentiment mining with different levels such as user level sentiment analysis, document level and as well as sentence level. In sentence level the sentence is inspected to find the direction of words to determine the opinion mining. The sentences are analyzed to determine sentence level opinion mining. The result shows the classification of sentiment as negative, neutral and positive.

While at document level the document is analyzed to find out the opinion of the whole document. In user level sentiment the comments of user on social media are analyzed to find out the user having same opinion on the topics.

2.1.1.1 Document level sentiment analysis

Whole document polarity is figure out by analyzing that whole document. Reviews of one product can be hold in one text file. The system estimates the score negative and positive comments for the product. Hence determine the opinion about the product. The opinion about Single product is determined by using the document level sentiment. The purpose and advantage of opinion mining at document is to find that overall opinion is getting about one entity. On the other hand the main drawback is that we cannot find out opinion about different entity [17].

2.1.1.2 Sentence level sentiment analysis

In this level of sentiment analysis the sentence is analyzed to find either the sentence shows positive, neutral or negative opinion. The Neutral opinion is determined as has no meaning. Sentence level sentiment analysis is closely related to the classification of subject and entity in the sentence. Subjective sentence express information related to the subjective point of views.

The main mission of sentence level opinion mining is to determine subjective and objective classification of the sentence [19].

2.1.1.3 Entity and Aspect Level Sentiment analysis

The object level sentient mining discovers the actual opinion of the people about the product. Entity level had done perfect analysis of the people point of views. However the sentence level and document level analysis does not analyze what actually people say about a product. Aspect level analyze the opinion rather than on the construction of the sentence. Aspect level consists on the target opinion. Considering the opinion target help to understand the problem related to the sentiment. For example “although the service is not that great, I still love this restaurant” has positive sentiment but this is not entirely positive. The sentence is negative about the service of the restaurant but give positive vibes about the restaurant.

The main theme of this analysis is to find out the entity level sentiments on the products. Take one more example “The phone’s quality is good, but its battery life is short”. The analysis of sentiment explains the negative and positive point of phone entity. Two aspects of the entity are evaluated such as quality of call and battery life. Call quality shows positive sentiment where battery life shows negative sentiment. Thus entity based sentiment convert unstructured text into structured ones. The obtained information is used to analyze quantitative and qualitative attributes [19].

2.2 Sentiment analysis Approaches

The way of finding out classification of opinion as positive or negative is known as sentiment analysis and opinion mining. Labeling of sentiment word manually is considered as time consuming process. Two popular techniques are used to obtain the sentiment analysis automatically.

1) Weight words are used as lexicon

2) Machine learning approach to automate sentiment lexicon

3) Linguistic based approach

In lexical approach used list of words or corpus to determine the polarity of the words [20]. Machine learning approach determines the sentiment by training the data set using algorithm and done classification using features and done testing to check accuracy [21]. Linguistic approach is used to check the direction of the text using semantic approach like negation, anatomy and phrases of words [22].

2.2.1 Lexicon Based Sentiment analysis

The lexicon based approach utilize dictionary of words with associated opinion words. This method match words with the dictionary to find out the value of scoring. Lexicon technique does not require any dataset to be train. While in machine learning method data is pre process or train.

The sentiment of the sentence and document is determines by combining the polarities of the single word present in the document. Combining the score it delivers sentiment of the whole sentence and document. Predefined list of words are utilized by this approach and each word has a sentiment in the list.

Lexicon based approach used following types of methods:

1. **Dictionary Based methodology**

   Positive opinion words and negative opinion words are identified by utilizing lexicon based dictionary such as Word net dictionary.

2. **Corpus Based methodology**

   Large amount of words are collected. Synthetic pattern based words are collected and words related to other opinion are found within the context.

The methodology of Lexicon based sentiment analysis is shown in figure 1.

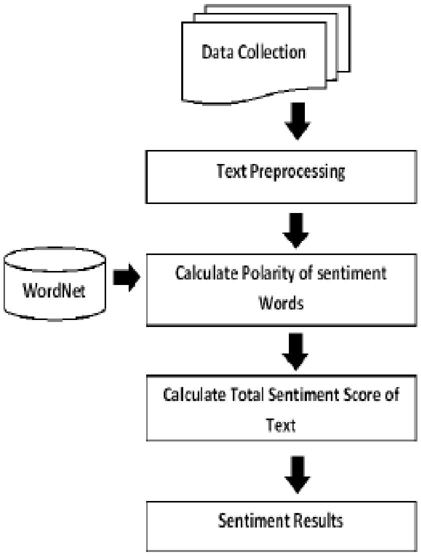

Data Collection
Text Preprocessing
WordNet
Calculate Polarity of sentiment Words
Calculate Total Sentiment Score of Text
Sentiment Results

Figure 2.1: Lexicon based sentiment analysis [23]

2.2.2 Comments-oriented Sentiment system

In this study focuses is done on the news comments. Comments independent evaluation is analyzed with multifocal features extraction. Comment oriented features extraction provides challenge for the polarity classification. To make the analysis more reliable and accurate the inclusive comments are filtered out. Opinion focus detection algorithm automatically detects the entire implicit algorithm. A taxonomy lexicon is build that evaluated the focus of each comment. As a result polarity and strength is obtained through the algorithm. Focus classification is obtained through opinion detection algorithm [23].

To prevent noisy data the spam comments should also be detected and removed from the data. 250 news items are used to build such lexicon. The lexicon contain structured linguistic dictionary. Taxonomy structure of features and objects were used that presents opinion in the news. The focus detection and sentiment analysis module are used in the system. Both modules use Lexicon approach for the analysis of the comments. The modules were comparing with the sentiment analysis specific modules. The focus was to address the issue of anaphora and abbreviation [23].

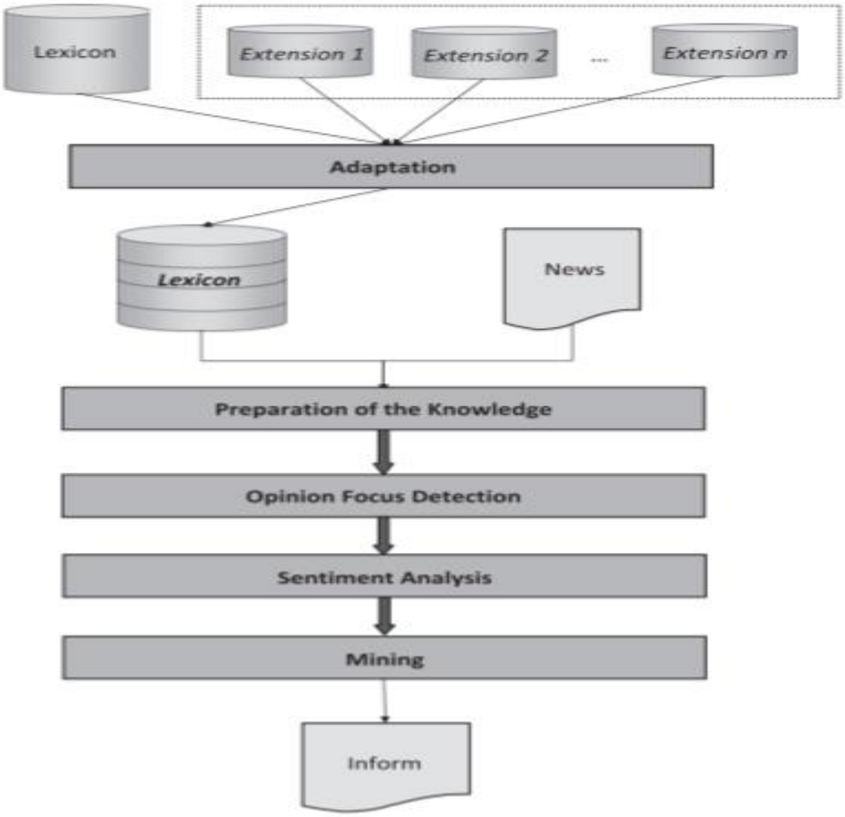


Figure 2.2: Comment oriented Sentiment Analysis algorithm [23]

2.2.3 Domain specific sentiment lexicon

Domain specific problems are address using sentiment analysis lexicons. Annotated document is used to extract the sentiment lexicons at document level. Bayesian estimation based algorithm is present. Document level annotation is turn by the algorithm that is calculated at the word level annotation. For annotation the problem of annotation is address at document level to minimize the efforts. To train the Bayesian estimator labeled documents are required. Labeling all the documents in this procedure require a lot of time. An active learning approach at the document level is introduced that will reduce inefficient labeling efforts. Active learning approach is a closed loop that point outs the important data which influence the decision boundary. Sampling the document is done through the active learner.

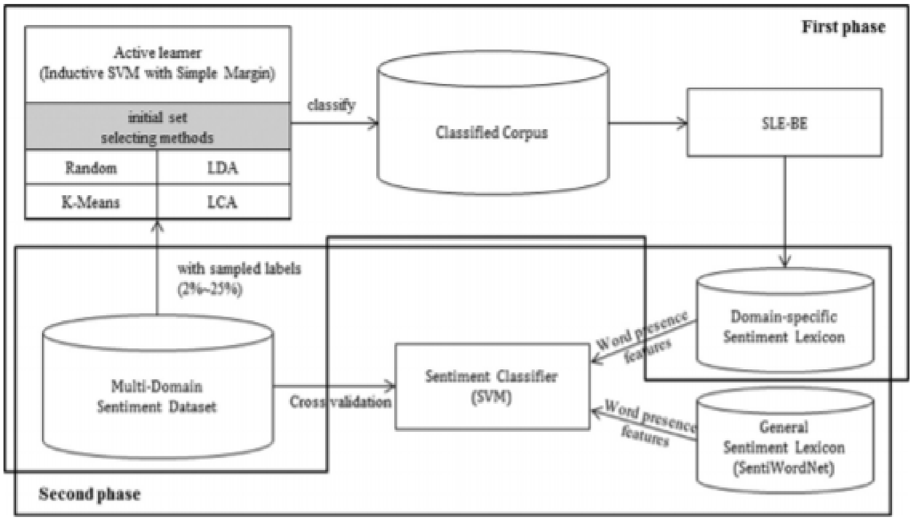


Figure 2.1: Domain specific sentiment architecture [24]

Active learning technique determines the boundary by identifying the most important data point. Hence find out the most important document by determining the classification boundary. Active learning provides remarkable result regarding F1 score. The active learning and proposed algorithm are used only for N-gram. Actually active learning provides better feature selection in classification of sentiment [24].

2.2.4 Domain-Specific Social Media Texts

Corpus-based lexicon method is used that extract sentiment word that are based on the language and content domain. Three language resources are applied by this method such as dictionary, use of seed lexicon and use of corpus. Use of corpus indicates the important domain and language specific sentiment. The developing corpus must be relevant to the language and content domains. This methodology focuses on the generation of new sentiment lexicon. The purpose of seed sentiment lexicon creation is to determine the polarity of the new words after establishing the relation between candidate and the seed in the lexicon. To eliminate the noise from the lexicon the dictionary is used. Dictionary not only removes the noise but also correct the spelling mistakes in social media contents. The methodology contains two phases these are recognition of the sentiment and extraction of the sentiment.

a) Developing corpus is used to extract the candidates
b) Seed lexicon is determine to find the relation between sentiment words and the candidate
c) Removal of irrelevant word from the seed is done
d) Repeat the step 2 and 3 until no repetition of word is done [25]

2.2.5 Micro blog-specific Chinese sentiment lexicon

An effective word detection method is introduced that improves the coverage of the sentiment lexicons by detect the popular words used by user. This method not only covers the distribution of words over messages but also cover the word distribution over users. New user words are detected and automatically add in the dictionary. The detection of new words redefines the previous words in subsequent iteration. As new words are automatically detected hence the performance of word segmentation has improved. To develop high quality sentiment lexicon a unified framework is introduced which comprises three types of sentiment knowledge frames [26].

1) Word sentiment knowledge calculates score of the words and emotions.
2) Sentiment similarity knowledge represents word similarities from all messages
3) Existing sentiment lexicons are determine to extract prior knowledge

Dataset contain more than 17 million messages. Sentiment lexicon can outperform in classification of sentiment polarity, subject detection at both messages level and sentence level [26].

2.2.6 Multi-lingual support for lexicon-based sentiment

Most of the sentiment lexicon deals with only one specific language that is English. However, growing amount in data on web demands multi-lingual system for determining sentiment. Text is automatically translated into the reference language. Then the translated text is analyzed. Mapping technique is proposed which map the reference sentiment for the target language sentiment. In this way words meaning are analyzed rather than the analysation of the whole lexicon. In order to generate language specific sentiment lexicon an approach is introduce which propagate sentiment from the set of words from each separate language sentiment lexicon. Sentiment lexicons are consequently used in determining the sentiment analysis techniques. The main impact of the technique is the use of mapping sentiment technique. Mapping technique constructs relationship between reference language and the target language. The result and effectiveness of mapping method is comparing with an existing translation method [27].

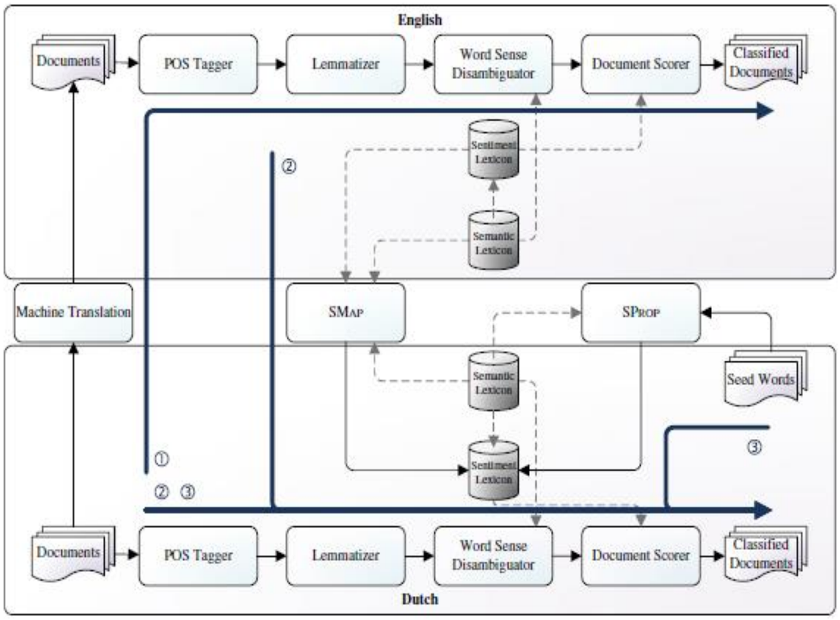


Figure 2.2: Multi lingual architecture [27]

The above framework supports two languages such as English as well as Dutch. In first step the messages in Dutch language are translated into English language. Then approach involves determining the sentiment analysis of the Dutch language. The proposed technique well performed in order to classify the sentiment as positive and negative [27].

***N*-gram lexicon**

The performance of products can be evaluates with the help of sentiment analysis by reviewing the comments of the user of the product. When there is a large amount of data to be handling then lexicon based sentiment approaches are used rather than machine learning because of difficulty in training the large amount of data. Most of the lexicons approach used sentiment scores along with the unigrams. Sentiment n-grams showed improvement in result when combing the features of negations and intensifiers. Sentiment n-gram with such that features is not publicly available.

Different methods are used to calculate the lexicon such as automatically creation of sentiment, manually creation and semi manually creation of sentiment. Manually creating a lexicon can be expensive in terms of both time and cost.

To handle large amount of data which contain numeric and textual data an automatic sentiment analysis is required. Automatic Senti n-gram calculates score by considering lists of negations and semantic unigrams. The extracted n-grams scores are compare with that of manually annotators by human to verify statistically equivalent [28].

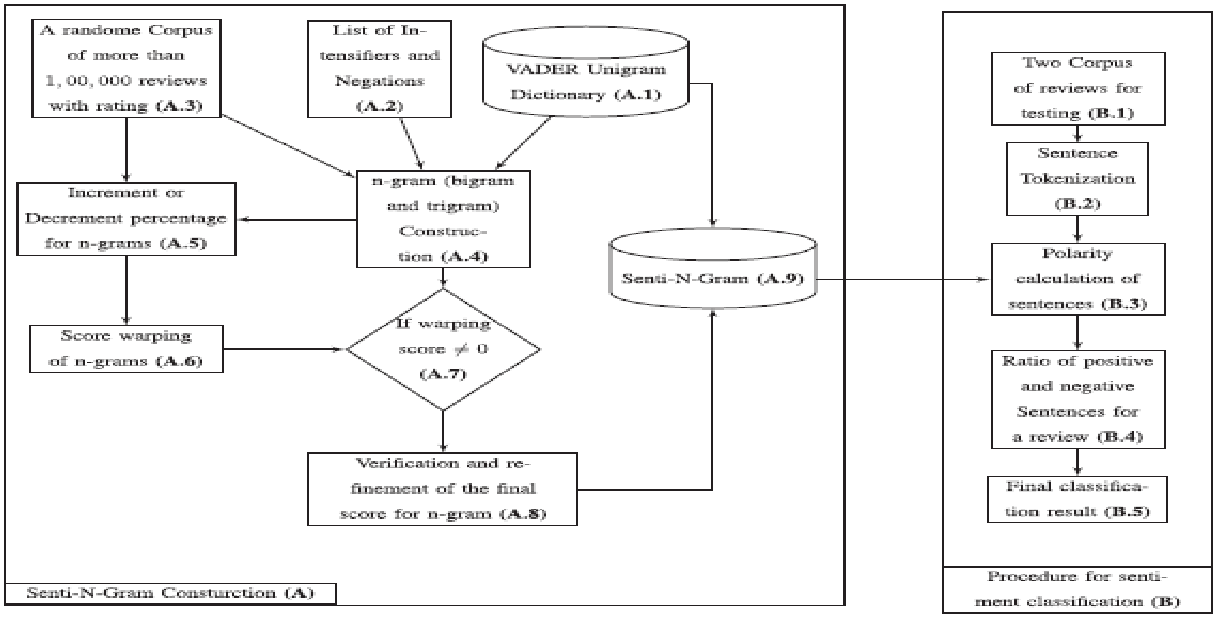

A randome Corpus of more than 1, 00, 000 reviews with rating (A.3)
List of Intensifiers and Negations (A.2)
VADER Unigram Dictionary (A.1)
Increment or Decrement percentage for n-grams (A.5)
n-gram (bigram and trigram) Construction (A.4)
Senti-N-Gram (A.9)
Score warping of n-grams (A.6)
If warping score ≠ 0 (A.7)
Verification and refinement of the final score for n-gram (A.8)
Senti-N-Gram Consturction (A)
Two Corpus of reviews for testing (B.1)
Sentence Tokenization (B.2)
Polarity calculation of sentences (B.3)
Ratio of positive and negative Sentences for a review (B.4)
Final classification result (B.5)
Procedure for sentiment classification (B)

Figure 2. 3: Framework of Senti-N-Gram [28]

Bigram or trigram scores are adjusted by alteration of the original Uni-gram score. This is done with the help of increment and decrement in the respective value. Bigram and trigram scores are range from “-5 to +5”.

The score of bigrams and trigrams are refining by eliminating all zero score of bigrams and trigrams. The score of final n-grams are verifying by following example: Score of “very good” should be higher as compare to “good” because word “very” amplifies well [28].

2.2.7 A Domain Transferable Lexicon for Twitter Sentiment

For the customer to express sentiment twitter messaging service become famous day by day. Capture the comments of customers accurately has become a challenge for the analyzer. Machine learning and traditional dictionary based approach are used for twitter sentiment analysis. Feature selection is the major challenge for machine learning technique. This specific challenge of feature selection is address and proposed novel technique for sentiment analysis of twitter data.

Lexicon based twitters specific data set is very small in n number. Twitter specific data set is domain transferable. Twitter specific data set utilization can improve result in terms of accuracy. In order to increase the feature matrix density and to solve the problem twitter specific set improve the efficiency. Positive, negative and neutral are the three classes used by most of twitter specific techniques. This study uses five classes for the classification of the twitter data by adding very negative and very positive. One of the most important attribute of domain transferable is also consider. Hierarchical approach was used for feature generation. Multi phases are shown in hierarchical approach. Feature grouping and Meta feature attributes are used for feature selection. This approach for sentiment analysis show increase in accuracy for accounting tweets.

2.2.8 Multi-tier Sentiment Analysis

One of the important areas of data mining is text mining. Text data are converted into numerical data sets in order to create the link with the database for the sake of identifying similarities. To present relative information and to presents hidden information, text mining plays an important role. Take out the sentiment from the text such as message or comment from the post is the way of gaining information. Sentiment analysis is done to extract sentiment on the product through customer feedback. Text mining is helpful in determining the useful information in order to increase the business share in the market. Sentiment analysis provides more accurate feedback of the customer on different products.

This study proposed the multi-tier classification system for the purpose of sentiment classification. Well known machine learning algorithms such as (Random forest, Naive Bayes, SVM and SGD) are used to implement the multi-tier system. Classification is done into five levels such as “negative, very negative, positive, very positive and neutral”.

There are seven levels of multi-tier classification system. These are as follows

a) Collection of Data set

b) Preprocessing of Data

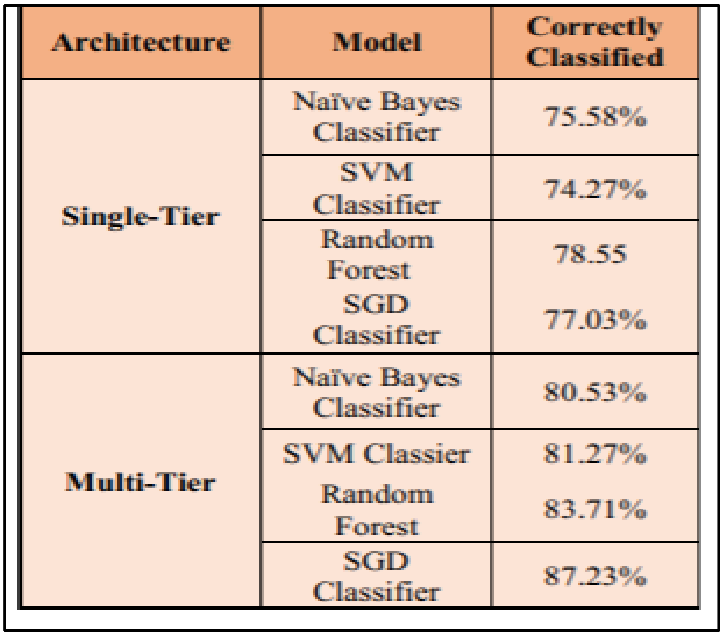
c) Dataset Training

d) Selection of features

e) Training the classifiers with training Dataset

f) Testing the Data

g) Accuracy measurements



Architecture	Model	Correctly Classified
Single-Tier	Naïve Bayes Classifier	75.58%
	SVM Classifier	74.27%
	Random Forest	78.55
	SGD Classifier	77.03%
Multi-Tier	Naïve Bayes Classifier	80.53%
	SVM Classier	81.27%
	Random Forest	83.71%
	SGD Classifier	87.23%

Figure 2.4: Single and multi tier architecture [29]

The above table show that accuracy improved in terms of number of increasing the trees. This increase has a limit and stop at a certain point and not showing any increase even when the number of trees increases. From above table it is showed that multi-tier architecture showed improved result in term of accuracy as compare to single tier. SGD classifier provided the best accuracy result in the proposed multitier architecture [29].

2.2.9 Ontology and SVM Classifier

Most of the Sentiment analysis techniques involved the lexicon based methodology by checking the similar words from the dictionary or list of seed words. Lexicon based technique is not efficient in terms of classification of sentiment analysis. Therefore Khin proposed the technique that is the combination of ontology and supervised machine learning techniques. Opinions and features are extracted from the comments to enhance the task of classification. Feature level sentiment classification is used in the propose method. The approach is divided into three main tasks that were POS tagging, extraction of domain related features and classification of sentiment. POS tagger labels the words as noun, verb, adjective etc. Brill tagger is used for POS tagging. This tagger not only tags the list of word but also sense of word tagging. The list of tagger is used to extract the features. To extract the domain related features, domain specific ontology is used in the proposed methodology. The main purpose to use ontology is to give domain specific information to the developers for the sake of information exchange. Existing hierarchical based taxonomy methodology are not enough for retrieval of information. For the formation of semantic structures and data analysis Formal concept analysis method is used to determine the conceptual structures in data sets and to identify the dependencies among the data set. Sentiment classification is done with help of linear SVM [30].

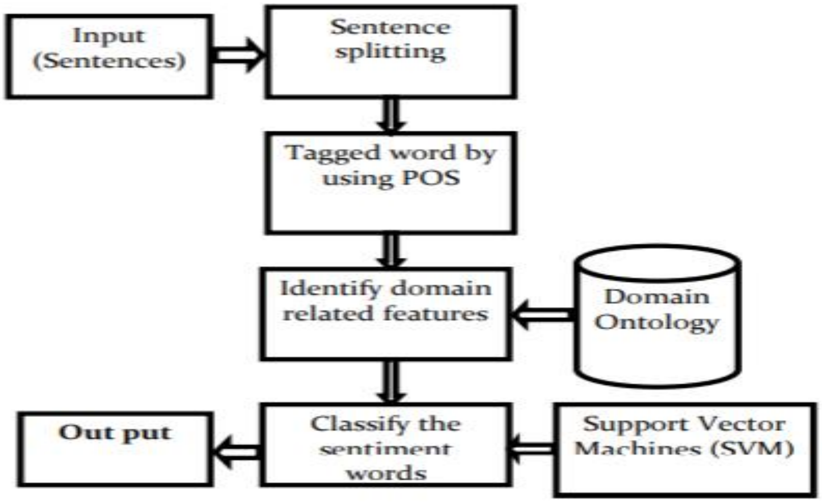

Input (Sentences)
Sentence splitting
Tagged word by using POS
Identify domain related features
Domain Ontology
Classify the sentiment words
Support Vector Machines (SVM)
Out put

Figure 2. 5: Ontology classifier [30]

2.2.10 Hotel Reviews

E-Commerce has got into as part of our life with the availability of high speed internet. Information on internet is available in unstructured format. Through internet E-Commerce has spread so widely to the huge extent. It is difficult for the customer to make decision on the basis of the product review. To make the decision customers read all the comments and finally analyze to make decision. Text mining plays important role in determining the decision. Text classification consists upon testing and training data set. Training data set contain features attributes, target value and class labels. Support vector machine check the target value in the testing data set. To categorize traditional data set, SVM shown to be produced highly effective result.

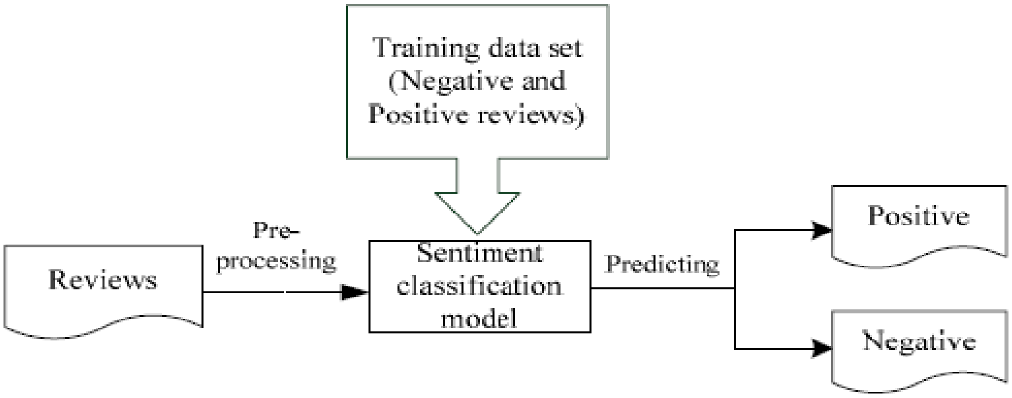

Training data set (Negative and Positive reviews)
Reviews
Pre-processing
Sentiment classification model
Predicting
Positive
Negative

Figure 2. 6: Supervised machine learning algorithm flow chart [31]

The data set contains 4000 hotel reviews. The data set contain half positive and half negative reviews. Data set divides into four equal sized folds. Uni gram based features are extracted from the data set. Recall, precision and F-score are used to measure the performance of the system. Result produce by TF-IDF is more effective as compare to frequency [31].

2.3 Supervised Learning Based Approach

One of the popular concepts in machine leaning is Aspect base sentiment analysis. Aspect base analysis work on the principle of determining sentiment analysis using training data set to provides neutral, positive and negative comments of the user. Existing techniques used word level analysis. Aspect of the entities is identified in Aspect Based Sentiment analysis. The main aim is to provide summaries by listing all the overall polarity. Aspect based sentiment analysis used machine learning techniques. For training purpose support vector machine algorithm have been used. The proposed Aspect based analysis is consisting upon 3 parts. These areas follow:

a) Pre processing of Dataset

b) Feature extraction according to Domain

c) Generation of Dictionary

The above model extract the meaning of the enter word. Mostly language chunks are used as input data for the sake of sentiment analysis. Simple filters are used to pre processing the data. Punctuation marks, white space and word boundaries are removed and that will not ultimately affect the features o sentences [32].

2.3.1 Extraction of Features

Conversion of language sentences into features is the basic task of machine learning algorithm. Sentiment dictionary and domain specific feature are the two modules of feature extraction. Feature extraction module extract features from the available set of data. The process of feature extraction consists upon three main parts:

1. Lexical Feature Extraction
2. Domain Specific Feature Extraction
3. Sentiment Analyzer

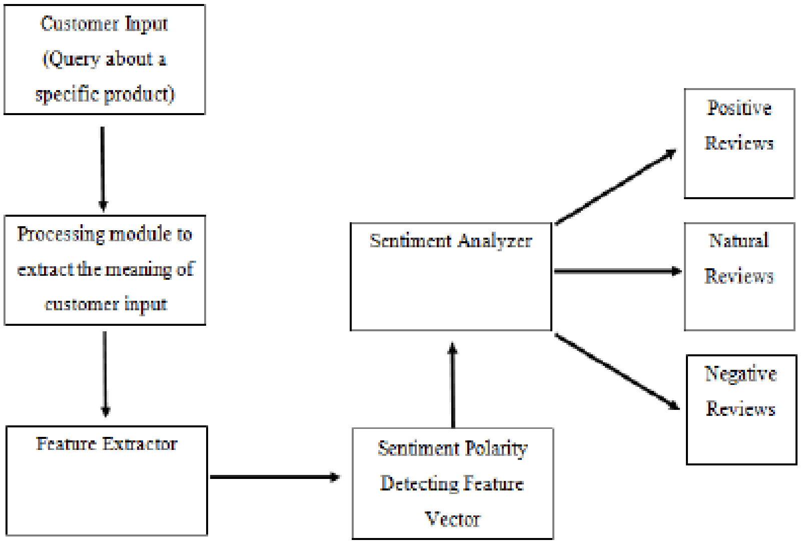

Customer Input (Query about a specific product)
Processing module to extract the meaning of customer input
Feature Extractor
Sentiment Polarity Detecting Feature Vector
Sentiment Analyzer
Positive Reviews
Natural Reviews
Negative Reviews

Figure 2. 7: Feature extraction flow chart [32]

The above diagram shows the architecture of the system. Domain specific features help out in determining the category of the product. Domain specific feature extraction help out to extract most relevant features about the product. The above figure shows that how domain specific analyzer converted un-labeled data set into domain specific clusters of words [32].

2.3.2 Polarity detection of Sentiment

The objective of the sentiment polarity detection is to detect the sentiment expressed towards a given aspect category in a given sentence. Given a set of pre identified aspect categories for a sentence, this phase aims at determining the polarity (positive, negative, or neutral) of each aspect category. Instead of multiple classifiers, such as implemented for the aspect category detection phase, but this research used to train a single classifier, to choose between positive, negative or neutral polarity.

The aim of sentiment analyzer is to determine the expressed sentiment in the respective category. The labels such as (positive, negative, neutral) are the output of the sentiment analyzer. Sentence is given to the feature extractor to convert the sentence into feature vector and classified sentiment as positive, negative and neutral. Machine learning algorithms are used for classification purposes.

Aspect based sentiment analysis provide analysis beyond the limit of word level and also provide semantic analysis of the input text through help of semantic networks. This method of sentiment analysis gives power to the other sentiment approaches. Sentiment expressed for each aspect is determined through aspect based sentiment analyzer. The ultimately aim is to provide list of similar aspect and the polarity. This ultimately effects on the result by providing better accuracy.

Aspect based sentiment analysis also improve the methodology of sentiment classification. For features extraction SVM, lexical dictionary and semantic features are used as a machine learning tool. Features such as POS tagging, unigrams, bigrams words cluster are used for classification.

Sentiment polarity classification is done with the help of multi-class classification system. Aspect based sentiment analysis focuses on SVM based algorithm for the analysis of system [33].

2.3.3 Critic learning through neural networks

Sentiment analysis helps in understanding the attitude of the people, emotions and attitude from large amount of data. In natural language processing sentiment analysis become most interesting research due to the rapid increment in data available on social media networking.

Through sentiment analysis textual features problem can be solved. For solving sentiment classification machine learning approaches are adopted and utilized. More ever deep earning techniques have been used to solve sentiment analysis problems. These approaches provide better result in terms of accuracy. Machine learning techniques easier the task of labeling the huge data set. On the other hand labeling data set by manually is difficult in terms of time consuming. Double negative structure and transition control are used as knowledge structures in sentiment analysis.

Studies have been conducted on sentiment analysis to address the issues in sentiment analysis. Deep learning approaches incorporate with the extraction of features from the huge data set. These techniques no doubt showed remarkable improvements but how ever suffer from two problems. Some time the prior knowledge given to the algorithm may be incorrect.

On the other hand these methods cannot have a scheme for judgment and cannot adopting the information. Second the addressed problem is the usage of simple rules such as clause and transition. Intelligent structure knowledge is not considered. Critic learning based deep neutral network architecture is proposed to address the drawback in finding classification of sentiment. The critic architecture consists upon two branches.

1. Branch one extract and count the number of features

2. Train the architect with given rules of information

Critic model adjust the text features by embedded within two branches. To utilize more knowledge such as structures of the text then there is need of determining more filters in the model. The contribution of research is as follows:

1. Critic framework is proposed which help in giving prior rule based knowledge

2. Sophisticated knowledge based rules are formed [34].

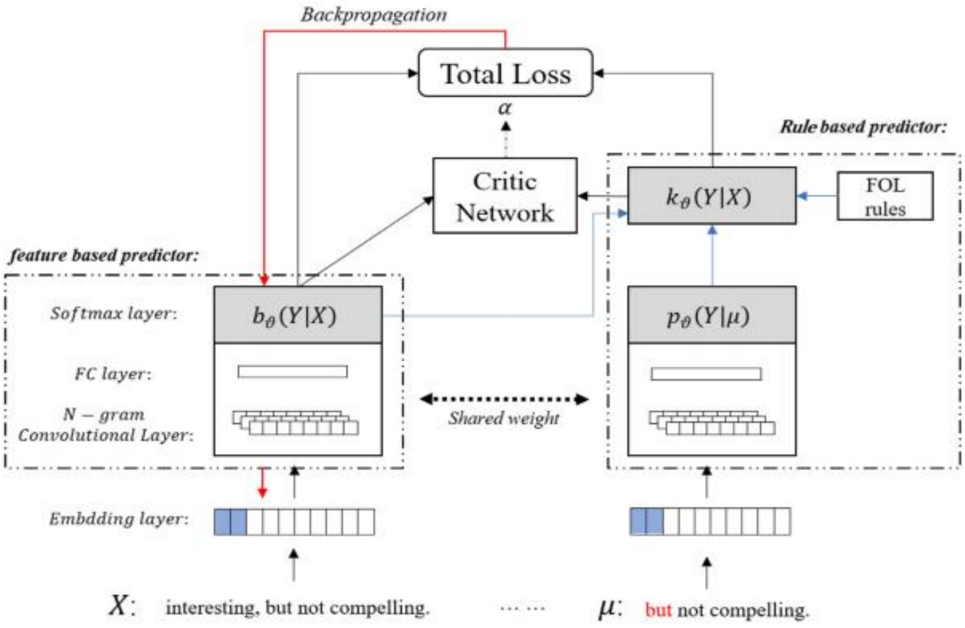


Figure 2. 8: Critical learning architecture [34]

Feature based predictor consist upon four layers. Embedding layer obtain embedding the vector for the words. N gram layer extracts the n-gram features from the data set with the help of different filters. We obtain k gram features by sliding the filters over the data. Max poling operation is applied on the features so that k gram features are extracted. In this way we obtain set of clusters to produce clusters. Each cluster is composed of embedding vectors.

Max pooling operator result is transform by FC appended layer. After conversion of the result then the soft layer is adopted to convert the connected layer into normalize distribution of probability [34]. The output is expressed as:

$$\boldsymbol{b\,\vartheta(y\ i \mid x\ i)}$$

2.4 Sentiment analysis for Urdu text data

People like to express emotions and present opinion in their own language. Urdu websites are becoming popular just like English Arabic and many other languages. Sentiment analyzer for other famous language like English, Chinese etc are not supportive for Urdu language. Sentiment analysis designed for English is not applicable for Urdu because of semantic and structural difference. Urdu language has complex morphological structure. With the passage of time Urdu online resource are becoming popular day by day. Developing Urdu lexicon is also challenge for the researchers.

In Urdu language handling of space is difficult task. The problem is that usage of space is not constant and that cause problem of space insertion or space omission problems. For example in Urdu word “انکا " is a combination of cardinal words but due to space omission problem this word is consider as individual word rather than two individuals words. Take another example of Urdu word “عقلمند “is handled as combination of two words after tokenization such as “عقل “and “ مند ” but in real the word “عقلمند “ is single word but due to space insertion problem this is consider as two words. In Asia Urdu language is one of the most spoken and understanding language. Urdu language is similar with Arabic in terms of script, and with respect to morphological structure it is similar with the Hindi language, despite all Urdu has its own different requirement as far as concerned with NLP. For handling of Urdu language there must be introduced system.

2.4.1 SentiUnits for Urdu Text classification

In [35] proposed a technique to handle Urdu is divided into two steps. These are annotation of Urdu text and then classify into sentiment classification. Polarity showed the orientation and intensity is showed by calculating the modifiers. The construction for lexicon can be divided into following parts:

1. Sentiment word phrases identification

2. Inflection or derivation identification

3. Grammatical rules identification

4. Identify intensities of modifiers

5. Multiple POS tags identification

6. Lexicon construction after Preprocessing

Senti Units are given polarity score after compression down with dictionary of words of classification. Annotation of SentiUnits should be done for better result.

2.4.2 Phrase level negation handling

In sentiment classification, negation plays an important role as the valence shifter to extract the sentences from the data set.

This paper [36] proposed the methodology of handling phrase level negation for the sentiment analysis of Urdu. Annotated lexicon based approach was used. Subjective phrases focused by this research known as SentiUnits. The SentiUnits include negation particles and adjective as the core items. The normalized text is tagged with POS tagger. Adjective phrase are chunk-ed out by shallow parsing and converted into SentiUnits. SentiUnits are analyzed for assigning the polarities calculations. The assigning polarity is treated with the combining effect the presence of negation in SentiUnits. Result of study showed that presence of negation has a high impact on miss classification. Implicit negation still needs an improvement.

2.4.3 Roman Urdu treatment for sentiment analysis

Normally users use native language for commenting. Native language is understandable for those people who speak that language. Comments in native language have very limit importance. It is useful if comments in native language make understandable for the people who only understand English language. The main aim of this search is to help the non native customer to get rating and information of different products from comments in Roman Urdu. Roman Urdu is a term used for the Urdu language written in Roman script.

Romangari is a word mostly used Hindi language. For posting the comments, people of India and Pakistan use Roman Urdu for this purpose. Somehow Hindi and Urdu are same languages with minor difference in writing script. Roman Urdu is a term used for the Urdu language written in Roman script. Similarly Romanagari is portmanteau of two words Romanand Devanagari. Romanagari is a term used for Hindi language written in Roman script. In Pakistan and India people normally use Roman Urdu and/or Romanagari for posting the comments. Urdu and Hindi are same languages however the difference between them is writing script and influence of other language . Urdu is influenced by Persian, Arabic and Turkish. Its writing style is in Arabic while Hindi is written in Devanagari script [4]. However both languages are much more similar and can be understandable by the people living in subcontinent (if it is written in Roman script). Let for example take a phrase posted in Roman Urdu “Iss mobile ka camera acha ha”. Its English translation will be like “The camera of this mobile is good”. The same phrase “Iss mobile ka camera acha ha” is fully understandable in Hindi. Our focus in this paper will be Roman Urdu [37].

2.5 Summary

After going through the studied literature, we analyze different strategies for the sentiment analysis of the text data. Systems are developed to conduct the sentiment upon English, Chinese and Hindi as well. Different systems are developed to identify positive and negative tweets. However no satisfactory research was done on Urdu language. Even Urdu language NLP and POS tagger are not as much develops as English language NLP and POS tagger tools exist.

Chapter 3
Methodology

"Research methodology is the systematic, theoretical analysis of the procedures applied to a field of study. Methodology involves procedures of describing, explaining and predicting phenomena so as to solve a problem; it is the 'how'; the process, or techniques of conducting research."(Kothari, 2004)

4.1 Problem statement

Simply stated, the problem explored in this thesis is the classification of Urdu news tweets by training Decision tree for the Identification sentiment as positive, negative and neutral. The classification of Urdu news tweet can be useful in many areas, especially for the people who speak Urdu.

For English language, sentiment analysis is taken out from last three decades. Most of the research related to sentiment analysis is done in English and Chinese languages. On the other side morphological rich languages such as Arabic, Turkish, and Urdu etc are ignored by the research community. Due to morphological operations, word level complexity is very high in these languages.

Urdu language has many different features which make it challenge for the researcher to done classification on it. Some of the challenges for Urdu language are lexicon intricacy, handling phrase level negation, dealing morphological complexity. These challenges are not address by the researcher and for this reason sentiment analysis is mostly done on English language. Urdu is the popular language of Asia and has a close relation with Indo-Aryan language. Urdu language contains compound characters to form the character shaped. The word meaning is depends on their position in the sentence. Urdu shares similarities among many languages such as it share script similarities with Persian and Arabic and has morphological similarities with Hindi language. Sharing similarities with other languages, even it has its own linguistic features. Techniques available for other languages are not applicable for Urdu language.

This thesis will address these challenges by exploring several variations of architecture selection, model search, regularization, and training paradigms within a deep learning context, with the aim of harnessing the expressive power of deep learning while avoiding over fitting.

4.1.1 Urdu language features

NLP community faces challenge to handle the Urdu languages. Here are some features which make it distinct from other language.

1) Morphological complexity of Urdu

Urdu is one of the morphological complex languages just like Arabic, Persian and Turkish language etc. For NLP community is quiet challenging for natural language processing and machine translation of morphological complex languages. Urdu possesses difficulty at word level such as reduplication, derivation and generation of multiple words from the root words. One of the challenge face by mostly is multiple root word generation for example the Urdu word “ علم ” is the root of the words “ معلومات” “ معلوم”, “ معلمہ”, “ معلم” معلمہ”, “ معلم”, “معلم”, “ عالمہ” , “ عالم ”. The Urdu language is blend of many languages that’s why it has influence of many other languages like English, Arabic, Persian, Turkish and Sanskrit.

2) Lexicon Intricacy of Urdu language:

Lexicon of the language is depends upon the in heredity from other languages that also help in making the vocabulary of particular language. As already discussed that Urdu language has blend features of many other language such as **“پوجا”** (worship, from Hindi), **“بدترین”** (worst, from English), “قمر” (moon, from Arabic).

3) *Context sensitivity of the Urdu script:*

Urdu language script has context sensitive. The alphabets of Urdu are considered as joiner and non joiners. The joiner job is to join the alphabet in different shape. This connectivity issue joiner raises the issue of word segmentation. Spaces in the word are not real boundaries of Urdu language.

4) Independent case markers of Urdu language:

The case marks play important role in grammatical functions. In other languages the case marks are attach through derivation and inflection. In Urdu case marks are treated independently because they are not attached with lexicon and depend upon lexicon. Case marks in Urdu have influence on the structure of the sentence. This issue also causes some grammatical issues and ambiguities in the sentence. Let’s take the example for the phrase “نام کے رنگوں” and “کے رنگوں نام” they both have same meaning but have different word order because Urdu text is independent case mark. The different words order is due the different word order [37].

Urdu language requires different set of methodology for sentiment analysis. The reason for different strategy is that it has different features such as complexity in morphology, complex grammatical rules and most important independent case marks. These issues demand sensitive type of methodology for the treatment of Urdu text.

4.2 Existing System Architecture for sentiment analysis of Urdu

The main aim of existing system for sentiment analysis and opinion mining is the use of lexicon based methodology and use of automated tools to determine the information of the subject such as attitude, behavior, opinions and the expression of the feeling in the document.

The most important issue is that existing system does not high light the method for the extraction of sentiment. The existing system focuses on specifically domain but not generic. The resources, human language and vocabulary must be relevant to the domain for getting the meaning full results. The main problem faced in existing system is the availability of Urdu text data set. It is difficult to handle the morphological issues of Urdu language Quantifying the selected words polarities, handling the negation, and the improvement in terms of system accuracy is also quiet challenging. Polarity computation result is also must be tangible for the reviewers of the customers.

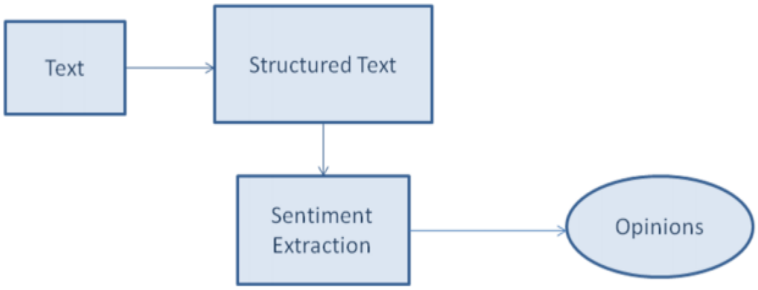


Figure 4.1: Existing system

4.3 Proposed system architecture

Proposed system methodology has two stages. First stage is pre processing of text and second step is classification of data set. The first step take out the data and pre process it for further operation. In the second step the process data is given to the classification algorithm for sentiment classification.

The pre processing of tweets include removal of hash tags at first step. Then removal of stop words is done. Part of speech tagging is done by the Urdu POS Tagging software.

For classification of pre processed data, feature extraction is done. The extracted features set are given to the machine learning algorithm for training. Then test data is given for testing. Decision tree is used as a machine learning algorithm.

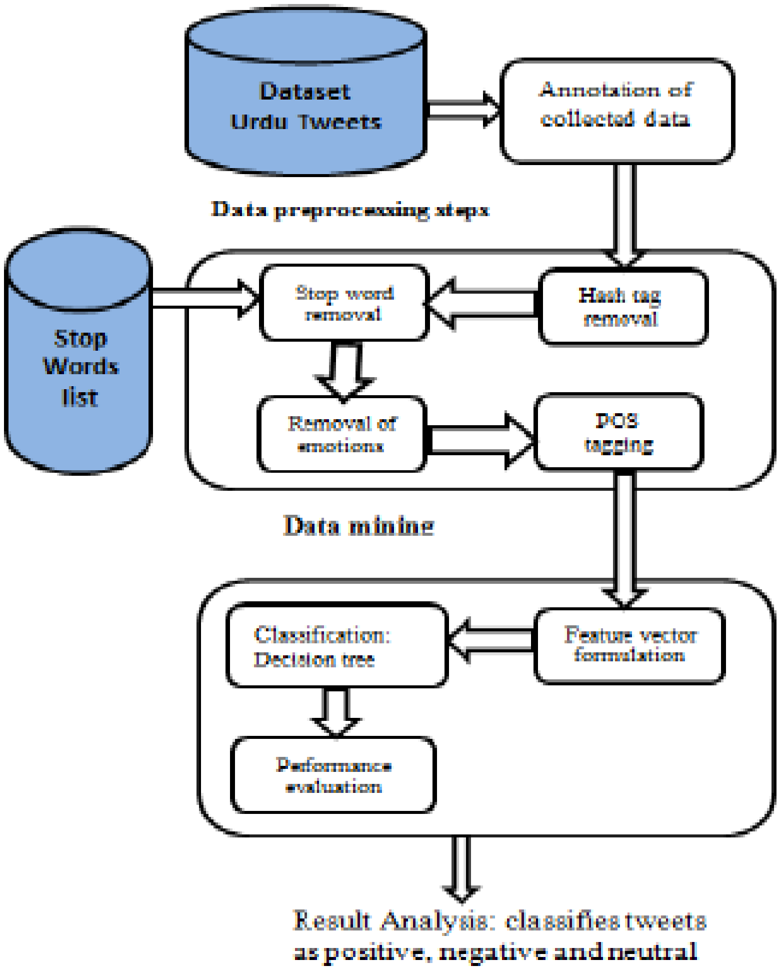


Figure 4.2: Proposed methodology frame work

4.3.1 Natural Language Processing (Nlp)

For the classification of sentiment analysis many pre processing steps are require for structuring the nonstructural text for the purpose of extracting features from them. NLP is the computer science field. The interaction between human language and computer is build up through Artificial Intelligence. So NLP relate to the connection of human with machine.

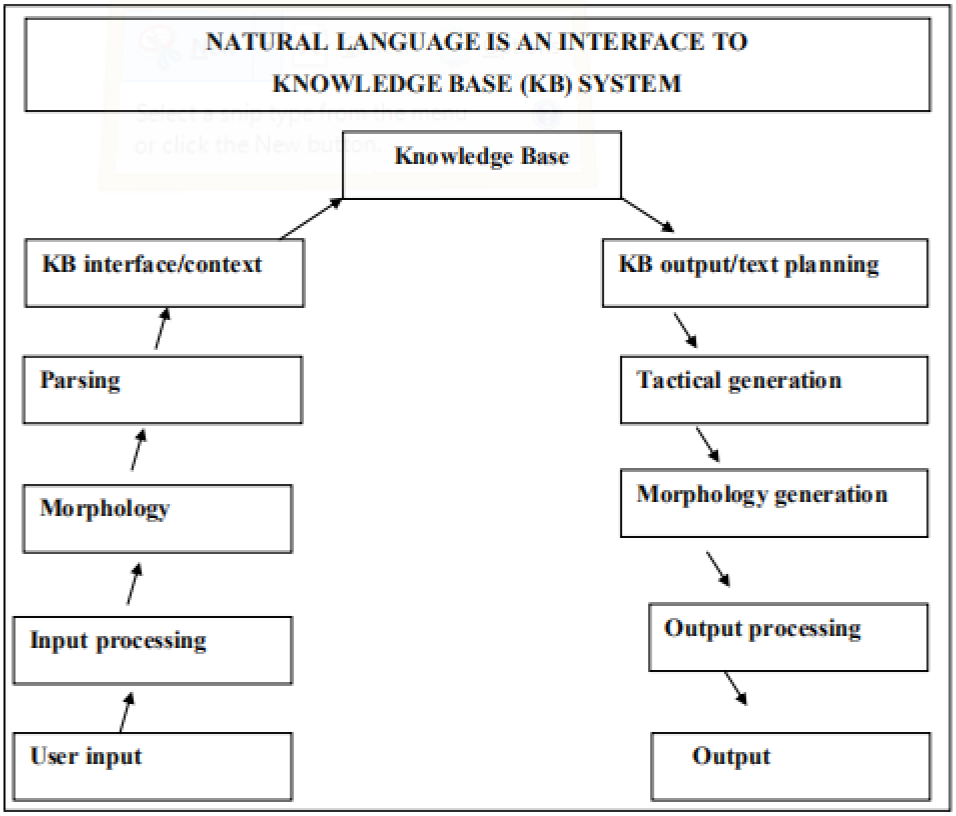


Figure 4.3: NL interface to the base knowledge

The main challenges of NLP are the following:

1) Understanding of Natural language

2) Extracting the meaning of human language

4.3.2 Applications of NLP:

✓ Grammar and spelling checking from the text

✓ Retrieval of information

✓ Classification of document

✓ Segmentation of text

- ✓ Understanding of e-mail
- ✓ Machine translation

NLP includes tokenization, segmentation of words, POS tagging, Stemming and parsing of the words.

The most fundamental step for NLP is tokenization. It divided a document or sentences into tokens. These tokens contain words or phrases. In English it is very east task to tokenize the words by spaces and take some consideration such as look upon entities.

Tokenization also remove stop words such as “the”, “is” etc these words are remove because these words provide very little amount of information. Famous tools for English word tokenization are Open NLP Tokenizer and Stanford Tokenizer.

For other languages which do not have explicit words such as Chinese and Japanese, tokenization is not a easy task as it is easy for English language. Word segmentation is mostly require in English language. Some approaches like “Conditional Random Fields” are applied to these languages for outperform result. Now word segmentation and Deep learning methods are applied to the Chinese word segmentation.

For Urdu language NLP is important because of its unique nature and different morphological nature. It has also importance around the globe due to its millions of speakers. Inflected nature of Urdu attracts the writers for its poetry.

Linguistic resources are not available for Urdu language which makes it as a challenging task for the researcher to perform NLP. Urdu is also famous for free word language as it has one word for same meaning in the sentence.

In Urdu text the problem of space leads to the one of the issue for constructing NLP. For example "انکب"the Urdu word is the combination of two words but the machine is considered as a single word because of space omission problem. For NLP Urdu lexical resource play important role.

In Pakistan the field of linguistic only focuses on teaching of English and on social linguistic areas. There is not much work done even for the Urdu Corpus. There has very limited capabilities in area of linguistic. It is necessary to provide facility to the people who speak and write Urdu language by minimizing the difference between new technology and common man. Urdu NLP systems are requiring for classification of Urdu text data.

For Urdu NLP initiative is taken up by one of the linguistic resource project that is "Essential Urdu Linguistic Resources project" is working on building up linguistic resource and tools by analyzing grammatical and semantic studies.

4.4 Tweets Data set collection

Collection of tweets data set for this project is not an easy task. No doubt Urdu blogs are available but on the other hand no label data is provided in Urdu language. Urdu corpus data is not available because very numerous amount of work done in Urdu language.

On the other hand rich amount of data set is available in English as much research for sentiment analysis is done in English language. For example "Senti Strength" is an English sentiment lexicon that is available publicly for the researcher.

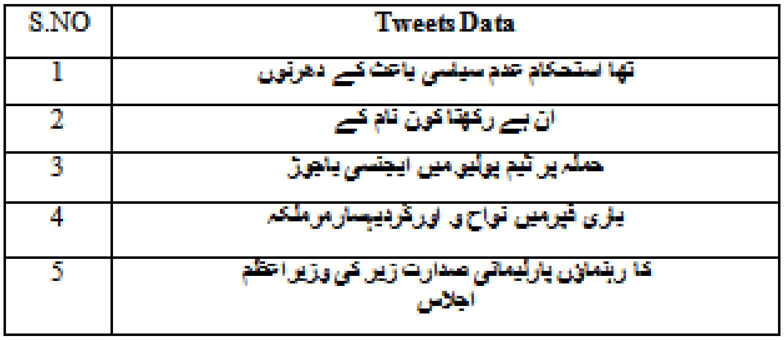
Similarly data set is available for Spanish as well as Arabic and Persian languages. For Urdu language little researcher work for building the data set. But that data set was very limited and not used by many researchers.

Twitter API is used for this research for the collection of data set. Twitter API contains two streams such as Streaming API and REST API for extraction of tweets. Both are important as they help full in collection of tweets. Streaming API provide support for the collection of training data set. While REST API is utilize for the collection of testing data set. Here we discuss the basic features of twitter API:

- ✓ Twitter API contains four main parts. These are “Tweets, entities, places and users”
- ✓ Twitter API has some restriction such as restriction on call in this way protect twitter from abusive cases
- ✓ If request is not correct than it return an error because it is based on HTTP that needs specific HTTP method
- ✓ API has some specific parameters for handling request such as paging generation and restrictions of library.

For acquiring data set in standard form News websites are chosen for this purpose. Table 1 show the data set.

Table 4-1 : Tweets Data



S.NO	Tweets Data
1	تھا استحکام عدم سیاسی باعث کے دھرنوں
2	ان بے رکھتا کون نام کے
3	حملہ پر ٹیم پولیو میں ایجنسی باجوڑ
4	بازی فیرمیں نواح و اورکزدیہسار مرملکہ
5	کا رہنماؤں پارلیمانی صدارت زیر کی وزیر اعظم اجلاس

Table 4-2: Collected Data set

Training Data	Testing Data
600 tweets	400 tweets

4.5 Urdu news tweets data set annotation

Annotation of data set is the method of making the data label for the purpose to use in machine learning. For the process of annotation, data can be in any form that is easily understandable for the human and as well as for machine such as

- ✓ Text data (in Chinese, English , Urdu or in any language)
- ✓ Image data set
- ✓ Audio and video data set
- ✓ Data set in tabular form etc

The annotation of data is quiet not an easy task. For machine learning it is necessary to annotate the data set in order to train the machine learning classifier. Simple task of annotation for machine learning are:

- ✓ Classification of text such positive, negative in terms of sentiment analysis
- ✓ Ratings of the collected data set

For practice the machine learning level, annotation of data set is quiet difficult task. Annotation of data is the one of best way to train the machine learning algorithm. It is easy as compare to tuning the data again and again. So as a result this becomes a key level.

.

The data sets of tweets were collected through twitter API. The collected dataset consist upon 1000 tweets total out of which of 600 tweets were taken as training data set and remaining 400 tweets were treated as testing data set.

The classifier is train with training dataset which is annotated with help of expert language annotators. Before annotation no pre processing is done on the data. By removing the stop words from tweets then it would be difficult for the annotator to understand the proper meaning of the tweet. Hence it would overall effect the annotation of the data.

Urdu tweets are annotated with help of Urdu language expert. For this purpose two annotators are arranged. The annotators annotate the tweets as positive, negative and neutral. Neutral sentences annotated on the basis of presence of objectivity sense. If two annotators are not agree then the tweet is consider again if it is difficult to make decision then the tweet is discarded from the data set. The tweets are annotated as positive if it contains positive word. Negative tweets are annotated on the basis of presence of negative words. As explain earlier tweets which are difficult for the annotator to predict it for annotation are eliminated from data.

3. 6 Data preprocessing

Dataset preprocessing is one of the data mining methodologies. Data preprocessing engage in conversion of raw data into understandable form. Raw data has some flaws like it is incomplete, inconsistent and have some lacking behavior and consider as noisy data. The main objective of data pre processing is to resolving these types of problems.

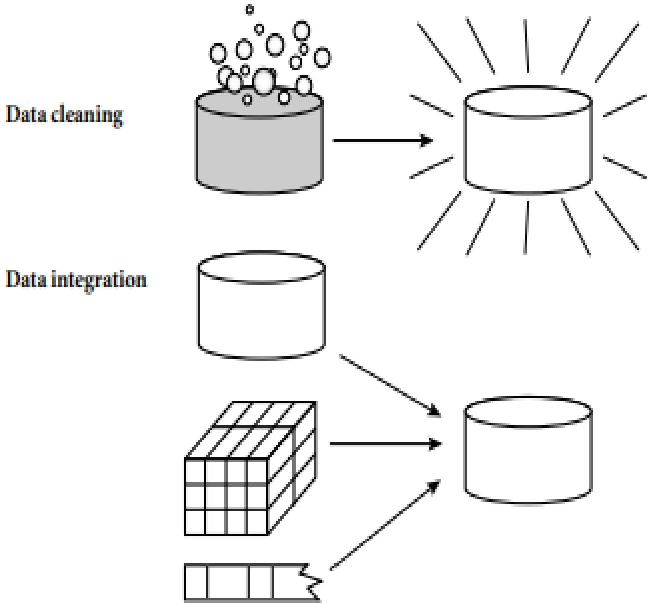


Figure 4.4 : Pre processing architecture

Data preprocessing techniques includes number of techniques. To remove the noise and to reduce data inconsistency and redundancy in data then data cleaning is used for this purpose. To merge the data from different data sources then data integration technique is used which merge data in the data ware houses. Data size can be reducing by minimizing the redundancy, elimination and clustering through data reduction features. Normalization of data is done in order to improve the efficiency of the data mining algorithm. So normalization is done to convert data into smaller range between 0 to 1. All above techniques work together to produce the result. As a result data cleaning produce transformed result by eliminating the noisy data [46]. The dataset was annotated by real annotators and then it is preprocess. In preprocessing stop words are removed after data annotation process. After preprocessing feature are extracted from data set related to sentiment.

.

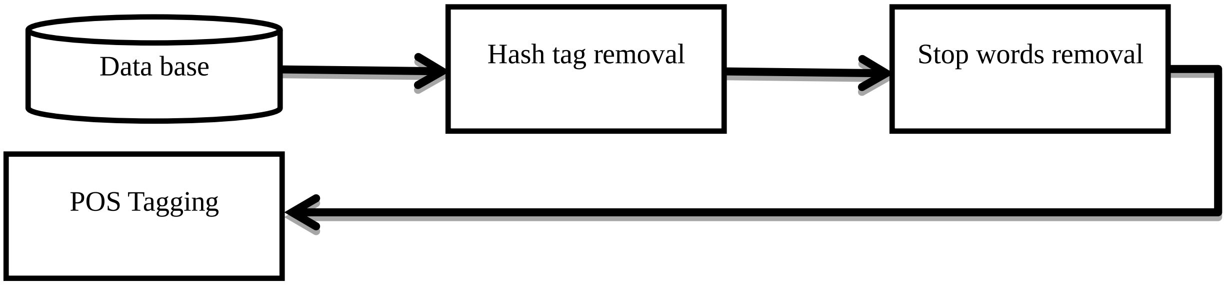


Figure 3.4 Flow chart of pre processing

4.5.1 Removal of Hash tag

Twitter is becoming one of the most powerful communication tools all around the world. Millions of people are posting tweets about what is happening in the world. Hash tag in tweet is known as a keyword or a phrase with the symbol "#" in the tweet. The symbol # can be inserted in place in the tweet. Popular and trendy topics are propagated by millions of people through use of hash tag. Hash tag creates communication with the similar interested users. Many user of twitter use hash tag for organization the post. Let's take an example

#Islamabad اصلی شوارمہ کیسا ہوتا ہے؟

Regular expression is used to remove the hash tag from the tweets. Hence in this way noise is removed from the data set.

4.5.2 Stop words removal

Stops words are important words in any language as they are also known as completing sentence words. Without stop words sentence completion is not complete.

In NLP stop words are removed for the purpose of efficient feature extraction method.

Stop words are removed after the annotation of the data set is done.

List of stop words is prepared for the removal from the text. Removal of stop words also results in accurate result of feature extraction. Stop words are removed by using look-up approach from the maintain list.

Some stop words are shown below in table 3.

Table 4-3: Urdu Stop words

باً	خب	اطراف
لوسہ	کب	کرتے
لگتی	کون	کئے
ہوچکے	کل	آج
پہلے	لگے	کر

4.5.3 POS tagging of the tweets

The major challenges of POS tagging is mapping the part of speech to the specific word. This method of tagging is known as POS tagging. Part of speech tagging is also known as grammatical tagging of the words. Sentiment analysis of the text is determined by the using of nouns and adjective. Noun expresses the name of person place and thing where adjective qualifies the noun. The tagged words are known as effective words. List of Part of speech are:

- ✓ Noun
- ✓ Verb
- ✓ Adjective
- ✓ Adverb
- ✓ Pronoun
- ✓ Preposition
- ✓ Conjunction

- ✓ Interjection

NLP is not done in Urdu language as a result no POS tagger is available. On the other hand lots of POS taggers are available for English language. Natural language processing Society in Pakistan has take initiative by arranging conferences in order to motivate researcher towards the building of NLP structure for Urdu language.

Recently interest has been shown towards the processing of Urdu language. There is no coordination between individual researchers in Pakistan. For NLP of Urdu it is essential to build collaboration between researchers so that they can share ideas regarding Urdu language. For POS tagging we integrate online Urdu POS Tagger in Visual studio. Urdu POS tagger is freely available at "Urdu Summary Corpus" that has 88.7% accuracy.

Urdu Tag list is provided in the below table. The tagging set includes 12 categories of POS.

Table 4-4 : Urdu Tag List

Categories	**Types**	**POS Tag**
1. Nouns	1.1 Common NN 1.2 Proper NNP	NN NNP
2. Verb	2.1 Main Verb Infinitive 2.2 M Verb Finite	VBI VBF
3. Auxiliary	3.1 Aspectual 3.2 Progressive 3.3 Tense 3.4 Modals	AUXA AUXP AUXT AUXM

4.Pronoun	4.1 Personal 4.2 Demonstrative 4.3 Possessive 4.4 Relative Demonstrative 4.5 Relative Personal 4.6 Reflexive 4.7 Reflexive	PRP PDM PRS PRD PRR PRF APNA
5. Nominal Modifier	5.1 Adjective 5.2 Quantifier 5.3 Cardinal 5.4 Ordinal 5.5 Fraction 5.6 Multiplicative	JJ Q CD OD FR QM
6. Adverb	6.1 Common 6.2 Negation	RB NEG
7. Adposition	7.1 Preposition 7.2Postposition	PRE PSP
8. Conjunction	8.1 Coordinate Conjunction Subordinate Conjunction 8.3 SCK 8.4 Pre-sentential	CC SC SCK SCP
9. Interjection	9.1 Interjection	INJ
10. Particle	10.1 Common	PRT

	10.2 Vala	VALA
11. Symbol	11.1 Common 11.2 Punctuation	SYM PU
12. Residual	12.1 Foreign Fragment	FF

4.6 Feature vector creation

The features extraction play important role in classification of sentiment such as positive, negative and neutral. The features are actually the terms that affect the efficiency of the time duration for the construction of the model. So feature extraction is one of the tasks important for the efficiency of training the data in classification. Features extraction also removes noisy features from the datasets. In this study, the feature vector is composed of the following features:

- ✓ identification of number of positive words count
- ✓ number of negative words count
- ✓ presence of negation
- ✓ Use of POS tags.

Twitter specific features such as emotions and hash tags are not presents in all the tweets. So twitter specific features are removed from the tweets during the data preprocessing stage. After preprocessing tweets are treated as simple text. Number of positive word list, number of negative word list and negation are treated as different features in feature vector. After that the search was done to identify parts of speech from tweets. The POS tagging is done with help of one of available online Urdu POS tagger. In this research adjective were taken as effective words. The adjective given positive sense given score as +1 while negative sense given as -1.

4.6.1 Positive word count

In sentiment analysis the no of positive words play an important role. Count of positive words has a great significance while categories tweet as positive, negative and neutral. If tweets contain more than two positive words then it is label as positive tweets. Some of positive words of Urdu are as follows:

Positive Words	حکومت؛ مسلم لیگ ن؛وزیر اعظم؛نواز شریف؛ پاکستان؛ن لیگ؛ حکومت پاکستان؛وفاق؛وفاقی حکومت؛ حکومت پنجاب

Algorithm to calculate count of positive words list

- ✓ The input tweet data is taken
- ✓ Through regular expression test tweet words are match with positive words list
- ✓ No of match words are count using word pattern

4.6.2 Negative word count

Count of negative words has a great significance while categories tweet as positive, negative and neutral. If tweets contain more than two negative words then positive words then it is label as negative tweets. Some of positive words of Urdu are as follows:

Negative Words	اپوزیشن؛اپوزیشن لیڈر؛پی ٹی آئی؛پیپلز پارٹی؛عوامی تحریک؛ تحریک انصاف؛لانگ مارچ؛سول نافرمانی؛متحدہ اپوزیشن؛احتساب

Algorithm to calculate count of negative words list

- ✓ The input tweet data is taken
- ✓ Through regular expression test tweet words are match with positive words list

- ✓ No of match words are count using word pattern

1. Handling of Negation

Negation words are handled in such a way that if negation words come before the positive words then the tweet is label as negative. If negation words come before the negative word then it is consider as positive word.

4.6.3 POS tagging words

Adjective and verb are known as effective words. In feature vector generation we only consider the adjective words. After POS tagging of the tweets, the algorithm only extract adjective and eliminate the rest of part of speech.

4.7 Sentiment Classification of Tweets

Feature vector comprises four features; classifications of tweets as positive, negative. Classification algorithm categorizes the tweets as negative, positive, and neutral. The machine learning algorithms divided in two classes such as supervised machine technique and unsupervised machine learning technique.

We rely on supervised machine learning classification technique .the classifier is trained by providing the training data set which is labeled in advance. The classification technique used C 4.5 decision tree which classify the tweets.

For detection of sentiment from the news tweets, we used decision tree (C45 Algorithm) classifier. Decision tree is the data mining tool that has tree like features such as graph of decision including outcomes of event. It is one of the algorithms of data mining that consist upon control structure. Applications of decision tree include:

- ✓ When there is an objective to achieve maximum profit
- ✓ When there are number of actions to be taken
- ✓ Events become out of control
- ✓ Uncertainty in the data

The C 4.5 algorithm is a data mining algorithm that is used as a classifier. C 4.5 algorithm is used to take a decision based on training data set.

4.7.1 C 4.5 Algorithm advantages

- ✓ Algorithm use single pruning pass method to eliminate over fitting
- ✓ Algorithm can handle continuous as well as discrete data
- ✓ Algorithm can also handle incomplete data issue

4.8 Summary

After getting the data related of tweets, datasets are developed and annotated using domain expert's knowledge. For detection of sentiment from the news tweets, after implication of preprocessing and feature extraction techniques, we used C4.5 Decision tree Machine Learning Algorithm during classification module. Dataset is used for training and cross validation purpose whereas testing data is used for testing purpose. Effectiveness of proposed methodology is discussed in next chapter.

Chapter 4

Results and Discussion

5.1 Introduction

We have discussed the step by step proposed methodology in detail by explaining all involved steps in detail. This chapter focuses on evaluating the result of proposed methodology. Also proposed methodology is compare with existing techniques and effectiveness of the result is analyzed.

5.1.1 Overview:

After going through all the steps of proposed methodology by using text available on twitter, is extract through Twitter API is used during learning activities of the algorithm. Tweets are selected and then provided to experts for annotation process. After attaining the annotated data, we have done preprocessing of the data. Outcome of each step is discussed in detail.

We have 500 samples of Urdu news tweets are collected after preprocessing in our data set1 which is an annotated data for the classification of tweets. We have labeled the tweets as 1 for positive-1 for negative and o for neutral.

We implemented the proposed methodology in visual studio using c#. Visual studio is an integrated development environment. Visual studio is used by programmers to develop websites, pages and even mobile apps.

5.2 Performance evaluation of classifier

Performance evaluation of the C 4.5 classifier was done through tenfold cross validation. We also calculate accuracy precision; recall and f-measure. The accuracy was calculated using the formula:

Accuracy= (NP/TN) where NP for Number of correct result and TN for total number of occurrences. Precision calculated using Precision = (Number predicted correct result/Number of predicted result). Recall calculated using Number predicted correct result/ Number predicted correct result. F-measure calculated using (2*Precision *Recall/ Precision + Recall).

The decision tree C45 was used for the classification of the Urdu news tweets as positive, negative and neutral.

Results of all three steps of proposed methodology are discussed below.

5.2.1 Preprocessing of text

5.2.1.1 Removal of hash tag and Stop word

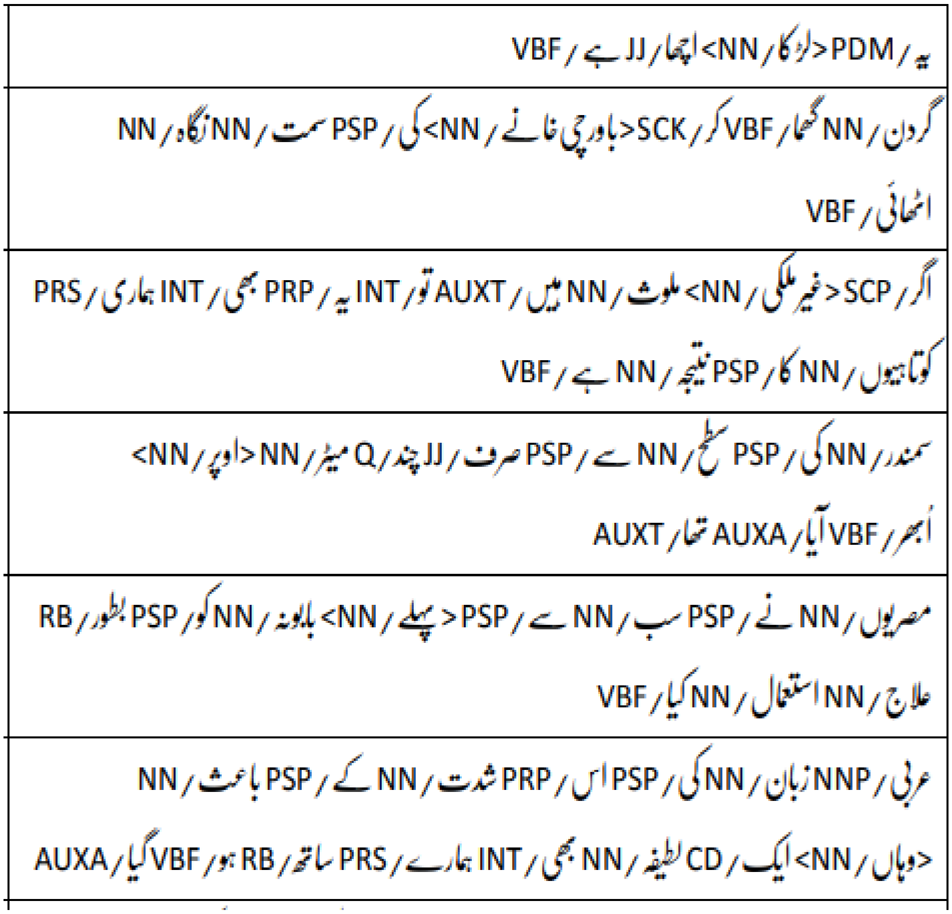
This step is applied on the 500tweets of dataset1 and single tweets which constitutes our second dataset. Tweets are passed to removal of hash tag (#) and stop word removal in which all the stop words are maintain in the list and look up technique is used. Along with this, words containing numbers and special words are also removed. Sample of obtained results are shown in table

Before pre processing	اصلی شوارمہ کیسا ہوتا ہے؟I #Islamabad
After pre processing	اصلی شوارمہ کیسا ہوتاI

5.2.1.2 *Part of speech tagging*:

Part of speech tagging is applied on tweets after removal of stop words. Different parts of speeches are identified for tweets. Part of speech tagging is done in broader aspect of four main categories. Obtained results are shown in below

Table 5-1: POS tagging result



یہ/PDM <لڑکا/NN> اچھا/JJ ہے/VBF
گردن/NN گھما/VBF کر/SCK <باورچی خانے/NN> کی/PSP سمت/NN نگاہ/NN اٹھائی/VBF
اگر/SCP <غیر ملکی/NN> ملوث/NN ہیں/AUXT تو/INT یہ/PRP بھی/INT ہماری/PRS کوتاہیوں/NN کا/PSP نتیجہ/NN ہے/VBF
سمندر/NN کی/PSP سطح/NN سے/PSP صرف/JJ چند/Q میٹر/NN <اوپر/NN> ابھر/VBF آیا/AUXA تھا/AUXT
مصریوں/NN نے/PSP سب/NN سے/PSP <پہلے/NN> بابونہ/NN کو/PSP بطور/RB علاج/NN استعمال/NN کیا/VBF
عربی/NNP زبان/NN کی/PSP اس/PRP شدت/NN کے/PSP باعث/NN <وہاں/NN> ایک/CD لطیفہ/NN بھی/INT ہمارے/PRS ساتھ/RB ہو/VBF گیا/AUXA

5.2.2 Feature Extraction:

After preprocessing steps, feature extraction method is applied on the final generated tweets. Tweets along with its features are shown in table 4.2.

5.3 Classification

5.3.1 Training and testing Phase

We have use Decision tree c4.5 model for training and testing of the model. Implementation is done in visual studio using c#.

5.3.2 Evaluation

Training data set is given to the classifier model to test the test data. Hence evaluation is done in this manner. Attained accuracy is shown through presentation of the confusion matrix. N in confusion matrix represents total number tweets data set.

TP--True positive: Actual True and Predicted as True.

TN-- True negative: Actual False and Predicted as False.

FP--False positive: Actual False and Predicted as True.

FN-- False negative: Actual True and Predicted as False.

5.3.3 Detection Accuracy:

A) For Training Data set

As a result of testing phase, labels are given to the tweets as positive, negative and neutral.

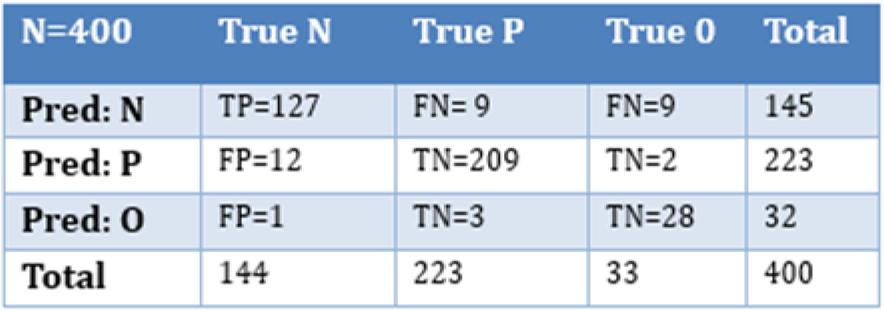
Total tweets = 1000

Positive tweets = 331

Neutral Tweets =153

Negative Tweets = 516

Table 5-2 : Confusion Matrix



N=400	True N	True P	True 0	Total
Pred: N	TP=127	FN= 9	FN=9	145
Pred: P	FP=12	TN=209	TN=2	223
Pred: O	FP=1	TN=3	TN=28	32
Total	144	223	33	400

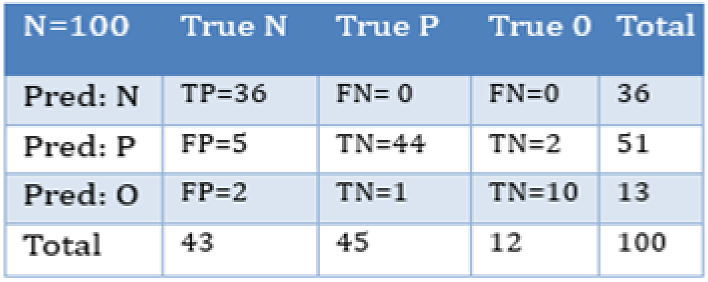
B) Cross Validation Results of training data

Using Formula Accuracy = (TP+TN)/500---------- (4.2)

Accuracy = 91% accuracy of the proposed methodology

Confusion matrix of training dataset is shown in table 4-5.

Table 5-3: Cross validation



N=100	True N	True P	True 0	Total
Pred: N	TP=36	FN= 0	FN=0	36
Pred: P	FP=5	TN=44	TN=2	51
Pred: O	FP=2	TN=1	TN=10	13
Total	43	45	12	100

5.4 Comparison with Existing Techniques:

As discussed earlier that very numerous work has been done on sentiment analysis of Urdu tweets. Research paper [44] developed system to identify positive and negative tweets from the tweets data set using the supervised machine learning technique in year 2018. In this papers author uses supervised learning technique to achieve desired results. These papers focus on features such as number of positive and negative words. We considered this research paper as state of the art work.

Authors use's the accuracy as performance evaluator such as Precision, Recall and F-measure. In our proposed methodology, we have focused on the extraction of feature vector. These features help to determined relevant and specific information about why tweets are positive or negative. Table 4-11 depicts that when compare to the existing technique our proposed methodology outperformed in term of using supervised machine learning technique.

Table 4-5: Comparison with existing techniques

Methodology	**Technique used**	**Accuracy**
[44]	Lexical based approach	87.36%
Our proposed Methodology	C4.5 Decision Tree Learning	91% using dataset2

5.5 Summary:

In this chapter, results of proposed methodology are discussed in detail. Results of pre-processing, feature extraction and classification are shown separately. Accuracy is calculated to evaluate the identification of difficult words in a text document. 97% aggregate accuracy is achieved according to the results which are generated for evaluation as a result of 10 folds cross validation. On training dataset, achieves 97% accuracy and on testing dataset2 achieves an accuracy of 92%. This is due to the fact that these may not exist in our training data and computed as an error.

Chapter 5

Conclusion and Future Work

6.1 Conclusion

After discussion of related work on Urdu text data, proposed methodology of sentiment analysis, and results of proposed methodology using decision tree in detail, Now we are able to conclude our research work and result of proposed methodology. In this chapter, overall research done in this thesis is concluded and future work is discussed for further enhancement in the result or in methodology. Sentiment analysis is mostly done in English language from last 3 decades. English corpus is available online and many POS tagging tools are also available. Urdu language is ignore from the researcher because of the morphological issues and challenges.

Urdu language has many different features which make it challenge for the researcher to done classification on it. Some of the challenges for Urdu language are lexicon intricacy, handling phrase level negation, dealing morphological complexity. These challenges are not address by the researcher and for this reason sentiment analysis is mostly done on English language. Urdu is the popular language of Asia and has a close relation with Indo-Aryan language. The basic purpose of this research is to done opinion mining in Urdu using supervised machine learning technique. The word meaning is depends on their position in the sentence. Urdu shares similarities among many languages such as it share script similarities with Persian and Arabic and has morphological similarities with Hindi language. Sharing similarities with other languages, even it has its own linguistic features. Techniques available for other languages are not applicable for Urdu language.

This research describes the systematize way of developing a proposed methodology. It describes each step involved in detail in implementation and importance of each step as well. It also describes how the data is collected; way of annotating the data set, feature extraction and apply the classification algorithm. The aim of our research was to implement a methodology/technique for determining the sentiment analysis of news tweets data. An efficient technique is achieved with significant increase in accuracy of difficult word identification. 97% accuracy is attained using labeled dataset which constitutes of 600 tweets and 92% accuracy is attained using testing dataset 2 which contain 400 tweets. This research show improved result in term of accuracy as compared to the previous research done on Urdu text data.

6.2 Future Work

Further research is needed to find more elaborated techniques, methods, architectures and its appropriate training algorithms. Decision tree as classifiers have here a high potential because they can compute with a high numbers of features. There are two main avenues for the extension of this research: application will be performed on more datasets, and other techniques will be applied for the purpose of feature extraction and classification.

6.2.1 Enhanced feature sets

Currently, we have used features based on need of sentiment analysis In future; enhancement to feature set can also be applied to enhance the overall performance of the technique for detecting the sentiment. Researcher shows interest to develop Urdu lexical data.

6.2.2 Improved NLP

English text is available in the form of free-text or unstructured formats. Before implementation of feature extraction and classification steps it is necessary to process unstructured data. For Urdu text Data no NLP is provided. There is need to done research for the development of Urdu NLP. Natural language processing tasks can be improved in future work for better results.

6.2.3 Availability of Annotated Data set

Currently, these techniques were only applied on 500 tweets data set. In order to show better, it is desirable to replicate this research on further datasets with different properties and features. The dataset in this thesis involved the classification of news tweets. This work could be extended to datasets of any sort of Urdu text data. Annotated data is not available so there is a need to develop annotated data set.

6.3 References

[1] Chikersal, Prerna, Soujanya Poria, and Erik Cambria. "SeNTU: sentiment analysis of tweets by combining a rule-based classifier with supervised learning." Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). 2015.

[2] Amjad, Kamran, et al. "Exploring Twitter news biases using urdu-based sentiment lexicon." *2017 International Conference on Open Source Systems & Technologies (ICOSST)*. IEEE, 2017.

[3] Bilal, Muhammad, et al. "Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques." *Journal of King Saud University-Computer and Information Sciences* 28.3 (2016): 330-344.

[4] Pang, B., and L. Lee. 2008. Opinion mining and sentiment analysis. Foundations and

Trends in Information Retrieval. Vol. 2(1-2), 1–135.

[5] Bowman, Q. Akram, A. Naseer and S. Hussain. Assas-band, an affix- exception-list based Urdu stemmer. Proceedings of the 7th Workshop on Asia Language Resources. Singapore. pages 40–47. (2009).

[6] Riaz, K.: Challenges in Urdu Stemming. Future Directions in Information Access, Glasgow(August 2007)

[7]Durrani, N., Hussain, S.: Urdu Word Segmentation. In: 11th Annual Conference of the

North American Chapter of the Association for Computational Linguistics (NAACL HLT

2010)   , Los Angeles, US (2010)

[8]Rashid, A., N. Anwer, M. Iqbal and Muhammad S. 2013. A Survey Paper: Areas,

Techniques and Challenges of Opinion Mining. Intl. J. Comp. Science. Vol. 10(6), 18-31.

[9] A Review of Feature Extraction in Sentiment Analysis Muhammad Zubair Asghar1 , Aurangzeb Khan2 , Shakeel Ahmad1 , Fazal Masud Kundi1

[10] Jones, Karen Sparck. "Natural language processing: a historical review." Current issues in computational linguistics: in honour of Don Walker. Springer Netherlands, 1994. 3-16.

[11] yu, H. and Hatzivassiloglou, V. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (2003).

[12] Narayanan, R., Liu, B. and Choudhary, A. Sentiment analysis of conditional sentences. In Proceedings of the 2009 Conference on Empirical Methods in Natural

Language Processing (Singapore, 2009). Association for Computational Linguistics, 180–189.

[13] akob, N. and Gurevych, I. Extracting opinion targets in a single-and cross-domain setting with conditional random fields. In Proceedings of Conference on Empirical Methods in Natural Language Processing

(2010).

[14] Popescu, A.-M. and Etzioni, O. Extracting product features and opinions from reviews. In Proceedings of Conference on Empirical Methods in Natural Language Processing (2005).

[15] Agarwal, Basant, Namita Mittal, and Erik Cambria. "Enhancing sentiment classification performance using bi-tagged phrases." *2013 IEEE 13th International Conference on Data Mining Workshops*. IEEE, 2013.Das, Dipankar, and Sivaji Bandyopadhyay. "Emotion analysis on social media: natural language processing approaches and applications." *Online Collective Action*. Springer, Vienna, 2014. 19-37.

[16] Haddi, Emma, Xiaohui Liu, and Yong Shi. "The role of text pre-processing in sentiment analysis." *Procedia Computer Science* 17 (2013): 26-32.

[18] Lafferty, J., McCallum, A. and Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proc. 18th International Conf. on Machine Learning. Morgan Kaufmann, San Francisco, CA, 2001, 282–289.

[19] Liu, Bing. "Sentiment analysis and opinion mining." *Synthesis lectures on human language technologies* 5.1 (2012): 1-167.

[20] Thelwall, Mike, Kevan Buckley, and Georgios Paltoglou. "Sentiment in Twitter events." *Journal of the American Society for Information Science and Technology* 62.2 (2011): 406-418.

[21] Melville, Prem, Wojciech Gryc, and Richard D. Lawrence. "Sentiment analysis of blogs by combining lexical knowledge with text classification." *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009.

[22] Tan, Luke Kien-Weng, et al. "Sentence-level sentiment polarity classification using a linguistic approach." *International Conference on Asian Digital Libraries*. Springer, Berlin, Heidelberg, 2011.

[23] Moreo, Alejandro, et al. "Lexicon-based comments-oriented news sentiment analyzer system." *Expert Systems with Applications* 39.10 (2012): 9166-9180.

[24] Park, Sungrae, Wonsung Lee, and Il-Chul Moon. "Efficient extraction of domain specific sentiment lexicon with active learning." *Pattern Recognition Letters* 56 (2015): 38-44.

[25] Deng, Shuyuan, Atish P. Sinha, and Huimin Zhao. "Adapting sentiment lexicons to domain-specific social media texts." *Decision Support Systems* 94 (2017): 65-76.

[26] Wu, Fangzhao, et al. "Towards building a high-quality microblog-specific Chinese sentiment lexicon." *Decision Support Systems* 87 (2016): 39-49.

[27] Hogenboom, Alexander, et al. "Multi-lingual support for lexicon-based sentiment analysis guided by semantics." *Decision support systems* 62 (2014): 43-53.

[28] Dey, Atanu, Mamata Jenamani, and Jitesh J. Thakkar. "Senti-N-Gram: An n-gram lexicon for sentiment analysis." *Expert Systems with Applications* 103 (2018): 92-105.

[29] Moh, Melody, et al. "On multi-tier sentiment analysis using supervised machine learning." *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. Vol. 1. IEEE, 2015.

[30] Shein, Khin Phyu Phyu, and Thi Thi Soe Nyunt. "Sentiment classification based on Ontology and SVM Classifier." *2010 Second International Conference on Communication Software and Networks*. IEEE, 2010.

[31] Shi, Han-Xiao, and Xiao-Jun Li. "A sentiment analysis model for hotel reviews based on supervised learning." *2011 International Conference on Machine Learning and Cybernetics*. Vol. 3. IEEE, 2011.

[32] Pannala, Nipuna Upeka, et al. "Supervised learning based approach to aspect based sentiment analysis." *2016 IEEE International Conference on Computer and Information Technology (CIT)*. IEEE, 2016.

[33] Zhang, Bowen, et al. "Sentiment analysis through critic learning for optimizing convolutional neural networks with rules." *Neurocomputing* 356 (2019): 21-30.

[34] Zhang, Bowen, et al. "Sentiment analysis through critic learning for optimizing convolutional neural networks with rules." *Neurocomputing* 356 (2019): 21-30.

[35] Syed, Afraz Z., Muhammad Aslam, and Ana Maria Martinez-Enriquez. "Lexicon based sentiment analysis of Urdu text using SentiUnits." *Mexican International Conference on Artificial Intelligence*. Springer, Berlin, Heidelberg, 2010.

[36] Syed, Afraz Zahra, Muhammad Aslam, and Ana Maria Martinez-Enriquez. "Sentiment analysis of urdu language: handling phrase-level negation." *Mexican International Conference on Artificial Intelligence*. Springer, Berlin, Heidelberg, 2011.

[37] Daud, Misbah, Rafiullah Khan, and Aitazaz Daud. "Roman Urdu opinion mining system (RUOMiS)." *arXiv preprint arXiv:1501.01386* (2015).

[38] Rehman, Zia Ul, and Imran Sarwar Bajwa. "Lexicon-based sentiment analysis for Urdu language." *2016 sixth international conference on innovative computing technology (INTECH)*. IEEE, 2016.

[39] Hashim, Faiza, and M. Khan. "Sentence Level Sentiment Analysis Using Urdu Nouns." *Department of Computer Science, University of Peshawar, Pakistan*.

[40]. Usman, Muhammad, et al. "Urdu text classification using majority voting." *INTERNATIONAL JOUR NAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS* 7.8 (2016): 265-273.

[41]. Bilal, Muhammad, et al. "Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques." *Journal of King Saud University-Computer and Information Sciences* 28.3 (2016): 330-344.

[42]. Ali, S. Abbas, et al. "Salience Analysis of NEWS Corpus using Heuristic Approach in Urdu Language." *International Journal of Computer Science and Network Security (IJCSNS)* 16.4 (2016): 28

[43]. Amjad, Kamran, et al. "Exploring Twitter news biases using urdu-based sentiment lexicon." *2017 International Conference on Open Source Systems & Technologies (ICOSST)*. IEEE, 2017.

[44] Mukhtar, Neelam, and Mohammad Abid Khan. "Urdu sentiment analysis using supervised machine learning approach." *International Journal of Pattern Recognition and Artificial Intelligence* 32.02 (2018): 1851001.

[45] Riaz K (2012) Comparison of Hindi and Urdu in computational context. Int J Comput Linguist Nat LangProcess

1(3):92–97

[46] http://hanj.cs.illinois.edu/cs412/bk3/03.pdf

[47] Kim, Hak J. "Online social media networking and assessing its security risks." *International Journal of Security and Its Applications* 6.3 (2012): 11-18.

[48] Li, Guangxia, et al. "Micro-blogging sentiment detection by collaborative online learning." *2010 IEEE International Conference on Data Mining*. IEEE, 2010.

[49] Li, Guangxia, et al. "Micro-blogging sentiment detection by collaborative online learning." *2010 IEEE International Conference on Data Mining*. IEEE, 2010.

[50] Feldman, Ronen. "Techniques and applications for sentiment analysis." *Commun. ACM* 56.4 (2013): 82-89.

[51] Ghosh, Saptarshi, et al. "Understanding and combating link farming in the twitter social network." *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012.