Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 96 (2016) 1560 – 1570

Procedia Computer Science

20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2016, 5-7 September 2016, York, United Kingdom

Pre Processing of Twitter's Data for Opinion Mining in Political Context

Ratab Gull[a*], Umar Shoaib[a], Saba Rasheed[b], Washma Abid[b],Beenish Zahoor[b]

"aUniversity of Gujrat,Gujrat,Pakistan"

"bNational University of Computer and Emerging Sciences, Islamabad ,Pakistan"

Abstract

In the wake of political activism among youth in particular and the whole population in general, everyone is not only eager to share their political orientation but equally curious regarding the voice of the masses. As a known notion, the perfect orifice to this emerging need of political activism can be found on social media platforms, from where the numerous aspects of public opinion can be captured easily. These sites have begun to have a large impact on how people think and act. It is a known phenomenon that public opinion is the largest indicator of success and failure of political parties and is a direct reflection of the party's reign. Where increased sharing of public feedback has increased awareness and promoted accountability, it has also created chaos and confusion for many. Using Twitter, the most popular micro blogging platform, this paper aims to give a method to ease and smooth the task of opinion mining with the help of linguistic analysis and opinion classifiers, which will together determine positive, negative and neutral sentiments for the political parties of Pakistan. A method is provided which pre-processes the raw data of twitter and comparison of two classification techniques to classify this data. That will aspire to capture a snapshot of current political scenario to promote the spirit of accountability, self-analysis and improvement in among Pakistani politicians. Moreover, with this we aim to give general public an important consolidated voice in the realm of politics.

*Keywords:*Tweepy;Pycharm;Sementic;Opinion Mining

[*] Corresponding author. *E-mail address:* ratabgull@yahoo.com

doi:10.1016/j.procs.2016.08.203

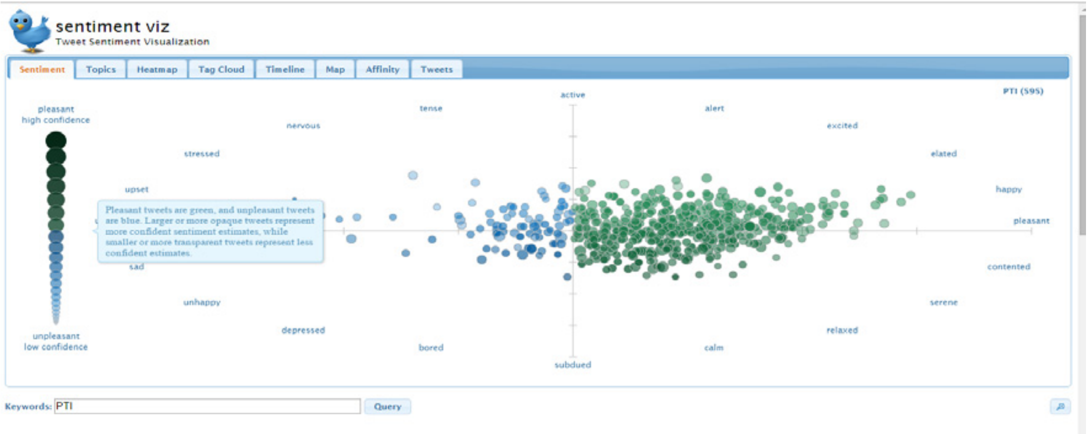
1. Introduction

General public these days react to political parties by the means of social media. Due to the large number and diversity of posts these important public reactions are not processed to formulate useful information that can be used to form a better picture of the public voice in the political scenario. To fill the information gap between public opinion and correct an accurate summary of a pool of opinion orientations of the country, it provides a unique opportunity to directly help a common man as well as politically invested individuals and experts seeking consolidated and easy access to social data.

Two of the most prominent implementations of opinion mining technology are Obama's Campaign and Indian Prime Minister Modi Campaign. Obama's Campaign took social media and the field of opinion mining to a whole new level, which resulted in the predictions of elections to be only 2.5 % off than the actual result. The significance went far beyond predicting the winners of the election. His approach amounted to a decisive break with 20th century tools for tracking public opinions. He revolutionized the field by introducing a new way of using social media to mine opinions. After this break the field of opinion mining came into the limelight and people began to realize the importance of opinions on social media and how they can be used to form analysis and predictions. Narendra Modi is said to be India's First Social Media Prime Minister. He has declared that social media is the direct form of information and said that it was social media that gave him the much needed local pulse. The need to utilize social media to mine opinions has never been this much. In today's world where everyone posts their opinions online, there is a great market to utilize these opinions to form proper information.

The people's opinions can be expressed linguistically in different forms called subjectively, emotions, evaluations, beliefs, sentiments and speculations. The sentiment analysis can be done from subjectivity detection both independently or dependently. The research in this area has recently started but the substantial growth in the online information in the recent years has proved the sentiment analysis a constantly growing area of research where the efforts of research community are concentrated. It is evident from the research published in the area of sentiment analysis that it is tedious task not just due to syntactic and semantic variations of language but also due to involvement and implicit assessments of objects or indirection extraction based on the emotions or attitudes of subjects.

Summary of some of the products developed for opinion mining is here. As a part of a study conducted by Department of Computer Science, NC State University an application by the name of Sentiment Viz was introduced (see Figure 1). This application uses twitter as its data source to classify tweets into different set of emotional categories. The tweets are extracted based on keywords**.**

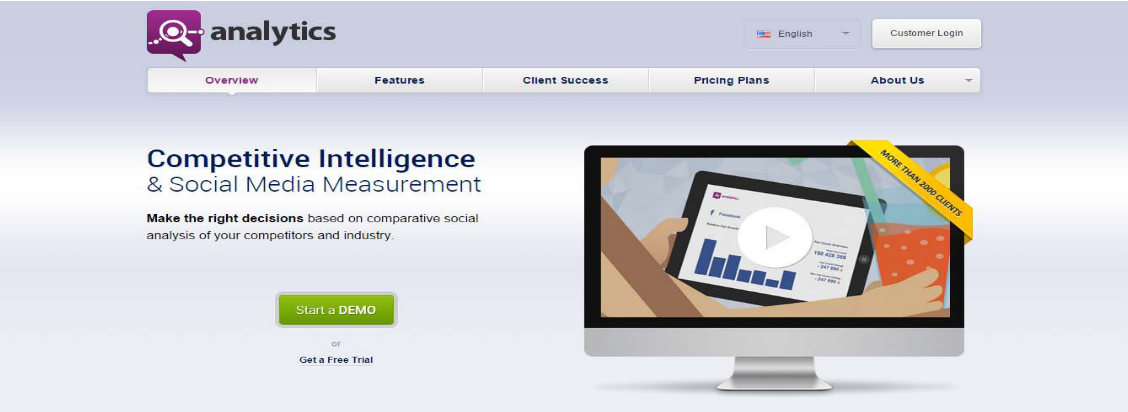


Pleasant tweets are green, and unpleasant tweets are blue. Larger or more opaque tweets represent more confident sentiment estimates, while smaller or more transparent tweets represent less confident estimates.

Keywords: PTI

Figure 1 Sentiment classification defined over the keyword "PTI"

The screenshot of the website shown in Figure 2 is a commercial system aimed for companies and think tanks looking for user input on different topics and entities. This website uses data from Facebook, Twitter and LinkedIn to formulae results in the mentioned domains.



Competitive Intelligence
& Social Media Measurement

Make the right decisions based on comparative social analysis of your competitors and industry.

Figure 2 Marketing page of social baker's analytics section

2. Literature Review

With a sharp increase in the users of social media sites, opinion mining has now become a very well researched topic among scholars. A publication by Pang and Lee, Foundations and Trends in Information Retrieval 2008, provides a high level overview of existing techniques and styles adopted for opinion mining. However, not many studies or development projects based on opinion mining have given much attention to micro blogging and structured political opinion orientation is almost an undiscovered field. Mentioned below are some major research and development that has been conducted in the domain of opinion mining over social media. The most important part of opinion mining is to determine opinion orientation of unit data that can be used to form an opinion summary. As the key determinants of an opinion are the use of opinion words, many researchers (Hatzivassiloglou, 1997) have tried to mine such words so that a semantic orientation can be built. A pool of specific dictionary words is used as seeds and their synonyms and antonyms are used to further streamline the technique. Moreover sentence level opinion classification has also been explored by (Kim, 2004). Furthermore, different studies have used a pool of different techniques based on the type of data at hand to mine opinions over different mediums. As determining positive and negative opinions are learning classification problem by characteristics. The researchers have investigated different classification techniques to achieve desired accuracy. Another approach of opinion mining is discussed by (SamanehMoghaddam & Martin Ester, 2013), where the opinion is classified into a quintuple consisting of target entity, aspect of entity, opinion holder, time when the opinion is expressed and orientation of opinion. The opinion is classified on the basis of sentiment lexicon but the disadvantage of using lexicon is that it is domain dependent. In another study conducted by (Wei, 2012) it was found that the opinions were classified into a holder, target, polarity and auxiliary. After that they also used lexicon classification. In an another research conducted by (SamanehMoghaddan & Martin Ester, 2012) opinions are classified into quintuples consisting of target entity, aspect of entity, opinion holder, time but in this they have used a sentiment orientation formula to classify the opinion. In a publication (Changhua Yang, 2007) the authors discuss using a web blog to mine opinions based on the emoticons assigned to blog posts to determine the mood of the users. Data was passed through a classier (SVM and Naïve Bayes) to mine rules and which then classified each unit into positive, negative and neutral categories. Built on a similar approach in a publication (Alec Go, 2009), Twitter has been used to obtain training data as the basis of opinion/sentiment classification. After testing a range of well-known classifiers, the authors determined that Naïve Bayes algorithm provides the most accurate (up to 81%) results on the test set. Another study on opinion mining was conducted by (Anwar & Rashid, 2013) in which they determined polarity of text and frequency distribution.

In this paper, we introduce a novel technique to do the opinion mining in political context. To the best of our knowledge this is the first study to introduce the analysis which provides the results (mined opinion of people) with respect to the region about any political party of Pakistan.

This process includes these steps Extracting customer feedback, POS tagging from natural language, Extract all features and their polarity, finding the polarity feature wise and overall polarity using Naïve Bayes probability and frequency distribution and Finding the polarity of each product with a total of polarity with frequency distribution by calculating the positive and negative comments.

This paper is organized into five sections. Section 1 provided the introduction about the topic while section describes the related work done in this area. In section 3, the methodology adopted to extract the information and perform sentimental analysis is discussed. In section the discussion about the study and results are illustrated. The conclusion and future work is shown in section 5.

3. Methodology

In order to extract the opinion first of all data is selected and extracted from twitter in the form of tweets. After selecting the data set of the tweets, these tweets were cleaned from emoticons, unnecessary punctuation marks and a database was created to store this data in a specific transformed structure. In this structure, all the transformed tweets are in lowercase alphabets and are divided into different parts of tweets in the specific field. The details about the steps adopted for the transformation of information are described in next subsections.

3.1 Data Extraction

It was the first step of this whole process. In this python was used as scripting language to extract data from twitter using a twitter API written in python named tweepy. The data was extracted using names of all the political parties, names of prominent party members and campaigns along with the major events occurred in that particular time frame to ensure that the tweets extracted are relevant and according to the requirements. Extracted dataset consisted of text of the tweet along with the date and time of the tweet.

3.2 Data Transformation

The information extracted from the tweets of selected dataset was in raw form. In order to use them, we had to clean and transform it into more usable structured dataset to ensure that our next phases of the process are smooth, easy efficient and effective. For this purpose, following steps shown in Figure 3 were adopted.

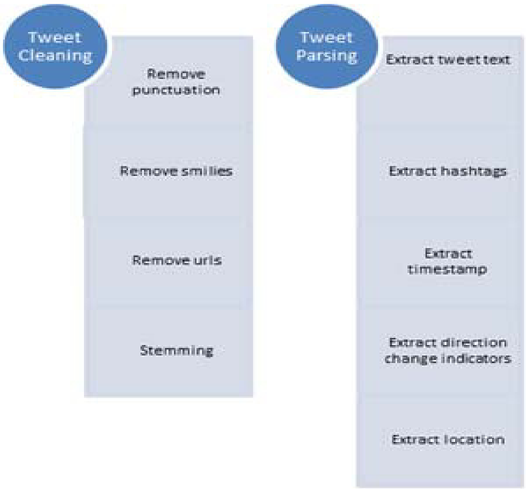


Figure 3 Visualization of data transformation process

3.2.1 Tweet cleaning

Tweet cleaning is the first step towards data transformation. This task consists of three subtasks to accomplish this process. The overview of the subtasks is given following.

- **Converting all Tweets into Lower case**
  We had to convert all tweets to lower case in order to bring the tweets in a consistent form. By doing this, we can perform further transformation and classification without having to worry about non-consistency of data. This task is done using python in Pycharm IDE with Lower () function. This function converted all alphabets in lower case and solved case sensitivity problem by making all data consistent in lower case.

- **Removing emoticons and punctuations**
  We had to remove emotions and punctuations because they were not needed in our analysis. One might ask why we removed emoticons. We removed them because when they were being extracted they appeared in the form of square boxes instead of proper emoticons. These garbage values of emoticons were removed by using "replace" keyword. Replace keyword was used in a way that it replaced all emoticon symbols and values with an empty space making data clear from useless emoticon mess.

- **Removing URL'S:**
  The very next step is to remove the URL'S, as they provide no information during analysis. URL's show links to other webpages and websites. These were of no use so from all tweets these were removed by using re.sub () in python, which replaced all sentences and sub parts of sentences started from http with blank spaces.

3.2.2 Tweet Parsing

After cleaning the data, the selected tweets are then parsed. This is the second and final step of data transformation and to get this achieved below explained steps were taken

- **Extracting Hash tags**
  As hash tags are the newest trends in voicing opinions and gaining public popularity, we have captured the essence of hash tags to help us determine the value of the tweet at hand. We know hashtags are those words

which start from number symbol (#) so extracted them using re.findall() from all tweets. This command extracted all words started from "#" symbol.

- **Extracting change of direction indicators:**
  We needed to save the words which can change meaning and context of the whole sentence like "and" "or" "but" etc. So we searched all those words and saved them in a separate column that will be used for better and accurate analysis.

3.3 Database Creation

To prepare storage for saving large amounts of streaming data and to be able to easily reuse required information the following structure of database was created to facilitate the steps involved in the next phases of this analysis process. Following figure 4 is representing different fields of the database designed for this process.

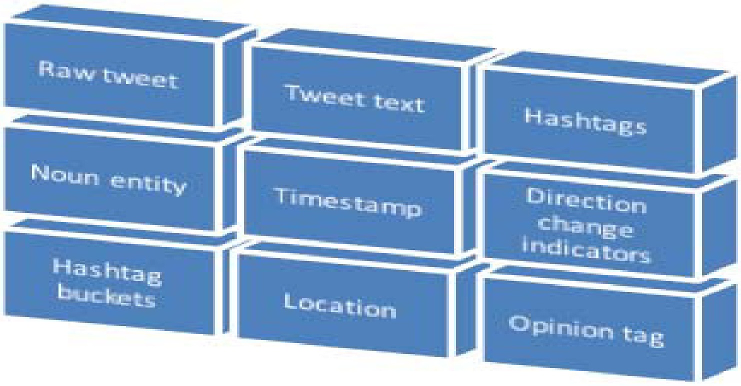


Figure 4 Structure of the database to store transformed information

In Figure 4 each block is representing a single field of the database. The first block of the figure is raw tweet; this column of the database contains the selected data in the form of raw tweets. It is exactly in the same form as it is extracted from twitter along with the date and time. After performing the data cleaning process it is forwarded to second phase, which is called Tweet text. It is the second column, which contains only tweet text cleaned from all emoticons punctuations and numbers only alphabets in lowercase. Hashtags column store all those words which were highlighted by the user and were appearing in tweets with the '#' symbol. All the nouns appeared in a tweet are stored in Noun entity field. Time and date of tweets are stored in Timestamp field. Direction change indicators are all those words which can change context of the whole text like "and", "or" etc. Hashtag buckets keep record of the quantity and impact of hash tags in the form of numbers. Location stores location of the tweet and the last block of the figure, which is Opinion tag, will store positive or negative opinion of the tweet after analysis at the end of whole process. After the database of the selected tweets is created the next step was collection and organization of the data set that can be used for training purpose. The details about this process are given in next section.

3.4 Collecting and organizing training data

The main sources of the training data were as follows:

- Selected, filtered and tagged twitter feeds covering major political opinions
- Word stat sentiment dictionary

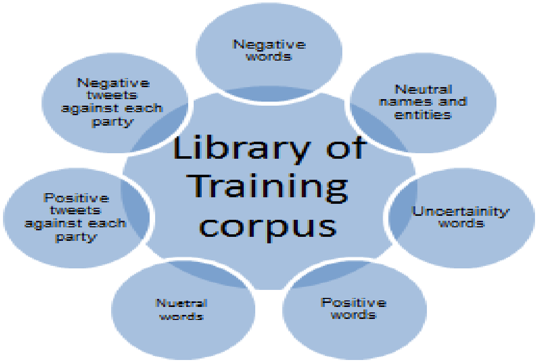

Negative words
Negative tweets against each party
Neutral names and entities
Library of Training corpus
Positive tweets against each party
Uncertainity words
Nuetral words
Positive words

Figure 5 Composition of training data

Forty per cent of twitter-extracted data was treated as first source of training data, which was used to make system learn for future analysis and predictions. That data was not used in its raw form instead it was transformed in to the specific format. The structure of the format is explained in above Figure 5. We gave labels to the tweets and tagged them on the basis of their polarity. The polarity could be positive and negative words used in them and what were the neutral entities, which were present in them.

Second source of training data was out self-built dictionaries of positive words, negative words, positive sentences, negative sentences and famous slangs of political parties.

Positive	• Positive indicator on topic
Neutral	• Neither positive nor negative indicators • Mixed positive and negative indicators • On topic, but indicator undeterminable • Simple factual statements • Questions with no strong emotions indicated
Negative	• Negative indicator on topic
Irrelevant	• Not English language • Not on-topic (e.g. spam)

Figure 6 Explanation of terms used in training data

The Figure 6 is giving details of the context and meaning of the taggers and labels used in training data. For example if a tweet has positive tag it mean it is giving positive opinion about some party. The neutral tag represents that this tweet neither gives a negative indication about some party or particular event nor have the positive opinion about any party. Negative means that tweet having this tag was discussing or expressing negative opinion about any political party and irrelevant told that this specific tweet is not in the political context. This data is garbage and not relevant to us.

3.5 Supervised sentiment taggers

To train our system, we used two different approaches and then compared their result to conclude which one performed better than the other. The general processes of classification of the tweets along with all the subtasks are described in Figure 7.

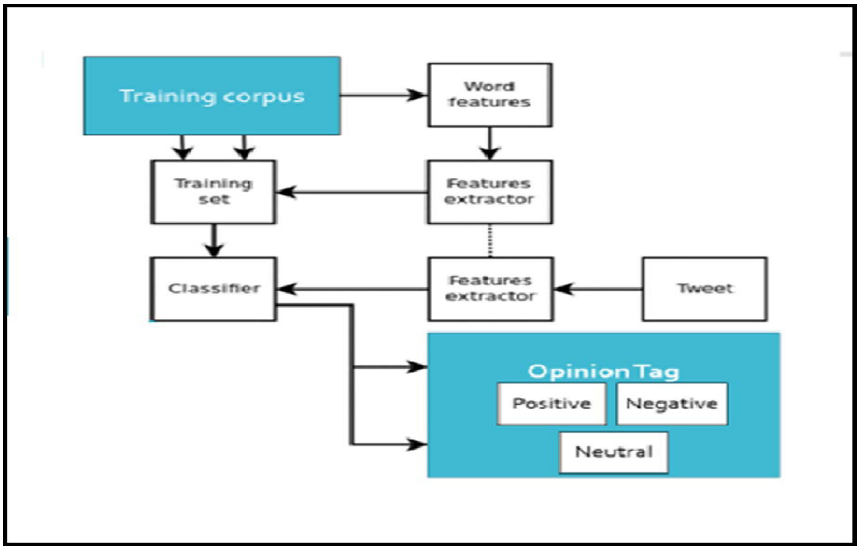
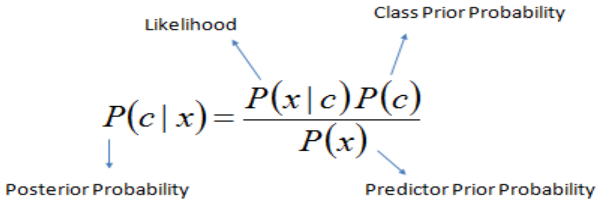


Figure 7 Details of the classification process

We first used Naive Bayes (McCallum Andrew & Kamal Nigam, 1998) approach, which used our training data's results to tag test data's tweets as positive negative or neutral. The Naive Bayes classifier uses the prior probability of each label, which is the frequency of each label in the training set, and the contribution from each feature. The algorithm operates on the following principle (see Figure 8):



$$P(c \mid x) = \frac{P(x \mid c) P(c)}{P(x)}$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Figure 8 Mathematical description of used Naive Bayes algorithm

The model of Naïve Bayes classifier was fed with the created trained data and after that classifier was trained to work on test data which was our remaining sixty per cent extracted data. After getting results from this classifier we would be able to use transformed tweet text's results as the basis of producing overall sentiment opinion.

The second approach was used to train system in comparison of Naïve Bayes was Support Vector Machines (Simon Tong & Daphne Koller, 2002). SVM's on the basis of features of the data predict a tweet either positive or negative and if it doesn't lie under these two tags declare it as neutral along with political party name about which tweet was posted. We used one verses rest multi category support vector machines method. In this method we trained one classifier for every political party on the basis of party's negative and positive slogans, leaders names hashtags were used in training. First we plotted all the data points in an *n* dimensional plane, where *n* was the total number of features which was in our case and then draw decision boundary to classify all points in different classes, where each class was representing a different political party. Following Figure 9 is showing pictorial presentation of this model.

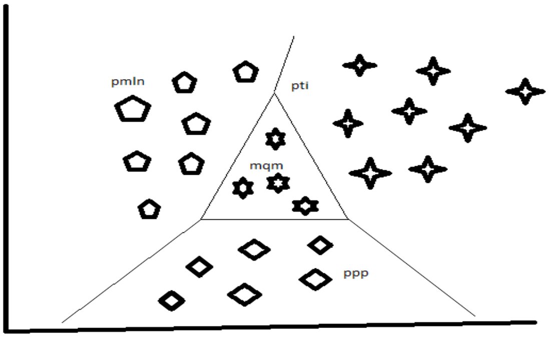


Figure 9 Working of SVM model

The training data was used to train this model, which became capable of classifying unseen test data after training. The results of this classification could be used further for opinion mining task or data analysis.

4. Results

The above stated model was developed using Python’s natural language processing for programming core, jQuery for statistical visualization and Sqlite database for storing data. Using this tool we extracted a total of eighty thousand tweets. Fifty thousand tweets were used as test data and thirty thousand tweets to train the model. The results we got on the basis of those fifty thousand test tweets are showed by pie charts and bar graphs, which are given in below Figures 10 and 11. The results showing different facts about political opinions in Pakistan like positivity of public in different cities towards Pakistan Muslim League (PMLN) are shown by this graph in below Figure 10.

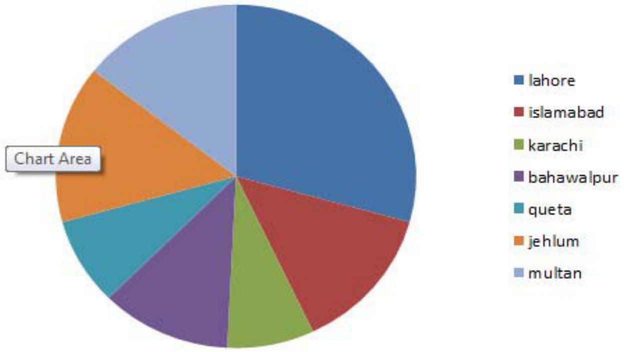


Figure 10 Visualization of positive opinion about PMLN is different cities

In the same manner another graph in following Figure 11 shows visualization of negative opinion of public for Pakistan Peoples Party (PPP) in different cities. It is showing that Lahore city have maximum negative opinion about PPP. This information is giving us insight about a lot of facts, for example that PPP have most vulnerability to their status in this city and needed to work hard in this city. Moreover, other parties can see this fact from a different angle that it is easier for them to get a lead as compared to PPP.

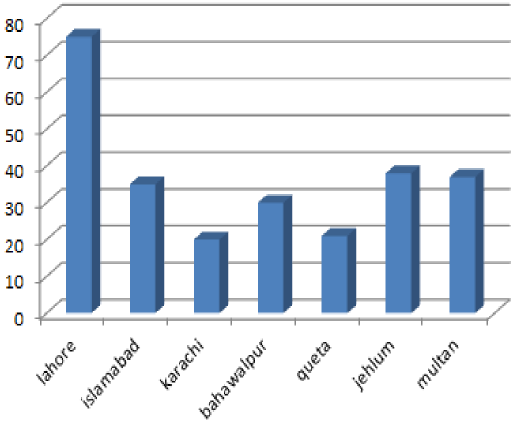


Figure 11 Magnitude of negativity about PPP in different cities

The comparison of results of both SVM and Naïve Bayes are shown in following Figure 12. This clearly shows that SVM is better than the Naive Bayes in terms of classification accuracy. Naive Bayes has an advantages in its simplicity with respect to SVM, while SVM is bit more complex in classifying the tweets, but overall in our case SVM perfomed better and gave better accuracy than Naive Bayes.

5. Conclusion and Future Work

In this paper pre-processing of raw data for a data analysis approach is presented that extracts quantitative and qualitative information from the social media text selected in the form of tweets in a specific date and time. The selected dataset is then transformed into more useful structured data. Using Twitter, the most popular micro blogging platform, the presented approach aims to complete the task of pre-process data for the purpose of opinion mining with the help of linguistic analysis and opinion classifiers which will together determine positive, negative and neutral sentiments for any given political party or any event during a specified timeline in Pakistan. This is an effective technique, which will aspire to convert raw data into useful transformed form to be used for the political scenario analysis, use for political scenario analysis, to promote awareness and improvement in systems. Moreover, with this we aim to give general public an important consolidated voice in the realm of politics. Main features of the presented technique are to extract tweets from Twitter along with their time date and location, then cleaning their form useless emoticons punctuations. After cleaning process, tweets are eventually stored in a database. Then, parsing on the whole text into useable datasets is performed with respect to the fields like hashtags, direction words and hashtag counts. Finally, two classification techniques SVM and Naive Bayes are compared for tweets classification. We concluded that SVM performed better in this case.

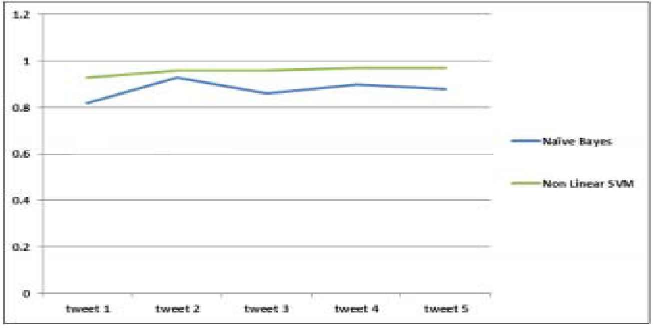


Figure 12 Comparisons of SVM and Naive Bayes

Future work can also be done on the extraction of data from other social media platforms like Facebook and Instagram. In the next phase the focus will also be given to multimedia data along with the textual information.

References

1. A. Go and L. Huang. Twitter sentiment analysis using supervision, in Proceedings of the Workshop on Languages in Social Media 2011; 3-38.
2. F. S. Gharehchopogh and Z. A. Khalifelu. Analysis and evaluation of unstructured data .in proceedings of 5th International Conference on Application of Information and Communication Technologies (AICT); 2011.p. 1-4.
3. C. D. Manning and H. Schutze. *Foundations of statistical natural language processing* MIT Press; 1999.
4. G.Mishne. Experiments with mood classification in blog posts. In 1st Workshop on Stylistic Analysis Of Text For Information Access 2005.
5. 5.Changhua Yang, K. H. Y.H. Emoticon classification using web blog corpora. In Proceedings of international conference on web intelligence 2007; 61-67.
6. Moghaddam, Samaneh, and Martin Ester. *Opinion mining in online reviews: Recent trends.* Tutorial at WWW2013; 2013.
7. Hatzivassiloglou,V. Predicting the Semantic Orientation of Adjectives. In Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics 1997; 174-181.
8. Kim, Soo-Min, and Eduard Hovy. "Determining the sentiment of opinions. Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics; 2004.
9. Naveed Anwar & Ayesha Rashid. Feature Based Opinion Mining of Online Free Format Customer Reviews Using Frequency Distribution and Bayesian Statistic. International Journal of Computers 2013; 81: 31-38.
10. Witten, Ian H., and Eibe Frank. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann; 2005.
11. H.Kumar, S.Jain Analyzing Delhi Assembly Election Using Textual Context of Social Media. In Proceedings of the *Sixth International Conference on Computer and Communication Technology* 2015; 78-85
12. Tong, Simon, and Daphne Koller. "Support vector machine active learning with applications to text classification." *The Journal of Machine Learning Research* 2 (2002) ; 45-66.
13. McCallum, Andrew, and Kamal Nigam. A comparison of event models for Naive Bayes text classification. *AAAI-98 workshop on learning for text categorization*. Vol. 752; 1998.