



Twitter Data Analytics – Sentiment Analysis of An Election

N. Saranya¹, A. V. Gukanesh², G. Karthick Kumar³, K. Karthik Raja Kumar⁴

¹Assistant Professor, ^{2,3,4} Student

^{1,2,3,4}Department of Computer Science and Engineering,

Sri Shakthi Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India

ABSTRACT

Social media is used to analyze political campaigns, stock market, movies, medicines, agriculture etc. Twitter a microblogging website where users read and write millions of tweets on a variety of topics daily. This project attempts to analyze the sentiment of the people for the election candidates based upon the live opinions and emotions. The focus of our project is to assign the polarity to each tweet that is whether the user expresses a positive or negative opinion. With the tweets that are extracted, we try to find how frequent their emotions change. We are also trying to classify and differentiate the sentiments of the people before and after the election based on the tweets they upload. The location of the twitter user is used to classify the geographical area which in turn helps to analyze the emotions of people of different areas. Our project uses the Naïve-Bayes approach in R language and R Studio for processing the textual data.

Keywords: *Twitter, R, Sentiment Analysis, Naïve Bayes Algorithm*

INTRODUCTION:

The social media sites which are used to make frequent posts in short publicly or in a group of users are called as the microblogging sites. These kinds of sites are hundreds in number. Therefore, there is a huge collection of data on these sites. The rate of data increases rapidly. These data express the moods and sentiments of the people in a large amount. The benefits of social media analytics include anticipation of business opportunities and competitive advantage. Cost efficiency is a major benefit of data analytics. Therefore, implementing the technology will result in

positive advantages for organizations in a number of industries. Twitter is one of the largest social media platform and a microblogging site where numerous users upload data. Twitter allows its user to read and post Twitter messages of size 140 characters. The Twitter messages are called as tweets. Nearly 656 Million tweets are tweeted in a day. Users' idea is expressed as tweets which in turn expresses their mood. Twitter makes these utterances to be available in a data stream, which can be mined using mining algorithms. In this paper, we discuss the analysis of the sentiment of the twitter users for an election based on the hashtags and emojis they use to tweet.

Sentiment Analysis:

Sentiment Analysis is the process of determining the self-indulgent feelings of the people, that is whether their opinion about something is positive, negative or neutral. Sentiment Analysis which is also known as Opinion Mining is used to derive the attitude of the Speaker. Sentiment Analysis of a social media has a variety of Applications such as marketing, reviewing, customer service etc., For example, movie reviews can be analysed and reports can be generated which can be used to decide how far the movie reached the people.

Working Model:

The system uses Naïve Bayes Approach for text categorization. For the categorization of the text, Naïve Bayes classifiers assume that the effect of a variable value on a given class is independent of the values of other variables. This assumption is called as conditional independence. In this paper, we propose an approach involving both dictionary-based and corpus-based techniques which finds the semantic

orientation of the sentiments in the tweets. Emoticons, neutralization, negation handling and capitalization is also considered as they are the huge part of the modern internet language. To uncover the sentiments, we will first extract the opinion words from the tweets and then we find out their orientation, that is to decide whether the sentiment word reflects the feelings of the user.

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

Likelihood $\rightarrow P(x|c)$ Class Prior Probability $\rightarrow P(c)$
 Posterior Probability $\leftarrow P(c|x)$ Predictor Prior Probability $\leftarrow P(x)$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

Fig 1. Navie Bayes Approach

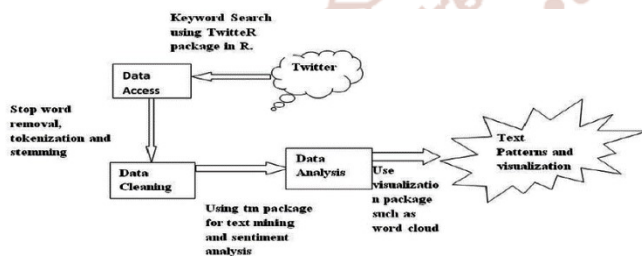


Fig 2. Working model

The following steps will brief the process of the proposed system which is discussed in this paper

1. Retrieval of tweets
2. Pre-processing of extracted data
3. Parallel processing
4. Sentiment scoring module
5. Output sentiment

1. Retrieval of tweets:

As Twitter is the most exaggerated part of social networking site, it consists of various blogs which are related to various topics worldwide. Instead of taking whole blogs, we will rather search on a particular topic and extract all the tweets related to that topic.

2. Pre-processing of extracted data:

After retrieval of tweets, Sentiment analysis tool is applied to raw tweets but in most of the cases results in very poor performance. Therefore, pre-processing techniques are necessary for obtaining better results.

i) Filtering:

Filtering is nothing but the cleaning of raw data. In this step, URL links (E.g. <http://twitter.com>), special words in twitter (e.g. “RT” which means Retweets), usernames in twitter (e.g. @kgk - @ symbol indicating a username) are removed and emoticons are replaced with special strings.

ii) Tokenization:

Tokenization is nothing but Segmentation of sentences. In this step, we will tokenize or segment text with the help of splitting text by spaces and punctuation marks to form a container of words.

iii) Construction of n-grams:

Set of n-grams can make out of consecutive words. Negation words such as “no”, “not” is attached to a word which follows or precedes it. For Instance: “I do not like jazz music” has two bigrams: “I do+not”, “do+not like”, “not+like jazz music”. So the accuracy of the classification improves by taking negation into account because it is a very common linguistic construction that affects polarity.

3. Parallel Processing:

Sentiment classifier which classifies the sentiments using Naïve Bayes Classifier where every database has hidden information which can be used for decision-making. The set of models are found by classification and this model can be used to predict the class of objects. It is a two-step process where the first step is Model Construction, in which the model is built from the training set and the second step is Model Usage which is used to classifying new data.

4. Sentiment Scoring Module:

The basic feature of this model is Polarity of the words. A dictionary which contains a list of English words and score which ranges from 1 to 3. The Scoring module is used to determine the sentiment of the textual data.

Polarity	Sentiment
1	Negative
2	Neutral
3	Positive

Table1. Polarity Values

The process involves the following steps:

5. Output the sentiment:

Page: 1602

The main objective of this paper was to describe and design system for classifying the sentiments of the people over a period of time using their tweets. It was developed using R and RStudio Big data processing technologies. We developed a set of analytical representation which helps the user to identify the data and can gain insights from it. We took a set of visualizations, implemented in shiny web applications.

REFERENCES:

1. Akshi Kumar and Teeja Mary Sebastian, "Sentiment Analysis on Twitter" IJCSI, Vol. 9, Issue 4, No 3, July 2012
2. G. Vinodhini, R. M. Chandrasekaran "Sentiment Analysis and Opinion Mining: A Survey" Volume 2, Issue 6, June 2012, IEEE paper.
3. Statistical Analysis with R Book by John M. Quick.
4. Web Application Development with R using Shiny by ChrishBeeley.
5. <http://dev.twitter.com/docs>.
6. <http://dev.twitter.com/overview/api>.
7. <http://www.r-bloggers.com/>
8. Go, R. Bhayani, L.Huang. "TwitterSentiment Classification Using DistantSupervision", Stanford University, Technical Paper, 2009
9. <http://apps.twitter.com/>
10. <http://www.bbc.com/news/world-asia-india>.
11. Use and Rise of social media as Election Campaign medium of India, Narasimhamurthy N, (IJIMS) 2014 vol.1 No.8 202-209.