International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 12 | Dec 2019 www.irjet.net p-ISSN: 2395-0072

SENTIMENT ANALYSIS OF TWITTER DATA ON GOVERNMENT INITIATED POLICIES

Umadevi Maramreddy[1], Srujana P[2], Krishna Vamsi P[3]

[1] *Professor, Department of CSE, Universal College of Engineering & Technology, Andhra Pradesh, India*
[2,3]*Student, Universal College of Engineering & Technology, Andhra Pradesh, India*

Abstract - *Sentiment analysis is the computational study of opinions, sentiments and emotions expressed in text. It's the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information from text. Sentiment analysis has a multitude of applications. It's used for social media monitoring, tracking of products reviews, analyzing survey responses and in business analytics. Twitter is an amazing micro blogging tool and an extraordinary communication medium.*

Sentiment analysis is in demand because of its efficiency – thousands of text documents can be processed for sentiment in seconds, compared to the hours it would take a manual human effort. A fundamental task in sentiment analysis is polarity detection: the classification of the polarity of a given text, whether the opinion expressed is positive, negative or neutral. So let us explore polarity detection using R, a programming language and software environment for statistical computing with a wide application in data analysis. The proposed work uses a supervised algorithm to build a classifier that will detect polarity of textual data and classify it as either positive or negative. It uses an opinionated dataset to train the classifier, data processing techniques to pre-process the textual data and simple rules for categorizing text as positive or negative which helps in getting overview of people in government initiative policies.

Key Words: *Sentiment, Polarity, Classification*

1. INTRODUCTION

In last couple of years of years, the social medium Twitter has become more and more popular. Since Twitter is the most used microblogging website with about 500 million users and 340 million tweets a day, it is an interesting source of information. The messages, or in Twitter terms the tweets, are a way to share interests publicly or among a defined group. Twitter distinguishes itself from other social media by the limited message size. The maximum size of 140 characters restricts users in their writing [1]. Twitter is therefore challenging their users to express their view in one or two key sentences. Because Twitter is widely adopted, it can be seen as a good reflection of what is happening around the world.

Among all that happens, the latest trends are most interesting for companies. The latest trends can be analyzed and when identified, reacted to. From a marketing point of view, these latest trends can be used to respond with appropriate activities, like product advertisements. Analyzing tweets can therefore be a goldmine for companies to create an advantage to competitors.

One interesting group are tweets expressing sentiments about products, brands or services. These messages contain an opinion about a specific subject. The sentiment of this opinion can be classified in different categories. An obvious example of three categories is the categories positive, neutral and negative. These categories have been studied a lot in the literature. The whole process of identifying and extracting subjective information from raw data is known as sentiment analysis. The sentiment analysis research field is closely related to the field of natural language processing. Natural language processing tries to close the gap between human and machine, by extracting useful information from natural language messages. In this research, the extraction of the sentiment from a tweet is studied.

In data analysis, algorithms have been developed which can be used to analyze data, with the goal to extract useful information. . Some widely used classification algorithms from the literature, such as Naive Bayes and Support Vector machines, will be applied and compared. This research will focus on the application of sentiment analysis to Twitter and comparing the performance of different classification algorithms on this problem. The main question of this paper is: ―How can twitter messages are accurately classified with respect to their sentiment? To answer this question, a sentiment analysis tool is implemented, which provides a framework for testing the quality of the algorithms. This tool provides a way to query the sentiment about popular policies of government like GST, Make in India etc. Make in India is an initiative launched by the Government of India to encourage multi-national, as well as national companies to manufacture their products in India. It was launched by Prime Minister Narendra Modi on 25 September 2014.

2. LITERATURE SURVEY

In recent years a lot of work has been done in the field of "Sentiment Analysis on Twitter "by number of researchers. In its early stage it was intended for binary classification which assigns opinions or reviews to bipolar classes such as positive or negative only.

In [3] proposed a solution for sentiment analysis for twitter data by using distant supervision, in which their training data consisted of tweets with emoticons which served as noisy labels. They build models using Naive Bayes, MaxEnt and Support Vector Machines (SVM). Their feature space consisted of unigrams, bigrams and POS. They concluded that SVM outperformed other models and that unigram were more effective as features.

In work [4] proposed a model to classify the tweets as objective, positive and negative. They created a twitter corpus by collecting tweets using Twitter API and automatically annotating those tweets using emoticons. Using that corpus, they developed a sentiment classifier based on the multinomial Naive Bayes method that uses features like N-gram and POS-tags. The training set they used was less efficient since it contains only tweets having emoticons.

A 3-way model for classifier [5] is developed to categorize sentiment into positive, negative and neutral classes. They experimented with models such as: unigram model, a feature based model and a tree kernel based model. For tree kernel based model they represented tweets as a tree. They arrived on a conclusion that features which combine prior polarity of words with their parts-of-speech (pos) tags are most important and play a major role in the classification task.

Twitter API [6] is used to collect twitter data. Their training data falls in three different categories (camera, movie, mobile). The data is labelled as positive, negative and non-opinions. They also eliminated useless features by using the Mutual Information and Chi square feature extraction method. Finally, the orientation of a tweet is predicted as positive or negative.

3. SENTIMENT ANALYSIS

In the proposed system, first we create a twitter application by logging into twitter website and obtain secret key and access token that are responsible for storing tweets[2]. From the twitter database, the user retrieves the necessary tweets in order to perform sentiment analysis. In order to classify the tweets, we use naïve Bayes or SVM classifier. Performing sentiment analysis using text is a difficult task. The text may be sarcastic that may lead to the sentiment FP. In the same way sentiment such as TP FN FP may also be present. Here, we can consider the automated tweets of size 2000 and perform sentiment analysis in R language the sentiment analysis includes text collection, pre-processing, analysis, validation. Advantages of proposed system is fast and good initial accuracy.

3.1. Introduction to Sentiment Analysis

Sentiment analysis can be defined as a process that automates mining of attitudes, opinions, views and emotions from text, speech, tweets and database sources through Natural Language Processing (NLP). Sentiment analysis involves classifying opinions in text into categories like "positive" or "negative" or "neutral". It's also referred as subjectivity analysis, opinion mining, and appraisal extraction. The words opinion, sentiment, view and belief are used interchangeably but there are differences between them.

Opinion: A conclusion opens to dispute (because different experts have different opinions)
View: subjective opinion
Belief: deliberate acceptance and intellectual assent
Sentiment: opinion representing one's feelings
An example for terminologies for Sentiment Analysis is as given below,
<SENTENCE> = The story of the movie was weak and boring
<OPINION HOLDER> =<author>
<OBJECT> = <movie>
<FEATURE> = <story>
<OPINION >= <weak><boring>
<POLARITY> = <negative>

Sentiment Analysis is a term that includes many tasks such as sentiment extraction, sentiment classification, and subjectivity classification, summarization of opinions or opinion spam detection among others. It aims to analyze people's sentiments, attitudes, opinions emotions, etc. towards elements such as, products, individuals, topics, organizations, and services.

3.2. R tool

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering …) and graphical techniques, and is highly extensible. R is an integrated suite of software facilities for data manipulation, calculation and graphical display.

One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.

R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.

The capabilities of R are extended through user-created packages, which allow specialized statistical techniques, graphical devices import/export capabilities, reporting tools etc. These packages are developed primarily in R, and sometimes in Java, C, C++, and FORTRAN. Some of the packages are:

twitteR: twitteR is an R package which provides access to the Twitter API. Most functionality of the API is supported, with a bias towards API calls that are more useful in data analysis as opposed to daily interaction.

ROAuth: Provides an interface to the OAuth 1.0 specification allowing users to authenticate via OAuth to the server of their choice.

RCurl: A wrapper for 'libcurl' Provides functions to allow one to compose general HTTP requests and provides convenient functions to fetch URIs, get & post forms, etc. and process the results returned by the Web server. This provides a great deal of control over the HTTP/FTP/... connection and the form of the request while providing a higher-level interface than is available just using R socket connections.

Sentiment: Analyses sentiment of a sentence in English and assigns score to it. It can classify sentences to the following categories of sentiments:- Positive, Negative, very Positive, very negative, Neutral or Sarcasm. For a vector of sentences, it counts the number of sentences in each category of sentiment. In calculating the score, negation and various degrees of adjectives are taken into consideration. It deals only with English sentences.

Plyr: A set of tools that solves a common set of problems: you need to break a big problem down into manageable pieces, operate on each piece and then put all the pieces back together.

ggplot2: A system for 'declaratively' creating graphics, based on "The Grammar of Graphics". You provide the data, tell 'ggplot2' how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details.

Wordcloud: creates word clouds of the results.

RColorBrewer: colour schemes for the plots and wordcloud

httpuv: Provides low-level socket and protocol support for handling HTTP and WebSocket requests directly from within R. It is primarily intended as a building block for other packages, rather than making it particularly easy to create complete web applications using httpuv alone.

4. TWITTER MINING FOR SENTIMENT ANALYSIS

In this proposed work we are working with three modules they are:

1. Authorization
2. Pre processing
3. Classification and graphical representation of result.

4.1. AUTHORIZATION

Step 1: Log on to https://apps.twitter.com/ Just use your normal Twitter account login.

Step-2: Click on the "Create New App" button. Choose your own application name, and your own application description. The website needs to be a valid URL.

Click "Yes, I agree" for the Developer Agreement, and then click the "Create your Twitter application" button. Go to the "Keys and Access Tokens" tab then look for the Consumer Key and the Consumer Secret. We will use these keys later in our R script, to authorize R to access the Twitter API. You will find the "Your Access Token" section. Click on the button labeled "Create my access token". Look for the Access Token and Access Token Secret. We will use these in the next step, to authorize R to access the Twitter API.

Step 3: Authorize R to Access Twitter

First we need to load the Twitter authorization libraries. We used the pacman package to I install and load my packages. The other packages we use are:

TwitteR: which gives an R interface to the Twitter API.
ROAuth: OAuth authentication to web servers.
RCurl : http requests and processing the results returned by a web server

Step 4: Click the "Authorize app" button, and you will be given a PIN. Copy the PIN to the clipboard and then return to R, which is asking you to enter the PIN. Paste in, or types, the PIN from the Twitter web page, then click enter. R is now authorized to run Twitter searches. You only need to do this once, but you do need to use your token strings and secret strings again in your R search scripts. Install the Sentiment Package Note that we only have to download and install the sentiment package once.

STEP 5: Create a Script to Search Twitter finally we can create a script to search twitter. The first step is to set up the authorization credentials for your script.

This requires the following packages:

TwitteR : which gives an R interface to the Twitter API

Sentiment: classifies the emotions of text plyr: for splitting text
ggplot2: for plots of the categorized results wordcloud: creates word clouds of the results

RColorBrewer: colour schemes for the plots and wordcloud httpuv: required for the alternative web authorization process

RCurl: http requests and processing the results returned by a web server

4.2. PRE PROCESSING

A tweet contains a lot of opinions about the data which are expressed in different ways by different users . The twitter dataset used in this survey work is already labeled into two classes viz. Negative and positive polarity and thus the sentiment analysis of the data becomes easy to observe the effect of various features. The raw data having polarity is highly susceptible to inconsistency and redundancy.

Preprocessing the data is done by cleaning and preparing the text for classification. Online texts contain usually lots of noise and uninformative parts such as HTML tags, scripts and advertisements. In addition, on words level, many words in the text do not have an impact on the general orientation of it. Keeping those words makes the dimensionality of the problem high and hence the classification more difficult since each word in the text is treated as one dimension [5]. To reduce the noise in the text should help improve the performance of the classifier and speed up the classification process, thus aiding in real time sentiment analysis. The whole process involves several steps: online text cleaning, white space removal, expanding abbreviation, stemming, stop words removal, negation handling and finally feature selection. Features in the context of opinion mining are the words, terms or phrases that strongly express the opinion as positive or negative. This means that they have a higher impact on the orientation of the text than other words in the same text.

4.2.1. Tokenize

Tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes input for further processing such as parsing or text mining.

4.2.2. Stemming

Stemming is the process of reducing inflected words to their word stem; base or root forms generally a written word form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root.

4.2.3. Filtering

Repeated words like good to show their intensity of expression are eliminated as they are not present in the sentiwordnet hence extra letters in the word must be eliminated. This elimination follows the rule that a letter can't repeat more than three times.

- Questions: Questions such like which, how, what etc. are not going to contribute to polarity hence in order to reduce the complexity, such words are removed.
- Removing Special Characters: Special characters like () {} [] etc. should be removed in order to eliminate discrepancies during assignment of polarity. For example "it's good" means if the characters are not removed may concatenate with the words and make those words unavailable in the dictionary.
- Removing Retweets: Many people may copy another person's tweets and retweet using a different account. This happens if he likes another user's tweet.
- Removing Urls: Generally Urls does not contribute to analysis of the sentiment in informal text e.g. "I have logged into www.ecstacy.com

4.3. CLASSIFICATION AND GRAPHICAL REPRESENTATION OF RESULT

Proposed work presents sentiment analysis to classify emotion function is from the sentiment package and "classifies the emotion (e.g. anger, disgust, fear, joy, sadness, surprise) of a set of texts using a naive Baye's algorithm. We will get a histogram between the number of tweets with each emotion that tells us whether the tweet is positive or negative. Finally, the words in the tweets, and create a word cloud that uses the motions of the words to determine their locations within the cloud.

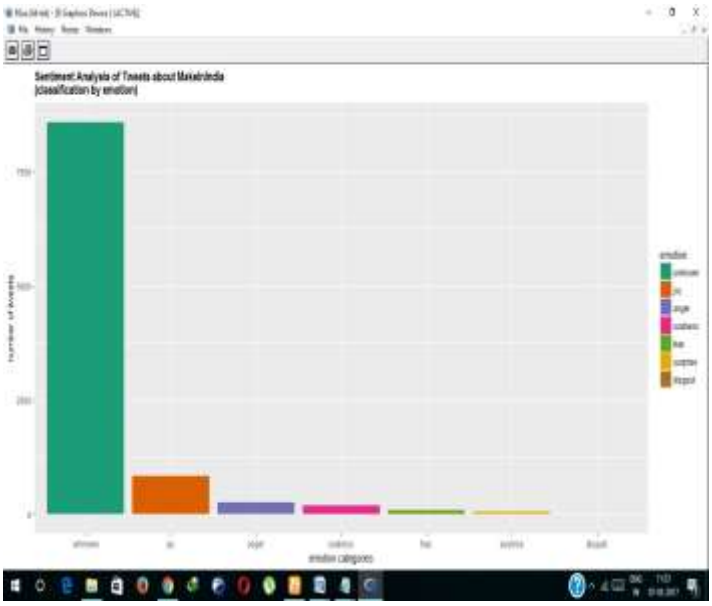

Fig -1: Sentiment Analysis of Tweets about Make in India, Classification by emotion

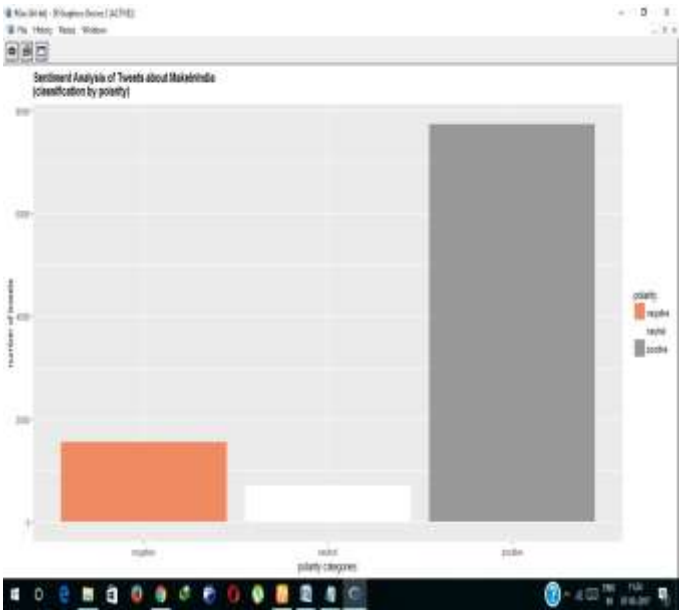

Fig- 2: Sentiment Analysis of Tweets About Make in India, Classification by Polarity.

Sentiment Analysis of Tweets about Make in India program using Bayes classifier is shown in Fig-1 and Polarity wise is shown in Fig-2. The same is represented as word Cloud in Fig-3.

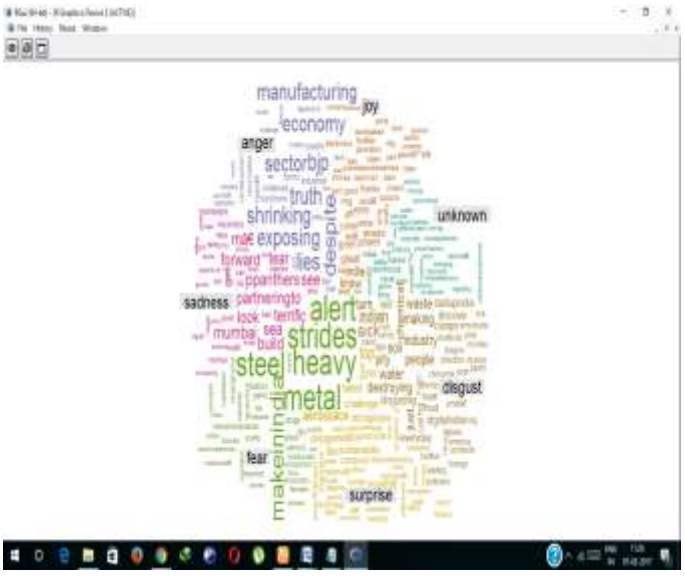

Fig- 3: Word Cloud of Tweets about Make in India

The sentiment analysis of tweets on GST is performed using SVM classifier and the results are in Fig-4 as classification by emotions and it is represented as classification by polarity in Fig-5.

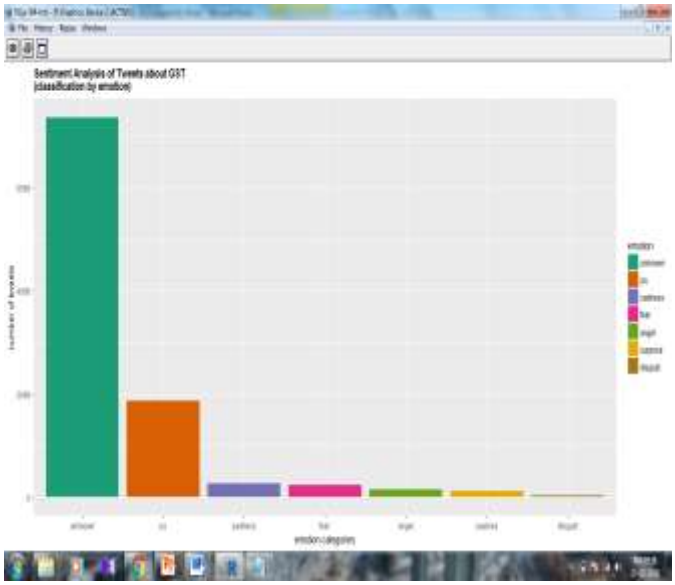

Fig- 4: Sentiment Analysis of Tweets about GST, Classification by emotion

Fig- 5: Sentiment Analysis of Tweets about GST, Classification by Polarity

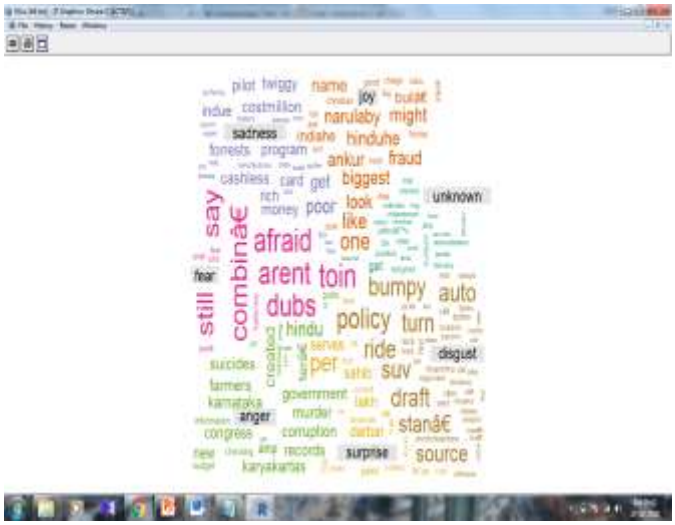

Fig-6: Word Cloud of Tweets about GST

These results are based on Naïve Bayes classifier for Make in India program and SVM classifier for GST are used. Based on these plots it is easy to conclude the sentiments of people about the policy is whether positive or negative. Further work is extended for accuracy of each classifier and to identify better classifier for sentiment analysis.

REFERENCES

[1] Kiplagat Wilfred Kiprono, Elisha Odira Abade, Comparative twitter sentimental Analysis based on Linear and Probabilistic Models, International Journal on Data Science and Technology, Volume 2, issue 4, july2016 Pgs:41-45,

[2] Shamanth Kumar, Fred Morstater and Huan Liu, Twitter Data Analytics, Springer, Aug 2013.

[3] Alecgo, RichaBhayani, Lei Huang, Twitter sentiment Classification using Distant Supervision, Processing 2009.

[4] Pak, Alexander and Paroubek, Patrick. Twitter as a Corpus for Sentiment Analysis and Opinion Mining, Volume 10, Proceedings of LREC, 2010

[5] Abbasi, A., France, S., Zhang, Z., and Chen, H. Selecting attributes for sentiment classification using feature relation networks. IEEE Transactions on Knowledge and Data Engineering, 23(3), 2011 pp. 447--462.

[6] Pablo Gamallo, Marcos Garcia, Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets, 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, Aug 23-24 2014, pp 171-175.

[7] Sindhwani, Prem Melville, Document-Word Co-Regularization for Semi supervised Sentiment Analysis , Vikas Business Analytics and Mathematical Sciences, IBM T.J. Watson Research Center, Yorktown Heights, NY 10598{vsindhw, pmelvil}@us.ibm.com.

[8] Caro Luigi Di, Grella Matteo. Sentiment analysis via dependency parsing. Comput Stand Interfaces 2012.

[9] Liu B. Sentiment analysis and opinion mining. Synth Lect Human Lang Technol 2012.

[10] Pang B, Lee L. Opinion mining and sentiment analysis. Found Trends Inform Retriev 2008; 2: 1–135.

[11] Pang NingTan, Michael Steinbach, Vipin Kumar Introduction to Data Mining, Pearson Edition.

[12] R for Beginners Text Book- Springer.

[13] Apoorva Agarva, BoyiXie, Ilia Vovsha, Owen Rambow and Rebecca, Passoneaua, Sentiment Analysis of Twitter Data, Proceedings of the Workshop on Languages Social Media in, Portland, Oregan, 2011, pages 30-38.