ISSN: 0975-766X
CODEN: IJPTFI
Available Online through *Research Article*
www.ijptonline.com

SENTIMENT ANALYSIS ON "EBOLA" OUTBREAK USING TWITTER DATA

A. Beryl Joylin, Aswathi T, Suma P, Nancy Victor
School of Information Technology and Engineering, VIT University, Vellore.
Email: beryljoylin@gmail.com

Received on 25-10-2016 *Accepted on 02-11-2016*

Abstract

Sentiment analysis is a process of identifying and extracting subjective information in source materials by performing text analysis and Natural Language Processing. It mainly focuses on determining the writer's attitude towards a particular topic is positive negative and neutral.In today's modern world, when a disease break out in a place, the micro-blogging site twitter will be flooded with tweets related to the disease. Performing sentiment analysis on the tweets related to the disease gives a insight about the impact of disease. The main objective of this work is to perform sentiment analysis on tweets related to the massive killer disease 'Ebola'. The tweets are classified into positive, negative and neutral by using Natural Language Processing and bag of words. Two bag of words, one containing positive words and another containing negative words are used for the classification purpose. The Natural Language Processing is performed on tweets to remove unnecessary characters and to filter out necessary words and these filtered words are compared with the bag of words to achieve classification of tweets. A graphical plot against the date of tweet and number tweets for positive, negative and neutral tweets provides a visualization about impact of the disease on people in their day to day life.

Key words: Sentiment analysis, Twitter, Natural Language Processing, Classification.

Introduction

Over a decade ago, all transactional systems were using relational databases. Relational databases has predefined data model and the storage of data in such systems is defined by fixed schemas. Applications like Twitter, Facebook, LinkedIn etc. started emerging between the years 2003 and 2010. The data generated by these applications are highly complex and unstructured. Also, the growth of this kind of data is exponential over time. Such exponentially growing data is termed as big data. This exponentially growing data can be structured or unstructured and cannot be handled by

any traditional database management systems. Structured data has a predefined data model whereas, unstructured data do not have any predefined data model. Twitter allows users to express their thoughts, opinions and random happenings in their lives in the form of tweets. Tweets generated increase exponentially with an average of 5700 tweets per second. Thus, tweets generated by twitter can be concluded as big data and it is highly complex and semi-structured/unstructured. Informational value that can be gained from the tweets are very little but, an important knowledge can be gained by aggregating millions of tweets. Various studies on twitter have demonstrated that a valuable insight about a social problem or an event or a population can be attained by performing analytics on tweets. One of the ways of getting a valuable insight about a social problem or an event is by performing sentiment analysis on tweets. Sentiment analysis is a process of extracting subjective information in source materials (example: tweets) by performing natural language processing and text analysis. Performing sentiment analysis on health related tweets will provide a deep intuition about the impact of the disease on the community affected by the disease. This paper mainly focus on extracting information related to a disease from tweets, perform sentiment analysis and make predictions such as best treatment available for the disease, best drug available for the disease, etc.

Related Works

Over the recent years, many studies have proved that sentiment analysis of twitter messages can bring out valuable predictions. Presently, sentiment analysis on Twitter data is done mainly on four major areas:

1. To Determine customer sentiments on a Product
2. To Analyse public health
3. To predict elections
4. To classify the movie reviews

When a new product is introduced in the market, the customers react to that product by posting tweets. These tweets were analysed to identify the trend of that product in the social media [1]. The count of the keywords like good, bad, excellent were used to make a trend analysis on a product. The timely graph plotted between trend and time enabled marketing people to make various decisions. Also, according to another recent research it is possible to detect irrelevant reviews by using sentiment analysis [2]. Public health related tweets is analysed to perform various tasks like tracking illnesses over times, measuring behavioural risk factors, localizing illnesses by geographic region, and

analysingsymptoms and medication usage [3]. According to another recent research, analysis of around 500 flu related messages over an 8 month period to forecast future influenza rates have obtained 95% correlation with national health statistics [4]. Generally, people's interest on a political party during elections is expressed in the form of tweets. Sentiment analysis of tweets related to a political election has enabled to predict the political party that wins. Analysis of over 100,000 messages containing a reference to either a political party or a politician enabled to predict the winning political party during 2009 German Federal Election [5]. It is very common that movie viewers start expressing their likes and dislikes as soon as they watch a movie. Classifying the movie reviews into positive, negative and neutral ones is another important area where sentiment analysis is done. Number of positive reviews reveal the the success of the movie[5]. Moreover, a proper scalable and efficient storage of data(tweets) is a basic and important requirement for sentiment analysis. The most simple way of storing data is by using Comma Separated Values(CSV) format. Many past researches have used different storage techniques like MongoDB, NoSQL databases is used for more scalable and improved storage [6].

Existing System

Initially, sentiment analysis on twitter data was performed to know the customer sentiments on a product. When people get excited about a product, they react about the product by posting tweets in Twitter. Such tweets about a product were extracted and sentiment analysis was performed to know the customer reactions on the product. This idea is applied in the field of health informatics and sentiment analysis is performed on tweets related to public health related information. Tweets related to public health are extracted and tasks such as tracking illness, analysing symptoms and medication usage has been performed.

Proposed System

Twitter users publicly post their personal information like 'I am affected by cough', 'This flu makes my life hard' etc. Collecting large number of such tweets related a particular disease and analysing those tweets helps in making conclusions like how the people react to that disease, best treatment available for that disease etc. In the proposed system, 'n' number of tweets related to a particular illness is extracted and sentiment level of each tweet is identified by comparing the words in each tweet with a set of positive and negative words. The sentiment level of tweets are identified by 3 values: +1(positive), -1(negative) and 0(neutral). The total number of positive tweets, negative tweets and neutral

tweets for each day is calculated and a graph between date of tweet and number of tweets for each of the sentiment levels is plotted. Thus how the people react to that illness and how the reaction changes over time can be identified. These sentiments are used to make certain predictions like the best treatment available that cured the disease, the best drug the cured that disease. In this paper, the disease used for case study is EBOLA.

System Architecture

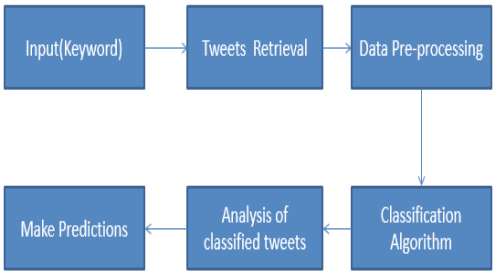


Figure no.1: System Architecture.

Figure no.1 shows the architecture of the proposed system. Tweets are extracted by giving an input keyword. The input keyword in this context is the name of an ailment. Tweets related to that ailment are retrieved, processed and classified based on sentiment levels. Sentiment level of each tweet is identified by comparing the individual words in the tweet with a set of positive and negative word. Each word in the tweet is given a sentiment score. The score can be +1(positive), -1(negative), and 0(neutral). The total score of each tweet is calculated and the tweets are classified as positive, negative, and neutral. These classified tweets are analysed to make valuable predictions.

Solution Methodology

The steps involved in implementing the proposed system are explained as follows:

A. Creating Twitter API

A twitter application is created by signing into twitter developers with twitter user id and password. After creating the application, a consumer key and consumer secret is generated. This generated consumer key and consumer secret is used for further authentication while extracting tweets [8].

B. Extract Tweets

R is a programming language used for data analysis. R has a standard set of packages which provides an interface to obtain authentication from twitter API and retrieve tweets. With the help of those packages, authentication is established and tweets are extracted. The extracted tweets are stored in Comma Separated Values (CSV) format [9] .

C. Create Word Cloud

A word cloud is generated from tweets by performing Natural Language Processing and text analysis. The word cloud is created for the purpose of quick visualization of most commonly cited words in a text [10], [11].

D. Compare Sentiments

Sentiment level of each tweet is identified by comparing the individual words in tweets with set of positive and negative words. Each word is given a score of +1(positive), -1(negative), or 0(neutral). The total score of the given tweet is determined by adding the score of individual tweets. Total number of positive tweets, negative tweets and neutral tweets for each day is calculated and a graph between date of tweet and number of tweets is plotted for positive, negative, and neutral tweets. This helps in identifying how the people react to an illness [12].

Results and Discussions

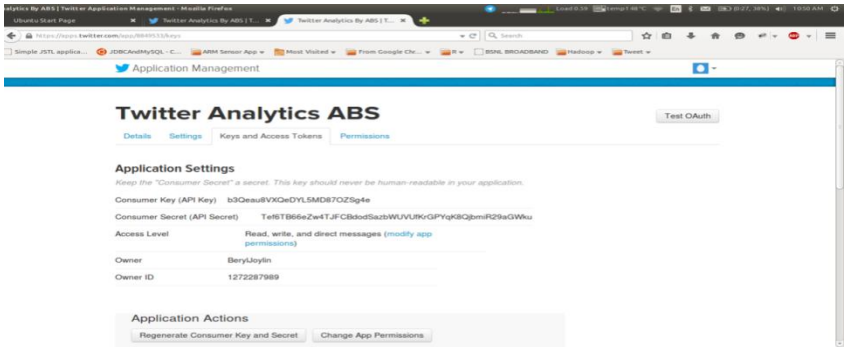


Figure no. 2: Twitter API

The Figure no. 2 shows the screenshot of the twitter API created. Consumer key and consumer secret obtained is shown in the Figure 2. This is used for the authentication during the extraction of tweets.

After extracting tweets, the tweets are stored in a CSV file. The CSV file contain fields such as date of extraction, time stamp of tweet created etc. The CSV file generated on extracting tweets is shown in Figure no. 3

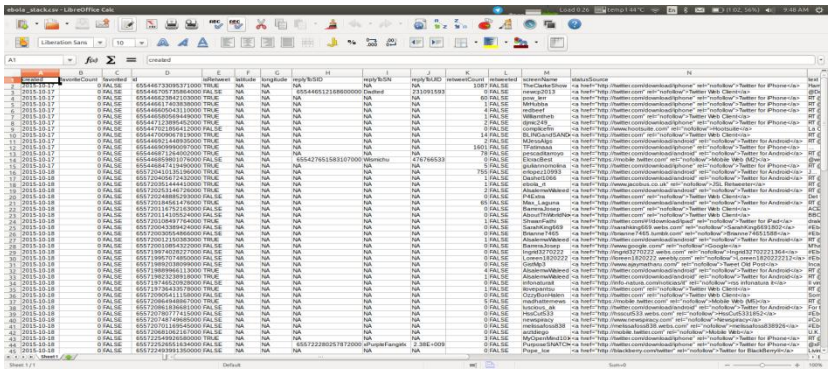


Figure no.3 Extracted tweets.

The Figure no. 4 shows the word cloud generated from tweets extracted. As the input keyword is EBOLA, the word that occur maximum number of times is Ebola. From the word cloud (Figure no. 4) it can be inferred that there are many cases of hospitalization, there are some people who are in tears, etc.

Figure no. 4: Word Cloud.

The Figure no. 5 shows the CSV file generated after assigning sentiment level to each tweets.

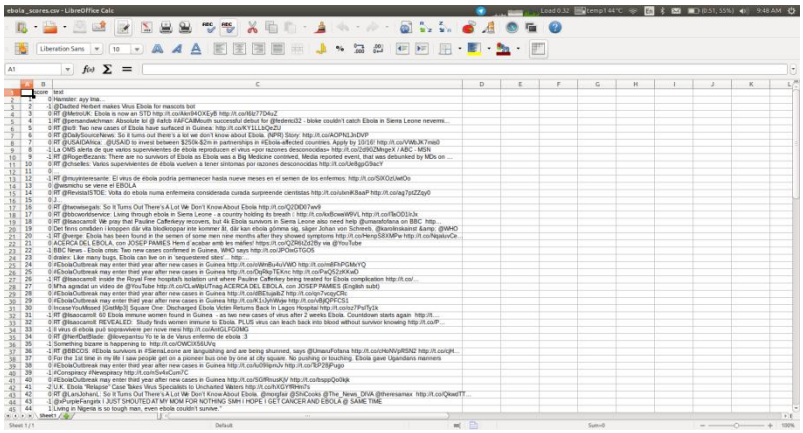
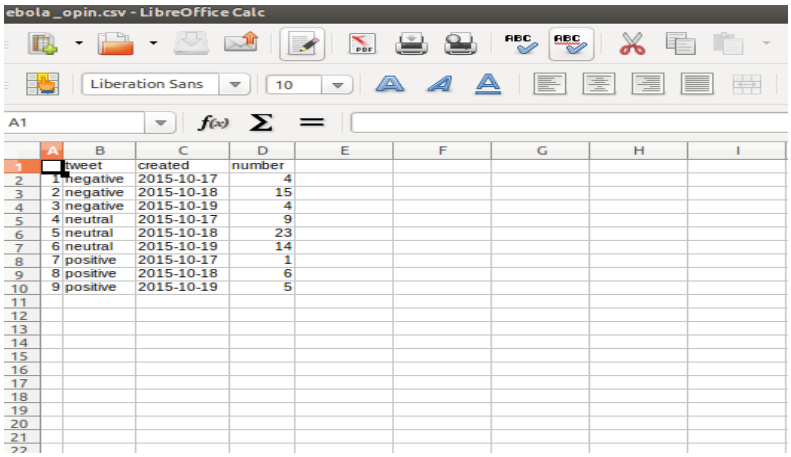


Figure no. 5: CSV File with sentiment level for each tweets.



	A	B	C	D
1		tweet	created	number
2	1	negative	2015-10-17	4
3	2	negative	2015-10-18	15
4	3	negative	2015-10-19	4
5	4	neutral	2015-10-17	9
6	5	neutral	2015-10-18	23
7	6	neutral	2015-10-19	14
8	7	positive	2015-10-17	1
9	8	positive	2015-10-18	6
10	9	positive	2015-10-19	5

Figure no.6: Number of Tweets under each sentiment level.

The total number of positive, negative and neutral tweets are calculated for each day. The CSV file with number of positive, negative, neutral tweets for each day is generated and the screen shot of this is shown in Figure no.6.

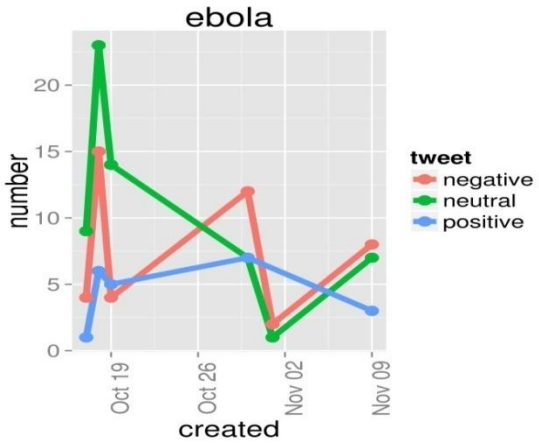


Figure no. 7: Graphical Representation of sentiment analysis on Ebola.

The graph between date of creation and number of tweets is plotted for each sentiment level. This is shown in Figure no. 7. This graph helps in getting an idea about the effect of Ebola on people.

Conclusion and Future Work

Tweets posted by the Twitter users convey a very little information. But, collecting millions of related tweets and analysing them will help in reaching valuable conclusions and predictions. A valuable insight about a disease can be attained by performing sentiment analysis on tweets. It helps in getting an idea about how the masses react to a disease over time. On the other hand, it also enables to make predictions such as best drug that is used to cure the disease, the geographical area which is most affected due to the disease etc. These predictions help people to access the best treatment, best drug etc. to a great extent.

This work can be extended to analyse information about many other diseases and make valuable conclusions on it. If a disease adversely affect a geographical area, the atmospheric data of that area can be linked with the analytics results of that disease. Thus, how the effect of a disease varies with atmospheric conditions could be identified. With these results, some predications such as how adversely disease can affect another area can be made.

References

1. Gaurav D, RajurkarR ,Rajeshwari M, Goudar G . *A speedy data uploading approach for Twitter Trend and Sentiment Analysis usingHADOOP*, India, 2015.

2. S Zol, P Mulay, "*Analyzing Sentiments for Generating Opinions (ASGO)-a new approach*", Indian Journal of Science and Technology, 2015, Doi no: 10.17485/ijst/2015/v8iS4/62327.
3. Paul M J ,Dredze M . *You Are What You Tweet: Analyzing Twitter for Public Health*, 2011, pp. 1-8.
4. CulottaC , Aa *A* . *Detecting influenza epidemics by analysing twitter messages* , USA, 2010, pp. 115-22.
5. TumasjanA ,SprengerT O , SandnerP G , Welpe I M . *Predicting elections with twitter: What 140 characters reveal about political sentiment*, 2010, pp. 1-8.
6. V. K. Singh; R. Piryani; A. Uddin; P. Waila, "*Sentimentanalysisof Moviereviewsand Blog posts*", Advance Computing Conference (IACC), 2013 IEEE 3rd International,Pages: 893 - 898, Doi: 10.1109/IAdCC.2013.6514345.
7. P. Parthiban, S. Selvakumar ,"*Big Data Architecture for Capturing, Storing, Analyzing and Visualizing of Web Server Logs* ",Indian Journal of Science and Technology,2016 Jan, 9(4), Doi no:10.17485/ijst/2016/v9i4/84173.
8. Twitter Analytics Using R Part 2: Create Word Cloud,https://www.credera.com/blog/business intelligence/twitter-analytics-using-r/, Date accessed:02/ 8/2016.
9. geoffjentry committed on GitHub,https://github.com/geoffjentry/twitteR, Date accessed: 10/08/2016.
10. Sentiment Analysis with "sentiment" sentiments. https://sites.google.com/site/miningtwitter/questions/sentiment/sentiment, Date accessed: 10/08/2016.
11. Sentiment Analysis on Twitter Data : Text Analytics Tutorialhttps://mkmanu.wordpress.com/2014/08/05/sentiment-analysis-on-twitter-data-text-analytics-tutorial/, Date accessed: 10/08/2016.
12. Twitter sentiment analysis with R,http://analyzecore.com/2014/04/28/twitter-sentiment-analysis/, Date accessed: 12/08/2016.