Twitter Crowd Mining and Data Fishing

Sherin Eliyas, R. Naveen, M. Siva Krishna

Abstract--- People use twitter as a medium to share their opinions, which in turn makes it a platform for analyzing public opinion. This kind of information in various fields can be used for further analysis. Text mining is used for classifying tweets into negative and positive statements depending on their purpose. Text messages are used to define mind set of large groups of people. From perspective of decision makers, precious information is provided by collection of text messages. In this paper, we use data collected from twitter and extract useful data from it using data mining tools. The results is represented as decision trees and this can be used to make further analysis or used to take decision for analyst. Here we evaluate or analyze the impact of tweets by using emotic icons for process.

Keywords--- Twitter, Crowd Mining, Data Fishing.

I. INTRODUCTION

In worldwide for market analysis, billions of dollars are spent. Data-driven decision is a necessary and powerful method for conducting business. Imagine how a business candidate will use this technique to analyze the data and increase his sale and a political party or a leader, uses this to increase his chances of winning in the election, without directly interacting with people or analyzing their views. This can be done using microblogging sites such as Twitter, Tumblr, Plurk, Pownce, and Jaikubu collecting people sentimental data and analyzing using public sentimental analysis. These micro blogging sites provides quick and easy way for people to analyze their opinion by collecting pool of data and by helping them to express their views keeping their opinion in mind. The information which is collected will be shared to their friend's circle and saved it in their own profile. This data gathered can be made as public, unrestricted or can be kept as private.

Subjectivity analysis, Opinion mining and sentiment analysis acts as a common goal of developing or applying data mining technology to execute collection of reviews, texts and opinions. Another aspect of developing is to do sentiment analysis for certain events, topics and for generating problem solving tools that can be used to classify and rank certain topics. These tools can be used to check and review events or specific movies and to give positive and negative impact of it and to define, whether it is in their favor or not. The growth on digital data is estimated to grow continuously by 40% annually. But, unfortunately, there is only less than 3% of it being analyzed in the year 2012. Analysis of Big data, which is also referred to as large amount of data, can be done using Data mining.

One variety or technique of data mining that can be used to extract information from organized data is Text Mining. Text mining focuses on textual data, this textual data is unstructured, and it is difficult to deal with algorithmically, though, text is commonly used for information transfer by the present society. Vast amount of textual data comes from email and social media. Public post real time messages and their opinion on variety of topics, creating it as a valuable platform for analyzing and tracking public opinion on current situation, including their opinion on the bus ticket price changes. Despite greater interconnectivity, shocks like food and water crises may not be immediately visible and cannot be easily traceable by traditional monitoring systems. So, it is often too late, and it is more expensive to respond. Timer in 1996 has examined that there is no country that can sustain its economic growth without first solving its food problem. The 2012 Economic Intelligent Unit had released Global Food Security Index has found that Indonesia's index on food security is below 50 on a scale of 0-100. One of the causes of this situation is due to increase in food commodities' prices. The average percentage of expenditure on food consumption is comparatively high in Indonesia i.e., 47.71% of the total income per capita spent for food consumption, change in the price of routine foods is a problem that needs to be examined.

II. EXISTING SYSTEM

There is body of research in micro blogging data especially for opinion mining. For those groups of persons who share a familiar interest such as music, movies, games, etc. opinion mining technique can be used. It is known that if the content gathered from tweet is brief, it contains required data to exhibit recognizable interests and characteristics. "Machine learning is an achievable tool to perform sentimental analysis using reviews of music or movie to a corpus" says Pang et al. Algorithms are of three types namely; Maximum entropy together with support vectors by naïve Bayes. Drawbacks of task compared to topic-based categorization were brought to light by this method. The work in Go et al is like technique used by Pang in which he used similar classifiers, which included microblogging data for large text movie reviewsfrom Twitter. These tools once applied for sentiment analysis cross the boundaries from 140 characters restricted tweets to longer text blocks. A neutral sentiment from corpora has been excluded from present paper. By keeping common emotic icons in twitter, search utility will collect accepted and rejected tweets.

Before training with the classifiers, the emotic icons are removed once the data required has been gathered.

Manuscript receivedJune 10, 2019.

SherinEliyas, MCA, School of Computing Science, Hindustan Institute of Technology and Science, Rajiv Gandhi Salai, Padur, Chennai, Tamil Nadu, India. (e-mail : Sherine@hindustanuniv.ac.in)

R. Naveen, MCA, School of Computing Science, Hindustan Institute of Technology and Science, Rajiv Gandhi Salai, Padur, Chennai, Tamil Nadu, India. (e-mail : Naveenraajj.r@gmail.com)

M. Siva Krishna, MCA, School of Computing Science, Hindustan Institute of Technology and Science, Rajiv Gandhi Salai, Padur, Chennai, Tamil Nadu, India. (e-mail : sivadeva33333@gmail.com)

Pak and Paroubek gathers collects information from Twitter, with the use of popular emotic icons such as smiley faces, sad faces and variations, and classifies it as negative tweets or positive. With help of newspaper accounts,impartial tweets are then collected to round out the corpora. Further analysis indicates issuance of letter frequencies in the group is normal. Naïve Bayes classifier is then applied to test posts. Analysis using bigrams gives the desirable output. As there will be huge collection of words for movie reviews users will spend more time thinking about the post which is posted, where users will be inclined to give screenshots of concept sent through cell phone. Theobservation which needs to be noted is that these gather frequent misspellings and amount of slang used by twitterians. It will lead to minor changes on any opinion analysis applied to micro blogging information. It is created to find positivity and negativity of a tweet using Corpus. While testing or training data no texts of neutral content will be included.

Grouping of Parameters:

- Used Classifiers
  - Naïve Bayes
  - Maximum Entropy
  - Support Vector Machine
- Use of emotic icons
  - Time of training information set
  - In time of testing information set
- Use of both unigrams and bigrams
- Word frequency versus presence
- Micro blogging vs Text data
- Word list size and source of negative/positive data
- Use of objective data
  - Data set - Training
  - Data set - Testing

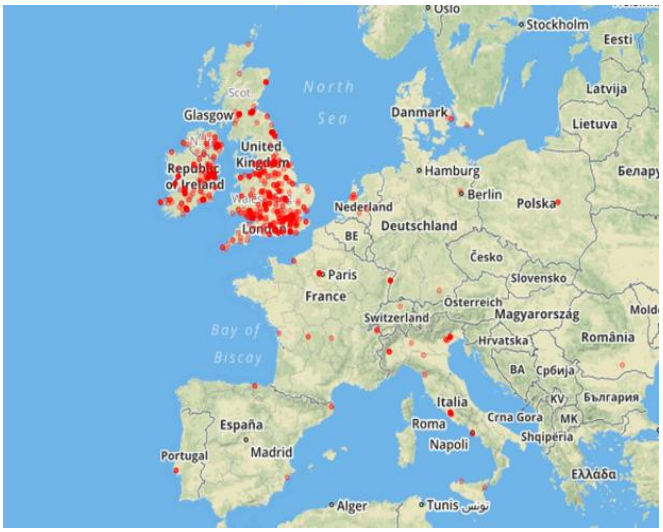
III. PROPOSED SYSTEM& RESULTS

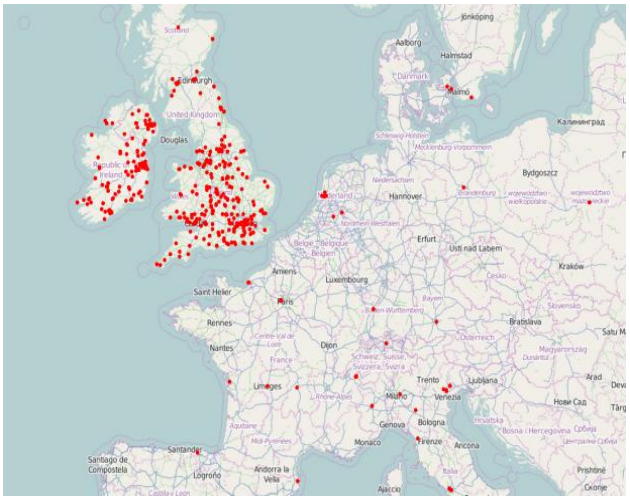
The process of extracting unstructured information from set of textual data is known as Text mining. The proposed system is that it is used to detect the following details from twitter

- Landmark:

It is used to detect the tweets location, which can be further used to detect landmark interests in that location.

Output

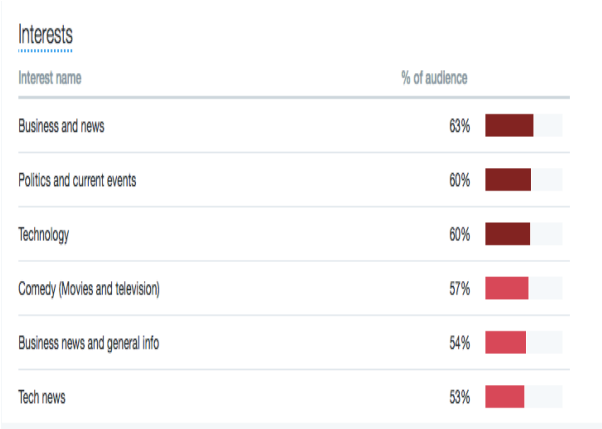




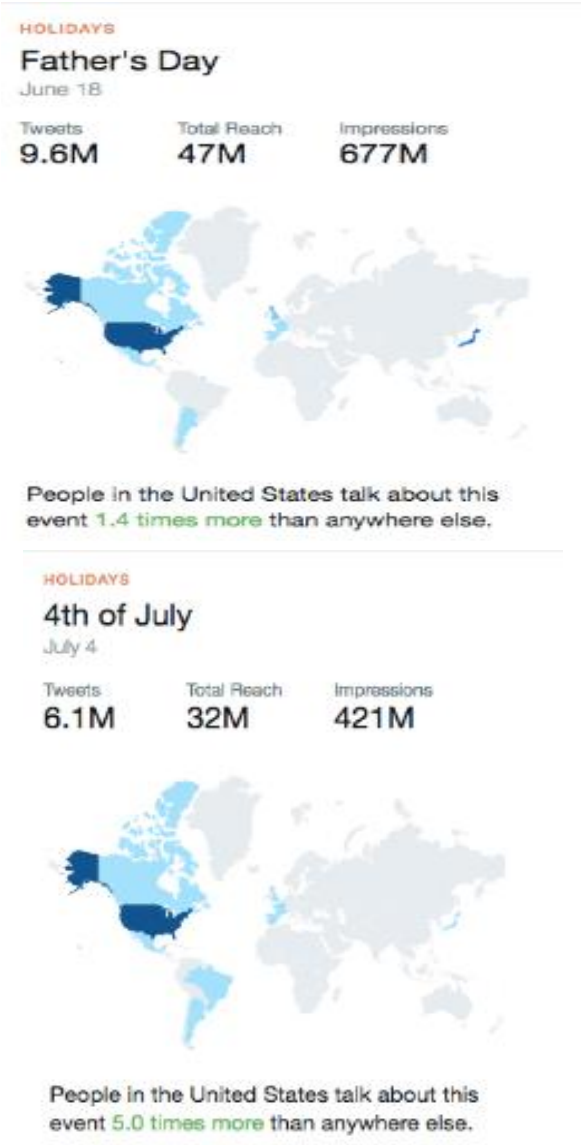
- Future Events Identification

By keeping tweets or data collected from different subjects, algorithms to predict future events will be developed.

Output

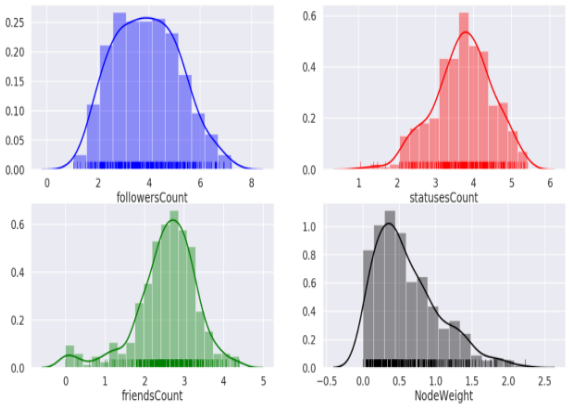


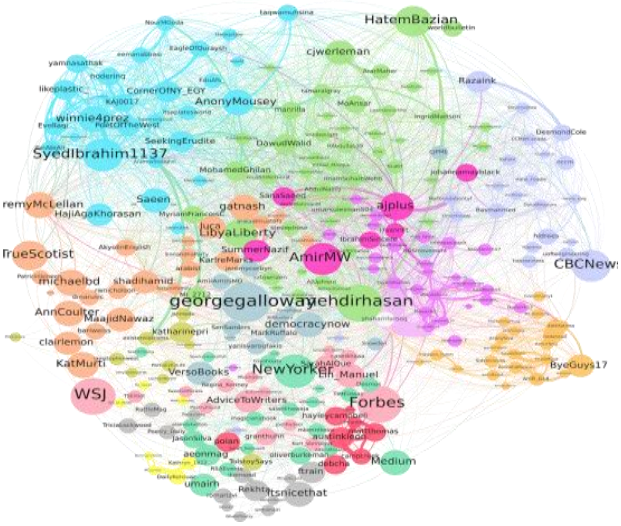
Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication



❖ Community Detection Measures:-
For improving Clustering results Hashtag, tweet and social connections will be used.

Output





Twitter was selected because of its reputation as the world's most popular micro blog.

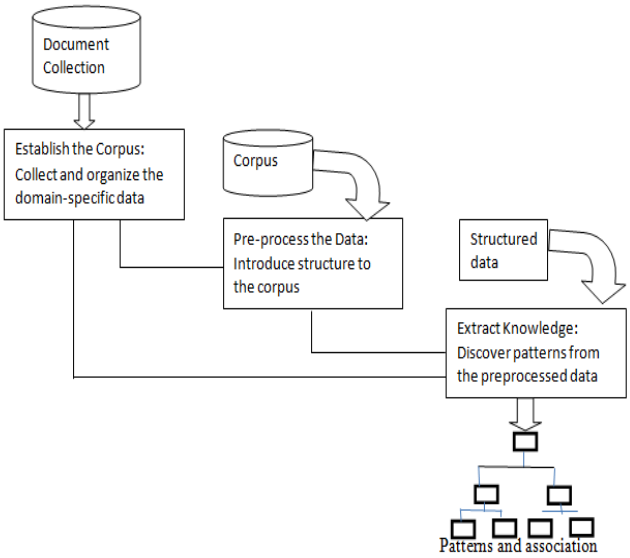


Figure 1: Text mining system diagram

Steps performed in the pre-processing stage along with the illustrations are as follows:

- *Tokenization*, which separates the text in the *tweet* into pieces of word called *token.*
- *Filtering*, that is the elimination of mention (@), hashtag (#), and RT (retweet) from the tweets.
- *Stemming*, that is taking root words by eliminating the word affixes and transforming them into their simplest form.
- *Case folding*, that is converting all lowercases to uppercase and vice versa.
- *Matrix*, this is a vector representing word tokens.

$$tf.idf(t,d) = tf(t,d) \cdot idf(t)$$

$$tf(t,d) = \sum_{i \in d}^{|d|} 1\{\mathrm{d}_i = \mathrm{t}\}$$

$$idf(t) = log \frac{|D|}{\sum_{d \in D}^{|D|} |\mathrm{t} \in \mathrm{d}}$$

Equationsproduce a term frequency-inverse document.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Before Pre-processing	@nav I look forward to read the study about Text Mining in the area of staple foods price changes topic
Tokenization	@nav I - look - forward - to - read - the - study - about - Text - Mining - in - the - area - of - staple - foods - price - changes – topic
Filtering	I - look - forward - to - read - the - study - about - Text - Mining - in - the - area - of - staple - foods - price - changes – topic
Stemming	I - look - forward - to - read - the - study - about - Text - Mining - in - the - area - of - staple - foods - price - changes – topic
Case folding	I - look - forward - to - read - the - study - about - Text - Mining - in - the - area - of - staple - foods - price - changes – topic
Matrix	Terms documents / i / Look / ... / topic 1st Tweet / 1 / 1 / ... / 1

IV. METHODS

Retrieve Tweets

With the help of twitter, we will collect data that is known as tweets in this document, from twitter database. Retrieving is nothing but collecting the necessary information or data that is required for further analysis. Here we will be collecting the data, which is stored in twitter for performing twitter data mining.

Cleaning the Data Collected

Once the required data has been gathered, the next step in the process is cleaning the data. Once the data has been collected, we should clean it i.e. the data gathered will be in different form and we should decide which format the data needto be for analyzing it. All the data should be in similar format for analysis, the collected data will or may contain the unnecessary data, which need to be filter out in these processes. Once the data is cleaned or it is free from errors, the next step can be performed.

Associationof Frequent Words

Once the data has been collected and cleaned, the next step will be association of frequent words. It is nothing but collecting the set of words which will fall into a single group. All the data which comes under one specific header will be collected or separated and it will be stored in a way that it can be easily understood and can be easily performed at the time of analysis

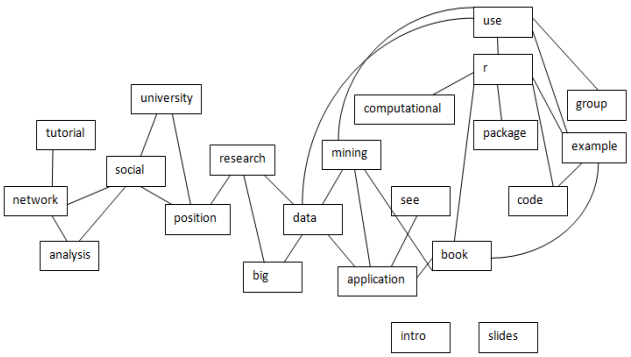


Figure 2: Frequent words Association

Word Cloud

It is nothing but collection of words in a subject or which falls under same domain. Word cloud will contain all the necessary for that process. In these the word size will differ from each other, in which size of each word will indicate its importance. In the below mentioned example data will have more importance compared to other words.

Grouping

Grouping is nothing but, the set of words which falls under one category will be grouped. Grouping will be done as the user or analyst defines. It will change from topic to topic as well as from project to project. It will be known for an analyst, how to group it.

Topic Modelling

Topic modelling, this is the place where the analysis of the words or mining of the data will happen. Data mining will be performed by using the words or searches, which the analyst wants to perform in.It is used for discovering the hidden sematic structures in a text.

V. CONCLUSION

The results of this study shows that by keeping data collected from Twitter can be used for further Analysis and investigations, which will tend increase the sales in business and to do sentiment analysis. This analysis can be further used to judge the behavior, activity and interest of the people by analyzing the data for that particular geographical area.

ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to my teacher SherinEliyas, who gave me the golden opportunity to do this wonderful project on the topic “Twitter Crowd Mining and Data Fishing”, which also helped me to in doing a lot of Research and I came toknow about so many new things I am really thankful to them.

Secondly I would also like to thank my parents and friends who helped me a lot in finalizing this project within time limit.

REFERENCES

1. I Chapter 10: Text Mining, in book R and Data Mining: Examples and Case Studies http://www.rdatamining.com/docs/RDataMining.pdf
2. I R Reference Card for Data Mining http://www.rdatamining.com/docs/R-refcard-data-mining.pdf
3. Free online courses and documents http://www.rdatamining.com/resources/
4. RDataMining Group on LinkedIn (12,000+ members) http://group.rdatamining.com
5. RDataMining on Twitter (2,000+ followers) @RDataMining
6. W. Fan and A. Bifet, “Mining big data: Current status, andforecast to the future,” SIGKDD Expolorations, vol. 14, pp. 1-5,2012..

7. I. H. Witten, “Text mining,” in The Practical Handbook of Internet Computing, M. P. Singh, Ed. Danvers, MA: Chapmanand Hall/CRC, 2005, ch. 14, pp. 314-341.
8. B. Amang and M. Sawit, KebijakanBerasdanPanganNasional:PelajaranOrdeBarudanOrdeReformasi, 2nd ed. Bogor,Indonesia: IPB Press, 2001.
9. BadanPusatStatistik. PersentasePengeluaran Rata-rata per KapitaSebulanMenurutKelompokBarang, Indonesia, 1999, 2002-2013.[Online]. Available:http://www.bps.go.id/tab_sub/view.php?tabel=1&daftar=1&id_subyek=05&notab=7.
10. R. Feldman and J. Sanger, The Text Mining Handbook: AdvancesApproaches in Analyzing Unstructured Data, New York, NY:Cambridge University Press, 2007.