International Engineering Journal For Research & Development
Vol.5
Issue 3

E-ISSN NO:-2349-0721

Impact factor : 6.193

TEXT CLASSIFICATION AND CLUSTERING OF SOCIAL DATA BY COMPUTATIONAL INTELLIGENCE APPROACH

[1]Miss Seema Sheikh , [2]Prof.Archana Vyas
Student (M.tech EXTC) [1], HOD(EXTC)[2], G.H.Raisoni University Amravati,India
seemaksheikh95@gmail.com[1] ,archana.vyasghru.edu.in[2]

ABSTRACT-

We have always enthusiastic to the trending technologies and also future science.However we should also known to the fact that earth is formed about 4.5 years billion years ago and first well equipped and brain develop man appeared 35000 years ago,But the journey from earth formed to now a days need statistical and probability distribution to store data and predict the things,opinion and many more.We are living in 21st Century,In which Human are very close to various devices, mobiles, laptops, tabs and other gadgets which generate huge volume of data and Microservices based web applications running on these have made it simpler for us to get any kind of data at any time and from any place daily. Social media platforms are also used for expressing our opinions for the products and services. The estimation and ranking of millions of the social site users can be collated to extract their perspective and sentiment towards any products or services and use that information for future market and business improvement or domain analysis.Hence the foremost thing is that to predict the things on the basis of data and analysis of behavior of products.In this paper, an open source approach is presented which we have collected and stored tweets from Twitter API and then pre-processed, analyzed,processed and visualized these tweets using R programming. To interpret sentiments of tweets we are utilizing a statistical tool, R programming. This sentiment analysis is based on text data retrieval from streamed web and then classifying human perspectives in eight distinct classifications of feeling (dislike, fear, anger, indication, sadness, trust,happy) and two unique sentiments (positive and negative). We present a new promotion vector for catalogue the tweets as positive, negative and extract human's opinion about products.

Keywords- Natural Language Processing (NLP), Sentiment Analysis,Twitter, Statistical Data and R Programming

I. INTRODUCTION

We have amount of good and best technologies in our hand and basically from day to day increasing Intelligent Testing and Artificial Intelligence,it is easy to acquire new technologies in our fingertip but very difficult to classify that which technologies are for which,which technologies are good and worst ,which technologies are giving good amount of products and bas products.Therefore people always find to classify the things,In order to get the best result and outcome.To classify the things ,We need only two things, Data and Technology that will bifurcate the substantial data by positive and negative way. Sentiment analysis is usually conducted at different levels from the coarse level to fine level. Analysis of emotions at a coarse level determines the emotion of the whole document and de fine level deals with an

analysis of emotions at a specific level. Sentence level emotion analysis comes in between these two.There are many researches on the area of sentiment analysis of user reviews. Previous researches show that the performances of sentiment classifiers are dependent on topics. Because of that we cannot say that one classifier is the best for all topics since one classifier does not consistently outperforms the other. Sentimental Analysis is a strategy to explore whether a gathered content is in positive, negative or neutral state. Essentially, it involves examining the emotions related with a piece of writing for any topic. Sentiment analysis is used to check the opinions, taste, views and interests of individuals by seeing diverse perspectives, for example, celebrity, politicians, foods, places, or some other topic. In sentimental analysis we usually categorize everyone's mood into different categories.Sentiment Analysis has three main levels. Following are the three levels of Sentiment Analysis A. Document level Sentiment Analysis B. Sentence Level Sentiment Analysis C. Aspect level Sentiment Analysis

II. LITERATURE REVIEW

In this paper, the authors describe the importance and application of opinion mining and sentiment analysis in social networks and the basic concepts, challenges, and extensive studies of different sections.

The authors describe the preprocessing steps applied to elaborate word bags from Twitter data and propose a subject-based sentiment analysis approach. The paper focuses on exploiting the impact of the default parameter of the topic modeling method.

In the other paper, the author presents an algorithm for converting "bulk data" available on social media (Twitter) into useful data and processing it according to our needs. Other benefits related to automated emotion analysis presented include topics that often differ from others in topics that are frequently expressed.All thoughts are drawn in real-time, giving time and full-time data based on previous responses to market changes, which makes it possible to plot trends over time using the R language on Twitter. The analysis obtained can be used to estimate the attitude of the people and to estimate the prevailing trends in the market and to estimate the areas of profit making.

The main purpose of the author or paper is to design a system of R and Hadoop which is big data processing technology for data analysis and visualization.They developed a set of analytical representations that help users to identify and gain insights from product, people, service and movie data, and they took a set of visualizations, implemented in shiny web applications that help integrate user interfaces with RHadoop.

The authors describe the use of sentiment analysis methods based on text-based information regarding health care. This information is ideally obtained from web sources. Sentiment analysis for health care identifies areas that are appreciated, criticized, proposed for improvement or reasoned after implementation.

According to Xing Fang and Justin Zhan the most fundamental problem in sentiment analysis, the sentiment polarity categorization for that he considered a dataset containing over 5.1 million product reviews of products belonging to four categories: beauty, books, electronics and home from Amazon.com .Previous papers in this field suggested removal of all objective content for sentiment analysis but here

instead, individual content is removed for future analysis.Inputs are reviewed that contain customer details, reviews, usability and rating.Ratings are considered to be an absolute truth for a more accurate analysis of the spirit of the review.The Max-entropy POS tagger is used to categorize the word into 46 tags.There is an extra Python program used specifically to speed up this process. As a result, there are a total of 25 million adjectives, 22 million adverbs and 56 million verbs known, which usually determine sentiment.No, not, such as rejected words are included in Adverbs while the Negative of Adjectives and Negations of Verb are specifically used to identify phrases.The algorithm also makes a list of phrases based on the occurrence.Below are the various classification models selected for classification: Naïve Bayesian, Random Forest, and Support Vector Machine.Although this paper addresses the problem of sentiment polarity categorization , it still faces many challenges and has its limitations.One such being the curse of dimensionality in feature vector formation which limits the number of dimensions and also forces to have the same number of dimensions.Performance of this approach is estimated by considering the average F1 score. Therefore, considering these limitations and improving accuracy and efficiency through them will benefit future work.

In this paper, Arijit Chatterjee and Dr. William Perrizo discusses how investor's bias affects market volatility. Sentiments were also analyzed on potential investor tweets and why they used Microsoft Azure over other sentiment analysis tools. Twitter is the largest social media platform and almost 500 million tweets have been created each time, with over 100 million active users a day.Some investors use Twitter daily to share their views on some of the ticker symbols, this paper discusses how these opinions of the investors affect the stock market.

III. METHODOLOGY OF PROPOSED WORK

This work typically involves many computational techniques (such as data and text excavation, natural language processing, etc.) and the complex analytical processes required to handle various data sources.

In addition, balancing the computational side and aesthetic side of the process using tables, charts, colors and other visual features is conducive to good data analysis and quick understanding of such data. In the process of extracting target results from the unstructured raw text that we extract from the web, the first is to identify the right datasource.
Pre-processing plays an important role in the first step of text extraction techniques and applications. This is one of the methods in the text mining process. In this article we discuss three packages in R language through which we can extract text on Twitter data.

Algorithm Steps will follow:

1. First we get a complex social data and are stored

2. After retrieving you have a tendency to rewrite the text, after that, the corpus desires a handful of transformations of changes, as well as removing the short letter with dynamic letters, punctuation / numbers, and the word stop.

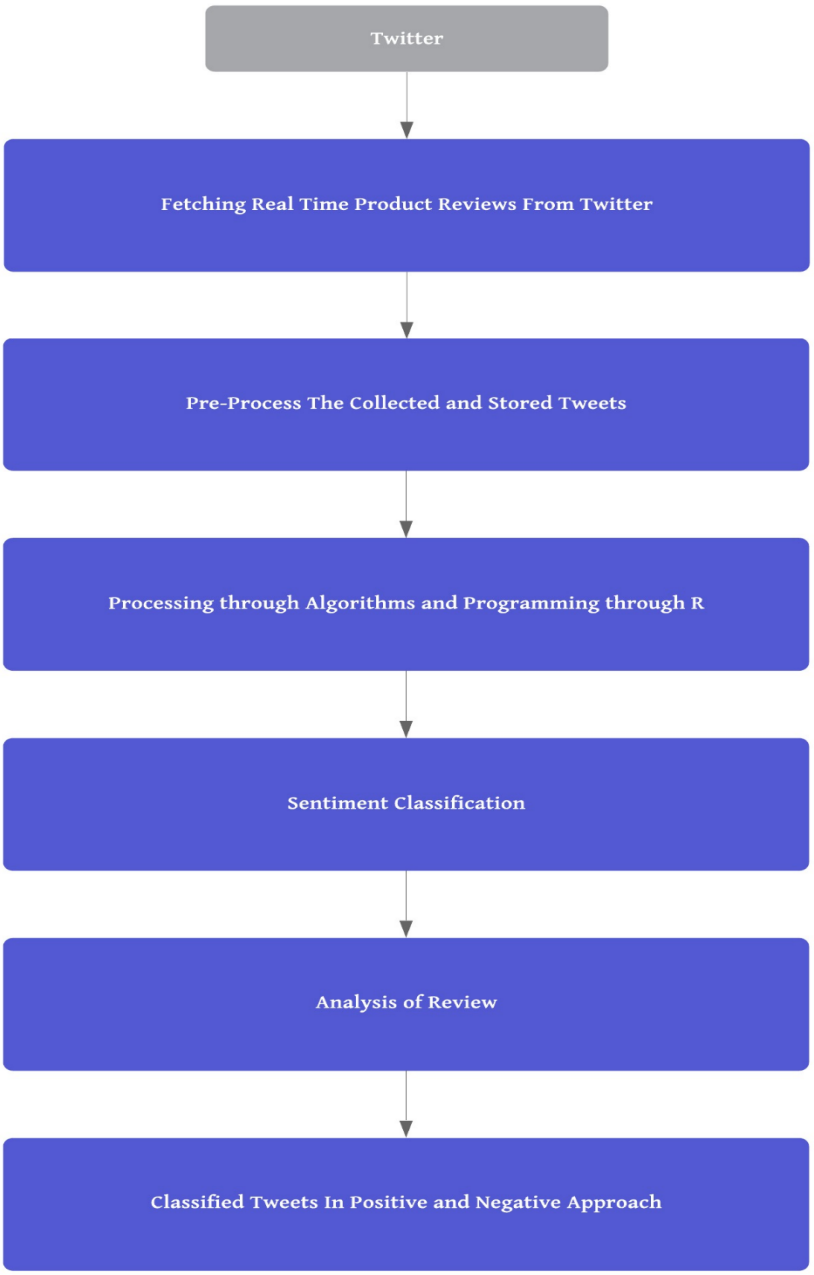
3.In most cases, the words have been scrambled to retrieve the original. For example, the "Example" and

"Examples" area unit each stemmed to "exampl". However, after that, one would like to complete the basics of their original form, so that the words look "normal".

4. After the replacement and stemming process is completed, we create a matrix document term. Depending on the confusion matrix, many text mining tasks can be performed, for example, clustering, classification, and association analysis.

5.With the help of a matrix, we can often identify words and their relationships between words.

6. After creating document-term matrix, we can see the output



Proposed Flow Diagram

IV. IMPLICATIONS

Twitter data is very useful in decision making as it provides many opinions on various topics. So text mining will take place on Twitter data and we are using computational techniques. In this paper we can analyze Twitter data, we can fetch twitter data on a particular topic and store it in R and before processing. Then we can apply several text mining steps on the twitter to pre-process the Twitter data and then we can analyze the preprocess data.

V. REFERENCES

1. Sudipta Roy, Sourish Dhar, Arnab Paul, Saprativa Bhattacharjee, Anirban Das, Deepjyoti Choudhury," Current Trends of Opinion Mining and Sentiment Analysis In Social Networks", International Journal of Research in Engineering and Technology, Volume 2, Special Issue 2, December 2013"
2. Pierre Ficamos, Yan Liu, "A Topic based Approach for Sentiment Analysis on Twitter Data", International Journal of Advanced Computer Science and Applications, Vol. 7, No. 12, 2016. '
3. Pooja Khanna, Sachin Kumar, Sumita Mishra, Anant Sinha, "Sentiment analysis: An approach to opinion mining from twitter data using R", International Journal of Advanced Research in Computer Science, Volume 8, No. 8, 2017. Pooja Khanna, Sachin Kumar, Sumita Mishra, Anant Sinha, "Sentiment analysis: An approach to opinion mining from twitter data using R", International Journal of Advanced Research in Computer Science, Volume 8, No. 8, 2017.
4. Shubham S. Deshmukh, Harshal Joshi, Pranali Pandhare, Aniket More, Prof.Aniket M. Junghare, "Twitter DataAnalysis using R", International Journal of Science, Engineering and Technology Research, Volume 6, Issue 4, April 2017.
5. M. Taimoor Khan, Shehzad Khalid, "Sentiment Analysis for Health Care", International Journal of Privacy and Health Information Management, 2015.
6. Onam Bharti, Mrs. Monika Malhotra, "Sentiment Analysis", International Journal of Computer Science and Mobile Computing, Volume 5, Issue. 6, pages 625 – 633, June 2016.
7. Xing Fang and Justin Zhan :"Sentiment analysis using product review data" Published in Journal of Big Data 2015
8. Alexander Pak, Patrick Paroubek. 2010, Twitter as a Corpus for Sentiment Analysis and Opinion Mini ng.
9. Alec Go, Richa Bhayani, Lei Huang. Twitter Sentiment Classification using Distant Supervision.
10. Jin Bai, Jian-Yun Nie. Using Language Models for Text Classification.
11. Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau. Sentiment Analysis of Twitter Data.
12. Fuchun Peng. 2003, Augmenting Naive Bayes Classifiers with Statistical Language Models