Rungta International Journal of Computer Science and Information Technology
3 (1&2), 2018, 1-4
http://www.publishingindia.com/ijcsit

Sentimental Analysis for Movie on Twitter

Tamal Dey[1*] and Kavya K. S.[2]
[1]Assistance Professor, Department of Computer Applications, PES University, Bangalore, Karnataka, India.
Email: tamal_dey@pes.edu
[2]MCA, Department of Computer Applications, PES University, Bangalore, Karnataka, India.
*Corresponding Author

Abstract: "Sentimental Analysis for Movie on Twitter" collects the emotions and feelings based on the online reviews sent via twitter in a public forum. The objective of this research work is to get the feedback from the people and their reactions, different emotional factors from live tweets. Based on the information collected calculations of different parameters is made and graphs have been drawn for pictorial assumption for the movie review.

This process involves identifying and categorizing opinions expressed in a piece of text in order to determine whether the writer's attitude towards a particular movie is positive, negative or neutral. The graph representation helps to the movie maker in decision making and analysing the facts occurred in current movie compute the budget for next movie in same category.

Keywords: Emotional factors, Feelings, Lexicon, Movie, Public forum, Social-media, Twitter, Word-cloud, Word-frequency, Word-stemming.

I. Introduction

Project Description

Product result is mainly important for every reputed company. The company can make a good name by improving their product results. This can be done by improving the quality of the product by using the feedback provided by user and getting the experience of current trends and makes the changes accordingly.

Sentimental analysis is a lexicon based natural language processing which helps in opinion mining extracting the responses in online social-media of the viewers. The study and survey of sentimental-analysis is helpful to know about the opinion, attitude of the teller about the context and emotions by the people.

In this project results achieved via word-cloud, word-stemming, word-frequency, etc. The word-cloud is nothing but combination of words from the tweets of tweeters and finalizes the result in small and large size words, large size word is most repeated and small size is less repeated. Word-stemming is to decrease the length of words and word frequency is to calculate the number of words.

This project gives programming interface to interact with the set of keys which are unique, it is needful to collect the data in Java Script Object Notation (JSON) format and map-reduce algorithm will run on Hadoop Distributed File System (HDFS).

i. Purpose

The main aim of this project is to perform sentimental analysis on social networking sites for e.g. handling twitter data to perform analysis in a single environment.

This project involves on development of an application with an user interface followed by data analysis and mapping or representing the analysed data in the form of graphs.

ii. Scope

The outcomes of this project is to produce some insightful emotions score of the public towards the subject such as positive or negative and related other eight emotions like *anger, fear, trust, sadness, joy, surprise, disgust and anticipation.*

The benefit of this project is to provide analysed report. So the experiments on large data sets need to be done, for example twitter data to do the analysis by taking movie reviews and responses about movies.

Movie reviews will be helpful for the film industry while giving the awards for the particular movie and it will help to increase the skills in the next version or in next upcoming movies for movie makers.

II. Literature Survey

A. Existing System

- Real data from the twitter holds several emotions of users on an object or product. Emotions of users may be a positive or a negative.
- In the traditional data base system, every tweet counts and tweets were collected and processed upon a definite product. The traditional data base system following

relational data sets and it is unfit for large data sets. The data we obtained all the time may or may not be structured. This is the reason in traditional data base system cannot process the large data sets. Since, generation of real data is very complex and huge in size these approaches will not be possible for processing.

- The unstructured data set is a combination of uneven fields, missed records, uninformative type. And also contains NA values it is also another hindrance for data set segregation. So, that the traditional system fails to overcome with every problems and fails to provide proper time efficiency.
- The sql is only applicable for smaller type of data sets. It is not suitable for large data sets.

B. Proposed System

- To come out of the traditional system problem, scholars suggest the big data technology, it is capable of handling each and every problems- also provides an optimal solution.
- The technology has capacity of handling large volume and unstructured real time data from twitter with variety of paradigm and it is important to apply this in many problems.
- The important features are to deals with variety, volume and velocity. These are the major problems in traditional database. Velocity is a major challenge for every domain. So, space efficiency is not at all considered as a big problem nowadays but we have to be aware of time efficiency, it is required for the retrieval of the data sets it should be complete quickly and moreover not giving chance for any delay. Retrieval of data quickly after the procession of data is important and complete the task for a huge and complex data items tends to be a major challenge here.
- Variety approach is very low compares to velocity and volume, another one which is included to these –veracity. Veracity is nothing but uncertainty.
- Analysis of data set is infinite because every developer works on the basis of set of requirements taken by users, customers or client.

III. 3R-Packages Used

A. TM

Text mining package is used to handling the text functionalities, manage the text documents.

B. Snowballc

Snowballc package is used to remove the repeated words, cleaning of data, stemming.

C. Wordcloud

Wordcloud package is used to create a cloud of so many words of data after stemming and lemmatization.

D. RColorBrewer

This package is just used to fill a colour to all graphs, to encast the beauty of the graphs.

E. Ggplot

This package is used to plot a variety of graphs, draw graphs in different ways in different colours.

F. Twitter

Twitter package is used to create a oauth connection for the twitter user using four Secret keys like consumer and access token keys.

G. Dplyr

Dplyr is used to convert the data set into proper rows and columns format while displaying.

H. Stringr

Stringr package is used to manipulate the each character of data.

IV. System Design

Block Diagram

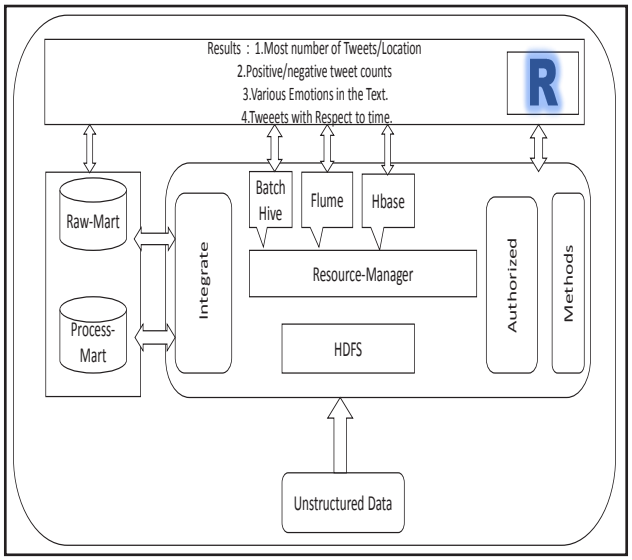

Fig. 1: Block Diagram of HDFS

In this architectural diagram, Raw-mart is a flume data and process-mart is a CSV file. Fetching up of unstructured data using flume and it has been saved in HDFS, after that Hadoop-YARN will act as a resource manager it holds the nodes. Run query on hive, TwitterAPI is authorized nobody can view or open the twitter user app without username or password. There is so many operations are going to plot the graphs using R. final result analysis contains the most number of tweets and locations, positive and negative tweet counts, emoticons on data set and tweets according to time.

V. Detailed Design

Process-Flow Diagram

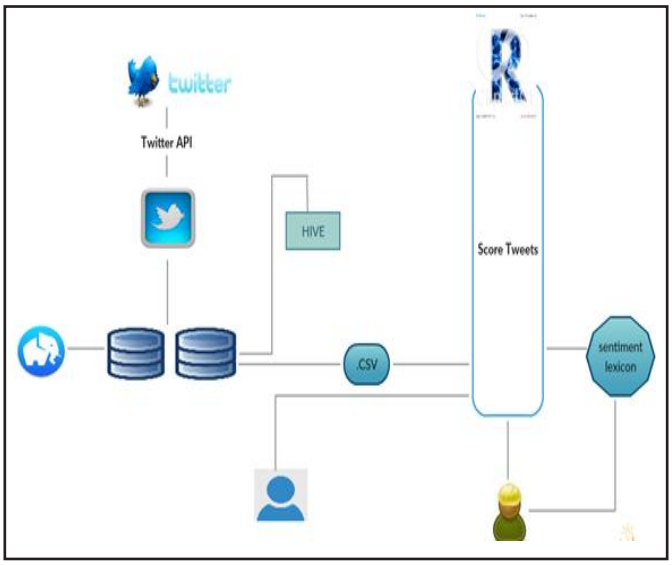

Fig. 2: Process-Flow Diagram

Process flow diagram is nothing but flow of each process of whole project. Create a API in TwitterAPI then fetch the data from the HDFS, write a query on that to fetch the required results. Convert it into CSV file, drop this in R, plot the graphs using lexicon based dictionary.

VI. Methodologies

A. Reading of Data Set

First of all fetch a flume data in hadoop eco-system; it will be saved in Hadoop distributed file system. Download the flume data save it in a system drive. This flume data will be in Json format, write a systematic query to convert this Json format data to CSV file and save it.

B. Cleaning of Data

After fetching of data and file converted to CSV format, start cleaning the punctuation tags in the data set.

C. Striping in the Data Set

Trim the tabs and white spaces between the words and sentences, it compress the data set size, and improve the beauty of the data.

D. Conversion of Cases

Transform all the text in the data set to lower case.

E. Removal of Words

Removing the stop words- it is nothing but removal of similar type words. It finds the similar words and remove, save the one copy of that word while doing Wordcloud. And remove the unrelated words.

F. Stemming

Do stemming and lemmatization on the text, remove a, and, for, is, of and also as, es, ing from the words.

G. Lexicon Dictionary

Use dictionary from hive and start creating our own dictionary containing 8 emoticons like sad, joy, anticipation, disgust, anger, trust, surprise and fear.

VII. Conclusion

The research work "Sentimental Analysis for Movie on Twitter" tends to get the emotions, response and feelings of public's for a particular movie by successfully completed with the output of top-twitters, top-tweeted locations, most tweeted days in April month, sentimental products and emotions of publics.

The analysis is useful in large type of data sets, easily we can fetches data sets from clouds, and also useful for final-people, production industries, products to get to know about their product results.

Sentimental analysis helps to know the industries services, opinions of users from several sides of regions.

VIII. Future Enhancements

In further this project can be extended up to get 3D graphs in the results.

Machine learning is the good research-area; we can implement the sentimental analysis in machine coding. Other algorithms can be used here instead of hadoop; map-reduce algorithm.

References

[1] Getting started with Hadoop tutorial, Cloudera. https://www.cloudera.com/developers/get-started-with-hadoop-tutorial.html

[2] R Tutorial - Learn R Programming - Programiz. https://www.programiz.com/r-programming

[3] https://www.packtpub.com/books/content/twitter-sentiment-analysis.

[4] trap.ncirl.ie/1842/1/annehennessy.pdf

[5] http://juliasilge.com/blog/Ten-Thousand- Tweets/

[6] https://www.youtube.com/watch?v=R650sOAq LUQ

[7] http://sunilgowda666.blogspot.in/2017/04/flume-installation-and-streaming.html

[8] https://sites.google.com/site/miningtwitter/questions/sentiment/analysis

[9] https://sites.google.com/site/miningtwitter/questions/sentiment/sentiment

[10] http://www.softwebsolutions.com/resources/sentiment-analysis-of-game-of-thrones-season-6-episode-1.html

[11] V. Jain, Sentiment Analysis of Tweets.

[12] Proceedings of the Workshop on Languages in Social Media. https://dl.acm.org/citation.cfm?id=2021109&picked=prox

[13] M. Kiruthika, S. Woonna, and P. Giri, "Sentiment analysis of twitter data," *International Journal of Innovations in Engineering and Technology*, vol. 6, no. 4, pp. 264-273, April 2016.

[14] B. S. Dattu, and V. Gore, "A survey on sentiment analysis on twitter data using different techniques," *International Journal of Computer Science and Information Technologies*, vol. 6, no. 6, pp. 5358-5362, 2015.

[15] V. Sahayak, Vijaya Shete, and A. Patnam, "Sentiment analysis on twitter data," *International Journal of Innovative Research in Advanced Engineering*, vol. 2, no. 1, pp. 178-183, January 2015.

[16] A. Kumar, and T. M. Sebastian, "Sentimental analysis on twitter," *IJCSI International Journal of Computer Science Issues,* vol. 9, issue 4, no. 3, pp. 372-378, July 2012.