International Journal of Future Generation Communication and Networking
Vol. 13, No. 3, (2020), pp. 2498–2505

Optimized Social media mining through systematic analysis

R Lavanya[1], V Thanigaivelan[2], B Prabhakaran[3], J Renuraman[4]
[1] Assistant Professor, SRM Institute of Science and Technology, Chennai
[2,3,4] Assistant Professor, SRM Institute of Science and Technology, Chennai
lavanya27382@gmail.com[1*], thanigailav@gmail.com[2]

Abstract

Analysis of organized information has seen huge achievement before. Though, scrutiny of huge measure shapeless information as film layout leftovers a testing region. Since YouTube information is made in an extremely colossal sum and with a similarly extraordinary speed, there is an enormous request to stock, course and painstakingly ponder this huge quantity of information to style it serviceable. The principle goal of this work is to show by utilizing Hadoop ideas, in what way information created since YouTube extracted and used to style focused on, continuous and educated choices. This work exploits SQL comparable inquiries keep running by utilizing HIVE to extricate important yield utilized for administration and analysis.

Keywords— *Big data; Data Analysis, Hadoop, Digital Marketing*

I. INTRODUCTION

Big Data- It characterizes Big Data as "an accumulation of informational indexes so expansive and complex that it ends up hard to process utilizing the accessible database administration apparatuses. The difficulties incorporate how to catch, minister, store, look, share, investigate and envision Big Data".[1] In the present condition, we approach more kinds of information. These information sources incorporate online exchanges, person to person communication exercises, cell phone administrations, web gaming and so forth. Enormous information like gathering informational collections which extensive and difficult in practical. It precise organized and unformatted information which develop extensive quickly not sensible by customary social database frameworks or ordinary factual apparatuses[16].

Hadoop- As associations are getting overwhelmed with enormous measure of crude information, the test here is that customary devices are ineffectively furnished to manage the scale and many-sided quality of such sort of information[2,3]. It is appropriate to address numerous analysis difficulties, particularly with enormous quantity of information thru an assortment of views. At its center, Hadoop is a structure for putting away information on substantial groups of item equipment — regular PC equipment that is reasonable and effortlessly accessible — and proceeding procedures alike that information. A bunch is a gathering of communicated PCs (known as hubs) which cooperate on a similar issue[4]. Utilizing systems of reasonable process assets to gain business understanding is the key incentive of Hadoop.

There have been noteworthy examinations on the client produced information in YouTube. Other than the substance shared by typical clients, YouTube has moreover presented the Partner Program , through which premium content owner who are persuaded by the advertisement revenues can transfer great copyrighted recordings, serving a significantly bigger client base [5]. Remarkable accomplice illustrations incorporate such mechanical mammoths as EA, ESPN, and Warner Brothers. An ever increasing number of independent companies and people have additionally joined forces with YouTube to profit by adapting their recordings, and their income has multiplied for a long time in succession. Machinima, a standout amongst the most famous YouTube accomplices, has additionally gotten noteworthy venture from Google to create all the more engaging recordings, additionally inferring the key part of YouTube partners [6,7].

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

Key terms in analyses are-

1. Data Mining- It is joining of measureable techniques. Utilizing capable scientific strategies connected to analyses information with developing that information. Then utilized to excerpt information and catch significant data utilized for building efficiency and effectiveness [15].
2. Data Warehousing- It is a databank. That sort of focal storehouse in gathering important data.
3. It has brought together rationale which diminishes the requirement with physical statistics amalgamation.
4. MapReduce- It is used in shortening huge volume of statistics as small valuable outcomes[8].
5. Hadoop-It helps in storing and retrieving data. It use HDFS Hadoop distribution file System[13].
6. Hive- It is database warehousing system used for querying and analyzing data[14].

II. RELATED WORKS

A. Data analysis

Stock data produce broad assortment of unformatted information, sort of information examined utilizing Hadoop structure. Securities exchange information investigation venture was directed with example 'NewYork StockExchange' informational index. Utilizing this stock information figured and planned as take care of stockpiling and preparing issues identified with a tremendous volume of information[9].

The dataset utilized as a part of this task was a comma isolated document (CSV) that contains the stock data, for example, day by day cites, stock opening value, stock most noteworthy cost, and so on the New York Stock Exchange. Utilizing Hive summons, a Hive Table was made. Once the table was made, the CSV information was stacked into the Hive Table. By utilizing the Hive select inquiries, Covariance for the stock dataset to the inputted year was computed. From the covariance comes about, stock intermediaries gave key proposals including the likelihood of stock costs moving the upward way or opposite course.

B. Sentiment Analysis of social data

Assumption investigation or supposition mining is characterized as sorting conclusions communicated on a web-based social networking stage about a given subject[10]. This venture was attempted to comprehend the remark author's state of mind towards a specific item or a given subject. As Association principle of information mining is utilized in all genuine utilizations of business and industry. Target of taking rfr44fApriori is to discover visit itemsets and to reveal the concealed data. This paper explains upon the utilization of affiliation rule mining in separating designs that happen every now and again inside a dataset and features the usage of the Apriori calculation in mining affiliation rules from a dataset containing violations information concerning ladies. With respect to this WEKA instrument is utilized for removing results. For this one dataset is taken from UCI archive and other information is gathered physically from the session court of sirsa to gather information on heart softening wrongdoings against ladies. The primary intention to utilize UCI is to initially check the best possible working of dataset and afterward apply Apriori on genuine dataset against wrongdoings on ladies which extricates shrouded data that what age bunch is in charge of this and to discover where the genuine guilty party is stowing away. In last the correlation is done between Apriori and Predictive Apriori Algorithm in which Apriori is preferred and quicker over Predictive Apriori Algorithm.

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

C. Predicting via Social Media

Diakopoulos et al. [12] was a researcher who demonstrated an association between 200 various steps to predict depressive illness before time. This study also expanded the ability to identify factors related to mental health on social media. The data might specify some of the feelings like guilty, self-hatred, being helpless, losing themselves, worthlessness, etc. The major contributions from this particular research are as follows: "They used a technique called the crowdsourcing for twitter users who have been diagnosed with clinical Major Depressive Disorders Major Depressive Disorders (MDD) using a screening test with the help of a tool called Center for Epidemiologic Studies Depression Scale (CES-D) (Center for Epidemiologic Studies Depression Scale) to determine different levels of depression of the crowd workers from Autobiographical Memory Test (AMT). Based on the results from tweets, they used several steps like user engagement, egocentric graph, emotion, depressive language use, linguistic style, and antidepressant usage to analyze users' behavior. After these steps were taken, they compared the attitude of the depressed users to that of a standard user, which indicated that the users with depression have gradually lowered their social activity, possessed greater negative emotions,

In this paper travel mishap is one of the pivotal zones of research in India [19]. An assortment of research has been done on information gathered through police records covering a restricted segment of roadways. The examination of such information can 17 just uncover data with respect to that divide just; yet mishaps are dissipated on thruways as well as on nearby streets. An alternate wellspring of street mishap information in India is Emergency Management explore Institute (EMRI) which serves and monitors each mishap record on each sort of street and spread data of whole State's street mishaps. In this paper, we have utilized information mining strategies to examine the information given by EMRI in which we first group the mishap information and further affiliation rule mining system is connected to recognize conditions in which a mishap may happen for each bunch [18].

As a feature of investigation, this task concentrated on conceivable deferrals and gave the yield in view of verifiable data sustained into the framework and addressed after inquiries: [17]

1. Are there any carriers which have altogether less postponements?

2. Which air terminal inside a similar metro territory offers minimal deferral to travelers?

3. How much does climate assume a part in flight delays?

The yield for the examination was that it makes a difference which aircrafts you go by for instance certain carriers performed superior to different carriers. Additionally, it was discovered that snowfall had a lot of effect in flight delays.

III. PROPOSED SYSTEM

A. Problem Definition

In proposed system, we use tweepy library file to fetch the twitter data.[6] We use text blob to carry out sentiment analysis to process the tweets and label them as positive, negative and neutral. Then the tweets are visualized on a pie chart to display the percentage of positive, negative and neutral tweets. Additionally, the accuracy of related disorder words such as anxiety disorder, depressive disorder, Attention Deficit Hyperactivity Disorder (ADHD), bipolar disorders are determined for positive, neutral and negative class. The features are:

1. This yields a good training based precision, results in a good recall and F-measure.

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

2. The computational time to fetch the tweets is reduced.

This results in good training based precision, recall, F-measure. The computational time to fetch the data from twitter is less. Sentimental analysis on linguistic features are usually of two types, Supervised Machine Learning based and Lexicon based. Here we apply the Lexicon based Sentimental Analysis..

B. Description

In this project we bring a particular channel's YouTube information utilizing YouTube API. We will utilize Google Developers Console and create a one of a kind access key which is required to bring YouTube open channel information. Once the API key is created, a .Net(C#) based console application is intended to utilize the YouTube API for getting video(s) data in view of an inquiry criteria. The content document yield created from the console application is then stacked from HDFS record into HIVE database. HDFS is an essential Hadoop application and a client can specifically collaborate with HDFS utilizing different shell-like orders bolstered by Hadoop. At that point we run queries on Big Data utilizing HIVE to separate the important yield which can be utilized by the administration for analysis. The project uses the YouTube Data API (Application Programming Interface) that permits the applications/sites to fuse works that are utilized by YouTube application to get and see data. The Google Developers Console is utilized to produce a one of a kind access key which is additionally required to bring YouTube open channel information. Once the API key is produced, a .Net(C#) based console application is intended to utilize the YouTube API for bringing video(s) data in light of a search criteria.

The text file document created from the console application is then stacked from HDFS (Hadoop Distributed File System) file into HIVE database. Hive utilizes a SQL-like interface to query information put away in different databases and record frameworks that incorporate with Hadoop. HDFS (Hadoop Distributed File System) is an essential Hadoop application and a client can straightforwardly cooperate with HDFS utilizing different shell-like commands supported by Hadoop.

C. System Architecture

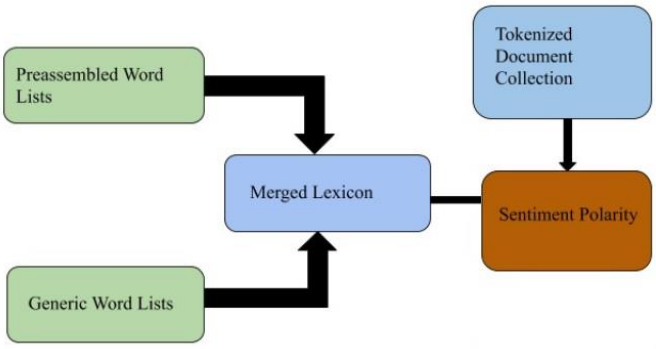

Fig. 1. Systematic view of social media data prediction

IV. RESULTS AND DISCUSSION

A. Modules Description

- ***YoutubeCategory.java***-It is MapReduce code to analyse the Youtube API data so that we can get top 5 videos of the category and description we want. The method for retrieving the Youtube data from the text file is YouTubeNamespace.CATEGORY_SCHEME.
- ***YoutubeUploader.java***-It is MapReduce code to analyse the Youtube API data so that we can get the top uploaders for the videos data we retrieved.

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

- ***YoutubeView.java***- It is MapReduce code to analyse the Youtube API data so that we can get the most viewed videos among the retrieved list of videos.
- ***Analyze.sh***-It is Shell-Script to run Hadoop Commands. It is used to execute merging and sorting command in the file.
- ***Getdata.sh***-It is a Shell-Script to copy data from server to HDFS. It is used so that we can store the data for further analysis.
- ***App.js***-It is Main Configuration File to run entire project. It changes Client server connection from AJAX to Socket.io.
- ***Searchapi.js***- Connect Youtube data API to fetch data into a file. It changes callbacks and data to be fetched.

B. Data Extraction module

One of the most popular social networks, Twitter, is a platform on which users usually share messages in the form of tweets. Twitter gives us permission to scrape and extract the data of any user using Twitter Application Program Interface(API) or Tweepy. The data is in the form of the tweets extracted from the user using twitter API credentials provided per user by Twitter themselves after we create a valid developer account. This is done by using the Tweepy library which authenticates the user to get access to publicly available tweets[6]. This module is executed by creating a GUI that has a input element to input the words , phrases or group of words that pertain to the required relevant tweets. Now the algorithm checks for the right authentication using the Tweepy library. Then the tweets relevant to the queried words or phrases are then extracted and stored in the database.The next part of the code uses two utility functions to clean the data and this is done by removing links and special words and then classsifying the sentimental type of the tweet using textblob's sentimental analysis method. Then we generate a textblob obj for the parsed tweets text to classify the tweets into different sentiments such as positive, neutral and negative. The parsing happens once the tweets are fetched.

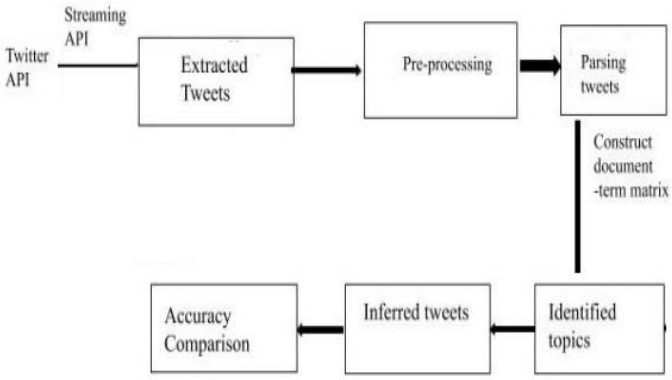

Fig.2. Categorizing Module view

Lemmatization cuts down words down to their root word. Using vocabulary and morphological analysis it transforms the tokens down to their root word[6]. Lemmatization is much more sophisticated and upgraded as compared to stemmer. Stemmer transforms independent word without any idea of the context to which the sentence is been spoken about whereas, in lemmatization the word is transformed taking it's meaning into consideration in the sentence.

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

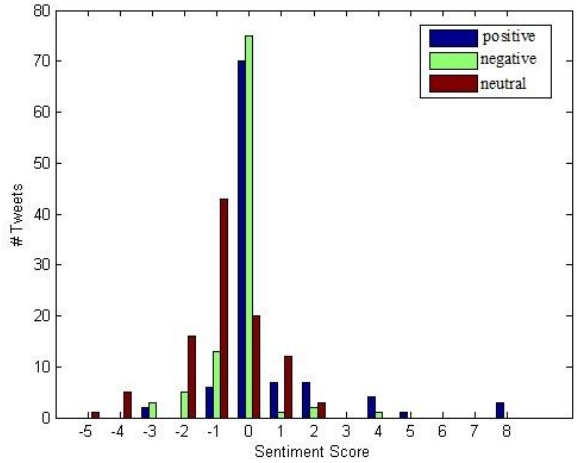

Fig.3. Tweet extraction score

Example: “better” is a good lemma. it is only captured by lemmatization and not stemming as it needs to refer the dictionary. Tweets which are fetched are categorized as positive and negative tweets based on the sentiments in it. It is called the polarity. This is done by sentimental analysis using TextBlob implementing Lexicon based sentimental analysis. In lexicon based sentimental analysis the words or tokens are assigned a polarity value between -1 to 1. 1.

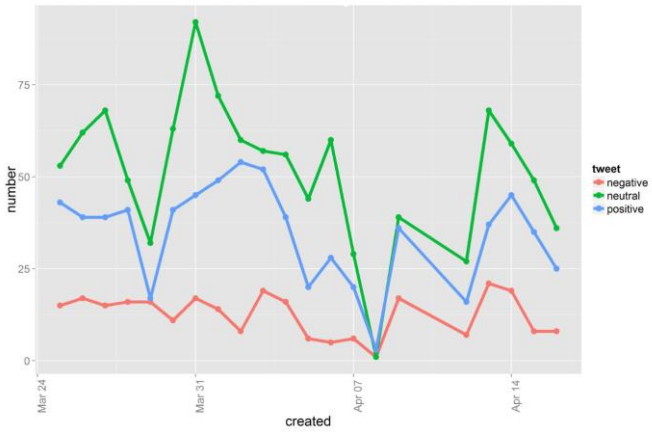

Fig.4. Twitter sentiment analysis with R

This is done by comparing each and every token from the sentiment lexicon which stores huge amount of words from the English dictionaries and has a particular sentiment value assigned to it from -1 to 1. Example words like nightmare has a negative value whereas words like happy has a positive value. The words which don’t portray emotions directly like time, water etc have neutral or numeric value as 0. Now the average of the summation of all the values are taken and a final value of the sentence is derived. This is either positive, negative or else neutral.

V. CONCLUSION

The task of big data analysis isn't just essential yet additionally a need. Actually numerous associations that have actualized big Data are acknowledging critical upper hand contrasted with different associations with no Big Data endeavors. The task is expected to dissect the YouTube Big Data and think of noteworthy bits of knowledge which can't be resolved otherwise. The output results of YouTube data analysis project demonstrate key bits of knowledge that can be extrapolated to other utilize cases also. One of the yield comes about depicts that for a particular video id, what number of likes were gotten. The quantity of likes - or "thumbs-up" - a video had has an immediate criticalness to the YouTube video's positioning, as indicated by YouTube Analytics. So if an organization posts its video on YouTube, at that point the quantity of YouTube

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

likes the organization has could decide if the organization or its rivals seem all the more noticeably, in YouTube indexed lists. Another output result gives us experiences on if there is an example of liking of interests for certain video class. This should be possible by examining the remarks check. For e.g., if the organization falls under 'comedy' or 'education' class, a significant dialog as remarks can be activated on YouTube. A remark investigation can additionally be directed to comprehend the mentality of individuals towards the particular video.

References

1. Wikipedia.org. 2016. Big Data. https://en.wikipedia.org/wiki/Big_data. [Online] February 2016. https://en.wikipedia.org/wiki/Big_data.
2. Datanami.com. 2016. Mining for YouTube Gold with Hadoop and Friends https://www.datanami.com.
3. 3pillarglobal.com. 2016. How to Analyse Big Data With Hadoop Technologies http://www.3pillarglobal.com.
4. Statista.com. 2016. Statistics and facts about YouTube. https://www.statista.com.
5. H. Garcia-Molina, J. D. Ullman and J. Widom. 2009. Database System Implementation: The complete book, 2nd edition. New Jersey: Prentice-Hall, Inc. 2009.
6. James Hong,Michael Fang, “Keyword Extraction and semantic Tag Prediction”.
7. Ming-Hung Hsu,Hsin-His Chen,”Tag Normalization and Prediction for Effective Social Media Retrieval.
8. Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. Psychological review, 111(4), 1036.
9. Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python. O’Reilly Media, Incorporated.
10. Budiu, R., Royer, C., & Pirolli, P. (2007). Modeling information scent: a comparison of LSA, PMI and GLSA similarity measures on common tests and corpora. In Large scale semantic access to content (pp. 314–332).
11. Chang, H.-C. (2010). A new perspective on Twitter hashtag use: Diffusion of innovation theory. Proceedings of the American Society for Information Science and Technology, 47(1), 1–4. Creative Commons Data Dump. (2011). Retrieved from http://data.stackexchange.com/about
12. Diakopoulos, N. A., & Shamma, D. A. (2010). Characterizing ebate performance via aggregated twitter sentiment. In Proceedings of the 28th international conference on human factors in computing systems (pp. 1195–1198).
13. Choudhury, M.D., G. M. C. S. and Horvitz, E. (2013). “Predicting depression via social media. Seventh International AAAI Conference on Weblogs and Social Media, Massachusetts.
14. De Choudhury, M., C. S. H. E. (2013). “Social media as a measurement tool of depression in populations.” Paper presented at the Proceedings of the 5th Annual ACM Web Science Conference.
15. Wang, X., Z. C. J. Y. S. L. W. L. B. Z. (2013). “A depression detection model based on sentiment analysis in micro-blog social network trends and applications in knowledge discovery and data mining.” Springer, 201–213.
16. Shardanand, U. and Maes, P.: Social Information Filtering: Algorithms for Automating "Word of Mouth". In: CHI ’95: Conf. Proc. on Human Factors in Comp. Sys. Denver, CO, 210-217. (1995).
17. R Lavanya, V Thanigaivelan, "Automated Investigation of Power Structures Annoyance Data with Smart Grid Big Data Perception”, International Conference on Sustainable Communication Networks and Application , Springer Lecture Notes on Data Engineering and Communications Technologies, 2019.
18. Littlestone, N. and Warmuth, M.: The Weighted Majority Algorithm. Information and Computation 108 (2), 212-261.

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

19. R Lavanya, V Thanigaivelan, " Big Data Analysis Applied for Short Term Solar Irradiance Forecasting", Indian Journal of Public Health Research & Development, 2019.

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC