IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

# Map Reduce concept based Sentiment Analysis Approach

**Bhavya Makkar[1*], Ayush Kaushik[2], Bhanu P. Lohani[3], Vimal Bibhu[4], Pradeep K.Kushwaha[5]**

[1, 2] Scholar, Department of CS&E, Amity University G.Noida, India
[3, 4, 5] Assistant Professor, Department of CS&E, Amity University G.Noida, India

Corresponding Author: bhavyamakkar1@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i4.924927 | Available online at: www.ijcseonline.org

Accepted: 18/Apr/2019, Published: 30/Apr/2019

***Abstract***— In the digital word there produce a large amount of data every second related to web based content or blogging data or the data generated from reviews. When we want to do the analysis of any data we need to know about the sentiments of the user who are directly or indirectly in the use of related data for this type of data processing we need sentiment analysis in fast manner, by the use of Map reduce architecture we split the related collected data into small clusters and analysis the data in very less time. Micro blogging locales have a great many individuals sharing their contemplations every day due to its trademark short and straightforward way of articulation. We propose and research a worldview to store the assessment taken away a prominent ongoing micro blogging administration, Twitter, spot clients present constant responses on and sentiments around everything. This paper mainly focuses on the concept of twitter sentiment analysis, here we have presented a system architecture how we can collect the data from the different sources and can process the data. We have focused the concept of Hadoop Map Reduce architecture for data processing in our research work. In the result section we have presented the analysis of sentiments collected from different source in tabular format as well as the graphical representation is given. A contextual analysis is introduced to represent the utilization and viability of the suggested framework.

***Keywords—Twitter,Sentiment anlysis,blogging,Hadoop,Map-reduce***

## I. INTRODUCTION

Continuous increment in wide-zone arranges availability guarantee inconceivably increased open doors for joint effort and asset sharing. Presently a-days, different informal communication locales like Twitter1, Facebook2, MySpace3, YouTube have picked up so much notoriety and we can't disregard them. They have turned out to be a standout amongst the most essential utilizations of Web . They enable individuals to fabricate tie up systems with this individuals in a light and convenient way and enable them to stake different sorts of data and to utilize a lot of administrations like picture distribution, web journals, wikis and so forth.

It is obvious that the coming of these continuous data organizing locales like Twitter have generated the production of an unequaled open accumulation of sentiments about each worldwide element that is of intrigue. Despite the fact that Twitter may arrangement for a fantastic channel for feeling creation and introduction, it presents more up to date and distinctive difficulties and the procedure is inadequate without proficient devices for examining those conclusions to assist their utilization.

"Even more starting late, there have been a couple of researches stretches out that apply end examination to Twitter corpora in order to remove generally speaking populace supposition with respect to political issues. In view of the extension of disagreeable and negative correspondence over long range casual correspondence goals like Facebook and Twitter, starting late the Authority of India enable to ease worries over banning of these desired locations where Web Users Continue to talk against any restriction on posting of substance. As published in Indian national paper [3] The Minister for Communications and Information Minister, proposed content viewing and control of relational associations like Twitter and Face book. This investigation did by us was helpful to use reassumed examination to gauge the straight forward attitude and perceive any upcoming restriction or negative influence on social media websites We strongly believe that restriction is not right way to seek after, the upcoming examples for research work for end mining in twitter can be used and connected for a numerous of useful applications in business (promoting knowledge; thing and institutions locate stepping and progress). This gave us direction to propose a model which gathers tweets on a specific subject/matter through the use of Twitter API and figures out the supposition presentation/score of each tweet."

 

"The vicinity of Sentimental Analysis of the tweets expects us to appreciate these feelings and emotions and distribute them into different classes like positive, negative, neutral. The most conclusive study has been done on audit destinations [4]. Audit destinations take things into matter with the estimations of number of items or moving pictures, along with these lines, specifying the actual space usableto exclusively business. The idea of investigation on Twitter posts is the stage in the field of estimation examination, as these posts will give us a more appropriate and more clear different conclusions and assumptions that can be clustered in with anything from the most recent mobile phone they purchased, movie they watched, political problems or the people perspective."

## II. RELATED WORK

Applying assumed investigation on Twitter is the up and coming pattern with specialists coming up with the logical preliminaries and its potential applications. The difficulties faced by novel to this problem zone are generally credited to the overwhelmingly casual tone of the small scale blogging. The procedure of reasoning the usability of micro blogging and more especially twitter as a corpus for opinion examination was given by Pak and Paroubek [5].

• Twitter contains huge number of posts and it gets increases every minute and up every day.

• The different people for expressing their decision about different focuses use Micro blogging stages.

• The social event of Twitter of individuals consists from normal customers to VIP customers, association members, government authorities, and even country presidents. It is possible to assemble content posts of individual from variety of social and emotional interests' get-togethers.

"The two Naive Bayes unigram models, two models Naive Bayes bigram and a Maximum Entropy model used to assemble tweets was executed by Parikh and Movassate [6]. They found in their study that the Naive Bayes classifiers worked much better than the old Maximum Entropy model could. The answer by using far away supervision was proposed by Go et al. [7], in which their actual readiness data which includes tweets with emoticons was used. Firstly, this strategywas exhibited by Read Ref. [8]. They manufacture new models using old models such as Naive Bayes, MaxEnt and Support Vector Machines (SVM). Their part space included bigrams, unigrams and POS. Great work has been done by Pak and Paroubek [5], request the tweets as target, positive and negative. Tweets were gathered as target or passion and a short time and after a period of time the dynamic tweets were assigned positive or negative was given by Barbosa et al. [9,10]. The part space used included features of tweets like retweet, hash tags, association, emphasis and objection stamps identified with features like before furthest point of words and POS of words."

"Batra and Rao[11] has utilized a dataset of tweets traversing from the month of June. The dataset contains about 60 million tweets. The element was removed for using the URLs and client labels were used to enlarge the elements found. A collection of about 20000 things for surveys that had been marked as positive or negative was used to prepare the model. Using this corpus the model figured the likelihood that a given unigram or bigram was being utilized in a positivesetting and the likelihood that it was being utilized in a negative setting. Twitter gathering information given by Firehouse, which gave all messages from each client progressively was used by Bifet and Frank [12]. They tried number of different things with three quick techniques that were appropriate to manage information streams: multinomial gullible Bayes, stochastic slope drop, and the Hoeffding tree.

## III. DATA CHARACTERISTICS

"Twitter is an informal communication and micro blogging administration that gives its clients a chance to post constant messages, called tweets. Tweets have numerous exceptional attributes, which ensnares new difficulties and shape up the methods for conveying conclusion investigation on it when contrasted with different spaces."

Following are the characteristics of tweets:

• Availability: The size of open data is large. Large number of people tweet in the open environment and appeared differently as compared in relation to Facebook thus making data even more freely available. The Twitter API contains collection of tweets for further planning.

• Writing strategy: The using of misguided spellings and using tweet slang that is normally understandable. As the messages are quick and short, people use short forms, ease spelling, and use emoticons and distinctive characters that pass on uncommon ramifications.

• Real time: Blogs are invigorated at longer between times of time as online diaries normally are longer and considering them requires some speculation. Tweets of course being obliged to 140 letters and are invigorated normally. This gives an all the all the more progressive feel and addresses the primary reactions to events.

• Topics: Twitter users post tweets about an extent of focuses on target goals which are proposed for a specific subject or reason. These differentiations from a part of past research, which focused on unequivocal spaces, for instance, film reviews.

• Message Length: The biggest length of a Twitter message is 140 characters. This is in connection to past inclination request ask about that focused on gathering longer messages, for instance which are used in Films, politics etc.

We now describe some basic terminology related to twitter:

 

• Emoticons: These are pictorial representation of outward appearances using symbols and letters. The goal behind emojis is to express the client's state of mind.
• Target: Twitter users use "@" image tag different clients on Twitter. Clients are naturally alarmed on the off chance that they have been referenced by someone.
• Hash labels: Twitter users use hash labels "#" for check points. It is utilized by Twitter clients to make their tweets more readable and more understandable for the other users.
• Special images: "RT" is utilized to demonstrate that it is a rehash of another person's prior tweet.

## IV. PROPOSED SYSTEM

The serious issues associated with enormous information are the accompanying:
•"The main test confronted is putting away and getting to the data from the substantial colossal measure of informational indexes from the bunches. We need a standard registering stage to oversee extensive information since the information is developing, and information stores in various information stockpiling areas in a brought together framework, which will downsize the tremendous information into sizable information for figuring. "
• The second test is recovering the information from the extensive web-based social networking informational collections. In the situations where the information is developing day by day, it's to some degree hard to getting to the information from the extensive systems on the off chance that we need to do explicit activity to be performed.
• The third test focuses on the calculation plan for dealing with the issues raised by the enormous information volume and the dynamic information attributes.
• This paper proposes three modules for finding and performing activity via web-based networking media informational collections.

The primary extent of the undertaking is to examining and bringing the Twitter IDs of those clients whose statuses have been rewetted the most by the client whose tweets are being dissected."The framework includes gathering the tweets from the informal organization utilizing the twitter API's. At that point second, this comprises of standard stage as Hadoop to fathom the difficulties of enormous information through Map Reduce structure where the total information is mapped to visit datasets and decreased to littler sizable information to simplicity of taking care of. Lastly incorporates examining the gathered tweets and bringing the Twitter IDs of those clients whose statuses have been retweeted the most by the client whose tweets are being dissected."

## V. SYSTEM ARCHITECTURE

• In the initial step I gathered the Twitter information (Tweets) utilizing API spilling of tokens and apache flume.
• Then in next step transferred the tweets into Hadoop Files Systems by HDFS directions. It incorporates moving of complete tweets of various clients to document frameworks. Lastly
• I connected Map Reduce Technique to discover the Twitter ID's of the most tweeted individuals
• After Performing the Map Reduce work, it recovers the came about information with the ids. Fig.1presents the architectural overview of the proposed system

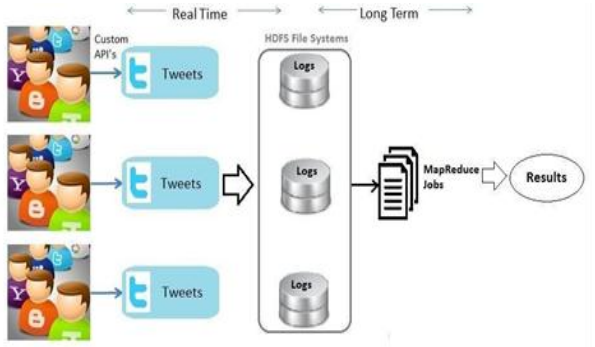


Fig. 1. System Architecture.

## VI. DATA COLLECTION & RESULT

We have collected the data for analysing the sentiments based upon three notation positive, negative and neutral emotions, then the result is analysed by the help of the above architecture given ,here we are presenting only the result part where, Table I shows the sentimental analysis results based on different queries including Movies, Politics, Lifestyle, and fake news. The bar chart, as shown in figure 2, illustrates the data based on the results we got form this step. If we run the program in different times we may get different results, small variance, based on the tweets we fetch. We run the program three times and these results are the average of the outputs.

Table.1.

| Query | Positive | Negative | Neutral |
|---|---|---|---|
| Movies | 77 | 11.6 | 19.5 |
| Politics | 28.2 | 13.5 | 64.3 |
| Lifestyle | 44.7 | 56.2 | 45 |
| Fake News | 19.5 | 74.1 | 11.4 |
| Justice | 37.8 | 17.4 | 47.5 |
| Humanity | 66.8 | 55.1 | 29.4 |

 

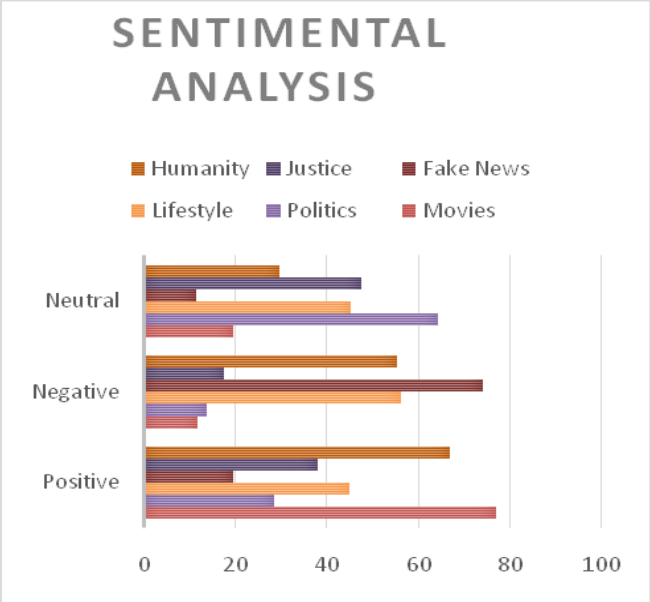


Fig2. Analysis part.

## VII. CONCLUSION

Conventional Enterprise Data Warehouses don't be able to stay aware of quickly expanding online networking information. With this framework, one can manufacture a dashboard to screen the notion of Twitter traffic around some random theme in close continuous (that is, with a deferral of 1-2 minutes), enabling you or your clients to exploit close constant Twitter opinion for business bits of knowledge or some other reason. "There are a few different ways to characterize and break down the web based life information, for example, facebook, Twitter and so on. Here anybody can perform distinctive tasks questions in these kind of information. In any case, the issue emerges when managing bigdata of a few kinds of unstructured information. Here it is explained by utilizing Hadoop and its bundles. What's more, we have done some examination on the tweets and the most number of tweet ids. "So it is inferred that handling time and recovering capacities are made simple when contrasted with other preparing and examining systems for a lot of information.

## REFERENCES

[1]L. Colazzo, A. Molinari and N. Villa. "Collaboration vs. Participation: the Role of Virtual Communities in a Web 2.0 world", International Conference on Education Technology and Computer, 2009,pp.321-325.

[2]nlp.stanford.edu/courses/cs224n/2011/reports/patlai.pdf

[3]National Daily, Economic Times: Articles. Economic Times .indiatimes.com,Collections

[4]K. Dave, S. Lawrence, and D.M. Pennock. "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews". In Proceedings of the 12th International Conference on World Wide Web (WWW), 2003, pp.519–528.

[5]A. Pak and P. Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010,pp.1320–1326.

[6]R. Parikh and M. Movassate, "Sentiment Analysis of User- Generated Twitter Updates using Various Classification Techniques", CS224N Final Report, 2009

[7]A. Go, R. Bhayani, L.Huang. "Twitter Sentiment Classification Using Distant Supervision". Stanford University, Technical Paper,2009.

[8]J. Read. "Using emoticons to reduce dependency in machine learning techniques for sentiment classification". In Proceedings of ACL-05, 43nd Meeting of the Association for Computational Linguistics. Association for Computational Linguistics,2005

[9]L. Barbosa, J. Feng. "Robust Sentiment Detection on Twitter from Biased and Noisy Data". COLING 2010: Poster Volume, pp.36-44.

[10]Bhanu Prakash Lohani, Vimal Bibhu, Ajit Singh, "Review of Evolutionary Algorithms based on parallel computing paradigm"SSRG International Journal of Computer Science and Engineering 4.6 (2017): 1-4

[11]S. Batra and D. Rao, "Entity Based Sentiment Analysis on Twitter", StanfordUniversity,2010

[12]A. Bifet and E. Frank, "Sentiment Knowledge Discovery in Twitter Streaming Data", In Proceedings of the 13th International Conference on Discovery Science, Berlin, Germany: Springer,2010, pp.1–15.

[13] V. Bibhu, P. K. Kushwaha and B. P. Lohani, "A review of security of the cloud computing over business with implementation," 2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH), Noida, 2016, pp. 192-198.doi: 10.1109/ ICICCS.2016.7542342

## Authors Profile

Mr. Bhavya Makkar is pursuing in B.Tech CSE at Amity Univeristy Greater Noida . He is presently working in the domain of Sentiment analysis and working with Hadoop.

Mr Ayush Kaushik is pursuing in B.Tech CSE at Amity Univeristy Greater Noida . He is presently working in the domain of Sentiment analysis and working with Hadoop.

Bhanu Prakash Lohani is an Assistant Professor in the Department of Computer Science and Engineering at Amity University Greater Noida . He is having 11 years of Teaching experience. His research area is Big Data Analytics & Parallel computing. He is a member of IEEE, IAENG & CSI.

Dr. Vimal Bibhu is Professor & Head of the Department in the Department of Computer Science and Engineering at Amity University Greater Noida . He is having 17 years of Teaching experience. His research area is Mobile Ad Hoc Networks , Cloud computing, Data Mining. He is a member of IEEE, IAENG, IET, CSI, ISTE.

Pradeep Kumar Kushwaha is an Assistant Professor in the Department of Computer Science and Engineering at Amity University Greater Noida . He is having 11 years of Teaching experience. His research area is Big Data Analytics & NIA. He is a member of IEEE, IAENG & CSI.

 