IMPLEMENTATION OF SENTIMENT CALCULATION ALGORITHM TO ANALYZE TWITTER DATA USING HADOOP

[1]Deepali Meshram , [2]Priyatai Mane, [3]Pranjali Mate, [4]Saroj Shambharkar, [5]Akanksha Surkar, [6]Anjum Shiekh
[1]Student, [2]Student, [3]Student,[4]Assistant Professor, [5]Student,[6]Student
[1]Information Technology Department, [2]Information Technology Department, [3]Information Technology Department, [4]Information Technology Department, [5]Information Technology Department, [6]Information Technology Department
[1,2,3,4,5,6]Kavikulguru Institute of Technology & Science, Ramtek, Nagpur, India

***Abstract**:* In today's generation where we are using large amount of data and working on that data which cannot be just specified by some quantity or it is even quite difficult to measure such a large amount of data. Companies like Google, Facebook, twitter, Microsoft have very large amount of data, this data in such a large quantity termed as big data. Handling such a large amount of data or governing such a large amount of data is difficult so the main objective behind this paper is to find such a algorithm/technique/method that can efficiently perform sentiment analysis on big data sets of twitter. So, to improve the scalability and efficiency it is proposed to implement the work on Hadoop ecosystem, a processing paradigm. The main focus of our paper is to find a technique that can efficiently perform sentiment analysis on twitter data considered as big data. Sentiment analysis is based on tweets done on twitter social networking website. Sentiment analysis is helpful and applicable to perform review and survey on the responses fetched from on line websites,social media, and health care material.

Index Terms **- Flume, Hadoop, Hadoop Ecosystem, HDFS, Sentiment Analysis.**

I. INTRODUCTION

Sentiment analysis is the analysis by treating computationally user or public opinion or sentiments received from any sources such as on line social networking websites and also subjectivity in an text especially text which can be obtained from the social networking websites. The sentiment analysis is used to know the opinion of users called as tweets with respect to the particular topic. Some cases it can be used for judgment like product success or failure in impressing the crowd when the new or updated version will be released to the market. Many peoples used is for business purpose,for publicity for advertisement.

Twitter is an on line social networking site which allows users to post real time short messages which is limited to 140-character those short messages which we called as "tweets". Registered users have access to read and post tweets, but unregistered users can only read the tweets. The alternate name can be used for Sentiment analysis is opinion mining can be defined as analysis of textual data specially data on collected from social media .

Sentimental Analysis is all about to get the real voice of people towards specific product, services, organization, movies, news, events, issues and their attributes. Sentiment Analysis includes branches of Computer Science like Natural Language Processing, Machine Learning, Text Mining and Information Theory and Coding. By using approaches, methods, techniques and models of defined branches, we can categorize our unstructured data which may be in the form of news articles, blogs, tweets, movie reviews, product reviews etc. into positive, negative or neutral sentiment according to the sentiment expressed in them.
Sentiment Analysis can be done on three levels given below:-

1. Document level
2. Sentence level
3. Aspect or Entity level

1. Document Level Sentiment Analysis: It is performed for the whole document and then decide whether the document express positive or negative sentiment.
2. Entity or Aspect Level Sentiment Analysis: At this level we performs fine-grained analysis. The goal of entity or aspect level Sentiment Analysis is to find sentiment on entities and/or aspect of those entities.

3. Sentence level Sentiment Analysis: It is related to find sentiment form sentences whether each sentence expressed a positive, negative or neutral sentiment Sentence level Sentiment Analysis is closely related to subjectivity classification. Many of the statements about entities are factual in nature and yet they still carry sentiment. Current Sentiment Analysis approaches express the sentiment of subjective statements and neglect such objective statements that carry sentiment.

II. Objective

To implement an algorithm for Sentiment analysis to determine the sentiments whether it is positive, negative or neutral towards a subject of interest, with high speed without compromising the actual data and accuracy. Sentiment analysis of the users who uses social networking websites.

III. Motivation

In today's competitive complex business world, the various aspects of business are intermingled. Change in one aspect has direct or indirect effect on the other aspect. Within an organization, this complexity makes it difficult for business leaders to rely solely on experience (or intuition) to make decisions. Need is to rely on data - structured, unstructured or semi-structured - to back up their decisions. Existing tools don't lend themselves to sophisticated data analysis at the scale the user requires. Tools like SAS, R programming, and Mat lab support the decisive analysis but tools are not designed for the massive datasets neither Data Base Management System nor Map Reduce can handle the data that are arrived at high rates. To bridge this gap the "Big Data" came into the scene. Big Data has given the organization a new way to analyze and visualize the data effectively. Big data can also be the key to actually deploying condition-based maintenance program and improve forecasting and scheduling of assets. The increasing dependence of businesses on technology ensures that the data will continue to grow at an even faster rate.

IV. Literature Survey

In paper" Big Data Sentiment Analysis using Hadoop", published in 2015 introduces an approach that can perform Sentiment Analysis quicker because vast amount of data needed to be analyzed. Also, it had to be made sure that accuracy is not compromised too much while focusing on speed. Sentiment Analysis on Big Data is achieved by collaborating Big Data with Hadoop. A dictionary of sentiment bearing words was used to classify the text into positive, negative or neutral opinion. Sentimental Analysis is all about to get the real voice of people towards specific product, services, organization, movies, news, events, issues and their attributes [1].

In paper "Sentiment Analysis on Twitter data", published in 2012 introduces an approach of applying sentiment analysis on Twitter is the upcoming trend with researchers recognizing the scientific trials and its potential applications. The challenges unique to this problem area are largely attributed to the dominantly informal tone of the micro blogging. Pak and Paroubek [2] rationale the use micro blogging and more particularly Twitter as a corpus for sentiment analysis. Micro blogging platforms are used by different people to express their opinion about different topics, thus it is a valuable source of people's opinions. Twitter contains an enormous number of text posts and it grows every day. The collected corpus can be arbitrarily large. Twitters audience varies from regular users to celebrities, company representatives, politicians, and even country presidents. Therefore, it is possible to collect text posts of users from different social and interest's groups. Twitter's audience is represented by users from many countries.

In paper "Literature Survey on Sentiment Analysis of Twitter Data using Machine Learning Approaches", published in 2017 mentioned Twitter is a micro-blogging website that has become increasingly popular with the network community. Users update short messages, also known as Tweets, which are limited to 140 characters. Users update their personal opinions on many subjects, discuss current topics and write about life events through tweets. This platform is favored by many users because it has no political and economic restrictions and is easily available to large number of people. As the number of users increase, micro-blogging platforms are becoming a place to find strong viewpoints and sentiment. People use twitter to forecast and analyze in a lot of different areas. For example, people have already forecasted the stock market success by using data from Twitter [3].

In paper" Twitter Data Analysis for Live Streaming By Using Flume Technology", published in 2018 introduces that system can generate the location details of the profile user to update in database. System configuration files are analyzed in the proposed system. The System can vary the data from different storage areas. The more importance of the proposed system is to automatic activation of Security profile depending on the context, in which the device is being used. They are very speedy. Poring over date to expand them from research takes time. They are appealing, visual depiction of data tends to have a contact and generates attention amongst the end users. It is easily understandable by viewers. Many of the social networks or any business websites can use word cloud to attract the public interests and make them to understand.

In paper," Effective Sentiment Analysis on Twitter Data using: Apache Flume and Hive", published in 2014 mentioned that they are going to use some coding techniques for crawling the data from the twitter where they can extract the data from the Twitter web pages by using some code that may be written either in JAVA, Python etc. For those they are going to download the libraries that are provided by the twitter guys by using this they are crawling the data that we want particularly. [4] After getting raw data they will filter by using some old techniques and also they will find out the positive, negative and moderate words from the list of collected words in a text file. All these words should be collected by us to filter out or do some sentiment analysis on the filtered data.[5],[6]. These words can be called as a dictionary set by which they will perform sentiment analysis. Also, after performing all these things and they want to store these in a database and coming to here they can use RDBMS[7] where they are having limitations in creating tables and also accessing the tables effectively.

V. Proposed Approach

The main focus behind this paper is to analyze the data which can be helpful in case of analytics of the behavior of the person which will help in feed for the specific person or thing. While working on the paper we kept eye on the speed as the analytics is to be performed on the large amount of data but while keeping focus on the speed we made sure that there should be no compromise in the accuracy. So, we tried to manage the balance of both the speed and accuracy with the use of sentiment analysis through Hadoop.

Basically, In this model we are focussing on an approach with which we can perform Sentiment analysis with high speed without compromising the actual data. During the analysis the primary focus is on accuracy as looking towards speed accuracy should not be compromised. We are using hadoop for mapping data into various machines where we are splitting the data into various modules. The analytics will help in analysing the behaviour and provide the feed according to the activity which will help in his interests.

In the analytics we have mainly categorized the sentiments using the value of the word we set. As the categorization of the word is already done as strong, weak and neutral helped to describe the nature of sentiment. Though the approach is dictionary-based approach which was used for categorization, we have not used any machines as the machine provides accuracy but they need to be directed for the actual purpose of work. Since we are using dictionary-based approach so there is a scope for the updating as if there are any changes or any type of modifications that can easily be implemented into the dictionary. After all the calculation one will easily be able to analyse the sentiments that one was expressing through the comment.

In the analysis we have considered about the subject field of the word as we have categorized as strong subject and weak subject as if the word describes positive sentiment about certain topic it will be categorized to strong and weak also same with the negative sentiment but if the word is nothing related to topic it will be sent to the neutral field.

VI. System Architecture

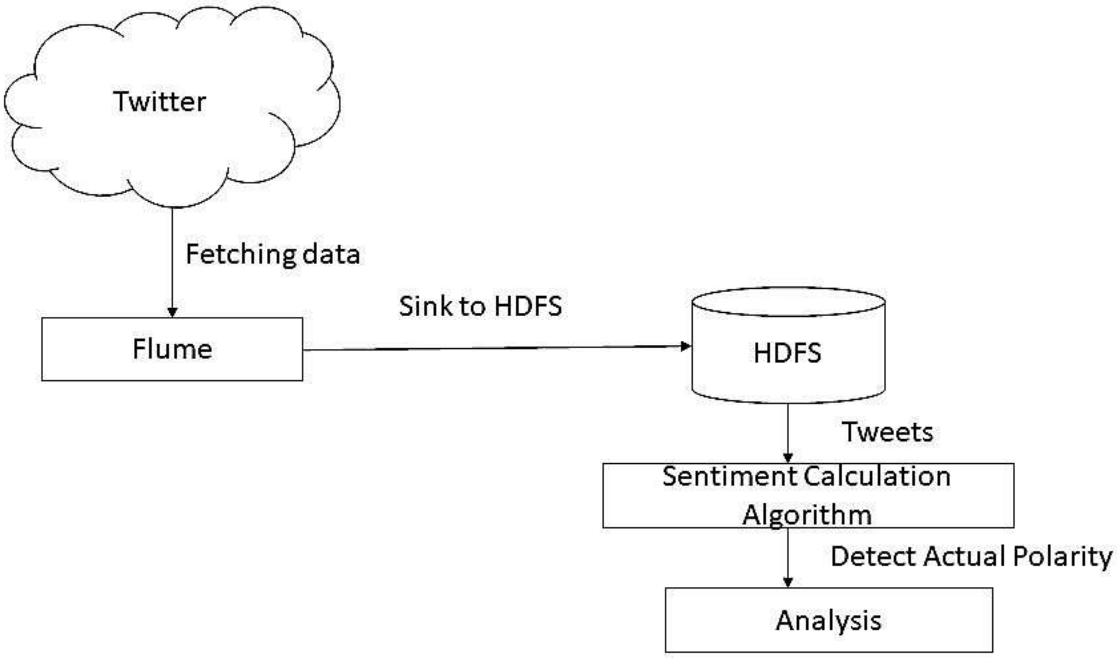


Fig.1 System Architecture

In System Architecture the first step collected the Twitter data (Tweets) using apache flume. Then uploaded the tweets into Hadoop Files Systems by (HDFS) commands. It includes moving of complete tweets of different users to file systems. In the analytics we have mainly categorized the sentiments using the value of the word set. As the categorization of the word is already done as positive, negative and neutral helped to describe the nature of sentiment. After performing categorization it retrieves the resulted data.

VII. System Description

Functional Requirement (Modules Description)

There are three modules comes under this proposed system.

7.1 Fetching data from Twitter

Using Flume, we can fetch data from various services and transport it to centralized stores (HDFS and HBase). Apache Flume is a tool or service ingestion mechanism for collecting aggregating and transporting large amounts of streaming data such as log files, events (etc...) from various sources to a centralized data store. Flume is a highly reliable, distributed, and configurable tool. It is principally designed to copy streaming data (log data) from various web servers to HDFS. A web server generates log data and this data is collected by an agent in Flume. The channel buffers this data to a sink, which finally pushes it to centralized stores.

In proposed approach we get the tweets from Twitter using the experimental twitter source provided by Apache Flume. We will use the memory channel to buffer these tweets and HDFS sink to push these tweets into the HDFS.

7.2 Preprocessing

We prepare the transaction file that contains opinion indicators, namely the adjective, adverb and verb along
Also, we the percentage of the tweet in Caps, the length of repeated sequences & the number of exclamation marks, amongst others. Thus, we pre-process all the tweets as follows:
a) Remove all URLs (e.g. www.xyz.com), Slang words (e.g. 2day#today), hash tags (e.g. #name), Stop words (e.g. a, a's, able, about) targets (@username), special Twitter words ("e.g. RT").
b) Remove all punctuations after counting the number of exclamation marks.

7.3 Sentiment Calculation Algorithm

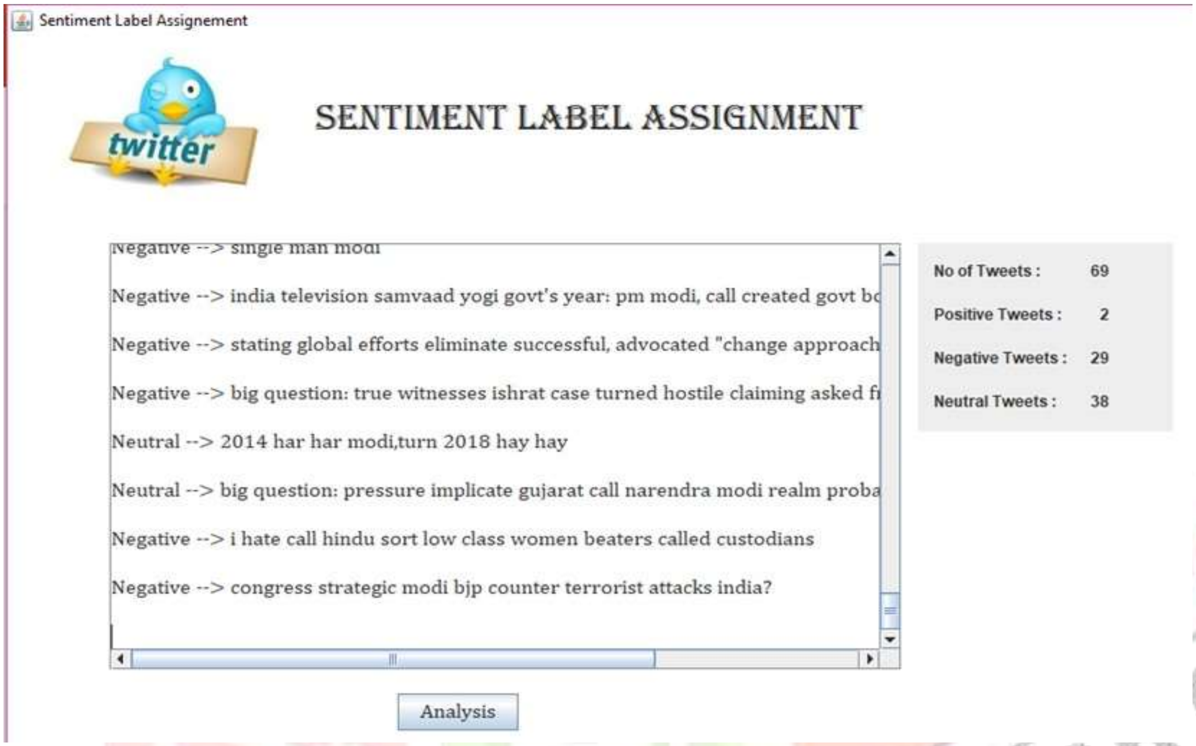
Sentiment calculation is done for every tweet and a polarity score is given to it. If the score is greater than 0 then it is considered to a positive sentiment and if less than 0 then negative else it is neutral.

```
Input: Tweets, SentWord_Dictionary
                    Output: Sentiment (positive, negative or neutral)
BEGIN
1) For each tweet Ti do the following
2) Initialize SentScore = 0;
3) For each word Wj in Ti that exists in Sentword_Dictionary.
 If polarity[Wj] = blind negation then Return negative.
Else
a. If polarity[Wj] = positive && strength[Wj] = Strongsubj then increment sentscore by 1.
b. Else If polarity[Wj] = positive && strength[Wj] = Weaksubj then add 0.5 to sentscore.
c. Else If polarity[Wj] = negative && strength[Wj] = Strongsubj then decrement sentscore by 1.
d.Else If polarity[Wj] = negative && strength[Wj] = Weaksubj then substract 0.5 from sentscore.
e. If polarity[Wj] = negation multiply sentscore by -1.

 If Sentscore of Ti >0 then Sentiment = positive.
Else If Sentscore of Ti<0 then Sentiment = negative.
Else Sentiment = neutral

4) Return Sentiment
5) END
```

VIII. RESULT

The purpose of this research was to device a method that can quickly compute the sentiments of huge data sets without compromising too much with accuracy. The proposed approach has performed very well in terms of speed and accuracy.

In figure 2,the result of algorithm shows total number of tweets is 69 out of which 2 positive tweets,there are 29 negative tweets and more number of neutral tweets which are equal to 38.Based on results shown in figure 1 the graph is plotted which clears the analysis of sentiments given by the people. In the graph shown in figure 3the analysis is done and a graph is plotted which indicates there are less number of positive tweets shown in red color ,then negative tweets shown in blue color are more as compared to negative tweets,and many people does not given any comments coming under neutral category tweets shown in green color.



Sentiment Label Assignement

SENTIMENT LABEL ASSIGNMENT

```
Negative --> single man modi
Negative --> india television samvaad yogi govt's year: pm modi, call created govt bo
Negative --> stating global efforts eliminate successful, advocated "change approach
Negative --> big question: true witnesses ishrat case turned hostile claiming asked fi
Neutral --> 2014 har har modi,turn 2018 hay hay
Neutral --> big question: pressure implicate gujarat call narendra modi realm proba
Negative --> i hate call hindu sort low class women beaters called custodians
Negative --> congress strategic modi bjp counter terrorist attacks india?
```

No of Tweets :	69
Positive Tweets :	2
Negative Tweets :	29
Neutral Tweets :	38

Fig.2 Polarity Score

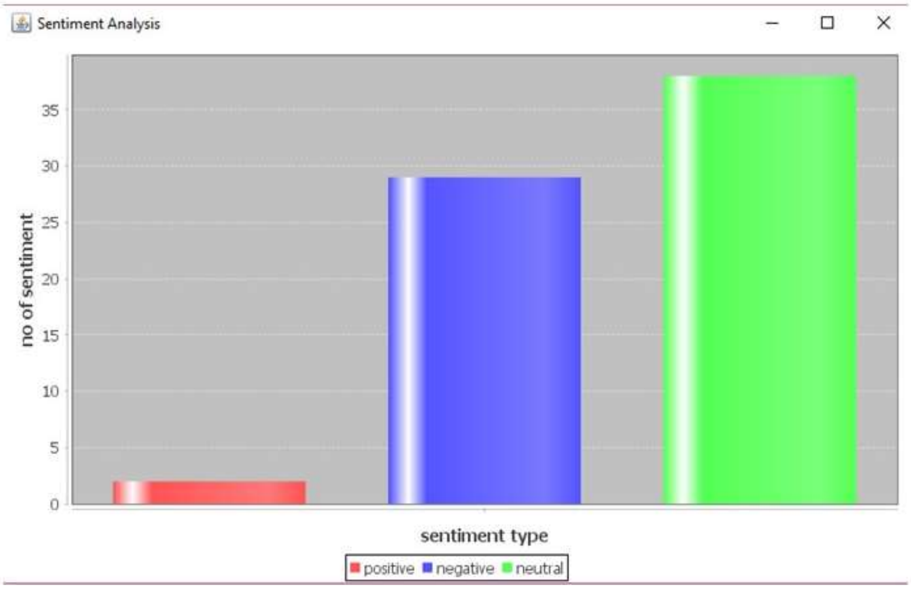


Fig.3 Analysis using Bar graph

IX. Conclusion

In this research, main focus was on performing sentiment analysis quickly so that big data sets can be handled efficiently. Hadoop was used to classify Twitter data without need for any kind of training. Our approach performed extremely well in terms of both speed and accuracy while showing signs that it can be further scaled to much bigger data sets with similar, in fact better performance.

Sentiment Analysis is being used for different applications and can be used for several others in future. The work can be further expanded by introducing techniques that increase the accuracy by tackling problems like implicit sentiments which still needs to be resolved properly. also, this work was implemented on a single node configuration and although it is expected that it will perform much better in a multimode enterprise level configuration, it is desirable to check its performance in such environment in future.

References

[1] Bing Liu, “Sentiment Analysis and Opinion Mining”, Morgan and Claypool Publishers, May 2012.p.18-19,27-28,44-45,47,90-101.

[2] 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.

[3] Phillip Tichaona Sumbureru. Analysis of Tweets for Prediction of Indian Stock Markets. IJSR 2013.

[4] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N paper Report, Stanford, 1-12.

[5] Tang, H., Tan, S., Cheng, X., A survey on sentiment detection of reviews, Expert Systems with Applications: An International Journal, v.36 n.7, p.10760-10773, September, 2009.

[6] Bahrainian, S.A., Dengel, A., “Sentiment Analysis using Sentiment Features”, In the proceedings of WPRSM Workshop and the Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, Atlanta, USA, 2013

[7] S. W. Ambler. Relational databases 101: Looking at the whole picture.www.AgileData.org, 2009.

IJCRT