ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

Twitter Sentiment Analysis Using Hadoop

Nishad D. Patil[1], Sayali C. Tingre[1], Swapnil S. Shivshetty[1], Kalyani S. Thorat[1], Sonali A. Muley[2]

UG Students, Department of Computer Engineering, MMIT, Lohagaon, Pune, India[1]

Professor, Department of Computer Engineering, MMIT, Lohagaon, Pune, India[2]

ABSTRACT: The ever increasing development in the field of computer science has lead to enormous research and application related to various fields, sentiment analysis has seen a huge development over the years various methods and systems have been proposed for its development. The study of sentiment can also be related to artificial intelligence as well, twitter is been the major source of data over the years hence the data collected from twitter for analysis is huge in volume so hence the existing system are incapable to handle this huge amount of data, hence we have proposed a system that is capable of handling this huge amount of data and generate accurate results for sentiments. The system has additional feature to compare sentiments of multiple hashtags as well.

KEYWORDS:Sentiment analysis; Data mining; opinion mining; Hadoop; Big data

I. INTRODUCTION

'Internet'-the term is a one which nearly every person in today's world is familiar with. It has touched the life of Mankind and constantly revolutionized his way of doing tasks and communicating with each other like no one ever imagined. The first workable prototype of the Internet came in the late 1960s with the creation of ARPANET, or the Advanced Research Projects Agency Network, followed by constant evolutions in technology of ARPANET and thereby launching the World wide web in 1991. Dynamically, the last two decades have reaped the benefits of the Internet like never before.

Today people cannot imagine their lives without the Internet. With the advent of the Internet came Social Media. Social Media is one of the most significantsource of information exchange of the century. Also, Online Social Networks have the ability to keep in touch with their friends and family and to share and express various types of information/ knowledge, news, educational content, business related information and many more. Moreover, even pictorial data and graphics are so easily exchanged. People of all ages use social media, mainly Twitter or Facebook to share their ideas, views and opinions of users related to on numerous topics like political, social, legal, economical etc.. Consequently, sentiment analysis of social media content may be of interest to different public sector organizations, especially in the security and law enforcement sector since such social networking platforms are open to all.

This paper mainly focuses on data generated from twitter[8] i.e. tweets from twitter. Data from twitter has been applied to address a wide scope of applications (e.g., political election prediction and disease tracking); however, no studies have been conducted to explore the interactions and potential relationships between twitter data and social events available from government entities.

Put into a larger social context, such research is important because government entities often work within limited resources to serve their constituents. The results from these analysis[8] can facilitate government entities and public service organizations to better understand the people they serve and the effect of their actions, as well as to identify potential issues in a timely manner. Our Approach is to download Twitter messages for a particular #hashTag and perform sentiment analysis i.e. to find positive sense, negative sense or neutral sense of that tweet using hadoop's mapreduce framework. Each #hashTag may have 1000 of users and new users are added every minute, these users express their opinion for that particular #hashTag in order to handle so many tweets we are using apache hadoop framework for analysis of the huge volume of data.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

II. RELATED WORK

In [2] sentiment[8] analysis has been the field of interest for many researches over the past few years, as the use of OSN is increasing day by day, many people now use OSN to view their opinion about a particular company, product or service. This data can be used for analysis and improving the companies sales.

Dmitry Davidov, Oren Tsur& Ari Rappoport[1]. Provided a supervised sentiment classification framework which is based on data from Twitter.. By utilizing 50 Twitter tags and 15 smileys used as sentiment labels, this framework avoids the need for labor intensive manual annotation, allowing identification and classification of the diverse sentiment types of short texts. They evaluate the contribution of different feature types of sentiment classification and show that their framework successfully identifies sentiment types of untagged sentences. They utilized 50 Twitter tags and 15 smileys as sentiment labels which allow them to build a classifier for dozens of sentiment types for short textual sentences. In their study they use four different feature types (punctuation, words, n-grams and patterns)[1] for sentiment classification and evaluate the contribution of each feature type for this task. They showed that their framework successfully identifies sentiment types of the untagged tweets.

Luciano Barbosa[2] provided a 2-step sentiment analysis classification method for Twitter, which first classifies messages as subjective and objective, it further distinguishes the subjective tweets as positive or negative. To better utilize these sources, he verified the potential value of using and combining them, providing an analysis of the provided labels, examine different strategies to combine these sources in order to obtain the best outcome; and, proposed a more robust feature set that captures more abstract representation of tweets, composed by meta-information associated to words and specific characteristics of how tweets are written.

SaschaNarr, Michael Hulfenhaus and SahinAlbayrak[3] provided examined a language-independent sentiment classification approach. They trained a classifier to label the sentiment polarity specifically of tweets. They used a semi-supervised emoticon heuristic to generate labelled training data. For any language, their approach requires only raw tweets of that language for training and no additional adjustments or intervention. They trained classifiers on tweets of 4 different languages: English, German, French and Portuguese. For their evaluation, they collected thousands of human-annotated tweets in these 4 languages using Amazon's Mechanical Turk2.

PreslavNakov,ZornitsaKozareva, Alan Ritter, Sara Rosenthal,SaraRosenthal,Theresa Wilson[4]. Proposed SemEval-2013 Task 2: Sentiment Analysis of Twitter, which included two subtasks: A, an expression-level subtask, A and B, a message level subtask. They used crowdsourcing on Amazon Mechanical Turk to label a large

Twitter training dataset along with additional test sets for Twitter and SMS messages for both subtasks. The primary goal of our SemEval-2013 task 2 has been designed for promoting research that will lead to a better understanding of how sentiment is conveyed through Tweets and SMS messages. Toward that goal, authors created the SemEval Tweet corpus, which contains Tweets (on both training and testing) and SMS messages (for testing only) of sentiment expressions annotated with contextual phrase-level polarity as well as an overall message-level[4] polarity

Anna Jurek, Yaxin Bi, Maurice Mulvenna[5] provided a lexicon based approach for analysing the sentiments of tweets on twitter. They have provided a algorithm that provides the intensity of the sentiments rather than the positive and negative label. They evaluated evidence-based combining function that supports classification process in cases when positive and negative words co-occur in a tweet. They have illustrated a case study of the relation in between sentiment of twitter post related to English defence league

Erik Cambria[6] provided approach for concept level sentiment analysis for automatic analysis of online opinions of various user's using natural language text by machines to go beyond the mere word level sentiment analysis of texts and provide approaches for opinion mining and sentiment analysis that enable's a more efficient passage from textual information to machine process able data node.

III. EXISITING SYSTEM

The present system used for sentiment analysis is a standalone system on a local native machine this system performs analysis on the base of hadoop itself it uses normal databases for storing of data from the twitter server, also it

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

lacks performance as calculating the sentiment on a normal machine without hadoop may utilize a lot of system resources and may result in the system failure as well when large amount of data is passed to these systems. The existing system uses only java for processing the data and calculating the sentiments of the particular tweet, as java process all the data on a single node the data processing takes a lot of time. The other drawbacks of the existing system are as follows,

- It takes lot of time for performing analysis in large amount of data
- It may result in system failure when a large amount of data is passed to the system
- Major functions and operations for sentiment analysis like stemming and NLP processing takes a huge amount of time.

In order to overcome these drawbacks we have proposed the following system with hadoop integration.

IV. PROPOSED SYSTEM

The proposed system is an advancement of an enhanced version of the previous system, the new system is hadoop integrated. The advance in hadoop has seen ever increasing development in hadoop we have used the hadoop mapreduce framework to implement sentiment analysis from the previous version of the system. The new system also calculates the emoticon senses and generates a hybrid score for give the overall sentiment of the tweet. The system has a self-defined dictionary for emoticon sense and also for stop word removal. The new system is expected to give better performance than the existing system. In our system we have used both approaches for comparing the results of both the system the first 3 tabs of our system are non hadoop tabs i.e. operations performed on these tabs don't use hadoop. Whereas the same operations are performed using hadoops mapreduce component from terminal, to check the difference in the execution time. The later 3 tabs are just tabs for visualizing the output of sentiment analysis in form of pie chart for the particular hashTag which also displays the total no of tweets for whom the sentiments are calculated, a graph chart for comparing the sentiments of 2 hashtags and the last tab shows the output file from which the sentiments are calculated.

V. SYSTEM ARCHITECHURE

The following diagram shows the overall architecture if the proposed system in detail:

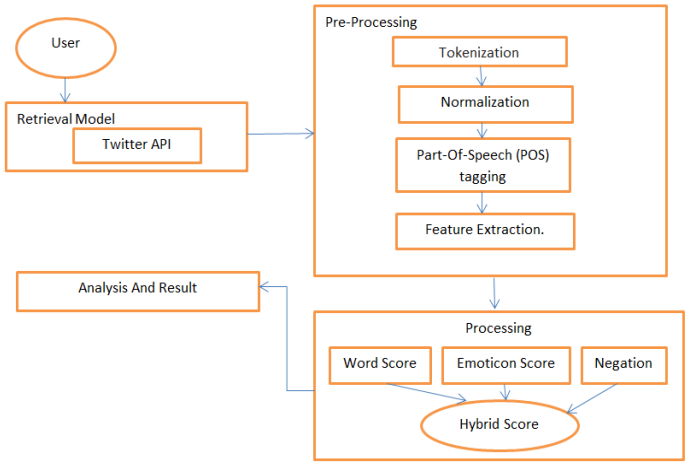

FIG 1:SYSTEM ARCHITECHURE

The overall system is divided into 4 modules for better and efficient development. Firstly all the tweets are fetched from twitter and stored in a .txt file which is used as an input for the system, the tweets are fetched and stored in a particular format as follows "username, screen name, text, created at, language, location, place" these are fetched using Twitter4j api.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

In the pre-processing phase the all the data that is not required for processing is removed i.e. stop word removal, URL removal, @ removal are performed and stemming is applied to the remaining part of the tweet, the preprocessing phase helps use to remove the unnecessary words before applying stemming for finding the root of the word, this helps us to increase the efficiency if the system and help reduce the cpu execution time.

After finding the root of the word we calculate the sense of the based on two approaches emoticon and word based later the results of both are merged and final result is displayed. For emoticon score there is a self-defined dictionary of emoticons. And for word based there is a dictionary with score for all the most popular adjectives. After all the sense are calculated the result are displayed.

The output is shown in a pie chart for a single hashTag and as a bar chart for comparing sentiments of multiple hashtags this is done with the help of the output of hadoop. the pre-processing phase is also done using hadoop mapreduce component wherein the operations of stop word removal, @ removal and URL removal are mapped and removed in the reducer phase. The hadoop framework also allows us to increase the efficiency of the system and helps use reduce the overall system time..

VI. EXPERIMENTAL RESULT AND ANALYSIS

The The system is expected to give accurate result for analysis of the sentiments in the form of pie charts and graphs the system uses two approaches to solve the problem using the normal approach and another using hadoop component. The one with hadoop component is expected to be more efficient and faster as compared to the normal system in comparison with large amount of data. The system was given 3 types of input file size small medium and large. The same inputs where passed and processed using normal approach and hadoop integrated approach. The following graph shows the comparison of both the outputs.

1. Small file less than 5mb(around 10 thousand tweets processed)

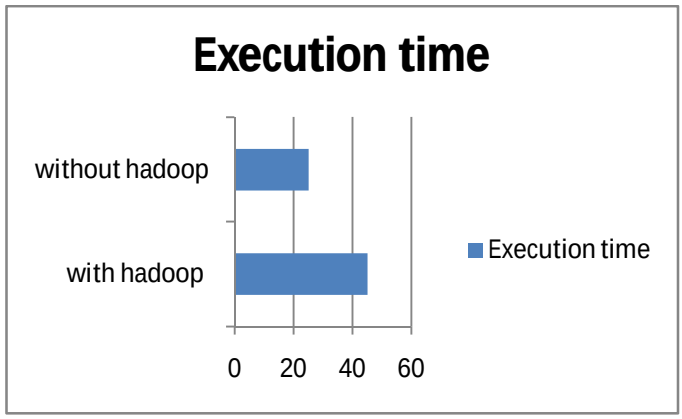

FIG 2:EXECUTION TIME FOR SMALLER FILES

From fig 2 i.ethe above graph we can conclude that for smaller files the normal system is quite capable and produces far better results than the hadoop based system as map reduce takes a lot of time for computations as it needs to initialize mappers and reducers and then combine them into one again for generating final result.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

2. Medium file 50-70mb(around 3 to 4 lakh tweets processed)

FIG 3:EXECUTION TIME FOR MEDIUM FILES

Fig 3 describes the execution of medium sized files the results are almost same as mapper and reducing take a fair amount of time so hence the result are almost similar time as its quite some data for normal system but hadoop system this data is quite small and it takes time due to the mappers and reducers.

3. Large file greater than 100 mb(more than 8 lakhs tweets processed)

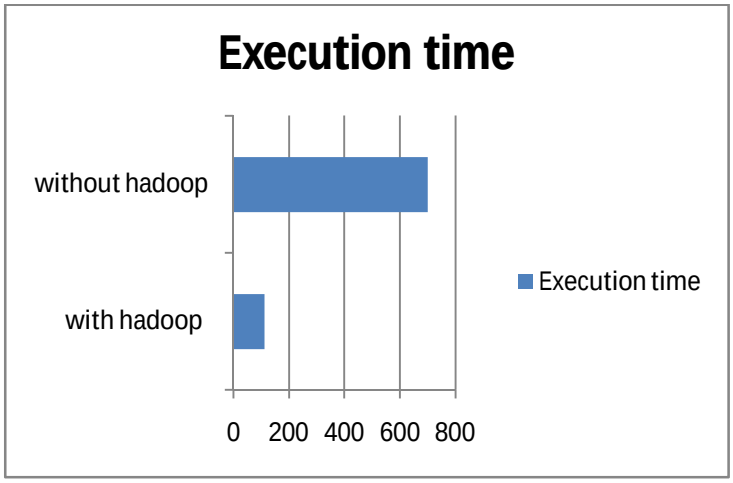

FIG 4:EXECUTION TIME FOR LARGER FILES

From fig 4 we can see that for files that are greater than 100 mb in size these files contain huge amount of data a txt file containing 100mb data is quite a large file for many system hence the time taken by normal system for computing these values is quite significant as comapred to hadoop based system, hence the time taken by hadoop to perform computations is very less.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

4. Big data greater than 1 GB (for multiple hashtags)

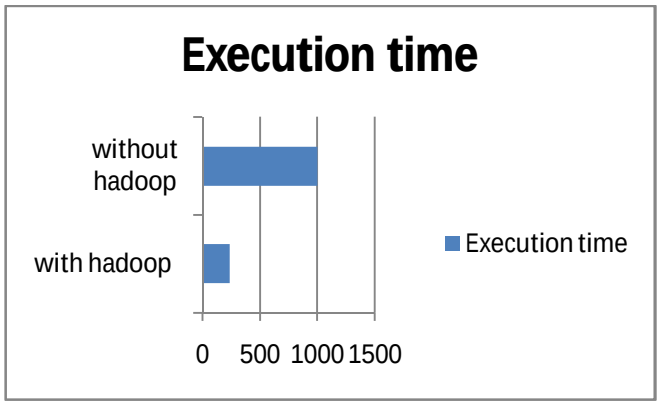

FIG 5:EXECUTION TIME FOR BIGDATA

Fig 5 describes a .txt file that is more than 1gb in size is really a huge file it may contain cores of tweets and the normal system will crash 9 out of 10 times during executing these dataset hence hadoop ecosystem is used. The hadoop ecosystem will not only prevent the system from crashing but it will also produce results faster and more effiecntly.

File size	With hadoop	Without hadoop
>5mb	25s	45s
50mb-70mb	62s	45s
100 mb	713s	60s
1gb	998s	257s

TABLE 1:EXECUTION TIME COMPARISION

Table 1 decribes and summarizes the results form above analysis, Since the data generated and fetched from twitter is in gbs or tbs hadoop will surely give the upper hand in execution from the local machines.the above table summarizes the execution time of various file size with and without hadoop integration. As twitter is the largest source of data generation in and around the globe hadoop integration will certainly benfot the analysis and provide is effecint and faster results.

VII. SCOPE OF THE SYSTEM

The same system base can be used by online shopping giants like flipkart and amazon for generating stars for the comments on their product rather than letting user enter the stars manual this will help users buy the desired product more precisely. The same system could be used by manufactures to get a review or opinion of particular product of the company to improve their product this will improve the company profits and sales.

VIII. FUTURE WORK

The field of computer science is an endless cycle of development and with every upgraded version there comes a better and more efficient technology. As this system uses hadoop framework, hadoop being a open source development framework sees the ever increasing development with it, in this system we have used hadoop map reduce framework for performing analysis and the hadoop jar has to be processed separately and output needs to be passed to the system, this same can be avoided by using SPARK[9][10] one of the latest component of hadoop framework using SPARK will enable us to reduce the execution time by 10x-100x as SPARK uses the concepts of RDD and the manual execution of the JAR file for map reduce framework would also be avoided. This will make the system more user friendly automated and faster and also more efficient.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

IX. CONCLUSION

Using the above steps we have implemented a system that process and gives us results of sentiment for a particular hashTag using hadoop and without hadoop as well. The system gives us output in the form of pie chart for single hashTag and as a bar graph for multiple hashTag , using this system we can do analysis of the most trending hashtags on twitter and also compare the same with similar hashtags.

ACKNOWLEDGMENT

The authors would like to thank their guide SonaliMuley and HOD P.M.Daflapurkar for their support throughout the year for their support. This work was carried out as an academic project for final year of engineering from Savitribai phulepune university.

REFERENCES

1. Dmitry Davidov, Oren Tsur& Ari appoport"Enhanced Sentiment Learning Using Twitter Hashtags and Smileys," Coling 2010: Poster VolAugust 2010, pp241–249,Beijing,.
2. Luciano Barbosa, "Robust Sentiment Detection on Twitter from Biased and Noisy Data"Coling 2010: Poster Vol, August 2010pp 36–44,Beijing,
3. Anna Jurek, Yaxin Bi, Maurice Mulvenna" Twitter Sentiment Analysis for Security-Related Information Gathering" 2014 IEEE Joint Intelligence and Security Informatics Conference pp 48-55
4. Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan, Jiajun Bu, Chun Chen, and Xiaofei He" Interpreting the Public Sentiment Variations on Twitter" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 5, pp 1158-1169, MAY 2014
5. Nishadpatil, Sayalitingre, Kalyanithorat, Swapnilshivshetty "A Survey Paper On Twitter Sentiment Analysis Using portat Stemming Algorithm" internationl journal of science and research VOL 4 Issue 10, pp 1707-1708 OCTOBER 2015
6. SatishGopalani, RohanArora" Comparing Apache Spark and Map Reduce with Performance Analysis using K-Means" International Journal of Computer Applications Vol 113 pp 8-11
7. Jayashri Khairnar1, Mayura Kinikar2" Sentiment Analysis Based Mining and Summarizing Using SVM-MapReduce" International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, pp 4081-4085.

BIOGRAPHY

Nishad Patil B.E. final year student from MMIT, lohgaon, Pune affiliated to Savitribai Phule Pune University. He is submitting this research work as a part of the B.E .project work

SayaliTingre B.E. final year student from MMIT, lohgaon, Pune affiliated to Savitribai Phule Pune University. She is submitting this research work as a part of the B.E .project work

Swapnil Shivshetty B.E. final year student from MMIT, lohgaon, Pune affiliated to Savitribai Phule Pune University He is submitting this research work as a part of the B.E .project work

Kalyani Thorat B.E. final year student from MMIT, lohgaon, Pune affiliated to Savitribai Phule Pune University. She is submitting this research work as a part of the B.E .project work.