International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 1 ǁ 2022 ǁ PP. 85-90

Analysis of Opinion Mining On Twitter Data Using Big Data Tools

Fatimah Zehra Islam
M. Tech. Scholar, All Saints' College of Technology, Bhopal, India

Zuber Farooqui
Professor, Dept. of CSE, All Saints' College of Technology, Bhopal, India

ABSTRACT*: With speedy innovations and growing web population, petabytes of data area unit beinggenerated each second. Process this monumental knowledge and analyzing may be a tediousmethod now-a-days. The quantity of information in period of time is growing rapidly. Nearly80% of the info is in unstructured format. Analysis of unstructured knowledge in period of timemay be a terribly difficult task. Existing traditional business intelligence (BI) tools perform bestonly in a pre-defined schema. Most of the real-time data are logs and don"t have any definedschema.. In this paper, a solution has been proposed that fetches real time twitterdata and stored into hadoop components. After storing, sentiment analysis has been performedon these data using big-data analytical tools like: Apache Flume, Apache hive and Apache pig.Finally, their performance comparison has been presented. Later comparison on the approach which hasbeen proposed and the approach which are existingwith the help of parameters, likeprecision, recall, F-measure and accuracy, and results show when the data is provided to theapproach proposed it gives, precision of 93.26, and accuracy of 91.73, and the same data when itis applied to the existing approach it show the precision of 89.56 and accuracy of 88.67 whichclearly shows that the proposed approach gives better outcomes than the existing approach*
KEYWORDS: *Apache Flume, opinion mining, Twitter Data, HDFS, Apache hive , Apache pig*

Date of Submission: 15-01-2022 Date of acceptance: 30-01-2022

I. INTRODUCTION

Micro blogging is a very famous and popularcommunication tool used among the Internet users [1]. Twitter is one of the big and largest social media sites which receive millions of tweets everyday on different and variety of important and trending issues. Users who post their tweets writeabout their condition, life, share opinions on variety of topics and discuss the hot and currentissues. These posts are then analyzed by Government, Elections, Business, Product review etc.for decision making. Sentiment analysis is therefore, one of the important areas of analysis oftwitter posts that can be very helpful in decision making. Social media has gained enormous popularity within marketing teams [2], and Twitter is aneffective tool for a corporation to get people excited about its new products launched. Twittermakes this easy to engage users and communicate straightly with them, and in turn, users will beable to provide word-of-mouth marketing for companies by discussing the products [3]. Givenlimited resources, and understanding it may not be able to speak with everyone that is the targetstraightly, marketing departments can be more effective by being selective about whom youreach out to rather than carrying out field surveys for acquiring feedback.

Performing and doing Sentiment Analysis on Twitter is more difficult than performing it forhuge reviews [4]. This is because the tweets are very small and short (only about 140 characters)and usually contain emotions, slangs, hash tags and other twitter exact jargon. For theimprovement of purpose twitter provides streaming API [5] which permits the developer anaccess to 1% of tweets tweeted at that time bases on the specific keyword. The object about thatthe sentiment analysis is done and performed on, is submitted to the twitter API‟s which doesadditional mining and provides the tweets related to only those objects.

Twitter data iscommonly unstructured example: using of abbreviations is very high. Also it sanctions the use ofemoticons which are direct pointers of the author‟s view on the subject. Tweet messages as wellas consist of a timestamp and the user name. This timestamp is useful for guessing the future rend application [6] of this project. User location if available can also help to gauge the trends indifferent geographical regions.

Sentiment analysis also recognized and known [11] as opinion mining [12]. Opinion mining ishelpful to companies to get business insights. The process of computationally identifying andcategorizing opinions spoken in a piece of text, particularly in order to define the user‟s attitudetowards a particular topic or a product. Sentiment Analysis is the process of detecting thecontextual polarity of text. In other words, it reflects that a piece of writing is positive, negative

or neutral [13].Sentiment analysis is enormously beneficial in social media monitoring [14] as it permits us togain an overview of the extensive public opinion behind certain topics. Social media monitoringtools like Brand-watch Analytics make that process quicker and easier than ever before.

Theapplications of sentiment analysis are powerful and broad. The capability to extract insights from social data is a practice that is being broadly accepted by organizations across the world toenhance the services provided by them. Changes in sentiment on social media have beendisplayed to correlate with changes in the stock marketToday the people are living in the world which is surrounded by 99% of data. There are differentmicroblogging sites where users express their visions about various products these sights andviews are nothing but opinions of people and it will go waste if it is not utilized in suitable wayso there is a need to use opinions of people in developing productivity, functionality of particularproduct, usefulness, application, technique or any entertainment resource. Hence, there is a requirement to develop a product which can analyze opinions of people. This product will beuseful in increasing market value of industries also satisfy needs of customers.There are various challenges in information filtering in micro-blogging environment. They are asfollows:

- Short texts: In Twitter, the text of a post is restricted to 140 characters. In terms of textclassification, short texts contain sparse data; therefore it is a challenge to classify them.
- Informal Language: Another challenge is of the informal structure of the language usedon Twitter. It contains slangs, abbreviations, stop words etc. So, it is important to identifykeywords and common words useful for text classification.
- Different Languages: Twitter is used by users around the world, therefore it containstweets in many languages.
- Identifying topics: It is necessary to identify relevant topics and filter out tweets withirrelevant topics.
- Constantly changing vocabulary: The vocabulary is constantly changing with newwords and phrases being added. So, there is a need for dynamic text classification system

The objectives for carrying out sentiment analysis can be as follows:

1. Content Retrieval: The large amount of data is collected using java Twitter streamingAPI.
2. Storage: This data is kept and stored in a certain format (HDFS: Hadoop Distributed Filesystem) therefore as to form key value pair that is needed to feed to mapper in mapreduce programming approach. The data which is stored in Hadoop Distributed File System.
3. Data Processing: Data collected over a period of time is processed by using hive and distributed processing software framework developed by Apache Hadoop and using mapreduce programming model and Apache hive frame work.
4. Analysis of Data: The output gained from reducer phase is analysed.
5. At the end outcome is gotten in the form of classified tweets that is Positive, Negativeand Neutral tweets.

The rest of thesis is organized as follows:-

Section 2:- outlines the related research background. And from studying the literature surveys.

Section 3:- presents the proposed algorithm and methodology.

Section 4:- describe the experiment and results, The resultsand analysis done on the twitter data, which is shown with the help of tables.

Section 5:- conclusion of the paper and future work is indicated.

II. RELATED WORK

Opinion mining is one of the most popular trends in today‟s world. Lot of research and literature surveys is being done in this sector. Bo Pang and Lillian Lee are pioneers in this field [18]. Current works in this field which uses a mathematical approach using algorithms for opinion polarity are based on a classifier trained using a collection of annotated text data. Before training, data is preprocessed so as to extract only the main content. Some of the classification methods have been proposed are Naïve Bayes, Support Vector Machines, K-Nearest Neighbors etc. Continuous research is being done to determine most efficient method for opinion mining.

Chawda et al. [19] describes that Big data analytics has attracted extreme interest from allindustry and academia recently for its effort to extract knowledge, information and wisdom formbig data. Big-data and cloud computing, two of the most significant trends that are defining thenew emerging analytical tools. Big data analytical capabilities using cloud delivery models couldease adoption for numerous industry, and most significant thinking to cost saving, it couldsimplify useful visions that could providing them with various kinds of competitive benefit.

Basaille et al. [20] have made a multi-paradigm framework by the name of SNFreezer which can

fulfil the requirements of tweet analysis and reduce the waiting time for those people whom theywant do research process on twitter data to employ storage system and computational resourcesto help large amount of data analysis.

Their main and basic approach for this paper which they have done is to associate concernsaboutdataharvesting, data storage, data visualisation and data analysis in a framework that helpinductive reasoning in the technical research.

Lai et al. [21] focuses on the communication of twitter which is using #hashtags especially onthe topic of Marriage Pour Tous that had happene in France in 2012-2013 which became thetopic and subject of the debate and controversy after which became very famous in societyIn this paper, they collected all of the tweets that had been marked and signed by the hash tag#marriagepourtous and then they applied variety types of analysis on those hashtags.

Doong et al. [22] investigates the difficulty of predicting twitter hash-tags popularity level adata set that contains of more than 18 million tweets including 748 thousand hash-tags have beenprepared by using Twitters rest API. Early adoption properties containing profile of tweetwriters and adoption time series are used to foretell a tag"s later popularity levelprécised by Twitter to make a list of current trending tags.

Bhardwaj et al. [23] describes about need of Big Data to make decision over complex problem. Big

Data is a term that refers and is said to collection of huge data sets comprising immense amountof data whose size is in the range of Petabytes, Zettabytes, or with high rate of growth, andcomplexity that make them hard to process and analyze using conventional databasetechnologies. Big-Data is produced and generated from different sources such as socialnetworking sites like Facebook, Twitter etc., and the data which is generated can be in variousformats like structured, semi-structured or unstructured format. For taking out valuableinformation from this vast amount of Data, new tools and techniques is a need of time for theorganizations to derive business benefits and to obtain competitive advantage over the market. In

this paper a comprehensive study of major Big -data emerging technologies by highlighting theirimportant features and how they work, with a comparative study between them is presented. Thispaper as well shows performance analysis of Apache Hive query for executing Twitter tweets inorder to calculate Map Reduce CPU time spent and total time taken to finish the job[24] .

Danthala et al. [25] present how of analyzing of massive knowledge like twitter knowledgevictimization Apache Hadoop that is able to method and analyzes the tweets on a Hadoopclusters. This additionally contains visualizing the results into pictorial depictions of twitter usersand their tweets.

Kumar et al. [26] economical Capabilities of process of massive knowledge victimization Hadoop Map cut back Proposes, many solutionsto the large knowledge drawback have emergedwhich incorporates the Map cut back surroundingschampioned by Google thatis currentlyaccessible ASCII text file in Hadoop. Hadoop distributed process, Map cut back algorithms andoverall design area unit a serious step towards achieving the secure edges of massive knowledge.

III. PROPOSED METHDOLOGY

Social media website is one of the popular media right now to share opinions or different oftopics and twitter is very popular social site to share everything related to opinions on differentof topics and discussions on current issues. These tweets generate the huge information relatedto different area like government, election, etc. millions of tweets are generated every day andwhich is very useful for decision making because everyone share their opinions and views onissues or variety of topics. Twitter sites receives petabytes of data every day and these data isnothing but a collection of tweets so these data are very important in real life to analyze differentscenario through which it helps us in decision making.

The analysis of twitter data gives realvision or different user opinions regarding what they think and to analysis these data provide abetter way for making any decision.For analyzing these huge and complex data requires a powerful tool, Hadoop is used that is an

open-source implementation of map-reduce, a powerful tool designed for deep analysis andtransformation of very large data

Algorithm Steps are as follow:

Step 1: users can share their opinions by posting a variety of tweets on twitter.

Step 2: all these tweets are stored in twitter database center, there are millions of tweets areposted every day on twitter which can generate petabytes of data which is stored on twitter datacenter.

Step 3: for analysis these huge and complex twitter data are needed which contains variety ofopinions posted by different users, flume is used to fetch these twitter data and store them intoHDFS, a twitter API is generated through which the real time twitter data is fetched from weband store them into HDFS.

Step 4: After storing these huge and complex twitter json data, an analyzing tool is needed toanalyze these complex data, for these hive is used which runs on top of the hadoop and takesinput from HDFS and its support SQL queries through which the data can be analyzed.
Step 5: Based on the analysis result from hive, the polarity of the tweets can be checked with theaid of polarity dictionary which contains a number of English words with their polarity from -5to +5 which indicates negative to positive and by joining these words polarity we can take adecision that which tweets are positive meaning and a negative meaning.

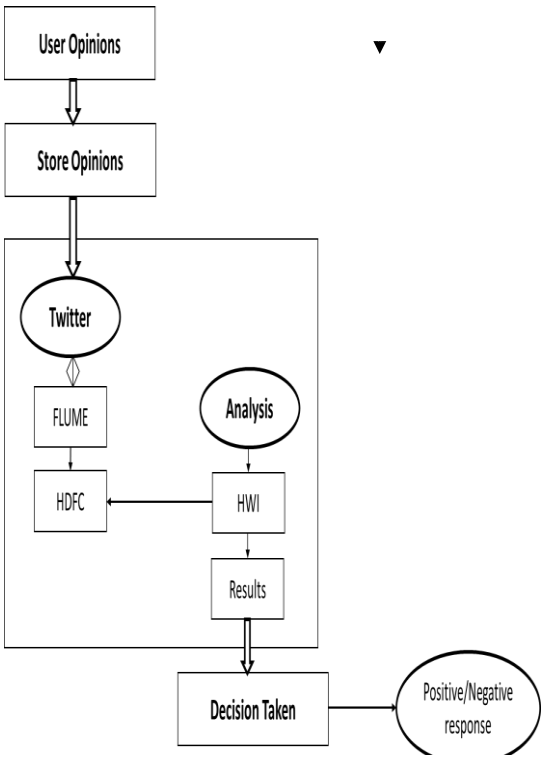


Figure 1 : Proposed methodology block diagram

IV .RESULT ANALYSIS

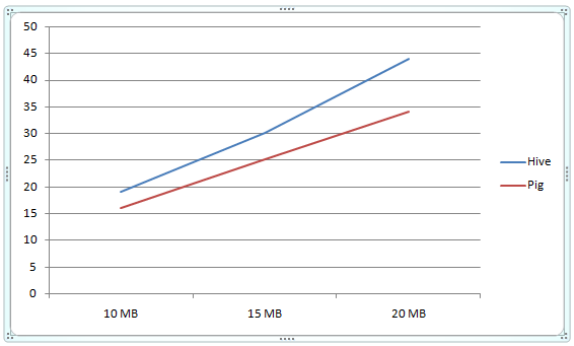
In this chapter experimental setup done, in this dissertation will be discussed. In this dissertation,real time sentiment analysis of Twitter data using HWI is done and method to find polarity oftweets is proposed. To accomplish this, all the experiments were performed using an i5-2410MCPU @ 2.30 GHz processor and 4 GB of RAM running ubuntu 14. As we have seen theprocedure how to overcome the difficulty which are facing in the existing problem that is shownclearly in the proposed system. So, to achieve this, the following method should be followed.

1. Creating Twitter Application
2. Getting data using Flume.
3. Analyzing using Hive Query Language (HQL)
4. Analyzing using Apache Pig

In section 2, a twitter application already created and fetching the twitter data sets from twitterdatabase using apache flume and these data sets are stored in HDFS, By default the format of thedatasets are SON (java script object notation) [23] data, and now these datasets are analyzedusing apache hive and apache pig

4.1 Analysis Performance Comparison

After analyzing the twitter data the polarity of tweets is gotten, in this thesis the compressionbetweenperformance of Apache pig and Apache hive is don for analyzing JSON data. For thisdifferent size of dataset is gotten on which the analysis can be performed using hive and pig.
The execution time taken by both the analytical tools on different size datasets are shown infigure 2.



Dataset size in (MB)

Figure: 2 Execution time taken by hive & pig

After getting the query execution time taken it said that pig performance for analyzing JSONdata is taking less time as compared to hive. in this result is shown that pig is best suitable foranalyzing JSON data, and pig is also best from generating less number of mapreduce job that‟sby its taking a less time as compared to hive. For this twitter data analysis pig is generating twomapreduce job and hive is generating five mapreduce job for analyzing twitter data, so it is saidthat pig is better in many parameters as compared to hive.

Table 1: Performance evaluation of this proposed work

Method	Precision	Recall	F-Measure	Accuracy
Existing	89.56	72.35	81.32	88.67
Proposed	93.26	75.54	85.14	91.73

V. CONCLUSION

Opinion Mining may be a terribly wide branch for analysis, a number of its necessary aspectshave been lined. an equivalent design may well be used for a spread of applications designed toseem at Twitter knowledge, like distinguishing spam accounts, or distinguishing clusters ofkeywords. In this the popularity of tweets can also be identified by which it can be said thatwhich tweet have a positive meaning or a negative meaning. In this paper the twitter data isfetched by using flume and store them into the HDFS and then these data are analyzed by usinghive and pig, The results and analysis done on the twitter data, which is shown with the help oftables, and diagrams, later the comparison is done between the tools on which thesentiment analysis has been done. And after that, this conclusion gotten that pig runs faster andworks in fewer map-reduce works compare to hive. Later comparison on the approach which hasbeen produced and the approach which are exiting, with the help of parameters, like precision,recall, F-measure and accuracy, and results show when the data is provided to the approachproposed it gives, precision of 93.26, and accuracy of 91.73, and the same data when it isproposed to the existing approach it show the precision of 89.56 and accuracy of 88.67 whichclearly shows that the approach proposed is better than the existing

Twitter API is simple to use and easily accessible. There are infinite possibilities on what we cando. There are some limitations on the number of queries and data that Twitter allows you toget every 15 minute. However an easy way around is that to get more access tokens. In the futuredirection is to fetch real time data from different sources and Hadoop tools such as Oozie couldbe considered for automation of the analysis steps.

REFERENCES

[1]. Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, andMore---Matthew A. Russell.
[2]. G. Szabo, and B.A. Huberman, “Predicting the Popularity of Online Content”,Communication of the ACM, 2010, 53(8), pp. 80-88.
[3]. R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, “Improving LDA topic models formicroblogs via tweet pooling and automatic labeling,” in Proceedings of the 36thInternational ACM SIGIR Conference on Research and Development in InformationRetrieval, ser. SIGIR ‟13. New York, NY, USA: ACM, 2013, pp. 889–892.

[4]. E. Cunha, G. Magno, G. Comarela, V. Almeida, M. A. Goncalves, and F. Benevenuto,"Analyzing the dynamic evolution of hashtags on twitter: a language-based approach," inProceedings of the Workshop on Language in Social Media (LSM 2011). Portland, Oregon:Association for Computational Linguistics, 2011, pp. 58–65.

[5]. "The Streaming APIs." Twitter Developers. N.p., n.d. Web. 23 Oct. 2014.

[6]. Y. Wang, J. Liu, J. Qu, Y. Huang, J. Chen, and X. Feng, "Hashtag graph based topic modelfor tweet mining," in Data Mining (ICDM), 2014 IEEE International Conference on, Dec2014, pp. 1025–1030.

[7]. H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a Social Network or a NewsMedia?", In: Proceedings of the 19th International Conference on World Wide Web, 2010,pp. 591-600.

[8]. McKinsey, Big Data: The Next Frontier for Innovation, Competition, and Productivity,McKinsey & Company, 2011, http: //www.mckinsey.com/.

[9]. Sagiroglu, S., &Sinanc, D, "Big data: A review", IEEE International Conference onCollaboration Technologies and Systems (CTS), 2013, pp 42-47.

[10]. K. W. Lim and W. Buntine, "Twitter opinion topic model: Extracting product opinionsfrom tweets by leveraging hashtags and sentiment lexicon," in Proceedings of the 23rdACM International Conference on Conference on Information and KnowledgeManagement, ser. CIKM ″14. New York, NY, USA: ACM, 2014, pp. 1319–1328.

[11]. K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop Distributed FileSystem," in the 26th IEEE Symposium on Mass Storage Systems and Technologies, pp. 1-10, May 2010.

[12]. K. W. Lim and W. Buntine, "Twitter opinion topic model: Extracting product opinionsfrom tweets by leveraging hashtags and sentiment lexicon," in Proceedings of the 23rdACM International Conference on Conference on Information and Knowledge Management,48ser. CIKM ″14. New York, NY, USA: ACM, 2014, pp. 1319–1328.

[13]. S. Li, G. Huang, R. Tan, and R. Pan, "Tag-weighted Dirichlet Allocation," in ProceedingsOf the 13th International Conference on Data Mining, ser. ICDM″13, vol. 0. Los Alamitos,CA, USA: IEEE Computer Society, 2013, pp. 438–447.

[14]. T.M. Saravanan and A. Tamilarasi, "Effective Sentiment Analysis for Opinion MiningUsing Artificial Bee Colony Optimization" in Hellenic Research: International Journal ofApplied Sciences, 828-840, 2016.

[15]. Bing Liu. Web data mining; exploring hyperlinks, contents, and usage data, chapter 11:Opinion Mining. Springer, 2006.

[16]. M. Cha, H. Haddadi, F. Benevenuto, and K.P. Gummadi, "Measuring User Influence inTwitter: The Million Follower Fallacy", In: Proceedings of the 4th International AAAIConference on Weblogs and Social Media, 2010.

[17]. O. Tsur, and A. Rappoport, "What″s in a Hashtag? Content Based Prediction of the Spreadof Ideas in Microblogging Communities". In: Proceedings of the 5th ACM InternationalConference on Web Search and Data Mining, 2012, pp. 643-652.

[18]. Mirko Lai, Cristina Bosco and Viviana Patti, Daniela Virone, "Debate on PoliticalReforms in Twitter: A Hashtag-driven Analysis of Political Polarization" in IEEE, 978-1-4673-8273-1/15, IEEE 2015.

[19]. Rahul Kumar Chawda, Dr. Ghanshyam Thakur, "Big Data and Advanced AnalyticsTools",IEEE2016, in Symposium on Colossal Data Analysis and Networking (CDAN).

[20]. IAN BASAILLE at el, "Towards A Twitter Observatory: A Multi-ParadigmFramework For Collecting, Storing And Analyzing Tweets", 978-1-4799-8710-8/16/$31.00 ©2016 IEEE

[21]. Mirko Lai, Cristina Bosco and Viviana Patti &Daniela Virone," Debate on PoliticalReforms in Twitter:A Hashtag-driven Analysis of Political Polarization", in 2015, 978-1-4673-8273-1/15/$31.00 c 2015 IEEE.

[22]. Shing H. Doong, "Predicting Twitter Hashtags Popularity Level", in 2016 49th HawaiiInternational on System Sciences, IEEE, DOI 10.1109/HICSS.2016.247

[23]. URL: http://www.json.org/ECMA-404 The JSON Data Interchange Standard.

[24]. Judith SherinTilsha S, Shobha M.S.," A Survey on Twitter Data Analysis Techniques toExtract Public, 06/Nov/2016-3:15 PMOpinion.", IJARCSE , Vol. 5, Issue 11 , Nov 2015 , 2277128X.

[25]. Manoj Kumar Danthala, "Tweet Analysis: Twitter Data processing Using Apache Hadoop", International Journal of Core Engineering & Management (IJCEM) Volume 1,Issue 11, February 2015, pp 94-102.

[26]. Praveen Kumar, Dr Vijay Singh Rathore," Efficient Capabilities of Processing of Big Datausing Hadoop Map Reduce", International Journal of Advanced Research in Computer andCommunication Engineering Vol. 3, Issue 6, June 2014, pp 7123-7126