Scholars Journal of Engineering and Technology (SJET) **ISSN 2321-435X (Online)**
Sch. J. Eng. Tech., 2017; 5(9):473-477 **ISSN 2347-9523 (Print)**

(An International Publisher for Academic and Scientific Resources)
www.saspublisher.com

Multilingual Sentiment Analysis on Twitter dataset using Naive Bayes Algorithm

Natasha Suri[1], Prof. Toran Verma[2]
[1]RCET, Bhilai Dept. of Computer Science and Engineering Bhilai, Chhattisgarh, India
[2]RCET, Bhilai Dept. of Information & Technology Bhilai, Chhattisgarh, India

***Corresponding author**
Natasha Suri

Article History
Received: 20.07.2017
Accepted: 26.07.2017
Published: 30.09.2017

DOI:
10.21276/sjet.2017.5.9.4

Abstract: Sentiment Analysis which frequently passes by the name opinion mining is one of the noticeable field in lots of research is going ahead because of its interminable application like online networking monitoring, product reviews and so on. Be that as it may, because of the noticeable utilization of social media the utilization of multilingual statements has turned out to be most basic as client tends to in their own particular safe place. These multilingual statements emerge due the utilization of more than one language to create a statement. Because of absence of clear grammatical structure it is exceptionally hard to discover correct sentiment out of it. This paper presents the analysis of sentiments of 4 languages tweets by applying Naïve Bayes algorithm. Our proposed method, effectively identify the sentiments of the users by utilizing their twitter walls comments and posts. We have analyzed a very famous movie, "Baahubali" with hash tag of "Baahubali2".
Keywords: NLP, Text Mining, Machine Learning, Multilingual Sentiment Analysis

INTRODUCTION

Opinion, reviews and comments of the people plays a very important role to figure out if a given populace is satisfied with the item, services and predicting their reaction on specific occasion of interest like review of a movie. These information are fundamental for opinion mining. Keeping in mind the end goal to find the sentiment of population retrieval of information from sources like Twitter, Facebook, Blogs are essential. Multilingual sentiment investigation turned out to be considerably more troublesome as the assets required are to be worked without any preparation [11,12]. Because of immense increment in the client created multilingual substance via web-based networking media and need in computerized system to identify it the Natural Language processing (NLP) community has attempting to grow new technique to manage this marvel and find hidden sentiment out of it. This paper essentially contains different strategy utilized for the multilingual sentiment analysis and its correlation [15,16].

The Existing Database is not ready to handle huge measure of information with in specified measure of time. Likewise this sort of database is constrained for handling of organized information and has a constraint when managing real time information. In this way, the traditional solution can't help association to manage and process unstructured information. With the utilization of Big Data advances like Hadoop is the most ideal approach to comprehend Big Data challenges. This help association to handle expansive measure of information in a systematic way.

Apache Hadoop and its Architecture

The Apache Hadoop programming library is a system that takes into account the distributed processing of extensive information sets crosswise over clusters of PCs utilizing simple programming models. It is designed to scale up from single servers to a large number of machines, every offering neighborhood computation and storage [13,14].

As opposed to depend on equipment to convey high-accessibility, the library itself is intended to recognize and handle failures at the application layer, so conveying an exceedingly accessible administration on top of a cluster of PCs, each of which might be inclined to failure [1].

Name Node and Data Node

Name node stores the data about Meta information which maps to the data node for real information. Data node contains the genuine information [17].

Data Replication

HDFS stores each file as an arrangement of blocks. These blocks are replicated to different racks on HDFS for adaptation to non-critical failure. The block size and replication variable can be configured from the configuration file of Hadoop [18].

Racks

Racks are the collection of data node. The data node which belongs to same system can be dealt with as one rack. On the off chance that one of the data node crashes, the replica of that data node which is available on another node begins moving to the failed data node [19,20].

MapReduce Architecture

Hadoop MapReduce is a software framework for executing tremendous measure of information i.e. terabyte data sets in parallel environment on large clusters (in a huge number of data nodes) which can be commodity hardware in a fault tolerant manner.

MapReduce jobs splits the input information set into different pieces of files which then are handled by the guide assignments in parallel form. The hadoop framework sorts the output of map phase which are then input to the reduce tasks. Both input and output files are stored on HDFS (Hadoop Distributed File System). The Hadoop framework has a duty of managing and scheduling tasks.

The MapReduce framework comprises of a si8ngle master JobTracker and one slave TaskTracker per cluster node. The master is responsible for scheduling the jobs' component tasks on the slaves, monitoring them and re-executing the failed tasks [2].

The execution of job begins when client submit the job to the job tracker with job configuration, which help to specifies map and reduce function and different parameters, for example, input and output way of data set.

Job Tracker

The Job Tracker is the service with Hadoop that farms out MapReduce tasks to specific nodes in the cluster, in a perfect world the nodes that have the information, or possibly are in a similar rack [1]. Client applications submit jobs to the Job tracker [2]. The JobTracker converses with the NameNode to determine the area of the information [3]. The JobTracker locates TaskTracker nodes with available slots at or near the data [4].

The JobTracker presents the work to the picked TaskTracker nodes [3]. The JobTracker is a state of failure for the Hadoop MapReduce services. If it goes down, all running jobs are halted.

Task Tracker

A TaskTracker is a node in the cluster that acknowledges tasks - Map, Reduce and Shuffle operations - from a JobTracker. Each TaskTracker is configured with an arrangement of slots; these demonstrate the number of tasks it can accept. At the point when the JobTracker tries to find some place to plan an assignment inside the MapReduce operations, it first searches for an empty slot on a similar server that has the DataNode containing the information, and if not, it searches for a vacant slot on a machine in the same rack [4].

LITERATURE SURVEY

One of the most common datasets exploited by many Corporations to conduct business intelligence analysis are event log files.

Jai Prakash Verma [5], designed a recommendation system which provides the facility to understand a person's taste and find new, desirable content for them automatically based on the pattern between their likes and rating of different items.

Subramaniyaswamy [6], focused on Unstructured Data Analysis on Big Data using Map Reduce. The proposed method will process the data in parallel as small chunks in distributed clusters and aggregate all the data across clusters to obtain the final processed data. The proposed method is enhanced by using the techniques such as sentiment analysis through natural language processing for parsing the data into tokens and emoticon based clustering. The process of data clustering is based on user emotions to get the data needed by a specific user. The results show that the proposed approach significantly increases the performance of complexity analysis.

Can Uzunkaya [7], focuses on Hadoop and its ecosystem and implementation Hadoop based platform for analyzing on collected tweets. The regarding analyzed results are transferred to graphical charts which is showed on a web page.

Manoj Kumar Danthal [8], proposed a model in which data is processed and analyzed using Info Sphere Big Insights tool which bring the power of Hadoop to the enterprise in real time. This also includes the visualizations of analyzing big data charts using big sheets.

Gaurav and Rajurkar [9], provide solution for speedy data downloading on HDFS by using source and sink (data ingestion) mechanism. The Hadoop is

flexible and scalable architecture. The proposed work is based upon the phenomenon of combination of open source software along with commodity hardware that will increase the profit of IT Industry.

Efthymios Kouloumpis [10], investigated the utility of linguistic features for detecting the sentiment of Twitter messages and evaluate the usefulness of existing lexical resources as well as features that capture information about the informal and creative language used in microblogging.

Rudy *et al.* [20], proposed a technique in light of a consolidated approach which included control based grouping, directed learning and machine learning. A 10 crease cross approval was completed for each example set. A cross breed characterization technique is utilized as a part of which a few classifiers cooperate. In the event that the primary classifier neglects to characterize then it is passed on to the following classifier. The procedure proceeds until the record is grouped or there is no other classifier left.

Zhu *et al.* [21], proposed an approach in light of fake neural systems to separate the archive into positive, negative and fluffy tone. The approach depended on recursive slightest squares back spread preparing calculation.

METHODOLOGY

The system architecture work flow is presented in fig-1. The following languages tweets we have considered:

- English
- Hindi
- Telgu
- Tamil

Firstly the tweets are downloaded via Twitter Achiever. It is stored into the text file. Then various analyses are performed on that dataset.

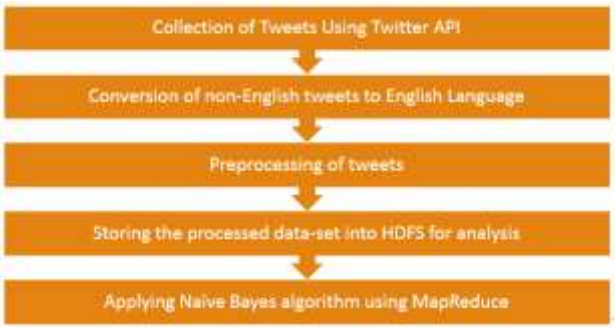

Fig-1: The Work flow of sentiment analysis framework

Downloding Twitter Dataset

The twitter dataset are downloaded via twitter achiever. After that it is stored in a text file for further processing,

Conversion of non-english langauge to english language

The conversion took place of non-English language. Here Hindi, Telgu, Tamil language are converted to English language via Google translator.

Processing of Tweets:

After conversion, the tweets are pre-processed. The preprocessing steps includes:

- Removal of special and unwanted symbol
- Removal of URL's
- Removal of White spaces
- Conversion of emoticons into its equivalent sentiment word.
- Removal of Hashtag.
- Removal of Username.

Storing into HDFS

After pre-processing the dataset is ready to be analyzed. The datasets with different languages are into HDFS so that map reduces function can use those dataset.

Apply Naïve Bayes Algorithm

The algorithm Naïve Bayes is implemented for the all language dataset. It process the each tweets word by word and store the score into HDFS. Naïve Bayes algorithm presents output into following format:

- Positive
- Very Positive
- Negative
- Very Negative
- Neutral

RESULT

The analysis of the twitter datasets are presented in this section. We have considered 4 languages for analysis. The intermediate and final results are presented below.

For intermediate output, we have considered the Tamil tweets of movie Baahubali2. Fig-2: to 4 shows the results.

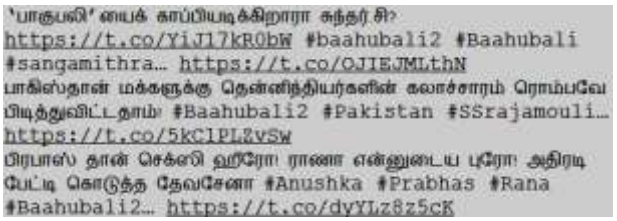
'பாகுபலி'யைக் காப்பியடிக்கிறாரா சுந்தர்.சி?
https://t.co/YiJ17kR0bW #baahubali2 #Baahubali
#sangamithra… https://t.co/OJIEJMLthN
பாகிஸ்தான் மக்களுக்கு தென்னிந்தியர்களின் கலாச்சாரம் ரொம்பவே
பிடித்துவிட்டதாம் #Baahubali2 #Pakistan #SSrajamouli…
https://t.co/5kClPLZvSw
பிரபாஸ் தான் செல்லி ஹீரோ! ராணா என்னுடைய ப்ரோ! அதிரடி
பேட்டி கொடுத்த தேவசேனா #Anushka #Prabhas #Rana
#Baahubali2… https://t.co/dyYLz8z5cK

Fig-2: snapshot of tweets downloaded from twitter

SundarCC to be able to save 'Pakubali'? Https://t.co/YiJ17kR0bW # baahubali2 #Bahahubali # sangamithra ... https://t.co/OJIEJMLthN
The culture of South Indian people has greatly appreciated the people of Pakistan! # Baahubali2 #Pakistan #Sarrajamouli ... https://t.co/5kC1PL8vSw
Prabhas's sexy hero! Rana my Pro! The interview given by Devasena #Anushka #Prabhas #Rana # Baahubali2 ... https://t.co/dyYLz8z5cK

Fig-3: snapshot of tweets converted to English

sundarcc to be able to save 'pakubali'? - # baahubali2 bahahubali # sangamithra ... -
the culture of south indian people has greatly appreciated the people of pakistan! # baahubali2 pakistan sarrajamouli ... -
prabhas's sexy hero! rana my pro! the interview given by devasena anushka prabhas rana # baahubali2 ... -

Fig-4: snapshot of pre-processed tweets

Analysis of results of 4 languages are presnted in fig-5 to 8.

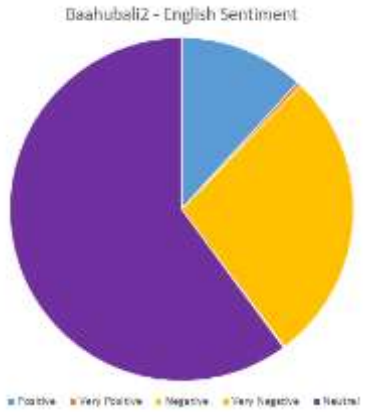

Fig-5: Shows the sentiment of English tweets

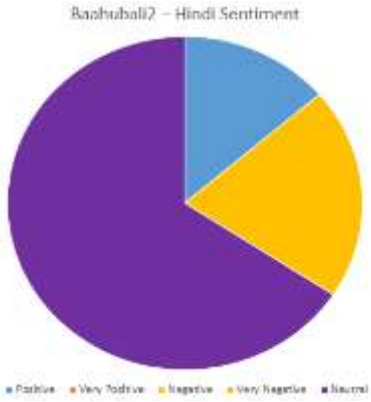

Fig-6: Shows the sentiment of Hindi tweets

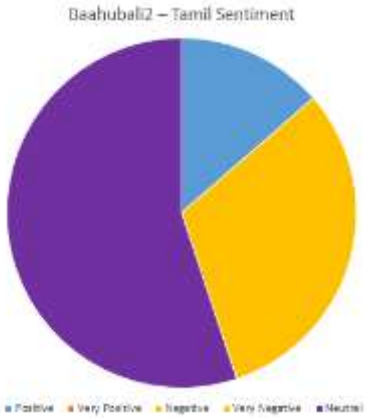

Fig-7: Shows the sentiment of Tamil tweets

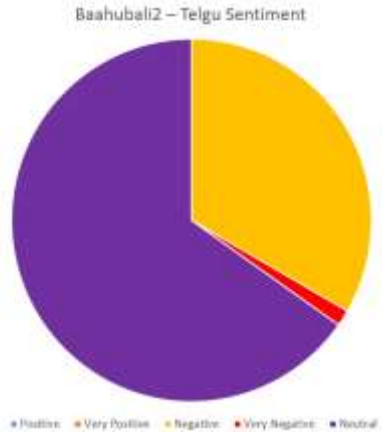

Fig-8: Shows the sentiment of Telgu tweets

CONCLUSION

From the analysis, we observed that the positive and very positive part of the tweets are heavily influenced by the noises presents in the dataset. The Proposed framework effectively demonstrate the relation of positive, very positive and neutral tweets. They are strongly correlated with each other. Hence for the movie Baahubali2 the sentiments of the users are around 73% and 27% for English language, 69% and 31% for Tamil language, 80% and 20% for Hindi language and finally 66% and 34% for Telgu language for positive and negative tweets respectively.

REFERENCES

1. Hadoop Introduction, https://hadoop.apache.org/
2. Apache MapReduce, https://hadoop.apache.org/docs/r1.2.1/mapredtutorial.html
3. JobTracker, https://wiki.apache.org/hadoop/JobTracker
4. TaskTracker, https://wiki.apache.org/hadoop/TaskTracker
5. Parveen H, Pandey S. Sentiment analysis on Twitter Data-set using Naive Bayes algorithm. InApplied and Theoretical Computing and

Communication Technology (iCATccT), 2016 2nd International Conference on 2016 Jul 21 (pp. 416-419). IEEE.
6. Subramaniyaswamy V, Vijayakumar V, Logesh R, Indragandhi V. Unstructured data analysis on big data using map reduce. Procedia Computer Science. 2015 Jan 1;50:456-65.
7. Uzunkaya C, Ensari T, Kavurucu Y. Hadoop ecosystem and its analysis on tweets. Procedia-Social and Behavioral Sciences. 2015 Jul 3;195:1890-7.
8. Danthala MK, Ghosh DS. Bigdata Analysis: Streaming Twitter Data with Apache Hadoop and Visualizing using BigInsights. International Journal of Engineering Research & Technology. 2015 May;4(05).
9. Rajurkar GD, Goudar RM. Notice of Violation of IEEE Publication Principles A Speedy Data Uploading Approach for Twitter Trend and Sentiment Analysis Using HADOOP. InComputing Communication Control and Automation (ICCUBEA), 2015 International Conference on 2015 Feb 26 (pp. 580-584). IEEE.
10. Kouloumpis E, Wilson T, Moore JD. Twitter sentiment analysis: The good the bad and the omg!. Icwsm. 2011 Jul 17;11(538-541):164.
11. Verma JP, Patel B, Patel A. Big data analysis: recommendation system with Hadoop framework. InComputational Intelligence & Communication Technology (CICT), 2015 IEEE International Conference on 2015 Feb 13 (pp. 92-97). IEEE.
12. Neri F, Aliprandi C, Capeci F, Cuadros M, By T. Sentiment analysis on social media. InProceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012) 2012 Aug 26 (pp. 919-926). IEEE Computer Society.
13. Shafer J, Rixner S, Cox AL. The hadoop distributed filesystem: Balancing portability and performance. InPerformance Analysis of Systems & Software (ISPASS), 2010 IEEE International Symposium on 2010 Mar 28 (pp. 122-133). IEEE.
14. Ren J, Lee SD, Chen X, Kao B, Cheng R, Cheung D. Naive bayes classification of uncertain data. InData Mining, 2009. ICDM'09. Ninth IEEE International Conference on 2009 Dec 6 (pp. 944-949). IEEE.
15. Dean J, Ghemawat S. Google Inc,". InMapReduce: simplified of the 4th Annual Symposium on Cloud Computing (SoCC'13) 2004 Dec.
16. Niu T, Zhu S, Pang L, El Saddik A. Sentiment analysis on multi-view social data. InInternational Conference on Multimedia Modeling 2016 Jan 4 (pp. 15-27). Springer, Cham.
17. Xu K, Liao SS, Li J, Song Y. Mining comparative opinions from customer reviews for Competitive Intelligence. Decision support systems. 2011 Mar 31;50(4):743-54.
18. Xia R, Zong C, Li S. Ensemble of feature sets and classification algorithms for sentiment classification. Information Sciences. 2011 Mar 15;181(6):1138-52.
19. Zhang Z, Ye Q, Zhang Z, Li Y. Sentiment classification of Internet restaurant reviews written in Cantonese. Expert Systems with Applications. 2011 Jun 30;38(6):7674-82.
20. Prabowo R, Thelwall M. Sentiment analysis: A combined approach. Journal of Informetrics. 2009 Apr 30;3(2):143-57.
21. Jian Z, Chen XU, Wang HS. Sentiment classification using the theory of ANNs. The Journal of China Universities of Posts and Telecommunications. 2010 Jul 1;17:58-62.