Advances in Computational Sciences and Technology
ISSN 0973-6107 Volume 10, Number 9 (2017) pp. 2789-2799

http://www.ripublication.com

Multilingual Sentimental Analysis on Twitter Dataset: A Review

Natasha Suri[1] and Prof. Toran Verma[2]

[1]Rungta College of Engineering and Technology
Dept. of Computer Science and Engineering, Bhilai, Chhattisgarh, India

[2]Rungta College of Engineering and Technology
Dept. of Information & Technology, Bhilai, Chhattisgarh, India

Abstract

Sentiment Analysis which frequently passes by the name opinion mining is one of the noticeable field in lots of research and 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 emerges 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. We survey various techniques which can be utilized to examine these multilingual proclamation accurately.

Keywords: NLP, Text Mining, Machine Learning, Multilingual Sentiment Analysis.

I. 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. 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.

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.

A. 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.

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].

i. 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.

ii. 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.

iii. Racks:

Racks are the collection of data node. The data node which belong 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.

B. 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].

II. LITERATURE SURVEY

One of the most common datasets exploited by many

corporations to conduct business intelligence analysis is event log files.

Jai Prakash Verma [5], designed a recommendation system which provides the facility to understand a persons 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 InfoSphere BigInsights 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.

Long-Sheng et.al. [22], consolidated the benefits of machine learning and data recovery methods utilizing a neural system based approach. B.Semantic Orientation The semantic introduction approach depends on unsupervised learning. It doesn‟t require any preparation keeping in mind the end goal to order the conclusion information. It is utilized to quantify how much positive or negative is the word‟s extremity.

Andrea et.al. [23], proposed a strategy in light of semi managed realizing, which presented a seed set and extended it later utilizing Word Net. The supposition was that the words with comparative introduction have comparable extremity.

Chunxu et.al. [24], proposed a strategy to perform estimation examination on substance whose logical data is not known ahead of time. In this strategy other related substance were utilized to remove the required relevant data and afterward utilized the data for deciding the introduction of the conclusion.

Ting-Chun et.al. [25], proposed an unsupervised learning calculation in view of grammatical form (pos) design. They utilized the conclusion expression as a question for an internet searcher and suppositions were anticipated in view of the list items.

Posse et.al. [26], utilized TF-IDF (term recurrence – reverse report recurrence) weighing for slant investigation. They utilized K-implies grouping on crude information, and afterward a voting component to additionally settle the bunching. Various executions of the procedure was connected to order the records into positive and negative gatherings.

Prabhu et.al. [27], utilized a basic vocabulary construct strategy in light of twitter information by distinguishing and separating opinions from hashtags and emoticons.

Fang et al. [28] received completely unique approach. They considered both universally useful vocabulary and space particular dictionary for deciding extremity introduction of conclusion words and encourage these dictionaries into regulated learning calculation, SVM. They found that broadly useful vocabulary performed exceptionally poor while space particular dictionary performed extremely well. The

framework characterized the conclusion in two stages: First the classifier is prepared to anticipate the perspectives and In Next the classifier is prepared to foresee the opinions identified with the angles gathered in step1.Their framework output around 66.8% precision.

Mudinas et al. [29] consolidated lexicon based and learning based ways to deal with build up an idea level feeling examination framework, pSenti. It used points of interest of both the methodologies and accomplished solidness and clarity from semantic dictionary and high exactness from an intensely regulated learning calculation. They removed conclusion words and considered it as elements in machine learning calculation. This cross breed approach pSenti accomplished an exactness of 82.30%.

TABLE 1. Various approaches for Sentiment Analysis

S. No	Ref. No	Technique Used / Mode of Operation	Data source	Approach to Analyze Twitter Data-set	Strength	Limitations
1	[5]	Hadoop/ MapReduce	Twitter dataset	Naive Bayes Algorithm	Performance of Naive Bayes algorithm increases by converting the emoticons by assigning its equivalent word.	Not consider multilingual language
2	[6]	Hadoop/ MapReduce, Mahout	Twitter dataset	Collaborative Filtering	Filter noise from data effectively	Long and complex process
3	[7]	Hadoop/ MapReduce, Hbase	Twitter dataset	Uses Traditional MapReduce Algorithm	execution time is less for processing twitter dataset	Need to use multilingual language
4	[8]	Hadoop/ MapReduce	Twitter dataset	Uses IBM BigInsights tool and MapReduce Algorithm for Streaming of data and visualization through BigSheets tool	Latency for processing is low	Need to use multilingual language
5	[9]	Hadoop/ MapReduce, Apache Pig, Hive, Apache Oozie	Twitter dataset	Mathematical Model and Aggregation Queries	Reduce the time delay and the cost	Inefficient for Real Time Analytics.
6	[10]	Standalone System	Twitter dataset	Hash Algorithm	Emoticons are considered and assigned polarity to each of them	Because of Standalone System, processing of data takes lot of time.
7	[11]	Hadoop/ MapReduce	Twitter dataset	Hybrid filtering	Takes less CPU execution time	Not efficient while working with small data set.
8	[12]	linguistic and semantic	1000 Facebook	Study takes also in account the data	automatically a sentiment polarity	Long and complex process

		approaches	posts	provided by Auditel regarding newscast audience, correlating the analysis of Social Media, of Facebook in particular, with measurable data, available to public domain.		
9	[13]	Hadoop/ MapReduce	exabytes of data	analyzes the performance of HDFS and uncovers several performance issues	evaluate tradeoffs between portability and performance in the Hadoop distributed filesystem.	Need to improve the performance of HDFS.
10	[14]	naive Bayes model	UCI datasets	Propose a novel naive Bayes classification algorithm for uncertain data with a pdf	Time complexity analysis and performance analysis based on experiments show that the formula-based approach has great advantages over the sample-based approach.	Need to improve the performance with respect to time complexity analysis
11	[15]	Hadoop/ MapReduce	Google's production web search service	Implementation of MapReduce runs on a large cluster of commodity machines and is highly scalable: a typical MapReduce computation processes many terabytes of data on thousands of machines.	The implementation makes efficient use of these machine resources and therefore is suitable for use on many of the large computational problems encountered at Google	redundant execution can be used to reduce the impact of slow machines
12	[16]	lexicon-based method	Twitter	The effectiveness of the correlation between different views is also studied using the widely used fusion strategies and an advanced multi-view feature extraction method.	Author showed that there are many inconsistent labels between the text view and image view in the collected tweets.	Not consider emoticons
13	[17]	SVM	Amazon customer reviews on product	Author proposed a novel graphical model to extract and visualize comparative relations between products from customer reviews, with the interdependencies among relations taken into consideration, to help enterprises discover potential risks and further design new products and marketing	Very high accuracy	Incapable of multiclass classification

				strategies.		
14	[18]	SVM	movie-review	A wide range of comparative experiments are conducted on five widely-used datasets in sentiment classification.	Robust to noise	Computationally expensive Slow
15	[19]	naive Bayes and SVM	written restaurant reviews	Standard machine learning techniques naive Bayes and SVM are incorporated into the domain of online Cantonese-written restaurant reviews to automatically classify user reviews as positive or negative.	Robust to noise	Incapable of multiclass classification

III. CONCLUSION

From all above review papers we have seen the influence of smaller scale blogging webpage twitter on the current trends and issues. The mining of twitter information helps any organization to examine the conduct of individuals on the premise of their opinion, reviews, audits, rating and so on. This sort analysis will definitely help any association to enhance their business productivity. The analysis of twitter information are done on different point of view like Positive, Negative and Neutral sentiment on tweets. Tweets can likewise be helpful in prediction of item product sales, quality of services offered by organization, feedback of users and so forth. Hence the tweets in the twitter assumes an essential part to analyze opinion of individuals.

In this paper, we did an investigation of different ways for multilingual sentiment analysis. We surveyed the machine learning and lexicon based approaches for multilingual sentiment analysis and joining a suite of existing methods, translation software can help analyze multilingual sentiment in a better way.

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] H. Parveen and S. Pandey, "Sentiment analysis on Twitter Data-set using Naive Bayes algorithm," 2016 2nd International Conference on Applied and Theoretical

Computing and Communication Technology (iCATccT), Bangalore, India, 2016, pp. 416-419.

[6] Subramaniyaswamy , Vijayakumar , Logesh, Indragandhi, ”Unstructured Data Analysis on Big Data using Map Reduce”, 2nd International Symposium on Big Data and Cloud Computing (ISBCC15).

[7] Can Uzunkayaa, Tolga Ensaria, Yusuf Kavurucub, ”Hadoop Ecosystem and Its Analysis on Tweets”, World Conference on Technology, Innovation and Entrepreneurship.

[8] Manoj Kumar Danthala, Dr. Siddhartha Ghosh, ”Streaming Twitter Data with Apache Hadoop and Visualizing using BigInsights”, International Journal of Engineering Research Technology (IJERT).

[9] Gaurav D Rajurkar, Rajeshwari M Goudar, ”A speedy data uploading approach for Twitter Trend And Sentiment Analysis using HADOOP”, HADOOP, 2015 International Conference on Computing Communication Control and Automation.

[10] Efthymios Kouloumpis, Theresa Wilson, Johanna Moore, ”Twitter Sentiment Analysis: The Good the Bad and the OMG!”, Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.

[11] Jai Prakash Verma, Bankim Patel, Ph D, Atul Patel, Ph D, ”Big Data Analysis: Recommendation System with Hadoop Framework”, 2015 IEEE International Conference on Computational Intelligence Communication Technology.

[12] Federico Neri, Carlo Aliprandi, Federico Capeci, Montserrat Cuadros, Tomas, "Sentiment Analysis on Social Media", 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

[13] J.Shafer, S.Rixner, A.L.Cox, “The Hadoop distributed filesystem: Balancing portability and performance", IEEE International Symposium on Performance Analysis of Systems & Software (ISPASS), 2010, pp.122-133.

[14] J.Ren, S.D.Lee, X.Chen, B.Kao, R.Cheng, D.Cheung, “Naive Bayes Classification of Uncertain Data", Ninth IEEE International Conference on Data Mining, 2009. ICDM ’09, pp. 944 – 949.

[15] J. Dean, S. Ghemawat, “MapReduce: simplified data processing on large clusters", The 6th Symposium on Operating Systems Design and Implementation, OSDI-04, USENIX Association, 2004, pp. 107 – 113

[16] Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik, “Sentiment Analysis on Multi-View Social Data”, Springer International Publishing Switzerland 2016.

[17] KaiquanXu , Stephen Shaoyi Liao , Jiexun Li, Yuxia Song, “Mining comparative

opinions from customer reviews for Competitive Intelligence", Decision Support Systems 50 (2011) 743–754.

[18] Rui Xia, ChengqingZong, Shoushan Li, "Ensemble of feature sets and classification algorithms for sentiment classification", Information Sciences 181 (2011) 1138–1152.

[19] Ziqiong Zhang, Qiang Ye, Zili Zhang, Yijun Li, "Sentiment classification of Internet restaurant reviews written in Cantonese", Expert Systems with Applications (2011).

[20] Rudy Prabowo, Mike Thelwall, "Sentiment analysis: A combined approach." Journal of Informetrics 3 (2009) 143–157.

[21] ZHU Jian , XU Chen, WANG Han-shi, "" Sentiment classification using the theory of ANNs", The Journal of China Universities of Posts and Telecommunications, July 2010, 17(Suppl.): 58–62

[22] Long-Sheng Chen, Cheng-Hsiang Liu, Hui-Ju Chiu, "A neural network based approach for sentiment classification in the blogosphere", Journal of Informetrics 5 (2011) 313–322.

[23] Andrea Esuli and FabrizioSebastiani, "Determining the semantic orientation of terms through gloss classification", Proceedings of 14th ACM International Conference on Information and Knowledge Management,pp. 617-624, Bremen, Germany, 2005.

[24] Chunxu Wu, LingfengShen, "A New Method of Using Contextual Information to Infer the Semantic Orientations of Context Dependent Opinions", 2009 International Conference on Artificial Intelligence and Computational Intelligence.

[25] Ting-Chun Peng and Chia-Chun Shih , "An Unsupervised Snippet-based Sentiment Classification Method for Chinese Unknown Phrases without using Reference Word Pairs", 2010 IEEE/WIC/ACM International Conference on Web Intelligence and intelligent Agent Technology JOURNAL OF COMPUTING, VOLUME 2, ISSUE 8, AUGUST 2010, ISSN 2151-9617 .

[26] Gang Li, Fei Liu, "A Clustering-based Approach on Sentiment Analysis", 2010, 978-1-4244-6793-8/10 ©2010 IEEE.

[27] PrabuPalanisamy, VineetYadav, HarshaElchuri, "Serendio: Simple and Practical lexicon based approach to Sentiment Analysis", Serendio Software Pvt Ltd, 2013.

[28] Ji Fang and Bi Chen, "Incorporating Lexicon Knowledge into SVM Learning to Improve Sentiment Classication", In Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), pages 94–100, 2011.

[29] A. Mudinas, D. Zhang, M. Levene, "Combining lexicon and learning based approaches for conceptlevel sentiment analysis", Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, ACM, New York,NY, USA, Article 5, pp. 1-8, 2012.