www.ijsrnsc.org

Available online at www.ijsrnsc.org

IJSRNSC

Volume-5, Issue-3, June 2017
Research Paper

Int. J. Sc. Res. in Network Security and Communication

ISSN: 2321-3256

Sentiment Analysis of Twitter Streaming Data for Recommendation using Apache Spark

Amit Palve[1*], Rohini D.Sonawane [2], Amol D. Potgantwar[3]

[1*]Department of Computer Engineering Sandip Institute of Technology and Research Centre, Nashik
[2]Department of Computer Engineering Sandip Institute of Technology and Research Centre, Nashik
[3]Department of Computer Engineering Sandip Institute of Technology and Research Centre, Nashik

**Corresponding Author: amit.palve@sitrc.org,*

Received 02nd May 2017, Revised 17th May 2017, Accepted 11th Jun 2017, Online 30th Jun 2017

Abstract—Twitter is free social networking micro blogging service. In that micro-blogging allows to registered members to broadcasting the short posts also called tweets. It can broadcast the tweets by using multiple platforms and devices. Twitter member replies to tweets. Existing system focuses on document level sentiment analysis that means they used Hadoop Framework for concerning moving or product reviews. In that system web pages or blocks on which posts are published therefore in that system complexity of document level opinion mining many efforts have been made towards the sentence level sentiment analysis. The existing systems classify the accuracy only one word. This process is time consuming due to documentation. In the system which we are devolve in that we used spark framework instead of Hadoop framework. Due to the use of Spark Framework, garbage or unclean data are removing. So that user gets better efficiency and less time required for processing, than earlier system.

Index Terms- Big data, classification, Map-Reduce, Spark, sentiment analysis, text mining, Twitter

I. INTRODUCTION

A large scale solution is presented in the build a sentiment lexicon and classifies tweets using a Map-Reduce algorithm and a distributed database model. [1].Here, the Map Reduce model we are brief describe the data processing in Map-Reduce is based on input data partitioning, the partitioned data is executed by a number of tasks in many distributed nodes and give the output of key value. That are exist two major task is called Map and Reduce respectively. Due to verify the complexity of document level opinion mining, many applications have been made towards the sentence level sentiment analysis. [2]This phrases and assigns to each one of them a sentiment polarity task of sentiment analysis for batch processing the existing system and real time sentiment analysis for computation. The tweets are important for analysis because data arrive at a high frequency and algorithms that process them have to do so under very strict constraint of storage and time. It will be given away how to automatically collect a quantity for sentiment analysis and opinion mining purposes and then perform quantity. Every one public tweets posted on twitter are freely available through a set of streaming APIs provided by Twitter. The sentiment classifier is constructed that is able to decide positive, negative and neutral sentiments. The hierarchical classification is already implemented on Hadoop and we are implemented on Spark.

II. REVIEW OF LITERATURE

Skuza et. al. discusses a possibility of making prediction of stock market basing on classification of data coming from Twitter micro blogging platform. Twitter messages are retrieved in real time using Twitter Streaming API. Tweets were collected over 3months period from 2nd January 2013 to 31st March 2013. It was specified in the query that tweets have to contain name of the company or hashtag of that name. Predictions were made for Apple Inc. in order to ensure that sufficiently large datasets would be retrieved only tweets in English are used in this research work. Reposted messages are redundant for classification and were deleted. Agarwal et. al. [1] also explored the POS features, the lexicon features and the microblogging features. Apart from simply combining various features, they also designed a tree representation of tweets to combine many categories of features in one succinct representation. A partial tree kernel [2-8] was used to calculate the similarity between two trees.

They found that the most important features are those that combine prior polarity of words with their POS tags. All other features only play a marginal role Barbosa and Feng [2] argued that using n-grams on tweet data may hinder the classification performance because of the large number of infrequent words in Twitter. Instead, they proposed using micro blogging features such as re-tweets, hashtags, replies, punctuations, and emoticons. They found that using these features to train the SVMs enhances the sentiment classification accuracy by 2.2% compared to SVMs trained from unigrams only. Speriosu et al. [4] constructed a graph that has some of the microblogging features such as hashtags and emoticons together with users, tweets, word unigrams and bigrams as its nodes which are connected based on the link existence among them. They then applied a label propagation method where sentiment labels were propagated from a small set of nodes seeded with some initial label information throughout the graph. Bian, Jiang, Umit Topaloglu, and Fan Yu describe an approach to find drug users and potential adverse events by analyzing the content of twitter messages utilizing Natural Language Processing (NLP) and to build Support Vector Machine (SVM) classifiers. Due to the size nature of the dataset (i.e., 2 billion Tweets), the experiments were conducted on a High Performance Computing (HPC) platform using Map Reduce, which exhibits the trend of big data analytics. David F. et. al. propose an open framework to automatically collect and analyze data from Twitters public streams. This is a customizable and extensible framework, so researchers can use it to test new techniques. The framework is complemented with a language-agnostic sentiment analysis module, which provides a set of tools to perform sentiment analysis of the collected tweets. Lin, Jimmy, and Alek Kolcz presents a case study of Twitters integration of machine learning tools into its existing Hadoop-based, Pig-centric analytics platform. The core of this work lies in recent Pig extensions to provide predictive analytics capabilities that incorporate machine learning, focused specifically on supervised classification. In particular, the authors have identified stochastic gradient descent techniques for online learning and ensemble methods as being highly amenable to scaling out to large amounts of data. In contrast to other linguistic approaches the authors adopt a knowledge-poor, data-driven approach. It provides a base-line for classification accuracy from content, given only large amounts of data. Tare and Mohit proposed strategy that uses Apache Hadoop framework, an open source java framework, which relies on Map Reduce paradigm and a Hadoop Distributed File System (HDFS) to process data. The proposed Map Reduce strategy for classification of tweets using Nave Bayes classifier relies on two Map-Reduce passes. They have used the Twitter4j library to gather tweets which internally uses twitter REST API [9,10,11,12,13,14].

III. SOFTWARE REQUIREMENT SPECIFICATION

A software requirements specification (SRS) is a comprehensive description of the intended purpose and environment for the proposed work under development. The SRS fully describes what the proposed work will do and how it will be expected to perform. A software requirements specification (SRS) is a illustration of the intentional motivation and environment for software under development. How software will work and what software is going to perform is described by SRS. The intention of SRS is to minimize the time required to achieve goals by the developers and also it reduces the cost of development. The SRS which is more efficient will show or define that how a particular application will be interacting with the hardware of the system, human users and other programs. Through SRS operating speed, availability, portability, maintainability, and speed of recovery are evaluated. IEEE (Institute of Electrical and Electronics Engineers) specification 830-1998 defines methods for SRS.

IV. SYSTEM OVERVIEW

In our proposed system for analyzing real time as well as on line data for real-time applications using sentiment analysis we have divided real time architecture into four parts, i.e., 1) Data streaming 2)Data Storage 3) Data cleaning 4)Data Tokenization 5)classification and 6) Recommendation. In these six units various algorithms or techniques will be implied on data for its analysis. In these four units various algorithms or techniques will be implied on data for its analysis. The functionalities and working of four units is as explained and shown in diagram below:

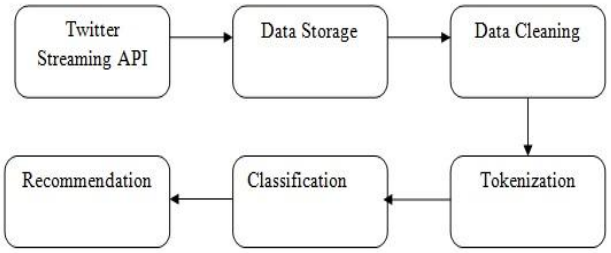

Fig 1. Proposed System

The real time extracting tweets using Twitter Streaming API we are need Twitter data For classification. Data streaming is purpose of we make use of API's twitter provides. Twitter provides the two API's; Sample Stream API1 and Filter

Stream API2. The difference between Sample Streaming API and Filter Streaming APIs are: Streaming API supports small and random sample connection and provides data in real-time. The Filter Stream APIs support keywords, user ID and location . In this phase, we can gather data for the online. In this phase,we can used a set of rule to remove the short tweets, non-English, same tweet sand garbage data. Remove Short tweets, we can send the text message used the short word that means d instead of them are remove. Remove non-English tweets, words in tweets are compared with common English words than other thresh- old are remove Remove similar tweets, tweet are compare the every other tweet most of tweets are similar than tweet are remove. In this phase, the process of breaking a string of text up into words, symbols and other meaningful elements. In this phase, we can classify the which tweets are positive or neg- ative. We can calculate the probability. The sentiment analyzing the tweets on batch processing and recently analyzing the tweets on real time computation. The hierar- chical classification is already implemented on Hadoop and we are implemented on Spark.

IV. SYSTEM ANALYSIS

Mathematical modeling is indispensable in many applications is successful in many further applications gives precision and direction for problem solution enables a thorough understanding of the system modeled prepares the method for better design of a system allows the efficient use of modern computing capabilities A mathematical model is a explanation of a system using mathematical concepts and words. The procedure of developing a mathematical model is termed mathematical modeling. Mathematical modeling is the art of translating problems from an application area into tractable mathematical formulations whose theoretical and numerical study provides insight, answers, and direction useful for the originating application. Following figure shows the mathematical model of system:

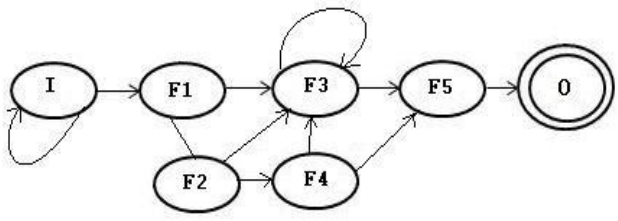

fig.2.State diagram

The Mathematical Model of system shown I is client input of API then proceed for further process. Data Storage process can be done i.e. Data gathering operation which filter tweets by tags and keywords. After that data cleaning process by removing the Unicode etc. Classification tweets into positive and negative neutral classes. The parameter explained in following section. Input (I) Parameter: I= [Dn] Where, I is a set of Input. Dn=Gather Data.Function(F) Parameter F=[F1,F2,F3,F4,F5] Where, F is a function for processing. F1= Filter Tweets. F2= Remove unicode(). F3= Classification neutral classes().F4= Calculations().F5= Filter class().Output (O) Parameter O=[ol] Where, O is the Output ol=Recommend Tweet. Process of the model is: 1.Twitter Sentiment Analysis which segment prediction can be done.2. Data Gathering Stage. 3.Preprpcessing perform on gathered data it remove the Unicode. 4.Tokenization. 5.Classification of positive and negative neutral classes. 6. Verify client and Retrieve encrypted file. 7. Recommend tweets to appropriate user.

VI . RESULT ANALYSIS AND COMPARISON

As per the experimental result analysis used the tweet in the sustem it gives the excellent result. Accuracy graph distinguishes between the existing system and implemented system by the percentage. Some tweet are used accuracyof result is improved compared to existing system.

	Precision	Recall	F1-Score	Support
Positive	0.88	0.90	0.89	7510
Negative	0.90	0.88	0.89	7490
Avg/total	0.89	0.89	0.89	15000

Table 1: Accuracy Analysis

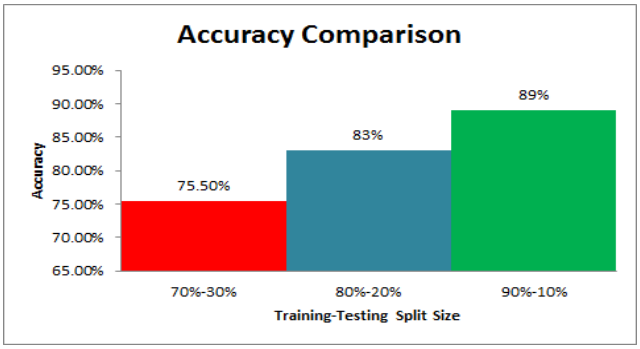

Fig.3.Accuracy Graph

		Positive	Negative
Actual	Positive	6772	738
	Negative	912	6578

Table No.2: Time Analysis

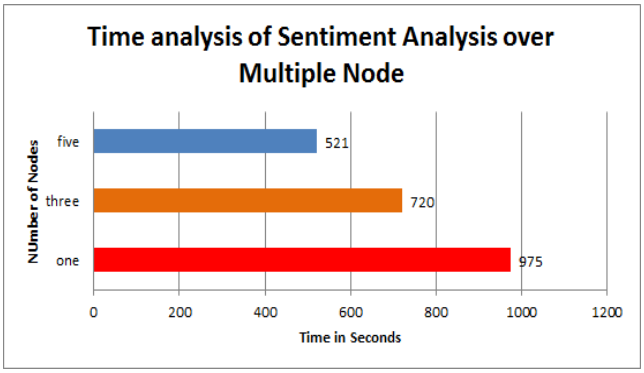

Fig.4.Time Graph

VII. CONCLUSION

The use of semantic features in Twitter sentiment classification and explored three different approaches for incorporating them into the analysis; with replacement, augmentation, and interpolation. It is proposed to stream real time live tweets from twitter using Twitter API, and the large volume of data makes the application suitable for Big Data Analytics. A method to predict or deduct the location of a tweet based on the tweet's information and the user's information should be found in the future.

ACKNOWLEDGMENT

This work is supported by SITRC(Sandip Institute of Technology and Research center), Nasik, Maharashtra, under the guidance of respected Prof Amit Palve and Dr. Amol D. Potgantwar, Head of Department Computer Engineering. This work would not have been possible without the enthusiastic response, insight, and new ideas from them.

REFERENCES

[1]. A. I. Baqapuri, S. Saleh, M. U. Ilyas, M. M. Khan, A. M. Qamar, "Sentiment classification of tweets using hierarchical classification", *2016 IEEE International Conference on Communications (ICC)*, Malaysia, pp.1-7, 2016.

[2]. Tumasjan A, Sprenger TO, Sandner PG, Welpe IM, "*Predicting elections with twitter: What 140 characters reveal about political sentiment"*, Icwsm Vol.10, Issue.1, pp.178-85, 2010.

[3]. R.V. Patil, S.S. Sannakki, V.S. Rajpurohit, "*A Survey on Classification of Liver Diseases using Image Processing and Data Mining Techniques*", International Journal of Computer Sciences and Engineering, Vol.5, Issue.3, pp.29-34, 2017.

[4]. Madnani N, "*Getting started on natural language processing with Python*", Crossroads, Vol.13, Issue.4, pp.5-9, 2007.

[5]. B. Pang, L. Lee, S. Vanity Nathan, "*Sentiment classification using machine learning techniques*", In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Vol.5, Issue.4, pp.79-86, 2002.

[6]. Y. Yamamoto, T. Kumamoto, A. Nadamoto, "*Role of emotions for multidimensional sentiment analysis of twitter*", In Proceedings of the 16th International Conference on Information Integration and Web-based Applications, USA,pp.107-115, 2014.

[7]. V. N. Khuc, C. Shivved, R. Namath, and J. Ramayana., "Towards building large-scale distributed systems for twitter sentiment analysis", .In Pro-ceedings of the 27th Annual ACM Symposium on Applied Computing, pages 459-464, 2012.

[8]. Dean J, Ghemawat S. "*MapReduce: simplified data processing on large clusters*", Communications of the ACM, Vol.51, Issue.1, pp.107-113, 2008.

[9]. L. Zhuang, F. Jing, X.-Y. Zhu, "*Movie review mining and summa-rization*", In Proceedings of the 15th ACM International Conference on Information and Knowledge Management, USA, pp.43-50, 2006.

[10]. W. Zhang, C. Yu, W. Men, "*Opinion retrieval from blogs*", In Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, USA, pp.831-840, 2007.

[11]. T. Wilson, J. Wienie, P. Ho Mann, "*Recognizing contextual polarity in phraselevel sentiment analysis*", In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, USA, pp.347-354, 2005.

[12]. T. Wilson, J. Wienie, P. Ho Mann, "*Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis*", Com put. Linguist, USA, pp.399-433, 2009.

[13]. Neetu Anand, Tapas Kumar , "*Text and Emotion Analysis of Twitter Data*", International Journal of Computer Sciences and Engineering, Vol.5, Issue.6, pp.181-285, 2017.

[14]. H. Yuh, V. Huitzilopitchli, "*Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences*", In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, USA, pp.129-136, 2003.

Authors Profile

Ms Rohini D. Sonawane has pursued Bachelor of Engineering in Computer from University of Pune in 2014. She is now pursuing Master of Computer Engineering from Savitribai Phule Pune University.

Prof.Amit H.Plave has teaching experience of 8 year with the expert area of work in IOT (Internet of Thing), data analytics and image processing & wireless sensor network.

Dr. Amol D. Potgantwar is a Associate. Professor of the Department of Computer Engineering, Sandip Foundation's, Sandip Institute of Technology and Research Centre, Nashik, Maharastra, India. The focus of his research in the last decade has been to explore problems at Near Field Communication and it's various application In particular, he is interested in applications of Mobile computing, wireless technology, near field communication, Image Processing and Parallel Computing. He has register patents like Indoor Localization System for Mobile Device

Using RFID & Wireless Technology , RFID Based Vehicle Identification System And Access Control Into Parking, A Standalone RFID And NFC Based Healthcare System. He has recently completed a book entitled Artificial Intelligence, Operating System, Intelligent System. He has been an active scientific collaborator with ESDS, Carrot Technology, Techno vision and Research Lab including NVIDIA CUDA, USA. He is a member of CSI, ISTE, IACSIT.

