REVIEW OF SENTIMENT ANALYSIS USING TWITTER AND BIG DATA

[1]Ms PAYAL S GADHAVE, [2]Prof M.D.NIRMAL,
[1]PG Student, Computer Department,PREC,Maharashtra,India.
[2]Asso. Prof. at Computer Department,PREC,Maharashtra,India.

ABSTRACT: In this paper we propse a joint classification for tweets using big data and and social networking site twitter. Today classification using tweets is generally done by splitting a tweet in to words and not the whole tweet is taken in to consideration so we thought of introducing a novel approach where tweets will be classified as whole and not in words. To enhance the concept we thought of using big data technology such as APACHE SPARK to help in classification as the tweets that are retrieved are in large numbers and not easy for a single machine to handle them. The analyzed information results that are returned will be assembled together on a single machine and the prediction returned are in the form of sentiment analysis as each tweet as a whole will have score and various sentiments will be attached with them.

Keywords: - Tweet, Segmentation, machine learning, Sentiment analysis, Sentiment classification, Big data.

INTRODUCTION

Today Social Networking Sites (SOCIAL NETWORKING SITES) have become an important part of our day to day life. We share a lot of personal data on these sites. They help us to make the world smaller and integrate like a small village with each other. There are many SOCIAL NETWORKING SITES available today and many more are increasing each day.

Thus a user uses many Social Networking Sites each day and communicate and share data with friends and family. This communication medium gave rise to complex structure whether a user really like the Social Networking Sites which he uses more or he needs another Social Networking Sites other than he uses more.

Thus one of the most famous SOCIAL NETWORKING SITES is TWITTER which is used to share data and post our thoughts and latest buzz upon the internet. The users using TWITTER have increased constantly in the recent years. So the analysis of this SOCIAL NETWORKING SITES may help in answering and predicting many answers.

This online social network is used by billions of people around the world to remainsocially connected to their friends, family members, andcoworkers through their computers and mobile phones .Twitter asks one question, "What's going on?" Answersmust be fewer than 140 characters. A status updatemessage, called a tweet, is often used as a message tofriends, family and colleagues. A user can follow other users; thatuser's followers can read her tweets on a daily basis. Auser who is being followed by another user need notreciprocate by following them back, which leaves the links of the network as directed. Since itslaunch on July 2006, Twitter users have increased dramatically.

Thus this kind of SOCIAL NETWORKING SITES can be used to predict and analyze the large amount of tweets generated and understand the sentiments behind each tweet whether it is positive, negative or neutral. So I took help of an IEEE paper where they have presented a system which helps in analyzing and helping in developing an application for the purpose of sentiment analysis.

LITERATURE REVIEW:

This chapter describes the fundamentals of tweet analysis. It helps in understanding various ideas put forward by various technical papers published by various pulishers.

BO PANG and LILLIAN LEE 2008, An important part of our information-gathering behavior has always been to find out what other people think. With the growing availability and popularity of an opinion rich resources such as review sites and blogs, new opportunities and challenges arise as people now can, and do, actively use information technologies to seek out and understand the opinions of others. The sudden breakout of activity in the area of opinion mining and sentiment analysis, which deals with the computational remedy of opinion, sentiment, and subjectivity in text, has thus occurred at least in part as a direct response to the growth of interest in new systems that deal directly with opinions as a first-class object. This survey covers techniques and approaches that promise to directly enable opinion-oriented information pursuing systems. Our focus is on methods that pursue to address the new challenges raised by sentiment aware applications, as compared to those that are already present in traditional fact-based analysis. We include material on summarization of evaluative text and on broader issues regarding privacy, manipulation, and economic impact that the development of opinion-oriented information-seeking services gives rise to opinion mining and sentiment analysis.

Bing liu 2012, Opinions are central to almost all human activities and are key influencers of our behaviors. Our faith and perceptions of reality, and the choices we make are, to a considerable degree, depending upon how others see and evaluate the world. For this reason, when we need to make a decision we often peruse out the opinions of others. This is not only true for individuals but also true for institutions. Opinions and its related concepts such as sentiments, assessments, attitudes, and emotions are the subjects of study of sentiment analysis and opinion mining. The inception and fast growth of the field coincide with those of the social media on the Web, e.g., reviews, forum discussions, blogs, Twitter, and social networks, because for the first time in history, we have a huge volume of opinionated data recorded in digital forms. Since early 2000, sentiment analysis has grown to be one of the most active research areas in NLP. It is also hugely studied in data mining, Web mining, and text mining. In fact, it has increased from computer science to management sciences and social sciences due to its importance to business, organizations and society as a whole. In recent years, industrial activities surrounding sentiment analysis have also increased. Many

startups have emerged. Many large corporations have built their own in house capabilities. Sentiment analysis systems have found their functions in almost every business, organizations and social domain. The goal of this book is to give an in-depth introduction to this appealing problem and to present a comprehensive survey of all important research topics and the latest developments in the field. Although the field deals with the natural language text, which is often considered the unstructured data, this book takes a structured way in introducing the problem with the idea of branching the unstructured and structured worlds and facilitating qualitative and quantitative analysis of opinions. This is important for practical applications. In this book, I first define the problem in order to provide an abstraction or solution to the problem. From the abstraction, we will naturally see its key problems and sub-problems

Jichang Zhao, Li Dong, Junjie Wu, and Ke Xu 2012, Recent years have witnessed the fiery development of online social media. Weibo, a Twitter-like online social network in China, has attracted more than 250 million users in less than three years, with more than 1000 tweets sent in every second. These tweets not only transmit the factual information, but also show the emotional states of the authors, which are very crucial for understanding user behaviors. However, a tweet in Weibo is awfully short and the words it contains evolve extraordinarily fast. Moreover, the Chinese corpus of sentiments is still very small, which prohibits the conventional keyword-based methods from being used. In wake of this, we developed a system called MoodLens, which to our best ability is the first system for sentiment analysis of Chinese tweets in Weibo. In MoodLens, 95 emoticons are graphed into four categories of sentiments, i.e. angry, disgusting, joyful, and sad, which deliver as the class labels of tweets. We then collect over 3.5 million labeled tweets as the corpus and train a fast Naive Bayes classifier, with an experimental precision of 64.3%. MoodLens also implements an incremental learning method to face the problem of the sentiment shift and the generation of new words. Using MoodLens for real-time tweets obtained from Weibo, several interesting temporal and spatial patterns are seen. Also, sentiment variations are well-captured by MoodLens to adequatly detect abnormal events in China. Finally, by using the highly efficient Naive Bayes classifier, MoodLens is able of online real-time sentiment monitoring.

Georgios Paltoglou and Mike Thelwall 2010, Most sentiment analysis approaches basically uses a support vector machines (SVM) classifier with binary unigram weights. In this paper, we find whether more sophisticated feature weighting schemes from Information fetching can enhance classification accuracy. We show that types of the classic tf.idf scheme adapted to sentiment analysis provide significant rise in accuracy, especially when using a sublinear function for concept frequency weights and document frequency smoothing. The techniques are tested on a huge collection of data sets and produce the best accuracy to our knowledge.

Georgios Paltoglou and Mike Thelwall 2010, Most of the sentiment analysis approaches use as baseline a support vector machines (SVM) classifier with binary unigram weights. In this paper, we explain whether more sophisticated feature weighting schemes from Information Retrieval can raise classification accuracy. We show that variations of the classic tf.idf scheme adapted to sentiment analysis add significant increases in accuracy, especially when using a sublinear function for term frequency weights and document frequency smoothing. The techniques are tested on a huge selection of data sets and produce the best accuracy to our knowledge.

PROBLEM STATEMENT:

Tweet analysis can be useful in understanding the sentiments behind a tweet which are in large numbers. The tweets generated can be useful in many walks of life from share trading to public health care. In real world many tweets may be fake and does have any meaning and sentiments behind it. With the use of Apache Spark we can speed up the process in analyzing the tweets.

The traditional system cannot predict and analyze the TWIITER using v and had some drawbacks such as : The system does not make good use of the TWIITER REST API. The system does not use google map with the TWITTER. The system is user dependanti.e the user has to analyze twitter by himself which is not possible. It cannot use machine learning algorithms. It cannot use JSON results. It cannot show the results properly. It cannot help in sentiment analysis of tweets. It cannot make efficient use of big data.

PROPOSED SYSTEM:

It proposes an effective use of APACHE SPARK to mine the data that is unstructured in the form of tweets. It proposes a machine learning algorithm to distribute data among the data nodes and again assemble the results to get a successful implementation.

It also proposes a natural language processing technique to handle the unstructured tweets and bring them in a structured format.it proposes the use of Twitter API to fetch the tweets according to keywords.

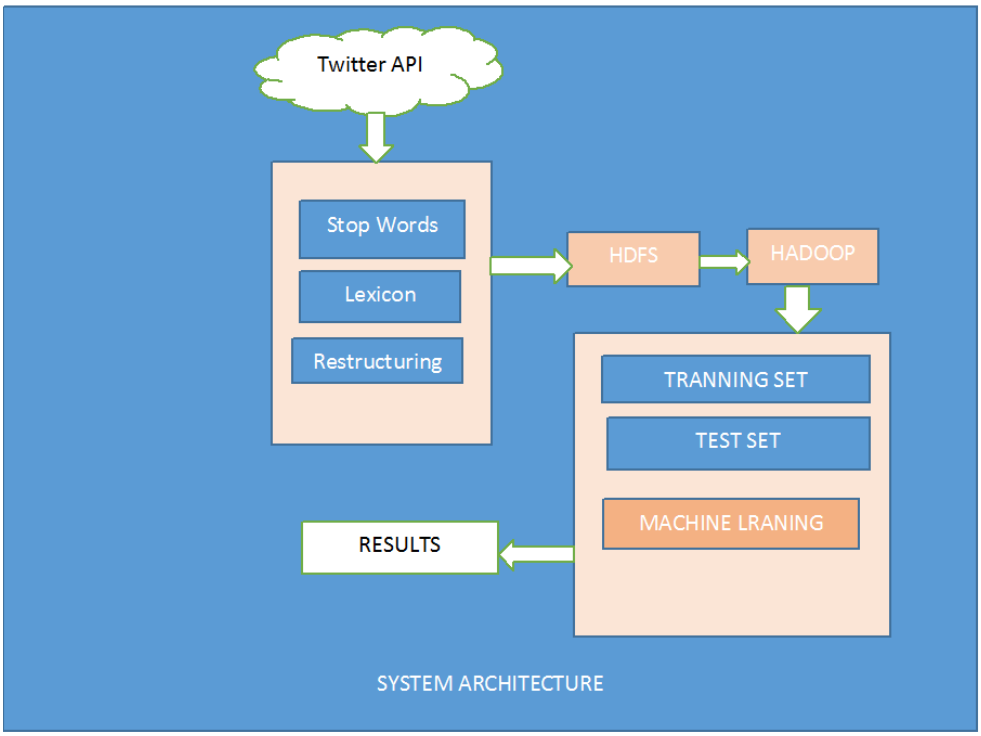
Advantages of proposed system

- It uses the Twitter Api to access the tweets according to keywords.
- It uses bag of words approach to remove the stop words that are not good for text mining.
- It makes effective use of Lexicon structure to refine the tweets
- It generates a test dataset automatically
- It makes effective use of hdfs
- It makes effective use to access and distribute the dataset
- It makes effective use of machine learning.

CONTRIBUTION IN THIS WORK:-

In the base paper the proposed system makes use of only a single machine where in our system we are proposing a distributed computing approach for sentiment analysis. In our system the dataset is split and distributed among data nodes to increase the speed of sentiment analysis using APACHE SPARK .

SYSTEM ARCHITECTURE/SYSTEM OVERVIEW



The following diagram show that the system architecture.

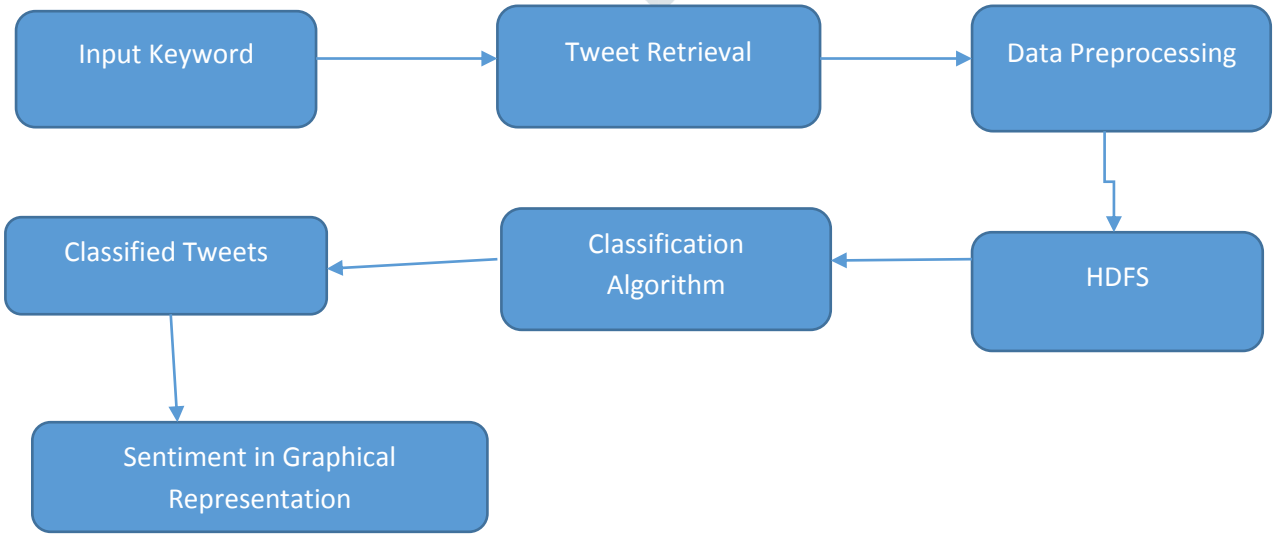
Twitter API: - Data can be collected from various sources like surveys, social networking sites, feedbacks, blogs, review etc. based on some keywords. Proper authorization is required if the data is to be fetched from social networking sites such as Facebook, Twitter etc.Different API are available to collect data from these sites. So to access twitter an application has to be created using twitter credentials which will generate a unique consumer key, consumer secret, access token and access token secret which will help us to access the tweets using Twitter API

Tweet Preprocessing: - In this phase the tweets that are fetched has to be preprocessed as many things in the tweets are not needed for text mining. So first the stop words are removed from the tweets. Then a lexicon grammar is used to shorten the tweet that will make more sense for sentiment analysis.

HDFS: - Then the tweets that are restructured are entered in to HDFS (HADOOP DISTRIBUTED FILE SYSTEM) . Then the test dataset which is generated is also put into HDFS.

ANALYSIS: - This phase if the main phase. Here the tweets and dataset that is being tracked by Then machine learning is applied on the dataset and results are returned as positive or negative whether the tweet has a positive sentiment or negative sentiment.The results are again assembled and shown as a whole.

SYSTEM ANALYSIS:



CONCLUSION:-

In this paper, we are developing novel sentiment analysis approach using TWEETER and APACHE SPARK together. The basic idea of the project is to use distributed computing in training and testing the machine learning classification using named nodes and data nodes together. We are going to assemble various predictions by machine learning algorithms together and view the results in three classes such as positive, negative and neutral according to the predictions returned by the system.

REFERENCES:-

[1] B. Pang and L. Lee, Opinion mining and sentiment analysis, Foundations Trends Inf. Retriev., vol. 2, no. 12, pp. 1135, 2008.

[2] B. Liu, Sentiment analysis and opinion mining, Synth. Lectures Human Lang. Technol., vol. 5, no. 1, pp. 1167, 2012.

[3] C. Havasi, E. Cambria, B. Schuller, B. Liu, and H. Wang, Knowledgebased approaches to concept-level sentiment analysis, IEEE Intell. Syst.,vol. 28, no. 2, pp. 001214, Mar.-Apr. 2013.

[4] C. D. Manning and H. Schtze, Foundations of Statistical Natural Language Processing. Cambridge, MA, USA: MIT Press, 1999.

[5] P. D. Turney, Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews, in Proc. ACL, 2002, pp. 417424.

[6] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, Lexiconbased methods for sentiment analysis, Comput. linguist., vol. 37, no. 2, pp. 267307, 2011.

[7] B. Pang, L. Lee, and S. Vaithyanathan, Thumbs up?: Sentiment classification using machine learning techniques, in Proc. EMNLP, 2002, pp.7986.

[8] J. Zhao, L. Dong, J. Wu, and K. Xu, Moodlens: An emoticon-based sentiment analysis system for chinese tweets, in Proc. SIGKDD, 2012.

[9] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, Learning word vectors for sentiment analysis, in Proc. ACL, 2011.

[10] G. Paltoglou and M. Thelwall, A study of information retrieval weighting schemes for sentiment analysis, in Proc. ACL, 2010, pp. 13861395.

[11] Y. Choi and C. Cardie, Learning with compositional semantics as structural inference for subsentential sentiment analysis, in Proc. EMNLP,2008, pp. 793801.