SENTIMENT ANALYSIS ON TWITTER

Pooja Dhotre[1], Yogeshwari Medhane[2], Shraddha Mandlik[3],
Abhishek Mansukh, Prof. J.V.SHINDE [5]

[1] *Student, Computer Engineering, LGNSCOE, Maharashtra, India*
[2] *Student, Computer Engineering, LGNSCOE, Maharashtra, India*
[3] *Student, Computer Engineering, LGNSCOE, Maharashtra, India*
[4] *Student, Computer Engineering, LGNSCOE, Maharashtra, India*
[5]*Professor, Computer Engineering, LGNSCOE, Maharashtra, India*

ABSTRACT

To propose a joint classification for tweets using big data and social networking site twitter. These days' classification techniques 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 unique approach where tweets will be classified as whole and not in words. To enhance and expedite the concept, we thought of using distributed processing 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 so distributed system is preferable. The analyzed information results that are returned will be assembled and interpreted 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 such as positive and negative.

Keyword - *Tweet, Segmentation, Machine learning, Sentiment analysis, Sentiment classification, NLP.*

1. INTRODUCTION:

These days SNS (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 SNS available today and many more are increasing each day.

Thus one of the most famous SNS 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 SNS may help in answering and predicting many answers.

This online social network is used by billions of people around the world to remain socially connected to their friends, family members, and coworkers through their computers and mobile phones. Twitter asks one question, "What's going on?" Answers must be fewer than 140 characters. A status update message, called a tweet, is often used as a message to friends, family and colleagues. A user can follow other users; that user's followers can read her tweets on a daily basis. A user who is being followed by another user need not reciprocate by following them back, which leaves the links of the network as directed. Since its launch on July 2006, Twitter users have increased dramatically.

Thus this kind of SNS 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 we thought of designing a

project to develop a system which helps in analyzing and helping in developing an application for the purpose of sentiment analysis.

2. Literature Survey

1. BO PANG and LILLIAN LEE in 2008:- 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.

2) Bing liu in 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.

3) Jichang Zhao, Li Dong, Junjie Wu, and Ke Xu in 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 Naïve Bayes classifier, with an experimental precision of 64.3.

4) Georgios Paltoglou/ Mike Thelwall in 2010:- Most sentiment analysis approaches use as base line a support vector machines (SVM) classier with binary unigram weights. In this paper, we explore whether more sophisticated feature weighting schemes from Information Retrieval can enhance classification accuracy. We show that variants of

the classic tf-idf scheme adapted to sentiment analysis provide significant increases in accuracy, especially when using a sublinear function for term frequency weights and document frequency smoothing. The techniques are tested on a wide selection of data sets and produce the best accuracy to our knowledge your research work Introduction related your research work Introduction related your research work Introduction related your research work Introduction related your research work Introduction related your research work Introduction related your research work Introduction related your research work Introduction related your research work Introduction related your research work

3. System Architecture

In proposed system first the tweets are accessed using Apache spark and Twiiter4j API. Then they are preprocessed and then machine learning is applied to it to get the sentiment.

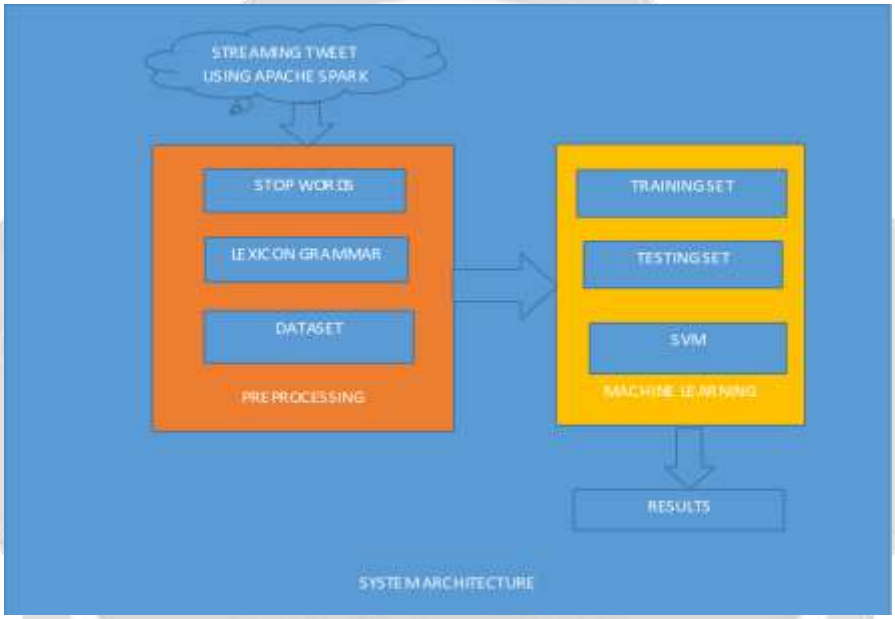

Fig -1: System architecture

3.1 Tweet access using Apache Spark Module

This module first initializes Apache spark over a local ip address then authentication is provided to twitter by creating a developer account and get authentication details for access. The apache spark will access clusters of twitter and access only the latest tweets.

3.2 Preprocessing Module

In this module the tweets will be accessed from text file and then the stop words removal is generated. The stop words which are not good for text mining are matched with an array of stop words and unnecessary words are removed from the tweets the preprocessed tweets are then stored in a separate text file.

3.3 Lexicon Grammar Module

In this module the preprocessed text file is accessed and AFFIN library is applied to it. It returns adjective, noun and pronoun of a word. We only access pronouns and shorten the tweet further. The shorten tweets are again stored in a separate text file.

3.4 Machine Learning Module

In this module first we design a training dataset with two classes positive and negative. Then a tweet is accessed and a test dataset is generated. Then a instance of SVM classifier is generated and training and testing dataset is applied to it. It returns the result in the form of parameters such as tp rate, fp rate, precision and recall. We take in to account precision if it is greater than 0.5 then the tweet comes under positive sentiment and if it is smaller than 0.5 the tweet comes under negative sentiment.

4. CONCLUSIONS

In this paper, we are developed novel sentiment analysis approach using TWITTER and APACHE SPARK together. The basic idea of the project is to download tweets using distributed computing i.e. Apache Spark and training and testing the machine learning classification using SVM algorithm. We are going to assemble the statistical results 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 as a whole.

5. 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 HumanLang. 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. 414.

[6] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, Lexiconbased methods for sentiment analysis, Comput. linguist., vol. 37, no. 2, pp. 267, 2011.

[7] B. Pang, L. Lee, and S. Vaithyanathan, Thumbs up?: Sentiment classification using machine learning techniques, in Proc. EMNLP, 2002, pp. 796.

[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. 138

[11] Y. Choi and C. Cardie, Learning with compositional semantics as structural inference for subsentential sentiment analysis, in Proc. EMNLP, 2008, pp. 793