International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 05 | May-2016 www.irjet.net p-ISSN: 2395-0072

Automatic Summarization Of Tweet From Social Media Using Tweet Classification and Clustering(Based On GPU).

ThoratPrashant,IchamVivek,Modi Malay,UmrekarPandurang

Student, Computer Department, Met's BkcIoeNashik, Maharashtra, India

---

Abstract-Social media is increased in presence and importance in society. A social network service consists of a representation of each user. Social networking sites allow users to communicate with people in the network by sharing thoughts, pictures, status, posts, activities and products. It has become one of the biggest forums to express ones opinion. The majority of earlier work in Rating Prediction and Recommendation of products mainly takes the star ratings of users on products. However, most reviews are written in a free-text format which is difficult for computer systems to understand, analyse and aggregate. The proposed system is able to collect useful information from the social website and efficiently perform sentiment analysis of the reviews on product. The work focuses on identifying the sentiment information from free form text reviews and using that information to rank the product. The sentiment of the user reviews is predicted using awelltrained effective Naive Bayes classifier parallel k means algorithm.System will produce tweets result in summarized form using sentiment analysis.

Keywords—Porter Stemmer, Parallel algorithm, GPU,Naive Bayes algorithms.

1. INTRODUCTION--

1.1 PROJECT IDEA

Increasing popularity of microblogging services such as Twitter, Weiboand Tumblr has resulted in the explosion of the amount of short-text messages. Twitter, for instance, which receives over 400 million tweets per day1has emerged as an invaluable source of news, blogs, opinions, and more. Tweets, in their raw form, while being informative, can also be overwhelming. For instance, search for a hot topic in Twitter may yield millions of tweets, spanning weeks. Even if filltering is allowed, plowing through so many tweets for important contents would be a nightmare, not to mention the enormous amount of noise and redundance that one might encounter. To make things worse, new tweets satisfying the filtering criteria may arrive continuously, at an unpredictable rate. One possible solution to information overload problem is summarization. Summarization represents a set of documents by a summary consisting of several sentences. Intuitively, a good summary should cover the main topics (or subtopics) and have diversity among the sentences to reduce redundancy.

Summarization is extensively used in content presentation, specially when users surf the internet with their mobile devices which have much smaller screens than PCs.Traditional document summarization approaches, however, are not as effective in the context of tweets given both the large volume of tweets as well as the fast and continuous nature of their arrival. Tweet summarization, therefore, requires functionalities which significantly differ from traditional summarization.In general, tweet summarization has to take into consideration the temporal feature of the arriving tweets.

Let us illustrate the desired properties of a tweet summarization system using an illustrative example of a usage of such a system.Consider a user interested in a topic-related tweet stream, for example, tweets about Apple. A tweet summarization system will continuously monitor Apple related tweets producing a real-time timeline of the tweet stream. a user may explore tweets based on a timeline (e.g., Apple tweets posted between October 22nd, 2012 to November 11th, 2012). Given a timeline range, the summarization system may produce a sequence of timestamped summaries to highlight points where the topic/subtopics evolved in the stream. Such a system will effectively enable the user to learn major news/ discussion related to Apple

without having to read through the entire tweet stream. Given the big picture about topic evolution about Apple, a user may decide to zoom in to get a more detailed report for a smaller duration (e.g., from 8 am to 11 pm on November 5th). The system may provide a drill-down summary of the duration that enables the user to get additional details for that duration. A user, perusing adrill-down summary, may alternatively zoom out to a coarser range (e.g., October 21st to October 30th) to obtain a roll-up summary of tweets. To be able to support such drill-down and roll-up operations, the summarization system must support the following two queries: summaries of arbitrary time durations and real-time/range timelines. Such application would not only facilitate easy navigation in topic-relevant tweets, but also support a range of data analysistasks such as instant reports or historical survey.

1.2 MOTIVATION OF THE PROJECT

Out of all the social networking platforms, Many people make use of twitter micro blogging site to express themselves in the limit of 140 characters. Twitter has had the provision of verified accounts since long and therefore the communication coming from these accounts looks more authentic and is successful in creating more buzz than the other social media platforms. so by analyzing tweets we can easily predict the sentimental analysis, trend analysis and volume analysis.

2. PROBLEM DEFINATION AND GOALS

2.1 PROBLEM STATEMENT

-Previous systems do not make use of parallel processing(GPU) so they require a lot of time for execution, hence increased time complexity.

-Previous systems are not able to provide accurate predictions.

-Previous systems are not able to represent the results graphically along with tweet locations.

2.2 GOALS AND OBJECTIVES

The main goals of the project are as follows:

-To analyze and draw meaningful inferences from the collection of tweets collected over the entire duration of elections.

-To check the feasibility of development of a classification model to identify the political orientation of the twitter users based on the tweet content and other user based features.

-To develop a system to analyse and monitor the election related tweets on daily basis.

-This system uses k-means algorithm with parallel processing(GPU),hence it requires less execution time and thus less time complexity.

-This system uses port stemmer and naïve bayes algorithm to increase the accuracy of predictions.

- This system will represent the prediction results and plot results geographically ongoogle maps.

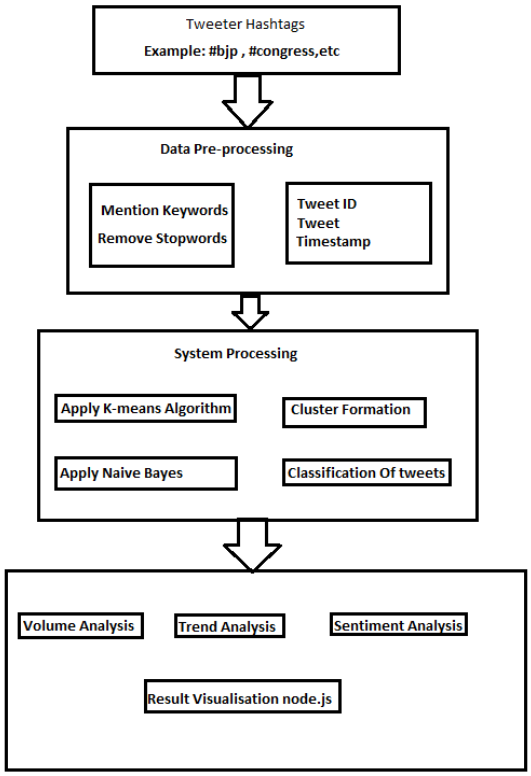
2.3 METHODOLOGIES OF PROBLEM SOLVING AND EFFICIENCY ISSUES

Methodology for data analysis: As we mentioned earlier, one of our research objectives is to collect big stream textual data and analyzethem.

Methodology for identifying sentimental orientation: To analyze and draw meaningful inferences from the collection of *tweets. To check the feasibility of development of a classification model to identify the sentimental orientation of the twitter users based on the tweet content and other user based features.

Methodology for system design : To develop a system to analyze and monitor the different issues related tweets

3.SYSTEM ARCHITECTURE --



4. SYSTEM DEATAILS AND ALGORITHMS

Algorithm 1: Porter Stemmer Algorithm

Input:

Let T be the set of downloaded tweets.

Output:

Processed tweets with all unwanted word, space and special character removal.

Algorithm 2: K-Means Clustering Algorithm

Initially, topic wise tweets are set as a canter of clusters. For every iteration, distance between center and sample is checked and sample is added to respective cluster. Distance between center and sample is measured using TF(Term Frequency) .Clusters are updated at every iteration. Based in TF(Term Frequency) weightage

Algorithm 3: Naive bayes Classifier Algorithm

Political orientation of users towards party, topics can be analysed from tweets. navebayes algorithm will be implemented to classify tweets into positive, negative and neutral classes.

Input : User Tweet.

Output : positive or Negative tweet Label is assign.

4.1 GPU FOR PARALLEL PROCESSING

A Graphical processor unit (GPU) , also called as visual processor unit.It is special electronic circuit designed to rapidly manipulate and alter memory to accelerate the creation of images in frame buffer intended for output to a display.In GPU processing of large block of data is done in parallel.GPU supports thousands of active threads.Modern GPU's are very efficient at manipulating computer graphics and image processing and their parallel structure make them more effective than general purpose CPU's.GPU are used in embedded system, mobile phones ,personal computers ,workstations and game console.

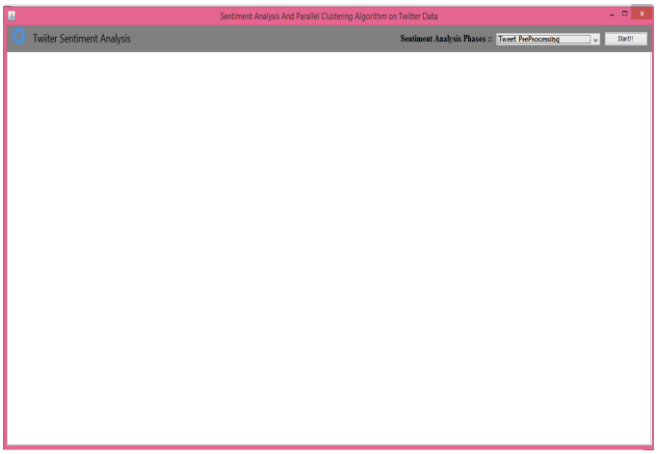


Fig. A:System Interface

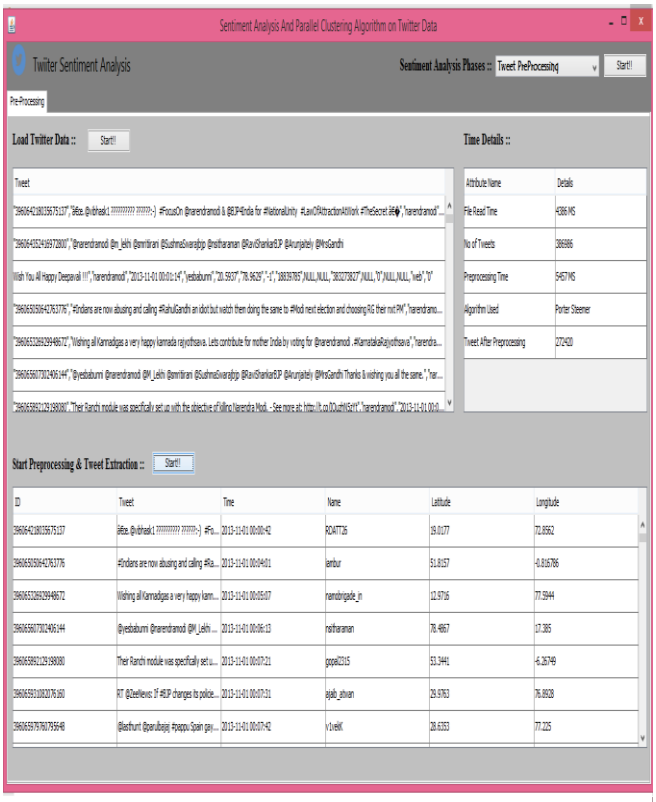


Fig. B: Pre-Processing

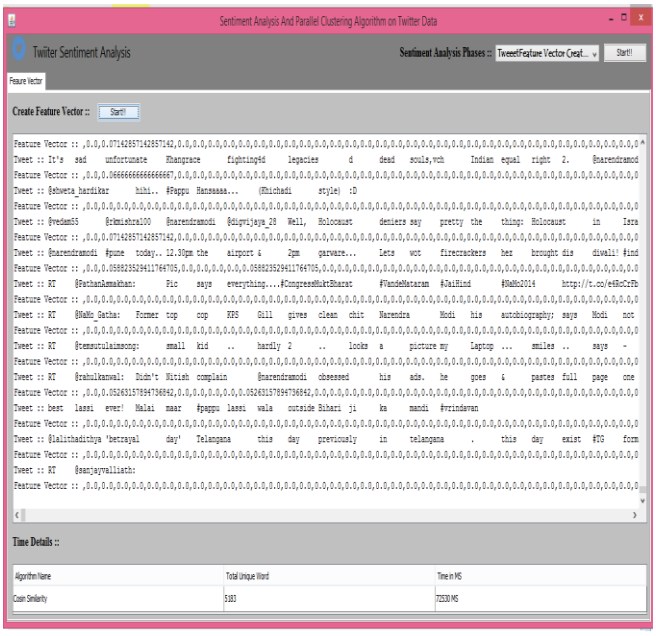


Fig. C:Feature Vector Creation

5.CONCLUSION--

A requirement of ABE with outsource decryption with verifiability is considered. Developing the original model of ABE with outsource Decryption. This ABE scheme with Verifiable outsource decryption and proven that it is secure and verifiable . Provided encrypted data is store in cloud and resilient access control . It eliminates Decryption on resource limited devices. This data flow is provide more secure connection between server and small devices .We more improve the data security process by ABE outsourced decryption technique .We use AES algorithm and Hellman Key Exchange technique for improving the security in data flow between server and small resource limited devices.

REFERENCES--

[1] P. S. Bradley, U. M. Fayyad, and C. Reina, Scaling clustering algorithms to large databases, in Proc. Knowl. Discovery Data Mining, 1998, pp. 915.

[2] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, A framework for clustering evolving data streams, in Proc. 29th Int. Conf. Very Large Data Bases, 2003, pp. 8192.

[3] D.Wang, T. Li, S. Zhu, and C. Ding, Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization, in Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2008, pp. 307314.

[4] Hull David A. and Grefenstette Gregory. A detailed analysis of English stemming algorithms.Rank Xerox ResearchCenter Technical Report.1996.

[5] A. McCallum, and K. Nigam, A comparison of event models for nave Bayes text classi_cation, Journal of Machine Learning Research, Vol. 3, 2003, pp. 12651287.

[6] Bollen, J., Mao, H., and Pepe, A. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In ICWSM (2011).

[7] Conover, M. D., Goncalves, B., Ratkiewicz, J., Flammini, A., and Menczer, F. Predicting the political alignment of twitter users. In Privacy, security, risk and trust (pas-sat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom) (2011), IEEE, pp. 192-199.

[8] B.Waters, Cipher text-policy attribute-based encryption: An expressive, e_cient, and provably secure realization, in Proc. Public Key Cryptography, 2011, pp. 5370.

[9] Zhenhua Wang, LidanShou, Ke Chen, Gang Chen, and SharadMehrotra: On Summarization and Timeline Generation for Evolutionary Tweet Streams

BIOGRAPHIES--

ThoratPrashantappearing for BE degree from the Department of Computer Engineering, MET's Bhujbal Knowledge City IOE, Nashik.

IchamVivekappearing for BE degree from the Department of Computer Engineering, MET's Bhujbal Knowledge City IOE, Nashik.

Modi Malayappearing for BE degree from the Department of Computer Engineering, MET's Bhujbal Knowledge City IOE, Nashik.

UmrekarPandurangappearing for BE degree from the Department of Computer Engineering, MET's Bhujbal Knowledge City IOE, Nashik.