Volume No:2, Issue No:6 (November-2016) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

Tweets Parallel Summarization Using TCP Protocol

P.Mangaveni
Final M.Tech (CSE) Student,
Department of CSE,
Vignan's Institute of Engineering for Women,
Visakhapatnam, Andhra Pradesh.

R.Rama Janaki Devi
Assistant Professor,
Department of CSE,
Vignan's Institute of Engineering for Women,
Visakhapatnam, Andhra Pradesh.

ABSTRACT:

Social media is one of the biggest forums to express opinions. Sentiment analysis is the procedure by which information is extracted from the opinions, appraisal and emotions of people in regards to entities, events and their attributes. Sentiment analysis is also known as opinion mining. Opinion mining is to analyze and cluster the user generated data like reviews, blogs, comments, articles etc. These data find its way on social networking sites like twitter, face book etc. Twitter has provided a very gigantic space for prediction of consumer brands, movie reviews, democratic electoral events, stock market, and popularity of celebrities. Short-instant messages, for example, tweets are being made and shared at an extraordinary rate. Tweets, in their crude structure, while being enlightening, can likewise be overpowering. The main objective of opinion mining is to cluster the tweets into positive and negative clusters.

An earlier work is based on continuous tweet stream summarization The proposed work is able to collect information from social networking sites like Twitter and the same is used for sentiment analysis. The processed meaningful tweets are cluster into two different clusters positive and negative using summarization technique such as tweet cluster probability. we propose a novel Parallel summarization framework using clusters Parallel summarization Frame work to aggravate the problem. In contrast to the traditional document summarization methods which focus on static and small-scale data set, Parallel summarization Frame work is designed to deal with dynamic, quick arriving, and huge scale tweet streams.

Our experiments on large-scale real tweets demonstrate the efficiency and effectiveness of our framework.

Keywords:
Tweet stream, Tweet Parallel summarization, timeline, information filtering.

1. INTRODUCTION:

The emergence of social media has given web users a venue for expressing and sharing their thoughts and opinion on different topics and events. Twitter, with nearly 600 million users and over 250 million messages per day, has quickly become a gold mine for organizations to monitor their reputation and brand by extracting and analyzing the Sentiments of the tweets posted by public about them, their markets, and competitors. Opinion mining uses some algorithm techniques to cluster the user opinions into positive and negative clusters. Tweets contain a wide variety of useful information from many perspectives about important events taking place in the world .The huge number of messages ,many containing irrelevant and redundant information ,quickly leads to a situation of information overload .This motivates the need for automatic summarization systems which can select a few messages for presentation to a user which cover the most important information relating to the event without redundancy and filter out irrelevant and personal information that is not of interest beyond the user's immediate social network. Earlier work is based on continuous summarization (tweet cluster vector).

Volume No:2, Issue No:6 (November-2016) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

Tweet cluster vector algorithm have been popularly used and proven its effectiveness in sentiment classification. The data structure called TCV for stream processing; these will produce continuous streaming in sequence order .So little bit filtering process going to be down .The large amount of data set will process like clusters on that cluster they can applies TCV. It's give data filtering slowly. It is highly depend on large amount of labeled data which results in time consuming and information filtering performance is slow. Based on the previous work tweet cluster probability method are proposed to overcome the problem of tweet cluster vector method which require large amount of data. In our proposed system we are come up with Tweet Clustering Probability (TCP).In this stream process we are Appling parallel Streaming techniques. Based on the parallel Stream process our data Clusters are filtering or sorting to get the unique data from the unstructured data set .we summarize the tweets by using Tweet cluster probability (TCP).

2. LITERATURE SURVEY:

2.1 Tweet summarization:

a novel framework consists of three major components discussed[1]. First, online tweet stream clustering method is used for cluster tweets using tweet cluster vector (TCV). Second, TCV-Rank summarization technique can be used for generating online summaries and historical summaries of arbitrary time durations. Third, topic evolution detection method can be used for monitoring summary-based tweet based on given timelines automatically from tweet streams.

2.2 Stream clustering:

Stream data clustering is a greater extent. BIRCH [2] clusters the data by using an in-memory structure called CF-tree instead of the original or actual huge data set. The thesis proposed another framework of clustering which stores only the important parts of the data selectively, and discards other parts of data which are waste. It is one of the most important stream clustering methods. It contains an online micro-clustering content and an offline macro-clustering component.

2.3 Real time information:

"Event summarization using tweets"[3], in this paper the key concept is discussed about real time streaming of information in search engines using leading search engines routinely displaying relevant tweets in response to user queries. Recent research has shown that a considerable fraction of these tweets are about "events", and the detection of novel events in the tweet-stream has attracted a lot of research interest. However, very little research has focused on properly displaying this real-time information about events.
Opinion mining or sentiment analysis refers to the application of natural language processing, computational linguistic and text analytics to identify and extract subjective information in source materials. Millions of people have primary focus on social media platforms to share their own thoughts and opinions in regards to their day to day life, business, celebrity, entertainment, politics etc.

2.4 Twitter corpus:

This thesis uses a dataset formed of collected messages from twitter. Twitter [4] contains a very large number of very short messages of 140 character created by the users of this micro blogging platform. The contents of the messages vary from personal thoughts to public statements [5]. Extracting the public opinion from social media text provides a challenging and rich context to explore computational models of natural language, motivating new research in computational linguistics.

2.5 opinion mining:

In opinion mining task documents and example are represented by thousands of tokens, which make the clustering problem very hard for many clustering system. In feature extraction [7], the original features converted to more compact new space. All the original features are transformed into new reduced space without deleting them but replacing the original

Volume No:2, Issue No:6 (November-2016) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

features through a smaller representative set.

3. PROPOSED SYSTEM:

Various techniques have been used to do sentiment analysis or opinion mining of tweets. But for parallel summarization of tweets live data has to be taken into consideration. In the proposed system data set will be instantly taken along with live information. A dataset is created using twitter posts of any given topic . As we know that tweets contains slang words and misspelling. So we perform a sentiment level sentiment analysis on tweets. This is done in three phases. In the first phase preprocessing is done .In this tweet data is minimizing. Second phase is streaming summarization of tweet data, in this tweet data is created of positive tweets, negative tweets and neutrals .Third phase is select and filter the tweet data into positive and negative tweets. Finally count the positive and negative tweets. In our proposed system we are come up with Tweet Clustering Probability (TCP).In this stream process we are Appling parallel Streaming techniques. Based on the parallel Stream process our data Clusters are filtering or sorting to get the unique data from the unstructured data set .we summarize the tweets by using Tweet cluster probability (TCP).

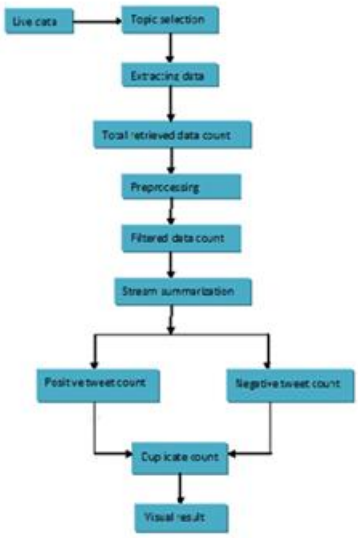

Fig. 3: System model for opinion mining on twitter data.

In the proposed system architecture (Fig. 3), it is shown that the social networking site twitter has been the source for data collection. Database is created using twitter posts of movie reviews and related tweets about those movies. The preprocessing is the major part of this architecture. The collected reviews are preprocessed and then words are extracted. Feature extraction are taken for absolute positive words like "wonderful", "awesome", "always" etc and negative words such as "never", "not", "hardly" etc. Feature vector is the vector of all high information words which appear in the document.

3.1 Data extraction:

Data extraction is to fetch the tweets using Twitter API v1.1 to collect the data from any given topic like movies, .politics etc. data will be collected according to local database storage .From the configuration page of app, it can also require an access token and an access token secret, Similarity to the consumer keys ,these string must also be kept private: they provide the application access twitter on behalf of an account.

Algorithm– I :

Tweets parallel summarization using TCP protocol

Input: Extracted data based on given topic

Output: Outcome in terms of clusters

1. T = { tweet topic}
   Extracted from twitter and store it in local data base.
2. T = Data Preprocessing (T). Perform data cleaning and normalization.
3. Parallel Stream summarization of processed data.
4. PSV= positive senti analysis of stream based on word set.
5. NSV= negative senti analysis of stream based on word set.
6. Removal of redundancy.
7. Clustering of Positive tweets and negative tweets.

Volume No:2, Issue No:6 (November-2016) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

3.2 Data cleaning and normalization:

In order to remove stop words and extract features we perform data cleaning and normalization. Preprocessing is the major part of this thesis. Preprocessing of data is the process of preparing and cleaning the tweets for clustering. Reducing the noise in the text should help to improve the performance of the clustering and speed up the clustering process. This project performed the following operations on tweets during cleaning and normalization.

1.Tokenization: given input as character sequence, tokenization is a task of chopping it up into pieces called tokens and at the same time removing certain characters such as punctuation marks.

2.Stop words removal: A stop-list is the name commonly given to a set or list of stop words. It is typically language specific, although it may contain words. Some of the commonly used stop words from English include"a", "of", "the", "I", "it", "you", "and", these are generally regarded as functional words which do not carry any meaning.

3.Internet Acronyms And Emoticons: in this process it convert internet acronyms like <3 to "love" or "gud" to "good" in order to make the meaning out symbols posted as part of the tweets.

4.Word Expansion: this system expands the famous acronyms as well. The expansion is considered for Standard English words. As an example AFAIK-As Far As I Know, LOL-lots of laugh, etc.

5.Repeated Words: if a word is being repeated in a tweet for more than two times consecutively, occurrences of the word had been limited to two occurrences. E.g. very very very very very good has been replaced by very very good. And a character like "loonnnnggggg" has been replaced by "long".

3.3 Feature extraction:

Features from tweets are extracted and this project uses the unigram, bigram and unigram + bigram (hybrid) feature extraction method. Hybrid features are taken for absolute positive words like "wonderful", "awesome", "always" etc and negative words such as "never", "not", "hardly" etc. As an example in following tweet user explains positive sentiment for target topic. This also ignores the future transitive verbs when followed by the query terms e.g. "Indian currency" does not make any value for prediction. This project also makes sure that generic terms like "Indian currency", is ignored during feature extraction to ensure exclusion of non-subjective data. The examples of feature words extracted from sample tweets are shown below

Table 1. Example showing tweets and senti words

Positive Tweets	Feature Words
Money policy, importance Indian currency and value	'importance', 'indian', 'currency'

Negative Tweets	Feature Words
Women allegedly dies shock finding currency withdrawal	'allegedly', 'dies', 'withdrawal'

Volume No:2, Issue No:6 (November-2016) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

Notes printed black listed companies supplied fakes	'Fake','black'

3.4 Feature selection:

Feature selection is used to make the clustering more efficient by reducing the amount of a data to be analyzed as well as identifying relevant features to be considered in clustering process.

Score (location)=$\sum_{i=1}^{n}$ sentiscore i \n

3.5 Feature vector:

The first step in modeling the document into vector space is to create a dictionary of terms present in the documents. To do this, we need to select all terms from the document and convert it to a dimension in the vector space. Tf-idf term frequency and inverse document term frequency value increases proportionally to the number of times a word appears in the tweet data set.

Let's take the data set below to define our domain space:

Train data Sets:

T1: The sky is blue.

T2: The sun is bright.

Test Data Set:

T3: The sun in the sky is bright.

T4: We can see the shining sun, the bright sun.

Now what we have to do is to create an index vocabulary (dictionary) of the words of the train data set, using the tex files T1 and T2 from the data set, we 'ill have the following dictionaries of all the words E (t) where t is the term:

E (t) =

If *t* is "blue"

If *t* is "sun"

If *t* is "bright"

If *t* is "sky"

Note that the terms like "is" and "the" were ignored as cited before. Now we can convert the test document set into a vector space, the first term of the vector represent "blue" term of our vocabulary, the second represents "sun" and so on. Now use the term-frequency to represent each term in our vector space; the term-frequency is nothing more than a measure of how many times the term present in our vocabulary E (t) are present in the document d3 or d4, this define the term-frequency as a counting function:

$$Tf(t, d) = \sum x€d(x.t)$$

This function counts the frequency of a word i.e. *tf (t, d)* returns is how many times that word present in the document, In this thesis we use a threshold value θ=2 if the word is appearing in the document more than two times that can be taken as a high information word and is taken into a dictionary making the dictionary of all high information words.

Where the fr(x, t) is a simple function defined as:

1, if x=t

Fr(x, t) =0, otherwise

Now it compares the documents term with high information word dictionary if the term occur in the document it returns the value 1 in the vector space else its return 0.its make the feature vector of 0's and 1's .

3.6 K-means clustering:

The k means algorithm is used as a last step in spectral clustering algorithm to plot the graph. K-means is the best known partitioning clustering algorithm due to its simplicity and efficiency. Given the data points and required number of k cluster (k is specified by the user), this algorithm iteratively partitions the data into k clusters based on distance function.

Volume No:2, Issue No:6 (November-2016) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

K-means algorithm summarizes as follow in the following Algorithm-3

Algorithm – 3: k-means algorithm

Input: Let X = $\{x_1,x_2,x_3,\ldots\ldots\ldots,x_n\}$ bet the set of data points and V = $\{v_1,v_2,v_3,\ldots\ldots\ldots,v_c\}$ bet the set of centers.

Output: clusters

Step 1: Randomly select 'c' cluster centers.

Step 2: Calculate the distance between each data point and cluster centers.

Step 3: Assign the data point to the cluster center whose distance from the cluster center is minimum of the entire cluster centers.

Step 4; Recalculate the new cluster center using: $V_i = (1/C_i)$ Where 'C_i' represents the number of data points in i-th cluster.

Step 5: Recalculate the distance between each data points and
new obtained cluster centers.

Step 6; if no data point was reassigned then stop, otherwise repeat step 3

4. EXPERIMENTAL RESULT:

This thesis provides experimental results to validate the usefulness of the results presented in previous sections. In this research first illustrate "tweets parallel summarization using TCP protocol" of feature vector with k-means algorithm. Using Eigen vectors to initialize k-means give better initial and final objective function values and better clustering results. Thus the theoretical connection between tweets parallel summarization using TCP protocol and k-means helps in obtaining higher quality results. This screenshot (Fig 1) shows the data extraction (Tweets files). These are the screenshots, which are taken during the project execution, and each screenshot shows project module working.

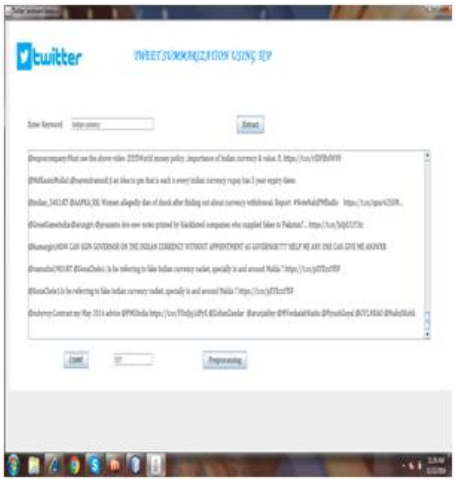

Fig 1.Data Extraction form.

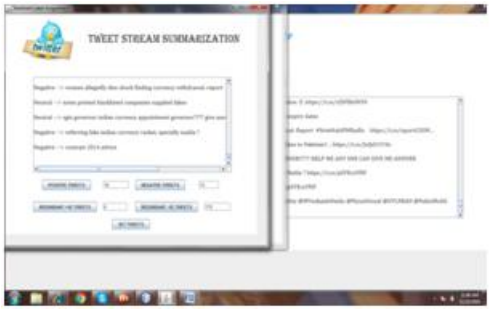

Fig 2.Preprocessing data.

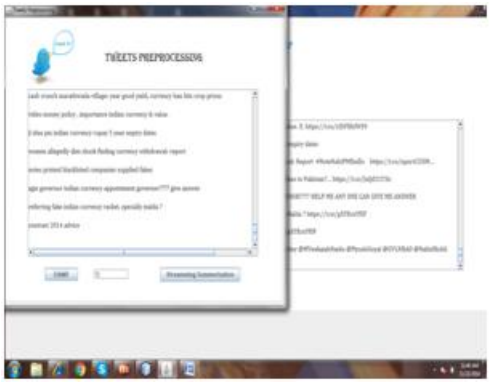

Fig 3.Streaming form.

Volume No:2, Issue No:6 (November-2016) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

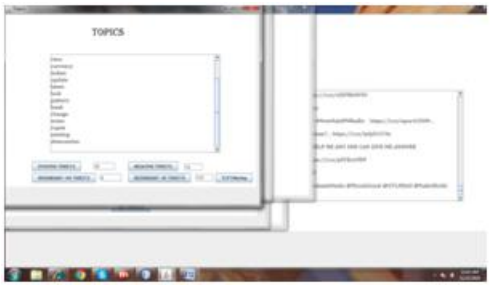

Fig4.Tokenizing process of positive and negative tweets.

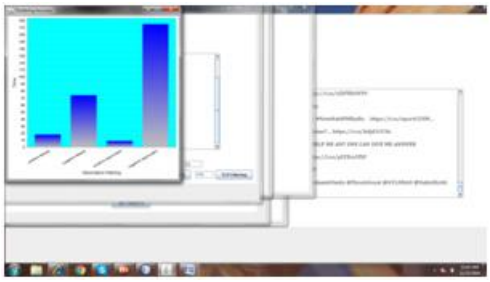

Fig 5.Visual representation of results.

5. PROCEDURAL EXECUTION AND RESULTS:

This thesis uses the dataset collected by twitter API containing tweets in English language, for the original terms and extended features This research achieves the best sentiment accuracy with Tweets parallel summarization using TCP protocol which also leads us to the best accuracy result. In this module the user enter the keyword of any topic .After click on the extract button and system load the tweet data .so tweet data is retrieved from twitter and calculate the number of lines in twitter data. the user enter the keyword then load the tweet data .So perform the tweet preprocessing and minimize the tweet data by removing unrelated information. In this the user preprocess the tweet data ,count the number of lines of minimizing the preprocessing data, then perform the streaming. The user selects the required or relevant tweet data among the extracted tweets. The system summarizes the positive tweets and negative tweets. And exhibits the visual representation of results in the form of bar chart.

6. CONCLUSION:

The thesis concludes that social network based behavioral analysis parameters can increase the prediction accuracy along with sentiment analysis in time. Twitter base social network provides the great platform in measuring the public opinion with the reasonable accuracy in any given topics like movie reviews ,politics ,etc with Tweet parallel summarization using TCP protocol based probability algorithm for sentiment analysis. In this thesis, a new opinion mining of twitter data using clustering probability technique is proposed that can solve the problem of domain dependency and reduce the need of annotated training data. This project main goal is to overcome the problem of clustering multiple files with unlabeled data and perform sentiment classification. Experimental results on extracted tweets demonstrate the graphical representation of obtained results.

7.REFERENCES:

[1] zhenhua wang, lidan shou,ke chen,Gang chen and Sharad Mehrotra,"On summarization and time line generation for evolutionary tweet streams" in IEEE transcation on knowledge data engineering ,2015,vol.27.

[2] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 1996, pp. 103–114.

[3]D.Chakarabarti, k.punera "Event Summarization using Tweets",in 2011,pp.66-73.

[4] Twitter as a Corpus for Sentiment Analysis and Opinion Mining, Alexander Pak, Patrick Paroubek. (2014) .

[5] Dewan Md. Farid, and Chowdhury Mofizur Rahman, "Mining Complex Data Streams:

Discretization, Attribute Selection and classification," Journal of Advances in Information Technology, Vol. 4, No. 3, August 2013, pp. 129-135.

[6] Saeys, Y, Inza, I & Larrañaga, P 2007, „A review of feature selection techniques in bioinformatics. Bioinformatics", vol. 23, no. 19, pp.2507-2517

[7] J.C.Gomez, E. Boiy, M.F.Moens. Highly discriminative statistical features for email classification. Knowledge and Information System, (2012), 31(1); 23-53

[8] Parikh and Movassate , Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques , Stanford University, 2009.