SENTIMENT ANALYSIS AND SUMMARIZATION OF TWEETS WITH TIMELINE GENERATION

[1]Omkar Patil-Karade, [2]Siddhant Gangakhedkar, [3]Shubham Darade,[4]Sarang Chandak
[1]Student, [2] Student, [3] Student, [4]Student
Department of Computer Engineering, Pimpri Chinchwad College of Engineering, SPPU, Pune, India.

Under the guidance of
Prof. Bodireddy Mahalaxmi
Professor, Department of Computer Engineering, Pimpri Chinchwad College of Engineering, SPPU, Pune, India.

Abstract: The popularity of social sites like Twitter, Facebook, and Instagram etc. is increasing rapidly. Huge number of tweets and short messages are being posted every single minute by the users from throughout the world, hence size of the data is very huge. As the data is incoming simultaneously from various sources, it is very critical and time taking to analyze and understand the meaning of data. Redundancy and noise is present in this data at a large extent. This must be removed by using Sentiment Analysis and Summarization of tweets to get the meaningful information. This will help users to get necessary information in short time. Sentiment analysis is contextual mining of text which identifies and extracts subjective information in source material. Initially through HTTP request-response, tweets get extracted from websites by searching keyword called Web Scraping. Preprocessing is done on extracted data in order to remove unwanted data such as stop words, suffixes, punctuation and conjunctions. Term Frequency (TF) and Inverse Document Frequency (IDF) is calculated for each word to know the frequency and importance of word in the document. K-Means clustering is performed on text data to get Tweet Cluster Vectors. The regular approaches of the summarization depend on the static or offline data. We removed this complexity of data and generated simple raw summarized text by using clustered data. Finally, the accuracy is calculated based on tweets and type of the result is displayed.

Keywords: Sentiment Analysis, Web Scraping, Term Frequency, Inverse Document Frequency. K-Means clustering.

I. INTRODUCTION

Extracting Sites such as Twitter have redesigned the way people find, share messages, and broadcast sensible information. Several organizations have been reported to generate and observation targeted Twitter streams to assemble and realize users' opinions. Targeted Twitter stream is typically constructed by filtering tweets with predefined variety criteria (e.g., tweets published by users from an environmental region, tweets that match one or more predefined keywords). Short-text messages such as tweets are being generated and shared at an unparalleled rate. Tweets, in their raw form, while being useful, can also be overwhelming. For both end-users and data analysts, it is a nightmare to plow through millions of tweets which have huge amount of noise and redundancy.

II. LITERATURE REVIEW

Table 2.1 : Literature Review

Sr. No	Studies	Algorithm / Method Used	Description
1.	A framework for clustering evolving data stream	Data stream clustering algorithm	TCVs are considered as potential sub-topic; for stream clustering, Clustream method is used. It includes online and offline micro clustering component.
2.	BIRCH: An efficient data clustering method for very large databases	BIRCH algorithm.	This paper presents a data clustering method named BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), and demonstrates that it is especially suitable for very large databases.
3.	Text stream clustering based on adaptive feature selection.	Text stream clustering	The paper mainly focuses on the problem of adaptive feature selection for clustering text stream. A validity index-based method of adaptive feature selection is proposed, incorporating with which a new text stream clustering algorithm is developed.

4.	A Probabilistic Model for Online Document Clustering with Application to Novelty Detection	Clustering algorithm	Use of non-parametric Dirichlet process prior to model the growing number of clusters, and use a prior of general English language model as the base distribution to handle the generation of novel clusters.
5.	LexRank: Graph based lexical centrality as salience in text summarization.	LexRank algorithm	In this model, a connectivity matrix based on intra-sentence cosine similarity is used as the adjacency matrix of the graph representation of sentences.
6.	Document summarization based on data reconstruction	DSDR algorithm	Summarize documents from the perspective of data reconstruction, and select sentences that can best reconstruct the original documents.
7.	On summarization and timeline generation for evolutionary tweet stream	TCV-Rank summarization algorithm	The paper mainly focuses on Tweet Cluster Vector (TCV), TCV Rank algorithm, Topic evolution. In which TCV used for making effective clustering of tweet.

We have studied the paper "A framework for clustering evolving data stream" (C.C. Aggarwal, J. Han, J. Wang, and P. S. Yu) in which TCVs are considered as potential sub-topic; for stream clustering, Clustream method is used. It includes online and offline micro clustering component. For recalling historical micro cluster, pyramidal time frame also proposed for random time duration. [1]

In "BIRCH: An efficient data clustering method for very large databases" Clusters the data based on an in-memory structure called CF-tree instead of the original large data set. They proposed a scalable clustering framework which selectively stores important portions of the data, and compresses or discards other portions. [2]

Also, we referred, "Text stream clustering based on adaptive feature selection" (L. Gong, J. Zeng, and S. Zhang) worked on a various service on the Web such as news filtering, text crawling, etc. It mainly focuses on topic detection and tracking (TDT). Clustering is used for analyzing text stream. [3]

In "A Probabilistic Model for Online Document Clustering with Application to Novelty Detection" in this paper we studied a probabilistic model for online document clustering. Nonparametric Dirichlet process prior to model the growing number of clusters, and use a prior of general English language model as the base distribution to handle the generation of novel clusters. [4]

For using function lexrank in TCV rank algorithm we have studied "LexRank: Graph based lexical centrality as salience in text summarization" (G. Erkan and D. R. Radev) in this paper lex ranking is calculated. Depending on the similar data graph is created. Lexrank is used for finding top ranked tweets among large data set. [5]

Also, in "Document summarization based on data reconstruction" proposed to summarize documents from the perspective of data reconstruction, and select sentences that can best reconstruct the original documents. [6]

Lastly, we have referred "on summarization and timeline generation for evolutionary tweet stream" we have referred Tweet Cluster Vector (TCV), TCV Rank algorithm, Topic evolution. In which TCV used for making effective clustering of tweet with the help of pyramidal time frame and tweet cluster vector, TCV rank summarization algorithm is used for generating online and historical summaries by evaluating top ranked function, depending upon top ranked tweets summarization is done. Topic evolution detection generates timeline by considering large variation of sub-topics in stream processing. [7]

III. CHALLENGES

Including data in the form of emoticons, images is difficult. Repetition of keywords in the summary. Increasing accuracy is one of the significant challenges in sentiment analysis. Most of the algorithms can be able to classify the dataset but failed in producing accuracy. This leads to the system unable to recommend further. The existing system produces lower accuracy, and the proposed approach must provide the accuracy more excellent than the existing system algorithms.

IV. MOTIVATION

As rapid growth in an internet, use of social media also increases. There are many social sites like Twitter, Facebook, Instagram etc. in which twitter has become one of the most popular social site for users to share information like text, audio, video etc. Short messages are being created and shared at massive rate. Twitter receives thousands of tweets per hour.

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, especially 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.

Sentiment analysis is one of the trending research studies undertaken by various researchers. Text reviews are generally consisting of various expressions, traditional words which are obtained through various researchers.

To identify the sentiment of the entire document, scoring algorithms can be used by classifying weighted phrases from the extracted text. Here users get an opportunity to express their individual opinion about particular topics. Every algorithm is efficient in their ways. Every algorithm has various advantages and disadvantages by performance, accuracy, and quality of training the data sets. Lots of researchers are doing in order to increase the accuracy of the data. This accuracy profoundly helps in the recommendation for the users who use the system. Accuracy is one of the essential measures in classification. This decides whether the algorithm functions properly or not for the dataset. Choosing an algorithm is highly depends on what kind of data we are using and how we are going to process and analyze it.

V. PROPOSED SYSTEM

We used both historical and online data for the dynamic summary generation and timeline generation in the proposed method. Because of online data extraction, we used Twitter API for getting maximum number of tweets (200 tweets). After extracting tweets in an efficient way, we proposed pre-processing module on data in order to remove unwanted words, phrases and symbols. In this way, the efficiency of further algorithms is already increased by providing accurate and necessary input to the algorithms. These further algorithms are K-Means clustering algorithm and TCV rank summarization algorithm which perform the main part of the system. Hence, the resultant summary is efficient with less repetitive words and no noisy data.

Existing algorithms only work on historical datasets which is less efficient as the latest data is not taken into consideration. It is highly challenging to preprocess data before clustering on the basis of TF-IDF value of words in a document. This problem is resolved using our proposed system. Proposed system as shown in figure 1 mainly consists of following modules:

1. Clustering module.
2. Summarization module.
3. Categorization module.
4. Timeline Generation module.

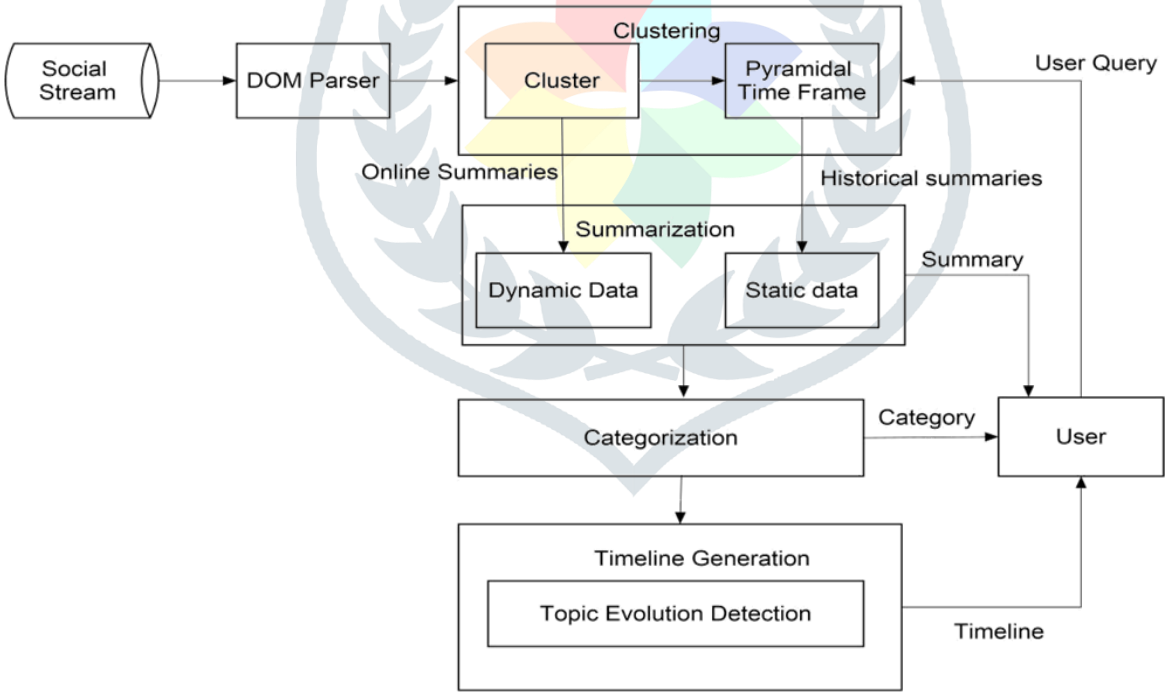

Figure 5.1 : System Architecture

5.1 Processing Proposed Modules

5.1.1 Clustering Module

This module uses K-Means clustering algorithm on extracted pre-processed data. It forms clusters of similar phrases considering rank of each word using its TF-IDF value. Cosine similarity is used for similarity matrix of words in the document. Output of this module is used for next phases like summarization and sentiment analysis.

5.1.2 Summarization Module

After applying clustering on data, summarization is done using TCV-rank summarization algorithm. Both individual summary of each cluster and file summary is generated in this module. Pre-processing of data is already done before this module. Hence, efficient summary is generated without repetition of words.

5.1.3 Categorization Module

In this module, from generated summary categorization is done on the basis of keywords. So user get to know about category of summary that what it relates to.

5.1.4 Timeline generation Module

The base of the timeline generation is topic evolution detection which uses online summaries and generates timeline. Topic evolution describes changes in subtopics by monitoring variation in stream clustering.

VI. ALGORITHMIC IMPLEMENTATION

6.1. K-Means

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable *K*. The algorithm works iteratively to assign each data point to one of *K* groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the *K*-means clustering algorithm are:

1. The centroids of the *K* clusters, which can be used to label new data.
2. Labels for the training data (each data point is assigned to a single cluster).

STEPS:

Step 1: Each document is represented as vector using the vector space model.
For example: TFIDF weight.

TFIDF: It stands for Terms Frequency Inverse Document Frequency, is a numerical statistic which reflects how important the word is to document in a collection or corpus.

a) Term Frequency: The number of times a term occurs in the document.
b) Inverse Document Frequency: Measure of whether a term is common or rare across all documents.

Step 2: Finding Similarity Score
Use Cosine similarity to identity similarity score of the Document.

Step 3: Preparing document cluster.

Step 4: Initializing cluster center.

Step 5: Finding closest cluster center.

Step 6: Identifying the new position of cluster center.

6.2. Summarization Algorithm

The main idea of summarization is to find a subset of data which contains the "information" of the entire set. Such techniques are widely used in industry today. Search engines are an example; others include summarization of documents, image collections and videos. Document summarization tries to create a representative summary or abstract of the entire document, by finding the most informative sentences.

STEPS:

Step 1: The first step would be to concatenate all the text contained in the articles.

Step 2: Then split the text into individual sentences.

Step 3: Find vector representation (word embedding) for each and every sentence.

Step 4: Similarities between sentence vectors are then calculated and stored in a matrix.

Step 5: The similarity matrix is then converted into a graph, with sentences as vertices and similarity scores as edges, for sentence rank calculation.

Step 6: A certain number of top-ranked sentences form the final summary.

6.3. Real time sentiment analysis

Sentiment analysis is contextual mining of text which identifies and extracts subjective information in source material, and helping a business to understand the social sentiment of their brand, product or service while monitoring online conversations. However, analysis of social media streams is usually restricted to just basic sentiment analysis and count based metrics. This is akin to just scratching the surface and missing out on those high value insights that are waiting to be discovered.

STEPS:

Step 1: Creating a Stream.
- a) Authenticate
- b) Build Stream0

Step 2: Data Cleaning

Tweets can contain many non-ASCII characters. Therefore, we need to sanitize it.

Step 3: Sentimental Analysis

Library used: TweetInvi.

Step 4: Produce output (Positive, Negative, and Neutral).

VII. RESULTS

- Proposed system helps to utilize user's time in a proper manner and increase their productivity and creativity.
- System is able to formulate which trend is going on around the world.
- System is able to analyze data with respect to particular topic using summary of the topic.
- System generated results are used for statistical study of other data.

VIII. COCLUSIONS AND FUTURE WORK

Thus, application will provide a way to generate efficient summary of text and helps in self and social development by providing idea about a topic in a faster way. It also detects human approach depending upon the data collected from various comments and also helps in identifying trends around the world which will helps in business process for decision making.

Further work will be based on the aspect level classification. Aspect level classification is something where the review will be specified rather than a standard classification. In aspect level classification, Features are specified and processed for the algorithms. For example, in the product review dataset, specified datasets such as electronics, books, mobile phones are taken, pre-processed and classified to get deeper accuracy, better performance. Thus, the review-based classification is highly developed these days and helps in the recommendation of the product to the users based on the reviews expressed by the user. We can also work on expressing emoticons used by the users in the review.

REFERENCES

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.

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.

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.

G. Erkan and D. R. Radev, *"LexRank: Graph-based lexical centrality as salience in text summarization",* J. Artif. Int. Res., vol. 22, no. 1, pp. 457-479, 2004.

Hull David A. and Grefenstette Gregory. *"A detailed analysis of English stemming algorithms"*. Rank Xerox ResearchCenter Technical Report.1996.

J. Zhang, Z. Ghahramani, and Y. Yang, *"A probabilistic model for online document clustering with application to novelty detection"*, in Proc. Adv. Neural Inf. Process. Syst., 2004, pp. 1617-1624.

L. Gong, J. Zeng, and S. Zhang, *"Text stream clustering algorithm based on adaptive feature selection"*, Expert Syst. Appl., vol. 38,no. 3, pp. 1393-1399, 2011.

N. A. Diakopoulos and D. A. Shamma, *"Characterizing debate performance via aggregated twitter sentiment",* in Proc. SIGCHI Conf. Human Factors Comput. Syst., 2010, pp. 1195-1198.

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

R. Yan, X. Wan, J. Otterbacher, L. Kong, X. Li, and Y. Zhang, *"Evolutionary time- line summarization: A balanced optimization framework via iterative substitution"* in Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2011, pp. 745-754.

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.

Z. He, C. Chen, J. Bu, C. Wang, L. Zhang, D. Cai, and X. He, *"Document summa- rization based on data reconstruction"*, in Proc. 26th AAAI Conf. Artif. Intell., 2012, pp. 620-626.