ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

Summarization of Social Media Data Using Topic Detection

Rajani D.Gavali[1], Prof. A.R.Kulkarni[2]

ME (CSE) Student, Department of Computer Science & Engineering, Walchand Institute of Technology, Solapur, Maharashtra, India [1]

Assistant Professor, Department of Computer Science & Engineering, Walchand Institute of Technology, Solapur, Maharashtra, India [2]

ABSTRACT: As we know in today's life Twitter, Facebook, Google plus are well known social sites now that everyone uses for different purposes. Many People have social accounts now days. Twitter is one of the growing social site that people are using for connecting, sharing with each other. There is dynamic flow of short text messages posted by many people called as tweets. It is difficult to do analysis of social media data which has large amount of noisy, informal text of tweet messages and stream data. As a result, people are unable to understand the current topics of discussion and because of time limitation it is impossible to read each and every tweet. To understand important information on social media, system will generate summary. In this paper, tweet stream of twitter is used and processed using Dynamic LM Classifier algorithm. This classification is useful in topic detection on social media.

KEYWORDS: Data mining, Summarization, Social media, Topic detection.

I. Introduction

A. MOTIVATION

Data mining has many techniques to find and analyse data from different sources and summarizing it into useful information. Usually the Internet provides large scale information than required. Even though filtering process is applied on text, processing and summarizing this text, short messages for human beings is difficult. A Social media provides best platform to people for communication and discussion, status update, way to express opinions regarding various topics. Twitter is one of the well-known social site on which people always search for trending topic, especially when surfing the internet with their mobile devices which have too small screen than computer screen. Twitter that enable their users to update, comment, and communicate with users in their social circle. Twitter users can compose and read short length 140-character messages called as tweets.Many People usually upload their daily routines, update status, other activities on social sites.

B. *PROBLEMS AND SOLUTIONS*

However while extracting information from social media such as the huge amount of fast arrived stream of data from Twitter, researchers and analysts face some problems such as language difficulty, short and unstructured message, some tweet contains only hyperlinks and so on. In the case of social sites, daily 500 million short tweets are posted on twitter in a variety of languages. It is difficult to get key contents of topics from social sites and becomes time consuming task to read millions of tweets and to find the reasonable tweets. Summarization is useful solution to this problem which helps to get key contents from large data set of social media. An effective summary includes the topic or subtopic evolved in the stream.

II. RELATED WORK

In the earlier research, various techniques were presented for summary generation Related to social media.
In this paper, system made use of new approach, a novel continuous summarization framework Sumblr is used which is designed to deal with dynamic, fastest arriving, and large amount scale of tweet streams of twitter. It normally consists of different major components. They proposed an algorithm for clustering which is online tweet stream clustering for the tweet stream and maintain statistics as tweet cluster vector (TCV). Another second approach is new summarization method, a TCV-Rank for generating online summaries and historical summaries of specific time durations. Third one is, they used good topic detection method, for monitoring variations in summary-based/volume-based to produce timelines from streams easily and automatically. [2].

In this research System is used and presented the real-time interaction of events such as earthquakes in Twitter. Also designedan algorithm for monitoring tweets to detect a target event. For detecting a target event, System made use of devise classifier of tweets streams which is based on special features such as keywords presents in a tweet, the total number of words, and their context. They also considered as every user of twitter can be a sensor and apply filtering called as aKalmanfiltering and particle filtering. Both are used for detection of location estimation in ubiquitous/pervasive computing [3].

In this work presented the algorithm for Automatic Twitter Topic Summarization is used called as Phrase Reinforcement algorithm. It is used to give summary of social site like Twitter which is accurate and meaningful. In this method they calculates occurrences of each word in every sentence of stream and then according to high word occurrence count, it builds a graph .Then it produced summary by combining or merging nodes present in the graph with the help of highest word occurrence count. They also made use of a front-end classifier for identifying trending topics within different general categories like sports, politics, and world events. After this, they were summarizing these topics in order to generate an automated real-time newspaper [4]

This paper presented algorithms for summarization of micro blog posts. Algorithm is used for collections of short posts or messages on specific topics that are available on the social site Twitter and displayed short summaries from the collections of messages on that specific topic. Here, goal was to produce summaries which are similar to human generated summary for the same collection of posts, messages on that particular topic. Also compare the summaries using summarizing algorithms with human created summaries and get the best results output [5]

In this work, they presented summarization method of Topics on social media Twitter of Tweet Ranking. For this they made use of the method Social Influence and Content Quality. In this work they performed segmentation of sub topic and gave summarization. Also they implemented content quality estimation, and generation of summary by removing redundancy of tweets related to that topics. [6]

In this paper, they argued that for highly structured and recurring events, like as sports, it is good way to use sophisticated techniques or methods to summarize the tweets. System also formalize the problem of summarizing event-tweets and presents a solution which is based on learning the underlying hidden state representation of the event via Hidden Markov Models. It also presents extensive experiments on real-world data. And shown that the model was significantly outperforms the some intuitive and competitive baselines [7].

This work addressed the problem of detecting or identifying new events from a collection of stream of Twitter post messages. To make event detection more feasible on web-scale, presented an algorithm which is based on locality-sensitive hashing technique which is able to overcome the limitations of old approaches, while maintaining competitive results. While retaining comparable performance, comparison with a state of- the-art system on the first story detection task shows that system achieved over order of magnitude speedup in processing time. Experiments of event detection on a collection of million Twitter posts show that news like celebrity deaths are the fast spreading news on social site Twitter [8]

In this work, they have made use of an algorithm which uses a simple frequency count method and in addition some NLP features to summarize the event specified by the user. This work of summarization algorithm also handles the so short, dissimilar, numerous, and noisy nature of tweets. This work will help users as well as researchers. [9]

III. METHODOLOGY

The following figure describes designed framework of the system. The goal of designed system is to generate summaries of different topics. The overall architecture of system is shown in figure 1.The system first fetches tweets from Twitter using Twitter API stream library.

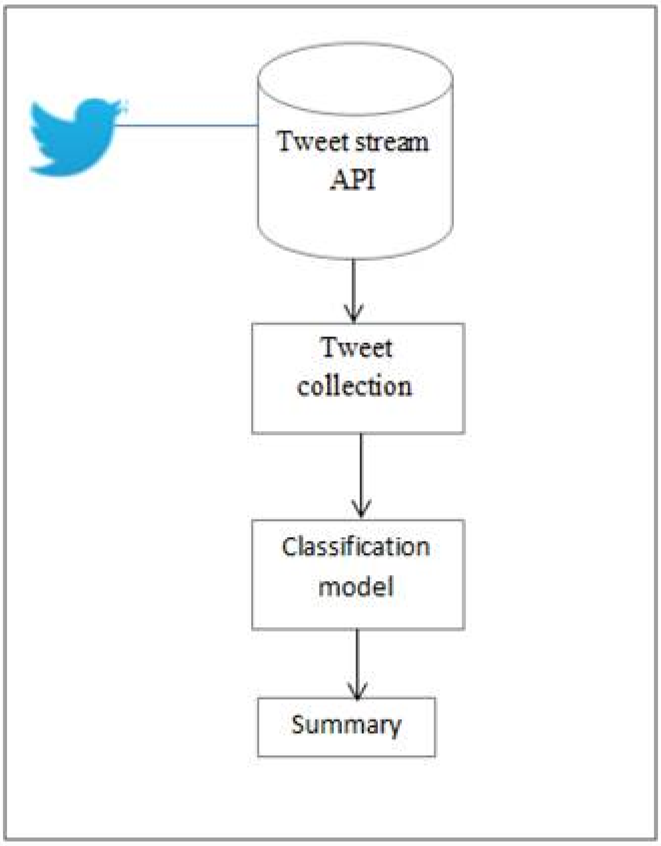

Figure 1 System Architecture

A. *TWEET COLLECTION MODULE:*

This module helps to fetch tweet stream from twitter by using twitter stream API. In this module, we used Twitter4j library for getting continuous stream of tweets. Tweet filter is used to remove links, non-English, symbols,re-tweet sentences. The filtered data is stored in a file system.

B. *CLASSIFICATION MODULE*

A Dynamic LM Classifier is a language model classifier.For the classification model Dynamic LM Classifier is used to classify tweets into general categories (sport, politics, technology, etc.) by using supervised learning method. This algorithm implements training and classification using lingepipe of Dynamic LM classifier. Training data is number of tweet files from the data set used by dynamic LM classifier for training. Training data is provided in tweet files and classified in to the general categories.

C. *ALGORITHM:*

- First initialize the classifier by using category array and n-gram size
- Training data is stored and organized into the category and it is read from the tweet files.
- After reading data, resulting data is used to train classifier for each and every category.
- Next for each category, read all testing data and execute the Dynamic LM classifier using lingepipe Andreturn the best category.
- This process is continued till the end of all the categories.

D. *TWEET SUMMARY:*

Tweet summary is the output of the system which provides summaries and contains the useful information to the user by using TF-IDF (Term Frequency-Inverse Document Frequency) method using keyword extraction. The TF-IDF identifies or assigns more weight to words that occur so many times within a data set and the least weight to terms that occurs less number of times.Summary describe what is currently discussed among public. Summary helps to understand what is happening during specific period.

IV. EXPERIMENTAL SETUP AND RESULTS

For making use of twitter data, the developer account is needed. With the help of twitter4j library system collects the tweet stream from Twitter site. Then stream of tweets is given to Dynamic LM classifier, the algorithm uses training data to classify tweets in to the specific categories for topic detection. Results of classification are in the form of general categories. Summary is actual output of the system.

For classification results, experimentation starts with collecting tweets on general topics such as sports, technology, and politics using Twitter Streaming API.Table 1 presents results sets which show the overall classifier accuracy for the classifier performance and Dynamic LM classifier can reach more than 60% of classification correctly.

Sr.no	Category	Total tweet count	Tweets used	Accuracy %
1.	Sport	300	195	65
2.	Technology	300	200	66
3.	Politics	300	190	63

Table 1: Classification results

For summarization result and evaluation, we used ROUGE toolkit which stands for Recall-Oriented Understudy for Gisting Evaluation. It determines the quality of a summary by comparing to the summaries generated by humans. The measures count the number of overlapping unit that is n-gram, word pairs, and word sequences for system generated summary and human generated summary. ROUGE-N is calculated as per given formula.

$$\text{ROUGE-N}=\frac{\text{count of matched n grams}}{\text{Number of n-grams in reference summary}}$$

For evaluating performance of summarizer, we collected the tweet stream using twitter stream API. Then filtration technique applied on tweet data set for removing non-English post, links, and symbols. From that set remaining tweets are stored in files. After this we asked to human volunteers to generate manual summaries by giving them tweet data set. Then we compared human created summary and system generated summary.

Here recall and precision are used for evaluation. In summaries the F-measure is average of precision and recall. Precision is refers to the ratio of matched tweets in reference summary and candidate summary. Recall is the ratio between the number of tweets correctly matched in candidate summary and the total number of tweets in the reference summary. It measures correctly matched tweets in the reference summary.

Document	Recall	Precision	F-measure
Cricket	0.7192982456140351	0.8282828282828283	0.7699530516431925
Football	0.7843137254901961	0.7741935483870968	0.7792207792207791
Technology	0.6129032258064516	0.5757575757575758	0.59375

Table2: Result table

Table above shows the comparison of system generated summaries with the human generated summaries. Here human generated summary taken as Reference summary and system generated summary as Candidate summary.

V. CONCLUSION AND FUTURE WORK

In this paper, we implemented Dynamic LM classifier and summarizing technique using TF-IDF method on tweet streams to provide useful summary to user with regard to topics. Summaries can help the user to get an overview of social sites quickly.The implemented work determines categories of the tweets and classifies them into different topics using Dynamic LM classifier that provides accuracy above 60% and then provides the summary of the tweets on topics which conveys information in few sentences by reducing the time for reading of the user. System eliminates short, dis-similar and noisy nature of the tweets using classification.
This work can be extended in future to provide multimedia feature like images along with summary to the end user.

REFERENCES

1. DehongGao, Wenjie Li, XiaoyanCai, Renxian Zhang and You Ouyang., “ Sequential Summarization of Twitter Trending Topics” ,IEEE/ACM Trans on Audio ,speech and language processing,vol.22,No 2,Feb 2014
2. Zhenhua Wang, LidanShouKe Chen, Gang Chen and SharadMehrotra., “On summarization And Timeline Generation for Evolutionary Tweet Streams”, IEEE Trans Knowl. Data Eng.,vol.27,No 5,May 2015
3. T. Sakaki, M. Okazaki, and Y. Matsuo., “Earthquake shakes Twitter users: Real-time event detection by social sensors”, in Proc WWW-10,2010, pp. 851–860
4. B. Sharifi, M.-A. Hutton, and J. K. Kalita, “Automatic summarization of Twitter topics”, in Proc. National Workshop Design Anal.Algorithms,2010.
5. B. Sharifi, M.-A. Hutton, and J. K. Kalita, “Experiments in microblog summarization”,in Proc. SOCIALCOM-10,2010, pp. 49–56
6. Y. Duan, Z. Chen, F. Wei, M. Zhou, and H.-Y. Shum, “Twitter topic summarization by ranking tweets using social influence and content quality”,Proc. Coling-12, pp. 763–780, 2012.
7. D. Chakrabarti and K. Punera, “Event summarization using tweets”, in Proc. AAAI-11, 2011.
8. S. Petrovic, M. Osborne, and V. Lavrenko, “Streaming first story detection with application to Twitter,” in Proc. ACL-10, 2010
9. Mr. Ganesh Mane, Mrs. Anita Kulkarni, “Twitter Event Summarization Using Phrase Reinforcement Algorithm and NLP Features”, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 5, May 2015.
10. FarisKateb,JugalKalita, “Classifying Short Text in Social Media: Twitter as CaseStudy”,Volume 111 - No. 9, February 2015

BIOGRAPHY

Rajani Dhananjay Gavali is pursuing her Master Degree of Computer Science & Engineering from Walchand Institute of Technology, Solapur, and Maharashtra, India. She has received Bachelor degree in Computer Science and Engineering from Solapur University. Her research interests include Data Mining, Summarization.

Prof. A.R.Kulkarni is an Assistant Professor in Department of Computer Science& Engineering, Walchand Institute of Technology, Solapur, Maharashtra, India.