ROAD TRAFFIC REAL TIME MONITORING THROUGH TWITTER STREAM

Pritesh Patil[1]
Department of Information Technology, AISSMS IOIT, Pune
Savitribai Phule Pune University, Pune, Maharashtra India

Abstract— Worldwide, the cities and the capitals suffer from a common problem –traffic congestion. It takes lot of time and money. The currently used tools for checking traffic are expensive and even have limitations. Since social networking sites, such as twitter, Facebook, etc. have spread in recent years becoming a new kind of real time information channels people share their information, emotions and opinions in terms of short messages. Twitter is one of the most popular social networking sites. The twitter data, i.e. tweets have a benefit over the other social networking sites that it is generally associated with some meta-data and have the 140 character constraint, thus allowing the news oriented data. This paper proposes a system that makes use of tweets to check out for traffic in real time. The system fetches tweet according to several search criteria, process them by using text mining techniques and finally classifies them using the SVM algorithm. The aim is to assign suitable class label to every tweet as related with an activity of traffic events or not. We use support vector machine as a classified model. Finally we can notify the user of presence of traffic by displaying the highlighted map and provide an alternate way.

Key words: Real-Time Monitoring, SVM, twitter stream.

I. INTRODUCTION

One of the biggest problem in modern life is the congestion due to the road traffic .Nearly all the cities from the world suffer from this problem. The time spent on these delays is far too much. To check this problem there are many tools like the sensors-cameras, radar and loops. Though these tools work well but there is a problem in using them. They require a high maintenance cost and they cover only a certain area of network Since the social networking sites like the facebook, twitter, google + have spread recently , becoming a new kind of real-time information channel. The era of digitalization that makes use of the smartphones and tablets, led to the easiness of use, and real-time nature these sites. People use these sites to express one self and to tell about some real time event.

In terms of social networking sites the messages are called as Status Update Messages (SUM). The SUM may contain, apart from the text, some additional-information like name of the user, timestamp, hashtags, and some mentions. If these SUMs are analysed properly they can serve as a good source of valuable information about an event or a topic. In fact, the SUMs are unstructured and irregular texts; they generally contain informal or abbreviated words, and misspellings. Most of the time the SUMs contain a huge amount of meaningless information, which needs to be filtered. It has been analysed that over 40% of all tweets are useless. So in order to get proper information from them we need to do the text mining over them. Text mining techniques are based has an idea that a document can be represented by the group of words contained in it [4]. At the time of text mining process, many operations are performed, depending on the goal, for example, the text filtering by means of specific keywords, and feature selection, i.e., reducing the number of features in order to consider just the relevant ones. Out of so many social networking platforms, we took into account Twitter because as nowadays it is the most popular micro-blogging service and it has a count of more than 600 million active users. Twitter even has more advantages over the other micro-blogging services. .Firstly the tweets are having a limit up to 140 characters, enhancing the real-time and news-oriented nature of the platform. Then the life-time of tweets is usually very short thus they provide the real-time events. Second, tweet can have meta-information that works as additional information. Next, Twitter messages are easily available, that is we can get a set of tweets for free of cost .Due to all the above mentioned reasons; Twitter becomes a good source of information for detection and analysis of the real-time event.

In this paper, we put up an intelligent system, based on text mining and machine learning algorithms, for real-time detection of traffic events from Twitter stream analysis. The system is built on the SOA i.e. service oriented architecture.

II. RELATED WORK

A. ET: Events from Tweets

Social media sites such as Twitter and Facebook have emerged as popular tools for people to express their opinions on various topics. The large amount of data provided by these media is extremely valuable for mining trending topics and events. In this paper, there is an efficient, scalable system to detect events from tweets (ET) [1]. It detects events by exploring their textual and temporal components. ET does not require any target entity or domain knowledge to be specified; it automatically detects events from a set of tweets.

B. Measurement and Analysis of Online Social Networks

This paper presents a large-scale measurement study and analysis of the structure of multiple online social networks. It examines data gathered from four popular online social networks: Flickr, YouTube, Live Journal, and Orkut. It traversed the publicly accessible user links on each site, obtaining a large portion of each social network's graph [2]. The results confirm the power-law, small-world, and scale free properties of online social networks. The data shows that social networks are structurally different from previously studied networks. Social networks have a much higher fraction of symmetric links and also exhibit much higher levels of local clustering.

C. Real-time Event Detection by Social Sensors

When an earthquake occurs, people make many tweet related to the earthquake, which enables detection of earthquake occurrence, simply by observing the tweets. As described in this paper, propose an algorithm to monitor tweets and to detect a target event. To detect a target event, we devise a classifier of tweets based on features such as the keywords in a tweet, the number of words, and their context [3]. We consider each Twitter user as a sensor and filtering. System detects earthquakes promptly and sends e-mails to registered users. Notification is delivered much faster than the announcements that are broadcast other news channels.

III. SYSTEM DESIGN

A. Fetch of SUMs and Pre-Processing

The first module, Fetch of SUMs and Pre-processing, extracts raw tweets from the Twitter stream, based on one or more search criteria (e.g., geographic coordinates, keywords appearing in the text of the tweet). Each fetched raw tweet contains: the user id, the timestamp, the geographic coordinates, a re-tweet flag, and the text of the tweet. The text may contain additional information, such as hashtags, links, mentions, and special characters. After the SUMs have been fetched according to the specific search criteria, SUMs are pre-processed. In order to extract only the text of each raw tweet and remove all meta-information associated with it; a Regular Expression filter is applied.

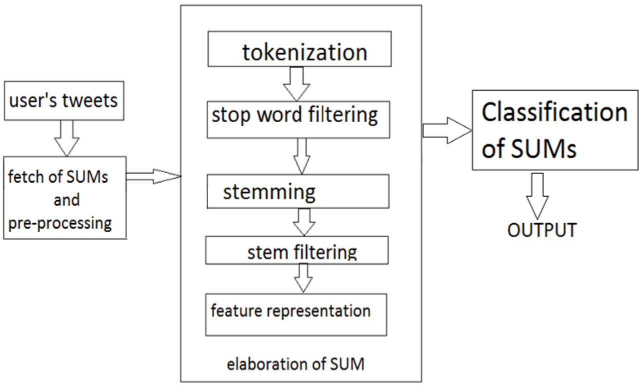

Fig. 1: System design

B. Elaboration of Sums

As we can see in the system design (Fig 1), the second processing module, Elaboration of SUMs, is devoted to transforming the set of pre-processed SUMs [5], i.e., a set of strings, in a set of numeric vectors to be elaborated by the Classification of SUMs module. For this text mining techniques are applied in sequence to the pre-processed SUMs. The text mining steps that are performed in this module are:

a) Tokenization is typically the first step of the text mining process, and consists in transforming a stream of characters into a stream of processing units called tokens e.g., syllables, words, or phrases. The tokenizer removes all punctuation marks and splits each SUM into tokens corresponding to words (bag-of-words representation). At the end of this step, each SUM is represented as the sequence of words contained in it.

b) Stop-word filtering consists in eliminating stop-words, i.e., words which provide little or no information to the text analysis. Common stop-words are articles, conjunctions, prepositions, pronouns, etc. Other stop- words are those having no statistical significance, that is, those that typically appear very often in sentences of the considered language (language-specific stop-words), or in the set of texts being analyzed (domain-specific stop- words), and can therefore be considered as noise.

c) Stemming is the process of reducing each word (i.e., token) to its stem or root form, by removing its suffix. The purpose of this step is to group words with the same theme having closely related semantics.

d) Stem filtering consists in reducing the number of stems of each SUM. In particular, each SUM is filtered by removing from the set of stems the ones not belonging to the set of relevant stems.

e) Feature representation consists in building, for each SUM, the corresponding vector of numeric features. Indeed, in order to classify the SUMs, we have to represent them in the same feature space.

C. Classification of SUMs

The third module, Classification of SUMs, assigns to each elaborated tweet a class label related to traffic events. Thus, the output of this module is a collection of N labeled tweets. The parameters of the classification model have been identified during the supervised learning stage. The classifier that achieved the most accurate results was finally employed for the real time monitoring with the proposed traffic detection system. The system continuously monitors a specific region and notifies the presence of a traffic event on the basis of a set of rules that can be defined by the system administrator. For this classification we are making use of the SVM i.e. the support vector machine algorithm.

1) SVM

SVM is a supervised learning model that analyzes data that can be used for classification and regression analysis and then over them builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. The Twitter centric features are based on the frequency of multiword hashtags with special capitalization
(e.g., #TooMuchTraffic). Because the clusters constantly
evolve over time, the features are periodically updated for old clusters and computed for newly formed ones. Finally, a support vector machine (SVM) classifier is trained on a labelled set of cluster features and used to decide whether the cluster (and its associated messages) contains real-world event information.SVM classifier filters the “noisy” tweets that are not related to road traffic events.

IV. CONCLUSION

The wide spread of the social networking sites has led to people using them to express their views and opinions about some event around them .in case of an emergency ,people need to know about the traffic on the way on which they are about to go. For example, during rallies, people who want to go fast can lookout for the alternate path. The tweet can be used as social sensor. They can help in locating traffic in a given region. The data mining techniques and the classification technique are used to identify the tweets. These tweets are related to the traffic and the location as requested by the user. At the end the users are provided with a map highlighting the route that they want.

References

[1]. Amina Madani, Omar Boussaid, Djamel Eddine Zegour and Algiers, Algeria, “What’s Happening: A Survey of Tweets Event Detection,” J. Internet Services Appl.,vol. 1, no. 1, pp. 7–18, 2010.

[2]. Parikh, Kamalakar Karlapalem., “ET: Events from Tweets,” Future Generat.Comput. Syst., vol. 25, no. 6, pp 599–616, Jun. 2009.

[3]. Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, Bobby Bhattacharjee., “Measurement and Analysis of Online Social Networks,” Proc. IEEE vol. 99, no. 1, pp. 149–167, Jan.
2011

[4]. M.W. Berry and M. Castellanos, “Survey of Text Mining” New York, NY, USA: Springer-Verlag, 2004.

[5]. Eleonora D’Andrea, Pietro Ducange, Beatrice Lazzerini, “Real Time Detection Of Traffic From Twitter Stream Analysis ”, IEEE.vol 11 , no. 5 , Jan 2016.