International Journal of Research in Science advanced Technology and management studies (IJRSTMS)

ISSN: 2459-425X • Website: www.ijrstms.com

Detection of Real-Time Traffic through Twitter Stream System

Dr Sai

1Siddhant College of Engineering,Savitribai Phule Pune University,Pune (Maharashtra), India.
2Adarsh Institute of Technology, Vita (Maharashtra),India
maheshbhosale65@gmail.com, ashishvankudre101@gmail.com

Abstract

Now day's social media networks have been recently used as a source of information for event detection with specific reference to road traffic activity occurrences and accidents or earthquake reporting system. Our paper, presents a real-time detection of traffic through Twitter Stream Analysis. Our proposed tool fetches tweets from twitter as per several searching criteria; processes these tweets by applying text mining methods and lastly evaluates classification of tweets. The goal is to give class label to each tweet updated by twitter, as related with a traffic or non-traffic. Our system performs real-time monitoring of street network and detection of traffic. We occupy our support vector machine as a classifier; furthermore, we accomplished an accuracy value of 90.75% by confront a binary classification issue (traffic versus nontraffic tweets). As well as we able to differentiate if traffic is caused by an external event or not, by solving a multiclass classification problem and obtaining accuracy value of 80.89%.

Keywords- twitter, twitter stream analysis, traffic event detection, tweet classification, text mining, social sensing.

I. INTRODUCTION

Twitter is good social communicator platform as well as bad tweets containing URLs for spam, phishing and malware distribution. Common twitter spam detection methods use account features such as the ratio of tweets consists of URLs and the account creation date or relation features in the twitter graph. These detection methods are in effective against feature fabrications also lexical features of URLs, URL redirection, HTML content, and dynamic behavior and consume much time and resources. Conventional suspicious URL detection schemes utilize number of features including lexical features of URLs, HTML content, URL redirection, and dynamic behavior. However, evading techniques such as time-based evasion and crawler evasion exist.

In our paper, we established a smart system, based on text computing and machine learning algorithms, for real-time traffic detection through Twitter stream analysis. After market study our system, has been designed and manipulated from the primary stage as an event-driven structure, built on a Service Oriented Architecture (SOA). The system exploits available technologies based on state-of-the-art techniques for text analysis and pattern classification. These technologies and techniques have been analyzed, tuned, adapted, and integrated to build an intelligent system.

We present a real time practical study, which has been performed for determining the most effective among different state-of-the-art approaches for text classification. The chosen approach was integrated into the final system and used for the on-the-field real-time base of traffic events. Many Twitter spam detection schemes have been proposed to fight with malicious tweets. These schemes can be partitioned into account feature-based, relation feature-based, and message feature based schemes. Account feature-based schemes use the distinguishing features of spam accounts such as the ratio of tweets containing URLs, the account creation date, and the number of followers and friends. Also, external users can easily fabricate these account features. The relation feature-based schemes focused on more complicated features that malicious users cannot easily fabricate such as the distance and connectivity apparent in the Twitter graph. Obtaining these relation features from a Twitter graph, however, requires a significant amount of time and resources as a Twitter graph is tremendous in size. Lastly the message feature-based scheme focused on the features of messages as word indicates. However, spammers can easily change the shape of their messages. A number of doubtful URL detection schemes have also been introduced. With reference to current approaches for using social media to extract useful information for event detection, we need to distinguish between small-scale events and large-scale events. Small-scale events (e.g., car crashes, traffic, local manifestations or fires) usually have a small number of SUMs related to them, belong to a precise geographic location, and are concentrated in a small time interval. Large scale events (e.g., earthquakes, tornados, or the election of a president) are characterized by a huge number of SUMs, and by a wider temporal and geographic coverage. So as per observation, due to less number of SUMs related to small-scale events, small-scale event detection is a non-trivial task. Several works in the literature deal with event detection from social networks. Many works deal with large-scale event detection, and only a few works focus on small-scale event. Detection of fires in a factory from Twitter stream analysis by using NLP methods and NB (Naive Bayes) classifier about small-scale event detection. In this project, we do focus on a these particular small-scale events, i.e., road traffic and we aim to detect and analyze traffic events by processing users' SUMs according to a certain area using Twitter as a source. With this aim, we propose a system which able to fetch, manipulate, and classify SUMs as related to a road traffic event or not.

In this project, we target on a particular small-scale event, i.e. road traffic and we aim to detect and analyze traffic events by processing users' SUMs belonging to a certain area and written in the Italian language. To this aim, we propose a system able to fetch, elaborate, and classify SUMs as related to a road traffic event or not.

II. PROPOSED SYSTEM

We focus on a particular small-scale event, i.e. Road traffic and we aim to detect and analyze traffic events by processing users' SUMs belonging to a certain area using Twitter as a source. So we propose a system which approach both multi-class classification and binary problems. As regards binary classification, we consider traffic-related tweets, and tweets not related with traffic. We use Multi-class classification, to split traffic-related class into two classes, namely traffic congestion or crash and traffic due to external event i.e. large scale event. We use hash tag principal in our system for this classification.

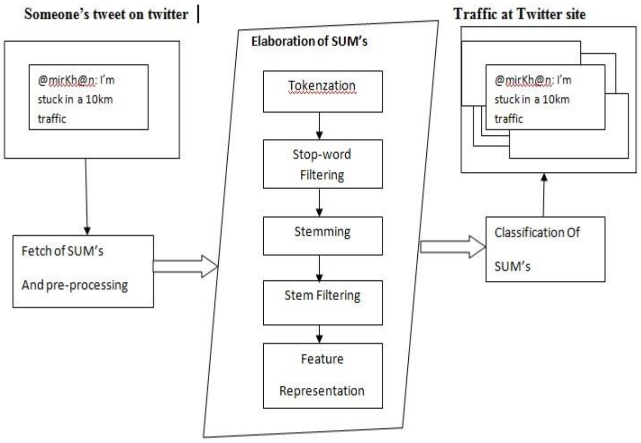

Fig. 1. System architecture for traffic detection from Twitter stream analysis

III. ALGORITHM

1) Fetch of SUMs and Pre-Processing: The first module, "Fetch of SUMs and Pre-processing", essence 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). So, each fetched raw tweet contains: the user id, the timestamp, the geographic coordinates, retweet flag, and the text of the tweet. The text contains additional information, such as hash tag's, links, mentions, and special characters. After SUMs have been fetched then these SUMs are pre-processed. In order to extract only text of each raw tweet remove all meta-information associated with it; a Regular Expression filter is applied.

2) Elaboration of SUMs: The second phase is, "Elaboration of SUMs", is devoted to transforming the set of pre-processed SUMs, i.e. set of strings, in a set of numeric vectors are elaborated by the "Classification of SUMs". Furthermore some text mining techniques are applied in sequence to the pre-processed SUMs. Some text mining steps performed in this module are described in detail below: 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 deletes all punctuation marks and divides each SUM into tokens corresponding to words. At the end of this step, each SUM is represented as the sequence of words contained in it. b) Stop-word filtering consists 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. This step is used to group words with the same theme having closely related semantics. d) Stem filtering consists in removing 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. 3) Classification of SUMs: The third module, "Classification of SUMs", assigns to each elaborated SUM a class label related to traffic events. Thus, the output of this module is a collection of N labeled SUMs. With the aim of labeling each SUM, a classification model is employed. 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 of 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 example, when the first tweet is recognized as a traffic-related tweet, the system may send a warning signal. Then the actual information of the traffic event may be sent after the identification of a several number of tweets with the same label.

IV. CONCLUSION

In this paper, we have proposed a system for real-time detection of traffic-related events through Twitter stream analysis. The system, made on a SOA, is able to fetch and classify streams of tweets and to notify the users of the presence of traffic events. Furthermore, the system is also able to differentiate if a traffic event is due to an external cause, such as football match, procession and manifestation, or not.

REFERENCES

a) *B. Chen and H. H. Cheng, "A review of the applications of agent technologyin traffic and transportation systems," IEEE Trans. Intell. Transp. Syst., vol. 11, no. 2, pp. 485–497, Jun. 2010.*

b) *F. Atefeh and W. Khreich, "A survey of techniques for event detection in*
c) *Twitter," Comput. Intell., vol. 31, no. 1, pp. 132–164, 2015.*

d) *Gonzalez, L. M. Bergasa, and J. J. Yebes, "Text detection and recognition*
e) *on traffic panels from street-level imagery using visual appearance," IEEE Trans. Intell. Transp. Syst., vol. 15, no. 1, pp. 228–238, Feb. 2014.*

f) *Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in Proc. 7th ACM SIGCOMM Conf. Internet Meas., San Diego, CA, USA, 2007, pp. 29–42.*

g) *P. Ruchi and K. Kamalakar, "ET: Events from tweets," in Proc. 22nd Int. Conf. World Wide Web Comput., Rio de Janeiro, Brazil, 2013, pp. 613–620.*

h) *T. Sakaki, M. Okazaki, and Y.Matsuo, "Tweet analysis for real-time event detection and earthquake reporting system development," IEEE Trans. Knowl. Data Eng., vol. 25, no. 4, pp. 919–931, Apr. 2013.*

i) *T. Sakaki, Y. Matsuo, T. Yanagihara, N. P. Chandrasiri, and K. Nawa, Real-time event extraction for driving information from social sensors," in Proc. IEEE Int. Conf. CYBER, Bangkok, Thailand, 2012, pp. 221–226.*