International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Detection of Real-Time Traffic through Twitter Stream Analysis

Mahesh V. Bhosale[1], Ashish Adinath Vankudre[2]
[1]Department of Computer Engineering, Siddhant College of Engineering, Savitribai Phule Pune University, Pune, India
[2]Assistant Professor, Department of CSE, Adarsh Institute of Technology, Vita (Maharashtra), India

Abstract— *Social networks have been recently emppIoyed as a source of information for event detection, with specific reference to road traffic activity congestion and accidents or earthquack reporting system. In our paper, we present a real-time detection of traffic from Twitter stream analysis. The system fetches tweets from Twitter as per a several search criteria; process tweets by applying text mining methods; lastly performs the classification of tweets. The aim is to assign suitable class label to every tweet, as related with an activity of traffic event or not. The traffic detection system or framework was utilized for real-time monitoring of several areas of the Italian street network, taking into consideration detection of traffic events just almost in real time, regularly before online traffic news sites. We employed the support vector machine as a classification model, furthermore, we accomplished an accuracy value of 90.75% by tackling a binar classification issue (traffic versus nontraffic tweets). We were also able to discriminate 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 malicious tweets containing URLs for spam, phishing, and malware distribution. Conventional Twitter spam detection schemes utilize account of features such as the ratio of tweets containing URLs and the account creation date, or relation features in the Twitter graph. These detection schemes are in effective against feature fabrications or consume much time and resources. Conventional suspicious URL detection schemes utilize several features including lexical features of URLs, URL redirection, HTML content, and dynamic behavior. However, evading techniques such as time-based evasion and crawler evasion exist.

In our paper, we propose an intelligent system, based on text mining and machine learning algorithms, for real-time detection of traffic events from Twitter stream analysis. The system, after a feasibility study, has been designed and developed from the ground as an event-driven infrastructure, 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 in order to build the intelligent system. In particular, we present an experimental 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 detection of traffic events.

In the existing system attackers use shortened malicious URLs that redirect Twitter users to external attack servers. To cope with malicious tweets, several Twitter spam detection schemes have been proposed. These schemes can be classified 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. However, malicious users can easily fabricate these account features. The relation feature-based schemes rely on more robust features that malicious users cannot easily fabricate such as the distance and connectivity apparent in the Twitter graph. Extracting these relation features from a Twitter graph, however, requires a significant amount of time and resources as a Twitter graph is tremendous in size. The message feature-based scheme focused on the lexical features of messages. However, spammers can easily change the shape of their messages. A number of suspicious 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., traffic, car crashes, fires, or local manifestations) usually have a small number of SUMs related to them, belong to a precise geographic location, and are concentrated in a small time interval. On the other hand, 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. Consequently, due to the smaller 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. Regarding small-scale event detection, the detection of fires in a factory from Twitter stream analysis, by using standard NLP techniques and a Naive Bayes (NB) classifier.

In this project, 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 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 and written in the Italian language. So that we propose a system which will able to fetch, elaborate, and classify SUMs as related to a road traffic event or not. The proposed system may approach both binary and multi-class classification problems. As regards binary classification, we consider traffic-related tweets, and tweets not related with traffic. We use Multi-class classification, to split the traffic-related class into two classes, namely traffic congestion or crash, and traffic due to external event. For this classification we use hash tag principal in our system.

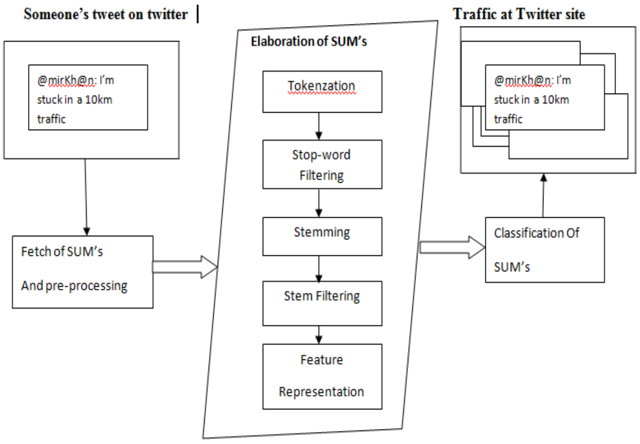

Fig. 1. System architecture for traffic detection from Twitter stream analysis.

III. Algorithm

1) Fetch of SUMs and Pre-Processing

The first module, "Fetch of SUMs and Pre-processing", 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 twee t). Each fetched raw tweet contains: the user id, the timestamp, the geographic coordinates, a retweet 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.

2) Elaboration of SUMs

The second processing module, "Elaboration of SUMs", is devoted to transforming the set of pre-processed SUMs, i.e., a set of strings, in a set of numeric vectors to be elaborated by the "Classification of SUMs" module. To this aim, some text mining techniques are applied in sequence to the pre-processed SUMs. In the following, the text mining steps performed in this module are described in detail:

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 SUMj 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.

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. To 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 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 example, when the first tweet is recognized as a traffic-related tweet, the system may send a warning signal. Then, the actual notification of the traffic event may be sent after the identification of a certain 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 from Twitter stream analysis. The system, built 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 discriminate if a traffic event is due to an external cause, such as football match, procession and manifestation, or not.

REFERENCES

[1] F. Atefeh and W. Khreich, “A survey of techniques for event detection in Twitter,” *Comput. Intell.*, vol. 31, no. 1, pp. 132–164, 2015.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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