TWITTER STREAM ANALYSIS FOR TRAFFIC DETECTION AND EARTHQUAKE

PAVAN.S.SONUNE[1] , SAMEER K.SHAIKH[2], RUSHIKESH.L.SONAVANE[3], PROF R.S.SHISHUPAL[4]

[1]Department Of Computer Engineering, Sinhgad Institute Of Technology ,Lonavala,
[2]Department Of Computer Engineering, Sinhgad Institute Of Technology ,Lonavala,
[3]Department Of Computer Engineering, Sinhgad Institute Of Technology ,Lonavala,
[4]Department Of Computer Engineering, Sinhgad Institute Of Technology ,Lonavala,

Pavan.sonune03@gmail.com,Rss.sit@sinhgad.edu, rushikesh.sonavane@gmail.com, Sameershaikh677@gmail.com

ABSTRACT — *Social networks have been recently empployed 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 monitoring system for traffic occasion detection from Twitter stream analysis. The system fetches tweets from Twitter as per a several search criteria; procedures 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 ofseveral areas of the Italian stree 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 classificatio model, furthermore, we accomplished an accuracy value of 95.75% 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 88.89%.*

Keywords- Traffic Event Detection, Tweet Classification, Text Mining, Social Sensing.

I. INTRODUCTION

Twitter is prone to 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[1][2]. These detection schemes are ineffective 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[3].

In this 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)[4]. 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[5][6][7]. Small-scale events[8][9][10][11]-[13] (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. LITERATURE REVIEW

1) What's Happening: A Survey of Tweets Event Detection

Author: Amina Madani, Omar Boussaid, Djamel Eddine Zegour and Algiers, Algeria

Twitter is now one of the main means for spread of ideas and information throughout the Web. Tweets discuss different trends, ideas, events, and so on. This gave rise to an increasing interest in analyzing tweets by the data mining community. Twitter is, in nature, a good resource for detecting events in real-time. In this survey paper, authors have presented four challenges of tweets event detection: health epidemics identification, natural events detection, trending topics detection, and sentiment analysis. These challenges are based mainly on clustering and classification. We review these approaches by providing a description of each one.

These last years have been marked by the emergence of microblogs. Their rates of activity reached some levels without precedent. Hundreds of millions of users are registered in these microblogs as Twitter. They exchange and tell their last thoughts, moods or activities by tweets in some words.

2) ET: Events from Tweets

Author: Ruchi Parikh, Kamalakar Karlapalem.

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, we build an efficient, scalable system to detect events from tweets (ET). Our approach 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.
The key components of ET are:
(1) an extraction scheme for event representative keywords
(2) an efficient storage mechanism to store their appearance pat- terns, and
(3) a hierarchical clustering technique based on the common co-occurring features of keywords.
Authors presented a scalable and efficient system, called ET, to detect real world events from a set of microblogs/tweets. The key feature of this system is the efficient use of con- tent similarity and appearance similarity among keywords, to cluster the related keywords. We demonstrate the effectiveness of this combination in our experiments. ET does not need any human expertise or knowledge from other sources like Wikipedia, and still provides very accurate results. ET is evaluated on two different datasets from two different domains and it yields great results for both of them in terms of the precision.

3) Measurement and Analysis of Online Social Networks

Authors: Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, Bobby Bhattacharjee.

Online social networking sites like Orkut, YouTube, and Flickr are among the most popular sites on the Internet. Users of these sites form a social network, which provides a powerful means of sharing, organizing, and finding content and contacts. The popularity of these sites provides an opportunity to study the characteristics of online social network graphs at large scale. Understanding these graphs is important, both to improve current systems and to design new applications of online social networks.

This paper presents a large-scale measurement study and analysis of the structure of multiple online social networks. We examine data gathered from four popular online social networks: Flickr, YouTube, Live Journal, and Orkut. We crawled the publicly accessible user links on each site, obtaining a large portion of each social network's graph. Our data set contains over 11.3 million users and 328 million links. We believe that this is the first study to examine multiple online social networks at scale. Our results confirm the power-law, small-world, and scale free properties of online social networks. We observe that the in degree of user nodes tends to match the out degree; that the networks contain a densely connected core of high-degree nodes; and that this core links

small groups of strongly clustered, low-degree nodes at the fringes of the network. Finally, the implications of these structural properties for the design of social network based systems.
Presented an analysis of the structural properties of online social networks using data sets collected from four popular sites. Our data shows that social networks are structurally different from previously studied networks, in particular the Web. Social networks have a much higher fraction of symmetric links and also exhibit much higher levels of local clustering. We have outlined how these properties may affect algorithms and applications designed for social networks.

4) Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors
Authors: Takeshi Sakaki, Makoto Okazaki, Yutaka Matsuo
Twitter, a popular micro blogging service, has received much attention recently. An important characteristic of Twitter is its real-time nature. For example, when an earthquake occurs, people make many Twitter posts (tweets) related to the earthquake, which enables detection of earthquake occurrence promptly, simply by observing the tweets. As described in this paper, we investigate the real-time interaction of events such as earthquakes, in Twitter, and 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. Subsequently, we produce a probabilistic spatiotemporal model for the target event that can find the center and the trajectory of the event location. We consider each Twitter user as a sensor and apply Kalman filtering and particle filtering, which are widely used for location estimation in ubiquitous/pervasive computing. The particle filter works better than other compared methods in estimating the centers of earthquakes and the trajectories of typhoons. As an application, we construct an earthquake reporting system in Japan. Because of the numerous earthquakes and the large number of Twitter users throughout the country, we can detect an earthquake by monitoring tweets with high probability (96% of earthquakes of Japan Meteorological Agency (JMA) seismic intensity scale 3 or more are detected). Our system detects earthquakes promptly and sends e-mails to registered users. Notification is delivered much faster than the announcements that are broadcast by the JMA.

5) Text Detection and Recognition on Traffic Panels From Street-Level Imagery Using Visual Appearance
Authors: Álvaro González, Luis M. Bergasa.
Traffic sign detection and recognition has been thoroughly studied for a long time. However, traffic panel detection and recognition still remains a challenge in computer vision due to its different types and the huge variability of the information depicted in them. This paper presents a method to detect traffic panels in street-level images and to recognize the information contained on them, as an application to intelligent transportation systems (ITS). The main purpose can be to make an automatic inventory of the traffic panels located in a road to support road maintenance and to assist drivers. Our proposal extracts local descriptors at some interest key points after applying blue and white color segmentation. Then, images are represented as a "bag of visual words" and classified using Naïve Bayes or support vector machines. This visual appearance categorization method is a new approach for traffic panel detection in the state of the art. Finally, our own text detection and recognition method is applied on those images where a traffic panel has been detected, in order to automatically read and save the information depicted in the panels. We propose a language model partly based on a dynamic dictionary for a limited geographical area using a reverse geo coding service. Experimental results on real images from Google Street View prove the efficiency of the proposed method and give way to using street-level images for different applications on ITS.

III. SURVEY OF PROPOSED SYSTEM

In our paper, we present a real-time monitoring system for traffic occasion detection from Twitter stream analysis. The system fetches tweets from Twitter as per a several search criteria; procedures 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 ofseveral areas of the Italian stree 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 classificatio model, furthermore, we accomplished an accuracy value of 95.75% 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 88.89%.

IV. MATHEMATICAL MODEL

Let S is the Whole System Consists:
S= {I, P, O}

I = Input.
P= Procedure.
O= Output.
I = {U, T, TS, url, Tk}.

1. Let U is set of number of twitter users in the system.
   U = {u1, u2,.....un}.
2. T is set of number Twitt or status update of twitter user.
   T= {t1, t2, t3…tn}.
3. TS is twitter streamer who analyzes the twits.
4. url is the URL of twitter user who have updated status.
5. Tk is the tokenization of SUM where, SUM is te Status Update Message of twitter user.

P = Procedure.
Step 1: The twitter streamer will collect all the urls from the SUM by users.

SU={ $t_{j1}^{T}, \ldots, t_{jh}^{T}, \ldots, t$}.

Where is the h[th] token and is the total number of tokens in SU.

Step 2: Filtering: In this step we perform tokenization of SUM and filtered the tokens and ignoring small meaning that is the word which don't have any information which is known as stop-word filtering.

Each SUM is reduced to a sequence of relevant tokens. We denote the j[th] stop-word filtered SUM as,

SUM = { $t_{j1}^{SW}, \ldots t_{jk}^{SW}, \ldots t$}.

Where t the k[th] relevant token and K_j, is with $K_j \leq H_j$, is the total number of relevant tokens in SUM

Step 3: Assigning labels to filtered Tokens:
In this step, system assigns a class label to each SUM related to traffic events. So at last there is collection of N labeled SUMs.
Step 4: In this the classifier that achieved the most accurate results by filtered tokens with labels was finally employed for the realtime monitoring with the proposed traffic detection system.
O = Output:
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.

V. SYSTEM ARCHITECTURE

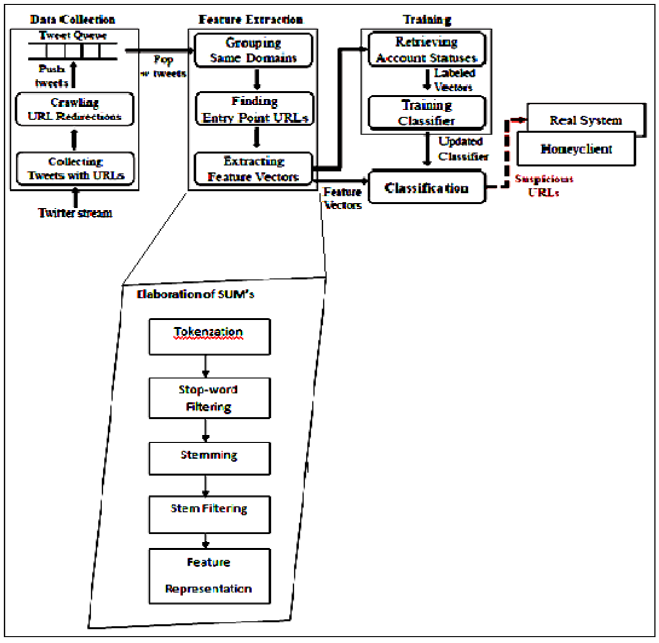

Fig.5.System Architecture

.

VI. CONCLUSION AND FUTURE WORK

Conclusion

1. In this paper, we have proposed a system for real-time detection of traffic-related events from Twitter stream analysis.

2. 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.

3. 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.

Future Scope:

As future work, we are planning to integrate our system with an application for analyzing the official traffic news web sites, so as to capture traffic condition notifications in real-time. Thus, our system will be able to signal traffic-related events in the worst case at the same time of the notifications on the web sites. Further, we are investigating the integration of our system into a more complex traffic detection infrastructure. This infrastructure may include both advanced physical sensors and social sensors such as streams of tweets. In particular, social
sensors may provide a low-cost wide coverage of the road network, especially in those areas (e.g., urban and suburban) where traditional traffic sensors are missing.

ACKNOWLEDGMENT

We might want to thank the analysts and also distributers for making their assets accessible. We additionally appreciative to commentator for their significant recommendations furthermore thank the school powers for giving the obliged base and backing.

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] A. 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] The Smarty project. [Online]. Available: http://www.smarty.toscana.it/
[5] 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.
[6] M. Krstajic, C. Rohrdantz, M. Hund, and A. Weiler, "Getting there first: Real-time detection of real-world incidents on Twitter" in Proc. 2nd IEEE Work Interactive Vis. Text Anal.—Task-Driven Anal. Soc. Media IEEE VisWeek," Seattle, WA, USA, 2012.
[7] J. Yin, A. Lampert, M. Cameron, B. Robinson, and R. Power, "Using social media to enhance emergency situation awareness," IEEE Intell. Syst., vol. 27, no. 6, pp. 52–59, Nov./Dec. 2012.
[8] 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.
[9] N. Wanichayapong, W. Pruthipunyaskul, W. Pattara-Atikom, and P. Chaovalit, "Social-based traffic information extraction and classification," in Proc. 11th Int. Conf. ITST, St. Petersburg, Russia, 2011, pp. 107–112.
[10] A. Schulz, P. Ristoski, and H. Paulheim, "I see a car crash: Real-time detection of small scale incidents in microblogs," in The Semantic Web: ESWC 2013 Satellite Events, vol. 7955. Berlin, Germany: Springer-Verlag, 2013, pp. 22–33.
[11] P. Agarwal, R. Vaithiyanathan, S. Sharma, and G. Shro, "Catching the long-tail: Extracting local news events from Twitter," in Proc. 6th AAAI ICWSM, Dublin, Ireland, Jun. 2012, pp. 379–382.