International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 04 -2016 www.irjet.net p-ISSN: 2395-0072

Real-Time Detection of Traffic From Twitter Stream Analysis

Sweety Kumari[1], Firdos Khan[2], Shekh Sultan[3], Ruchita khandge[4]

[1234] *B.E Student, Computer Engineering, Dr.D.Y.Patil Institute of engineering and technology, Maharashtra, India*

Abstract - ***Internet sites are source of info for event detection, with specific mention of the road traffic activity blockage and accidents or earth-quack sensing system. In this paper, we present a real-time monitoring system intended for traffic occasion detection coming from Twitter stream analysis. The system fetches tweets coming from Twitter as per a several search criteria; methods tweets, by applying textual content mining methods; last but not least works the classification of twitter posts. The goal is to assign suitable class packaging to every tweet, because related with an activity of traffic event or perhaps not. The traffic recognition system or framework was utilized for real-time monitoring of various areas of the street network, taking into account detection of traffic occasions just almost in actual time, regularly before on-line traffic news sites. All of us employed the support vector machine like a classification unit, furthermore, we accomplished a great accuracy value of ninety five. 75% by attempting a binary classification issue. All of us were also capable to discriminate if traffic is triggered by an external celebration or not, by resolving a multiclass classification issue and obtaining accuracy worth of 88. 89%.***

Key Words: *Social media; Traffic detection; Text mining; Privacy; Service Oriented Architecture (SOA), machine learning, Twitter stream analysis*

1. INTRODUCTION

Social media platforms are widely used for distributed information about the detection of events, such as traffic blocking, incidents, natural disasters (earthquakes, storms, fires, etc.), or other events. An *event* is defined as a real-world existence that happens in adefinite time and space [1], [7]. Generally traffic related events, people frequently share by means of an SUM information about the current traffic situation around them while driving. For this purpose, event detection from social networksis also often employed with Intelligent Transportation Systems(ITSs). ITSs afford, e.g., real-time information about weather, traffic congestion or regulation, or plan efficient (e.g., shortest, fast driving, least polluting) routes[4], [6], [8]. Event detection from social networks investigation is a more stimulating problem than event detection from traditional broadcasting like blogs, emails, etc. In fact, SUMs are unstructured and unequal texts, it holds informal or shortened words, mistakes or grammatical errors [1]. SUMs contain a huge amount of not useful or wordless information, which has to be clarified. According to Pear Analytics, it has been estimated that over 40% of all Twitter2 SUMs (i.e., *tweets*) is senseless with no useful data for the audience. For all of these reasons, in order to analyze the data coming from social networks or text mining techniques, We use to extract important data [18],of data mining, device learning, numbers, and Natural Language Processing (NLP).In this paper, we propose a brilliant system, based upon text mining and equipment learning algorithms, for current detection of traffic occasions from Twitter stream evaluation. The system, after having a feasibility study, offers been designed and created from the ground because an event-driven infrastructure, constructed on a Service Focused Architecture (SOA). The program exploits available technologies {centred |structured |established} on state-of-the-art processes for {textual content |text message} analysis and pattern {category |distinction}.These technologies and {methods |approaches |tactics} have been analysed, {fine-tined| configured |calibrated}, adapted, and integrated to be able to build the intelligent system. In particular, we present a great experimental study, which {offers |provides |features} been performed for {deciding} the most effective {amongst |between} different state-of-the-art approaches {intended for |to get pertaining to} text classification. The {selected picked} approach was integrated {in to | in} {the last | the ultimate| a final} system and {utilized |applied| employed} for the on-the-field {current | timely} detection of traffic {occasions| situations |incidents}.

2. OBJECTIVE

The vital objectives of proposed system is as follows:-
Design a real-time detection system for traffic analysis.
The aim is to assign suitable class label to every tweet, as related with an activity of traffic event
or not.
It performs a multi-class classification, which recognizes non-traffic, traffic due to congestion or crash, and traffic due to external events.
It detects the traffic events in real-time and
It is developed as an event-ambitious infrastructure, built on an SOA architecture

2.1 Related Work

Text mining means to the process of automatic extraction

of important information and knowledge from unstructured text. Text mining is a difference on a field called data mining [2], that tries to find interesting pattern from large databases. Text mining, also known as Rational Text Exploration, Text Data Mining or Knowledge-Discovery in Text (KDT), refers mostly to the process of extracting interesting and non-trivial information and knowledge from unstructured text. As most data (over 80%) is stored as text, text mining is supposed to have a high commercial potential value. Knowledge may be exposed from many sources of information, yet, formless texts remain the largest readily existing source of knowledge.Most of text mining methods are based on the idea that a document can be faithfully represented by the set of wordscontained in it (*bag-of-words* representation [2]).Data mining and device learning algorithms (i.e., support vector machines (SVMs), decision trees, neural networks, etc.) are applied to the documents in the vector space representation, to build classification,clustering or regression models.In this way we aim to support traffic and city administrations for managing scheduled or unexpected events in the city.

3. Existing System

In the existing system enemies use shortened malicious URLs that transmit Twitter users to external attack servers. To deal with malicious tweets, several Twitter spam revealing schemes have been proposed. These systems can be classified into explanation feature-based, relative feature-based, and message feature based schemes. Account feature-based systems use the individual features of spam accounts such as the ratio of tweets containing URLs, the account making date, and the number of groups and friends. However, malicious users can easily fabricate these account structures. The relation feature-based schemes depend onmore strong structures that harmful users cannot easily create such as the distance and connectivity obvious in the Twitter graph. Eliminating these relation structures from a Twitter graph, however, needs a significant amount of time and properties as a Twitter graph is excellent in size. The message feature-based system focused on the lexical structures of messages. However, spammers can easily modify the shape of their messages. A number of suspicious URL detection systems have also been introduced.

4. Proposed Framework

We propose intellectual scheme, based on text mining and machine learning algorithms, for real-time finding of traffic events from Twitter stream analysis. The scheme, after a feasibility study, has been designed and developed from the ground as an event-driven substructure, built on a Service Oriented Architecture (SOA). The system activities present technologies based on state-of-the-art systems for text analysis and pattern classification. These technologies and systems have been evaluated, tuned, adapted, and integrated in order to build the intelligent system. In particular, we present an investigational study, which has been performed for determining the most effective among different state-of-the-art methods for text classification. The chosen approach was included into the final scheme and used for the on-the-field real-time detection of traffic events.

4.2 Advantages of proposed system:

- It completes a multi-class classification, which identifies non-traffic, traffic due to congestion or crash, and traffic due to external events.
- It senses the traffic events in real-time and
- It is developed as an event-driven structure, built on an SOA architecture

5. System Architecture

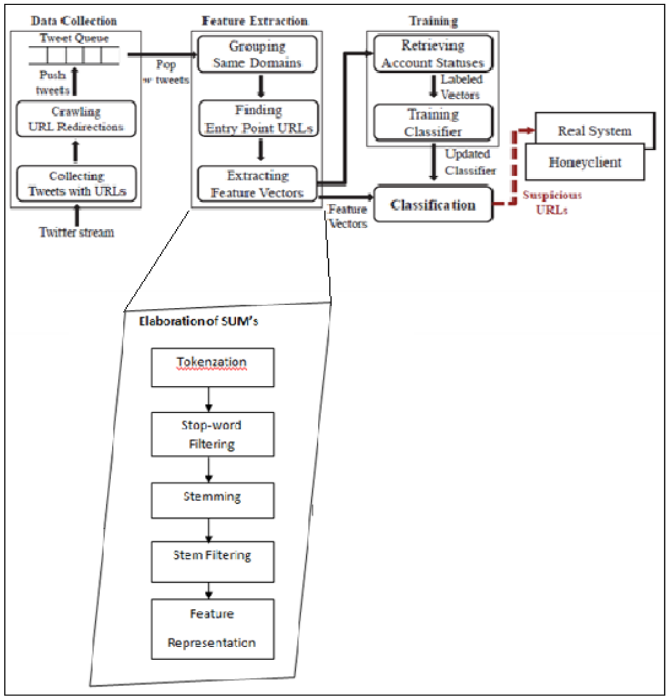

Fig 1: System overview

1) Fetch of SUMs and Pre-Processing:

The first module, Fetch of SUMs and Pre-processing, removes 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 re-tweet flag, and the text of the tweet.The text may contain additional information, such as hash tags, 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 routes to be elaborated by the "Classification of SUMs" module. To this aim, some text mining techniques are applied in classification to the pre-processed SUMs. In the following, the text mining steps implemented in this module are described in detail:

a) Tokenization is normally 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 SUMs denoted 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 arithmetical 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 analysed (domain-specific stop-words), and can therefore be considered as noise.

c) Stopping 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 collection words with the same theme having closely related semantics.

d) Stem filtering consists in decreasing the number of stems of each SUM. In specific, each SUM is filtered by eliminating from the set of stems the ones not going to the set of related stems.

3) Classification of SUMs:

The third module, Classification of SUMs assigns to each elaborated SUM a class label related to traffic events. Hence, the output of this module is a group of N labelled SUMs. To the aim of labelling 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.

6. 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.

$SUM_j^T = \{t_{j1}^T, \dots, t_{jh}^T, \dots, t_{jH_j}^T\}$.

Where t_{jh}^T is the hth token and $H_j$is the total number of tokens in SUM_j^T .

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 jth stop-word filtered SUM as,

$SUM_j^{SW} = \{t_{j1}^{SW}, \dots t_{jk}^{SW}, \dots t_{jK}^{SW}\}$.

Where$t_{jk}^{SW}$the kth relevant token and Kj, is with Kj ≤ Hj, is the total number of relevant tokens in SUM_j^{SW}

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

labelled SUMs.

Step 4: In this the classifier that achieved the most accurate results by filtered tokens with labels was finally employed for the real time 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.

7. Experiment and Result Analysis

In this system we are used three types of classes for SUM classification which are updated by user i.e traffic related, Non traffic releted and Traffic due to External event classification is done by using NaviBayes classifier

The first two calss traffic releted and non traffic releted is also called 2Dataset and whole classes i.e traffic related,Non traffic releted and Traffic due to External event is also called as 3Dataset.

In these section we perform classification of SUM by the applying of NB Classifier, SVM and Textmining Technique. Some source words are used to fetching the SUM which is releted to Traffic Event i.e traffic, busy, jam, crush, queue, stuck, slowdown, signal etc.

After classification of SUM it's place in it's desired class and our system send notification to suspeecious user to knowing him about traffic status. Some examples are show in below

Table 1

Text of tweet	Class
To all friends of Catania, crash on the bypass! Queue from Talegaon to Akurdi	Traffic
Crash on Mumbai highway! Traffic slowed	Traffic
I happy to see that this great player is well after crash he had	Non Traffic

Fig 2: 2Dataset class

Table 2

Text of tweet	Class
Ind-pak match stucked near the stadium a huge traffic	Traffic due to exernal event
Traffic is slowed due to PM is coming	Traffic due to exernal event
I happy to see that this great player is well after crash he had	Non Traffic

Fig 3: 3Dataset class

The proposed system is developed and tested for the realtime monitoring of some or several areas of the Pune road network, by means of the study of the Twitter stream upcoming from those areas. The main aim is to execute a continuous monitoring of frequently busy road and highway in order to sense possible traffic events in real-time or even in advance with reverence to the traditional news media.

9. CONCLUSIONS

We advancing explored the authentication as well as trust and reputation calculation and management of CSPs and SNPs, which are two very critical and hardly explored issues with respect to CC and WSNs integration. Further, we proposed a novel ATRCM system for CC-WSN integration. Discussion and analysis about the authentication of CSP and SNP as well as the trust and reputation with respect to the service provided by CSP and SNP have been presented, followed with detailed design and functionality evaluation about the proposed ATRCM system. All these demonstrated that the proposed ATRCM system achieves the following three functions for CC-WSN integration:

1) Verifying CSP and SNP to avoid spiteful impersonation attacks.

2) Calculating and handling trust and reputation regarding the service of CSP and SNP.

3) Helping CSU choose required CSP and assisting CSP in selecting appropriate SNP,

Based on

(i) The authenticity of CSP and SNP;

(ii) The attribute requirement of CSU and CSP;

(iii) The cost, trust and reputation of the service of CSP and SNP.

ACKNOWLEGMENT

It provides us great pleasure in presenting the preliminary project report on Real-Time Detection of Traffic From Twitter Stream Analysis. I would like to take this opportunity to thank my internal guide Prof. and project coordinator Prof., for giving me all the support and guidance I needed. I am really grateful to them for their kind support. Their valuable ideas were very helpful. I am also grateful to Prof., Head of Computer Engineering Department, Dr. D.Y Patil Institute of Engineering for his essential support, suggestions. In the end our special thanks to College Management and all Staff for providing several resources such as laboratory with all needed software platforms, continuous Internet connection, for Our Project.

References

Takeshi Sakaki, Makoto Okazaki, Yutaka Matsuo, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors," IEEE Trans. Services Comput., vol. 5, no. 4, pp. 564–577, Fourth Quarter 2012.

[1] AminaMadani, 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.

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

[2] Alan Mislove, MassimilianoMarcon, 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.

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

[4] T. Sakaki, Y. Matsuo, T. Yanagihara, N. P. Chandrasiri, and K. Nawa, "Real-time event extraction for driving Programming, Bangalore, India, Jan 2010, pp. 147-158.

[5] M. Yuriyama and T. Kushida, "Sensor-cloud infrastructure Physical sensor management with virtualized sensors on cloud computing," in Proc. 13th Int. Conf. Netw.,-Based Inf. Syst., Sep. 2010, pp. 1–8

[6] Chen and Harry H., "A Review of the Applications of Agent Technology in Traffic and Transportation Systems," Wireless Commun. Mobile Comput, vol. 14, no. 1, pp. 19–36, Jan. 2014.

[7] Álvaro González, Luis M. Bergasa and J. Javier Yebes, "Chen and Harry H.," ACM Trans. Sensor Netw., vol. 5, no. 2, Mar. 2009, Art. ID 10.

BIOGRAPHIES

This is a Firdos khan Taher khan.he was born in 1991 and completed his degree in computer engineering in 2016 at Dr D Y Patil institute of engineering and technology pune. He is contribute his exallence for this project to helping the peoples.

This is a Sweety Kumari. She was born in 1993 and completed her degree in computer engineering in 2016 at Dr D Y Patil institute of engineering and technology pune. She is contribute her best skills in these system

This is a Ruchita Ganesh Khandge. She was born in 1992 and completed her degree in computer engineering in 2016 at Dr D Y Patil institute of engineering and technology pune. She is contribute her best skills in these system

This is a Shaikh Sultan Shaikh Mahebub.he was born in 1992 and completed his degree in computer engineering in 2016 at Dr D Y Patil institute of engineering and technology pune. He is contribute his exallence knowledge for this project.