International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 10 208 - 213

Monitoring System for Traffic Analysis Using Twitter Stream

Chaudhari Bhavesh N.
Dept. of Computer (B.E)
BVCOE&RI, Anjaneri
Nasik, India
bnchaudhari2029@gmail.com

Dalvi Rajat R.
Dept. of Computer (B.E)
BVCOE&RI, Anjaneri
Nasik, India
dalvi.rajat@gmail.com

Divate Karishma A.
Dept. of Computer (B.E)
BVCOE&RI, Anjaneri
Nasik, India
divatekarishma23@gmail.com

Narkhede Charulata T.
Dept. of Computer (B.E)
BVCOE&RI, Anjaneri
Nasik, India
narkhedecharulata@gmail.com

Prof. Handore Sonali A.
Dept. of Computer
BVCOE&RI, Anjaneri
Nashik, India
narkhedecharulata@gmail.com

***Abstract*—**Social networks are often utilized as a supply of data for event detection like road holdup and automobile accidents. Existing system present a period of time observance system for traffic event detection from twitter. The system fetches tweets from twitter and then; processes tweets victimisation text mining techniques. Last performs the classification of tweets. The aim of the system is to assign the suitable category label to every tweet, whether or not it's associated with a traffic event or not. System utilized the support vector machine as a classification model. The projected system uses the system supported semi-supervised approach, which provides coaching victimisation traffic connected dataset. we have a tendency to propose a bunch approach for classification of the tweets in traffic connected and non- traffic connected tweets. We use a geometer distance to calculate the similarity between the tweets

*****

I. INTRODUCTION

Wikipedia defines a social network service as a service that "focuses on the building and confirmatory of on-line social networks for communities of individuals who share interests and activities, or who have an interest in exploring the interests and activities of others, and that necessitates the employment of package."

A report printed by OCLC provides the subsequent definition of social networking sites: "Web sites primarily designed to facilitate interaction between users who share interests, attitudes and activities, like Facebook, Mixi and MySpace." These social networking sites are used for the maintaining the social relationship, finding the users with similar interests. The message shared by user in social networks is called standing Update Message (SUM). Sum may contain, apart from the text, meta-information like timestamp, geographic coordinates, name of the user, links to different resources, hashtags, and mentions . Total thought of in a} very specific geographical region may provide the right information. Social networks and media platforms square measure being widely used as a provide of information for the detection of events, like traffic, incidents, and natural disasters. In this paper, our focus is on specific small-scale event road traffic. Our aim to sight and analyze traffic connected events by method users' messages belonging to a specific area and written inside English. We tend to propose a system able to extract, analyze SUMs as related to a road traffic event or not. Few papers are planned for traffic detection victimization Twitter stream analysis. Social networks will offer a variety of advantages to members of an organisation:

1. Support for learning: Social networks will enhance informal learning and support social connections inside teams of learners and with those concerned within the support of learning.
2. Support for members of an organisation: Social networks will doubtless be used my all members of an organisation, and not simply those concerned in operating with students. Social networks will facilitate the event of communities of follow.
3. Engaging with others: Passive use of social networks will offer valuable business intelligence and feedback on institutional services (although this might produce to moral concerns).
4. Ease of access to data and applications: the benefit of use of the many social networking services will offer advantages to users by simplifying access to different tools and applications. The Facebook Platform provides an example of however a social networking service may be used as an setting for different tools.

5. Common interface: A potential advantage of social networks is also the common interface that spans work / social boundaries. Since such services square measure typically employed in a private capability the interface and also the approach the service works is also acquainted, therefore minimising coaching and support required to take advantage of the services during a skilled context. This can, however, even be a barrier to people who want to own strict boundaries between work and social activities.

II. EXISTING SYSTEM

Recently, social networks and media platforms are wide used as a supply of knowledge for the detection of events, like hold up, incidents, natural disasters (earthquakes, storms, fires, etc.), or alternative events.

– Sakaki et al. use Twitter streams to sight earthquakes and typhoons, by observation special trigger-keywords, associated by applying an SVM as a binary classifier of positive events (earthquakes and typhoons) and negative events (non-events or alternative events).
– Agarwal et al. concentrate on the detection of fires during a manufacturing plant from Twitter stream analysis, by victimisation commonplace human language technology techniques and a Naive Thomas Bayes (NB) classifier.
– Li et al. propose a system, referred to as TEDAS, to retrieve incident-related tweets. The system focuses on Crime and Disaster-related Events (CDE) like shootings, thunderstorms, and automotive accidents, and aims to classify tweets as CDE events by exploiting a filtering supported keywords, spatial and temporal info, variety of followers of the user, variety of retweets, hashtags, links, and mentions.

Disadvantages:

– Event detection from social networks analysis may be a more difficult drawback than event detection from ancient media like blogs, emails, etc., wherever texts square measure well formatted.
– SUMs square measure unstructured and irregular texts; they contain informal or abbreviated words, misspellings or grammatical errors.
– SUMs contain a large quantity of not helpful or hollow info.

III. PROPOSE SYSTEM

In this paper, we have a tendency to propose associate degree intelligent system, supported text mining and machine learning algorithms, for time period detection of traffic events from Twitter stream analysis.

The system, once a feasibleness study, has been designed associate degreed developed from the bottom as an event-driven infrastructure, designed on a Service oriented design (SOA).

The system exploits on the market technologies supported progressive techniques for text analysis and pattern classification. These technologies and techniques are analyzed, tuned, adapted, and integrated so as to create the intelligent system.

In explicit, we have a tendency to present associate degree experimental study that has been performed for deciding the foremost effective among completely different progressive approaches for text classification. The chosen approach was integrated into the ultimate system and used for the on-the-field time period detection of traffic events.

In this paper, we have a tendency to target a specific small-scale event, i.e., road traffic, and that we aim to observe and analyze traffic events by process users' SUMs happiness to an exact space and written within the Italian language. to the current aim, we have a tendency to propose a system ready to fetch, elaborate, and classify SUMs as associated with a road traffic event or not.

To the simplest of our information, few papers are projected for traffic detection using Twitter stream analysis. However, with relevance our work, all of them target languages completely different from Italian, use completely different input options and/or feature choice algorithms, and think about only binary classifications.

Advantages-

– Tweets are up to one hundred forty characters, enhancing the period and news-oriented nature of the platform. In fact, the life-time of tweets is typically very short, so Twitter is that the social network platform that's best suited to review SUMs associated with period events.
– Each tweet is directly related to meta-information that constitutes extra info.
– Twitter messages are public, i.e., they're directly on the market with no privacy limitations. For all of those reasons, Twitter could be a smart supply of data for period event detection and analysis.
– Moreover, the projected system may work at the side of different traffic sensors (e.g., loop detectors, cameras, infrared cameras) and ITS observation systems for the detection of traffic difficulties, providing a affordable wide coverage of the road network, particularly in those areas (e.g., urban and suburban) wherever traditional traffic sensors are missing.
– It performs a multi-class classification that acknowledges non-traffic, traffic owing to congestion or crash, and traffic because of external events.
– It detects the traffic events in real-time; AND iii) it's developed as an event-driven infrastructure, engineered on AN SOA design.

IV. SYSTEM ARCHITECTURE

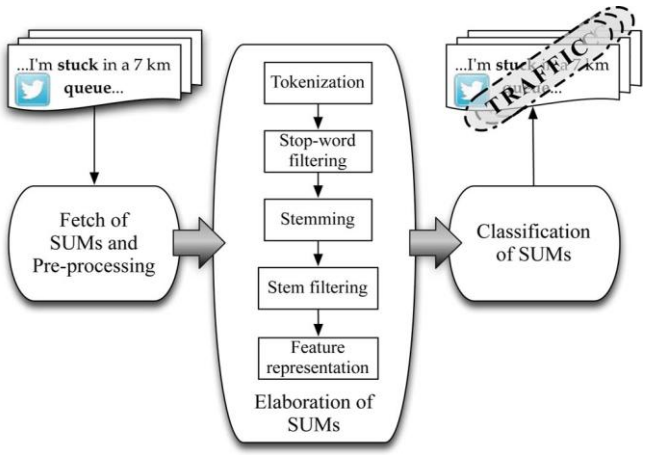

Figure 1: System Architecture

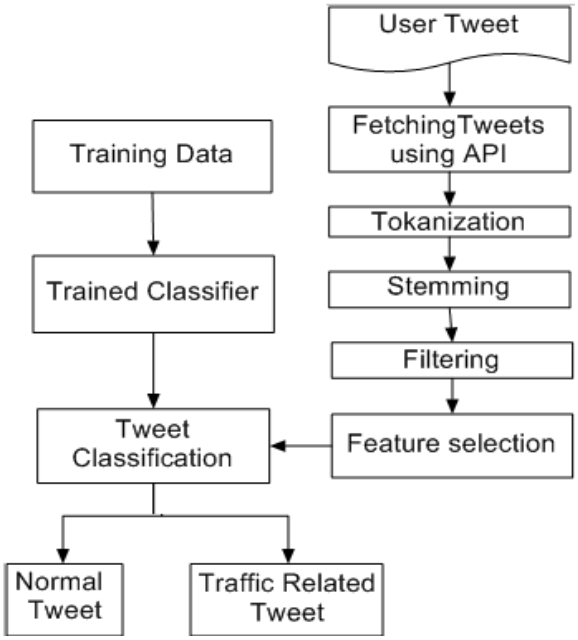

Figure 2: Flow of System Artchitecture

MODULES:

1. Fetch of SUMs and Pre-Processing
2. Elaboration of SUMs
3. Classification of SUMs
4. Setup of the System

1. Fetch of SUMs and Pre-Processing:

The first module, "Fetch of SUMs and Pre-processing", extracts raw tweets from the Twitter stream, supported one or additional search criteria (e.g., geographic coordinates, keywords showing within the text of the tweet). Every fetched raw tweet contains: the user id, the timestamp, the geographic coordinates, a retweet flag, and also the text of the tweet. The text could contain extra info, like hashtags, links, mentions, and special characters. During this paper, we took solely Italian language tweets under consideration. However, the system is simply tailored to deal with completely different languages. When the SUMs are fetched in keeping with the precise search criteria, SUMs are pre-processed. so as to extract only the text of every raw tweet and take away all meta-information related to it, a daily Expression filter is applied. More very well, the meta-information discarded are: user id, timestamp, geographic coordinates, hashtags, links, mentions, and special characters. Finally, a case-folding operation is applied to the texts, so as to convert all characters to character. At the top of this elaboration, every fetched total seems as a string, i.e., a sequence of characters.

2. Elaboration of SUMs:

The second process module, "Elaboration of SUMs", is dedicated to reworking the set of pre-processed SUMs, i.e., a group of strings, in a very set of numeric vectors to be careful by the "Classification of SUMs" module. to the current aim, some text mining techniques are applied in sequence to the pre-processed SUMs. within the following, the text mining steps performed during this module are delineate in detail: tokenization is often the primary step of the text mining method, and consists in reworking a stream of characters into a stream of process units known as tokens (e.g., syllables, words, or phrases). Stop-word filtering consists in eliminating stop-words, i.e., words which offer very little or no info to the text analysis. Common stop-words are articles, conjunctions, prepositions, pronouns, etc. alternative stop-words are those having no applied mathematics significance, that is, those who usually seem fairly often in sentences of the considered language (language-specific stop-words), or within the set of texts being analyzed (domain-specific stop-words), and might so be thought-about as noise. Stemming is that the method of reducing every word (i.e., token) to its stem or root kind, by removing its suffix. The aim of this step is to cluster words with a similar theme having closely connected semantics.

In the following, the text mining steps performed during this module are represented in detail:

 ISSN: 2321-8169
208 - 213

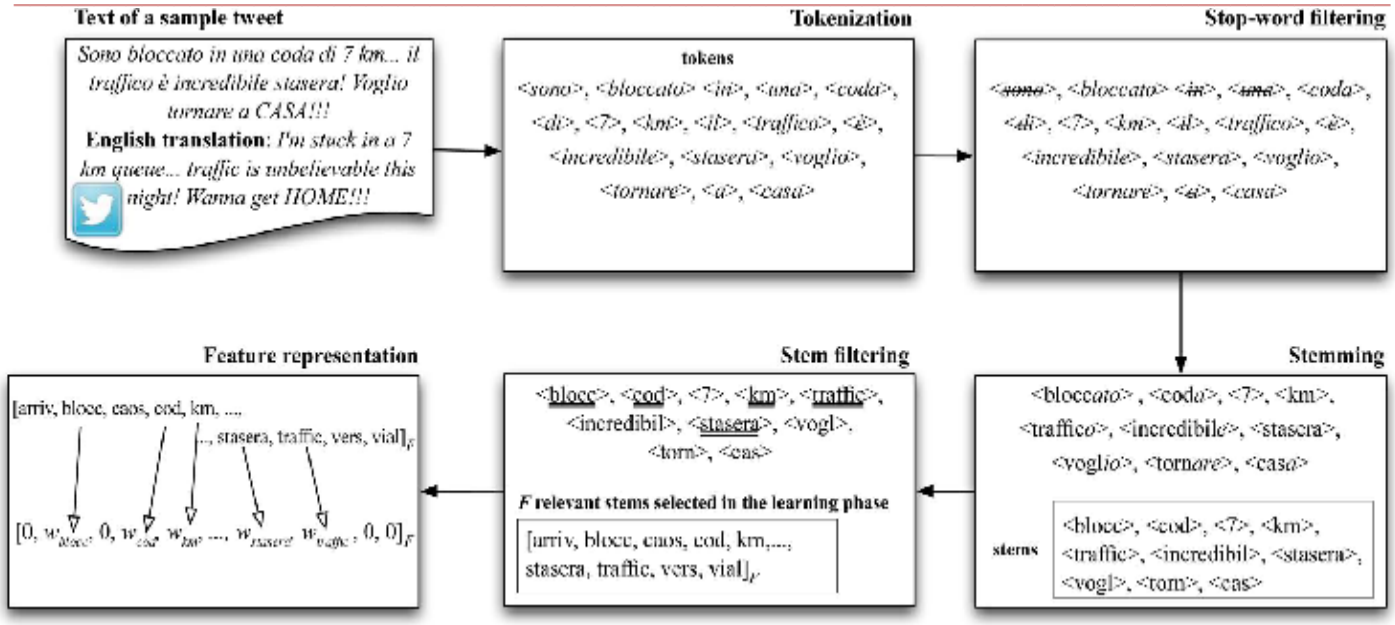

Figure 3: Steps of the text mining elaboration applied to a sample tweet.

a. tokenization is usually the primary step of the text mining process, and consists in remodeling a stream of characters into a stream of process units referred to as tokens (e.g., syllables, words, or phrases). Throughout this step, alternative operations are sometimes performed, like removal of punctuation and different non-text characters, and standardization of symbols (e.g., accents, apostrophes, hyphens, tabs and spaces). within the planned system, the tokenizer removes all punctuation marks and splits every add into tokens corresponding to words (bag-of-words representation

b) stop-word filtering consists in eliminating stop-words, i.e., words which give very little or no data to the text analysis. Common stop-words area unit articles, conjunctions, prepositions, pronouns, etc. different stop-words area unit those having no applied math significance, that is, those that typically seem fairly often in sentences of the thought-about language (language-specific stop-words), or within the set of texts being analyzed (domain-specific stop-words), and can so be thought-about as noise. The authors in have shown that the ten most frequent words in texts and documents of the English language area unit about the 20–30% of the tokens in an exceedingly given document.

In the planned system, the stop-word list for the Italian language was freely downloaded from the Snowball Tartarus web site half-dozen and extended with different circumstantial defined stop-words. At the tip of this step, each SUM is therefore reduced to a sequence of relevant tokens. We tend to recall that a relevant token may be a token that doesn't belong to the set of stop-words;

c) Stemming is that the method of reducing every word (i.e., token) to its stem or root type, by removing its suffix. The purpose of this step is to cluster words with constant theme having closely connected linguistics. within the planned system,the stemmer exploits the Snowball Inferno Stemmer seven for the Italian language, supported the Porter's formula.Hence, at the tip of this step every add is painted as a sequence of stems extracted from the tokens contained in it.

d) Stem filtering consists in reducing the quantity of stems of each SUM. Specifically , every add is filtered by removing from the set of stems those not happiness to the set of relevant stems.

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

For each SU M j SF we define the vector

$X_j = \{x_{j1} \ldots xjf, \ldots, x_{jF}\}$ where each element is set according to the following formula:

$$x_{jf} = \begin{cases} w_f & \text{if stem } \hat{s}_f \in SUM_j^{SF} \\ 0 & \text{otherwise.} \end{cases}$$

3. Classification of SUMs:

This module, assigns to each elaborated total a class label related to traffic events. Thus, the output of this module is a collection of N labelled SUMs. To the aim of labeling each total, a classification model is employed. The parameters of the classification model are well-known throughout the supervised learning stage. Actually, as a result of it are

mentioned, fully completely different classifications models are considered and compared. The classifier that achieved the foremost correct results was finally used for the realtime observation with the planned traffic detection system. The system endlessly monitors a particular region and notifies the presence of a traffic event on the concept of a bunch of rules that will be defined by the supervisor. as an example, once the first tweet is recognized as a traffic-related tweet, the system may send a alert. Then, the actual notification of the traffic event is also sent once the identification of an exact variety of Tweets with a similar label.

Proposed clustering Algorithm:
Input: training Dataset T, take a look at dataset D,
Output: Clustered Tweet set.

1. Method: At the beginning train the classifier using semi-supervised traffic related coaching dataset.
2. Fetch user tweets from loudspeaker account
3. Store in database
4. for each tweet in database
5. Calculate the similarity exploitation geometer distance with trained knowledge.
6. If(similarity > Threshold)
7. Add tweet to traffic connected tweet set
8. Else
9. Increase traditional tweet set.
10. End if
11. End for
12. Return classified tweets

4. Setup of the System:
As declared previously, a supervised learning stage is needed to perform the setup of the system. above all, we want to identify the set of relevant stems, the weights related to every of them, and also the parameters that describe the classification models. we use a group of Ntr labelled SUMs as training set. throughout the training stage, every sum is elaborated by applying the tokenization, stop-word filtering, and stemming steps. Finally, the tweets were manually labelled with 2 possible class labels, i.e., as associated with road traffic event (traffic), e.g., accidents, jams, queues, or not (non-traffic). additional very well, initial we tend to read, taken, and properly assigned a traffic class label to every candidate traffic class tweet.
Formula:

$$CS = \left(\bigcup_{j=1}^{N_{tr}} SUM_j^S \right) = \{s_1, \ldots, s_q, \ldots, s_Q\}.$$

CONCLUSION

In this paper, we have planned a system for real-time detection of traffic-related events from Twitter stream analysis. The system, designed on a SOA, is ready to fetch and classify streams of tweets and to notify the users of the presence of traffic events. Moreover, the system is additionally ready to discriminate if a traffic event is because of AN external cause, like soccer match, procession and manifestation, or not. We've exploited offered package packages and state-of the- art techniques for text analysis and pattern classification. These technologies and techniques are analyzed, tuned, custom-made and integrated so as to make the general system for traffic event detection. Among the analyzed classifiers, we've shown the prevalence of the SVMs that have achieved accuracy of ninety five.75%, for the 2-class drawback, and of 88.89% for the 3-class drawback, within which we've also thought of the traffic because of external event category. The most effective classification model has been utilized for real time observance of many areas of the Italian road network. We've shown the results of a observance campaign, performed in Sep and early October 2014. We've mentioned the capability of the system of detection traffic events virtually in real time, typically before on-line news internet sites and local newspapers.

REFFERENCES

[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] G. Anastasi *et al.*, "Urban and social sensing for sustainable mobility in smart cities," in *Proc. IFIP/IEEE Int. Conf. Sustainable Internet ICT Sustainability*, Palermo, Italy, 2013, pp. 1–4.
[5] A. Rosi *et al.*, "Social sensors and pervasive services: Approaches and perspectives," in *Proc. IEEE Int. Conf. PERCOM Workshops*, Seattle, WA, USA, 2011, pp. 525–530.
[6] 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.
[7] J. Allan, *Topic Detection and Tracking: Event-Based Information Organization*. Norwell, MA, USA: Kluwer, 2002.
[8] K. Perera and D. Dias, "An intelligent driver guidance tool using location based services," in *Proc. IEEE ICSDM*, Fuzhou, China, 2011, pp. 246–251.
[9] 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.

[10] 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.

[11] A. 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.

[12] 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.

[13] P. M. d'Orey and M. Ferreira, "IT'S for sustainable mobility: A survey on applications and impact assessment tools," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 2, pp. 477–493, Apr. 2014.

[14] K. Boriboonsomsin, M. Barth, W. Zhu, and A. Vu, "Eco-routing navigation system based on multisource historical and real-time traffic information," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 4, pp. 1694–1704, Dec. 2012.

[15] J. Hurlock and M. L. Wilson, "Searching twitter: Separating the tweet from the chaff," in *Proc. 5th AAAI ICWSM*, Barcelona, Spain, 2011, pp. 161–168.

[16] J. Weng and B.-S. Lee, "Event detection in Twitter," in *Proc. 5th AAAI ICWSM*, Barcelona, Spain, 2011, pp. 401–408.

[17] S. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, *Text Mining: Predictive Methods for Analyzing Unstructured Information*. Berlin, Germany: Springer-Verlag, 2004.