Mining Twitter for Road Traffic Event Detection

[1] Sujith M S, [2] Rahamathulla K
[1] Student, [2] Assistant professor, CSE Dept.
Government Enggineering College,Thrissur, Kerala,India
[1] sujithmsathyan@gmail.com

***Abstract:* — a road traffic event detection, provides information for emergency traffic control and management purposes. Twitter is rapidly emerging as a efficient tool for the contribution and spreading of information that has an immense value for increasing awareness of traffic incidents. In this paper, a system for road traffic event detection from tweet analysis is presented. The system fetches tweets from Twitter according to several search criteria, processes the tweets by applying text mining techniques and finally classifies the tweets. The aim is to assign the appropriate class label to each tweet, whether related to a traffic event or not. The class labels used are non-traffic, traffic due to congestion or crash, and traffic due to external events. A combination of multi-class Support Vector Machine (SVM) and Decision tree classification algorithm is being implemented to classify tweets that reflect road traffic conditions. This can possibly help the drivers and concerned authorities to identify the traffic conditions in specified places.**

***Keywords*—Social Network, twitter, traffic tweets, decision tree and SVM, multi-class classification.**

I. INTRODUCTION

Twitter is a popular social media tool which allows people to exchange information freely and instantly. It is widely used for real-time text information dissemination. Twitter represents an important first-hand source of information and can be used to get almost instantaneous information about events. This instant information can be used in traffic alerting, to inform relevant people and agencies about traffic events. Timely and accurate traffic alerting has significant benefits for many types of users : emergency services, road users, highways authorities, recovery agencies, police and others.

The user message shared in social networks is called tweets, and it may contain, apart from the text, meta-information such as time stamp, geographic coordinates (latitude and longitude), name of the user, links to other resources and hash tags. Several tweets referring to a certain topic or related to a limited geographic area may provide, if correctly analyzed, a great deal of valuable information about an event or a topic.

nowadays, social networks and media platforms have been widely used as a source of information for the detection of events, such as traffic detection, incidents and natural disasters (earthquakes, storms, wildfires, etc.). An event can be defined as a real-world occurrence that happens in a specific space and time. In particular, regarding traffic-related events, people often share information about the current traffic situation around them while driving or traveling by means of a tweets. Event detection from social networks analysis is a more challenging problem than event detection from traditional media like blogs, emails, etc., where texts are well formatted. In fact, tweets are unstructured and irregular texts, and they may contain informal or abbreviated words, grammatical errors or misspellings. Due to their nature, they are usually very brief, thus becoming an incomplete source of information. Furthermore, tweets contain a huge amount of useless or meaningless information, which has to be filtered. Traditional text mining techniques lose accuracy when applied to tweet mining. Among the challenges for tweet mining are: 1) limited information of fewer than 140 characters; 2)informal expressions and word variations caused by spelling errors, tweet slang and abbreviations. 3) data volume: only a tiny proportion of the whole tweets will be relevant to any given topic. Major challenge is to successfully identify and analyze the tweets accurately, instantly and automatically. The proposed system for real-time detection of traffic-related events from tweet stream analysis involves a real-time monitoring system for traffic event detection from twitter stream analysis. The system fetches tweets from twitter according to several search criteria, processes tweets by applying text mining techniques, and finally performs the classification of tweets. The aim is to detect and assign the appropriate class label to each tweet, as related to a traffic events or not.

II. RELATED WORKS

in the various approaches for using social media to extract useful information for event detection,events are classified into two categories, small-scale events and large-scale events. Large scale events (earthquakes, tornado, or the election of a president) are characterized by a huge number of tweets, and a wider temporal and geographic coverage. On the other hand, small-scale events (traffic, car crashes, fires, or local manifestations) usually have a small number of tweets related to them, belong to a precise geographic location, and are concentrated in a small time interval. Due to the smaller number of tweets 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. Twitter mining for traffic related tweets belongs to small scale event detection.

Regarding traffic event DETECTION,SHEN ZHANG[2] proposed an automatic incident detection, an intelligent transportation management system that provides information for emergency traffic control and management purposes. The approach used a combination of lda and document clustering, and allowed for semantic filtering of the incident-topic tweets regarding the topic distribution and spatial point pattern analysis was employed to investigate the spatial pattern of incident-topic tweets in a case study region in seattle, where a considerable clustering pattern was observed at different scales up to 600m. A distance-based spatial clustering algorithm was used to extract features from tweet point process, and seattle downtown area was chosen as a representative sample environment with feature points of high density, proving that it is possible to reliably detect clusters of tweets posted spatially close to traffic incidents.

D'ANDREA, E, DUCANGE, P, AZZERINI, B, & MARCELLONI[1] proposes a system for real-time monitoring system for traffic event detection by analyzing tweets. The system fetched tweets from twitter according to several search criteria; processed tweets by applying text mining techniques; and finally performed the classification of tweets. The aim was to assign the appropriate class label to each tweet, whether related to a traffic event or not. The system used support vector machine as a classification model, and achieved high accuracy value solving a binary classification problem (traffic versus non-traffic tweets). Multi class problem was also solved using this system by classifying into traffic caused by an external event or not.

NAPONG WANICHAYAPONG, WASAWAT PRUTHIPUNYASKUL,[3] proposed an approach for road traffic data extraction and classification in which traffic information was extracted from twitter using syntactic analysis and then further classified into two categories: point and link. Point information was associated with only one point e.g.(a car crash at a crossroad)and link information was associated with a road start point and an end point (e.g. A traffic jam between two squares). A dictionary was used in this approach, number of words in dictionary affecting the performance of the tokeniser. For classification, more place words gave more accurate classification.

Sakaki, takeshi[4] proposed a method to extract real-time traffic information using twitter as a type of social sensor. This was a new approach to acquire valuable information for drivers from social media. The system extracted driving information from social media using text-based classification methods. Because geographical coordinates are necessary to note where the driving information had occurred, it incorporated a method to transform geographically related terms into geographical coordinates. This method used SVM and extracted location information from each tweet using gps information, geo-location web services, a user-generated dictionary, and contextual information.

HASBY, MUHAMMAD, AND MASAYU LEYLIA KHODRA[5] proposed a method for optimal path finding based on traffic information extraction from twitter. The system extracted traffic information from twitter, and then used the extracted result as heuristic in finding the optimal route. The extraction process was conducted continuously to monitor traffic information. Path finding was done after receiving an input of start node and end node, and then the optimal route was found based on the traffic information from the information extraction process. Named entity recognition(ner) process was conducted by classification model. Traffic information extracted from tweets that use #lalinbdg hashtag and tweets from @lalinbdg account were utilized using information extraction techniques. This traffic information was employed later as a heuristic for path finding process.

Wang, d, al-rubaie, a davies & clarke[6] proposed an traffic status alert and warning system. In this approach traffic related tweets are classified using tweet-lda. When comparing proposed tweet-lda and SVM, tweet-lda worked fine with good accuracy. Gutierrez, c., figuerias,p., oliveira, p., costa, r., & jardim-goncalves[7], proposed an approach to integrate and extract tweet messages from traffic agencies in uk. Objective of this approach is to detect the geographical focus of traffic events. This method composed of several steps: tweet classification, event type classification, name entity recognition, geolocation and event tracking.

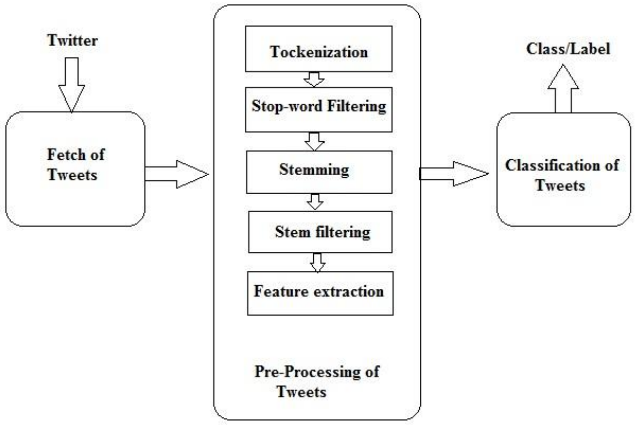


Fig 1. Design of system for classification of tweets

III. PROPOSED METHOD

In this method, several tweets are collected based on hash tags traffic, slowdown, accident, highway, road jam, queue, block, crash, road, huge traffic and combination of these tags . These help in describing the process to locate the relevant data and relevant twitter hash tags required for identifying traffic status(a twitter hash tag is a word beginning with a sign, used to emphasize or tag a topic). The work flow, which includes both qualitative analysis and data mining algorithms, is developed in order to improve the performance. Figure a illustrates the proposed system. The flow can be summarized in the following steps:

1. Data is collected from tweet content using tweet crawler.
2. A detailed data analysis is done.
3. The traffic condition in the specified area are categorized and a combination of multi-class classifiers is proposed which can be implemented by decision tree and support vector machine classification algorithm.
4. The performance of the proposed classifier can be assessed by comparing it with other state-of-the-art multi class classifiers. The classification algorithm is used to train a detector that could assist identitification of traffic condition.the result could help the drivers and authority to identify traffic condition and take necessary action to overcome risks.

A. *Data collection*

The irregularity and diversity of the languages results in a challenging task of collecting the data related to the road traffic. Twitter apis [16] were used to search tweets. The search process was exploring and it started by searching based on boolean combinations of possible keywords such as traffic, accident, jam, block, slow down, road, etc., and further expanded the keyword set and combined the boolean logic iteratively. Many of the irrelevant tweets with spam were discarded. The studies show that hash tag slowdown was the most popular hash tag which occurs most frequently, where the tweeples are used to tweet their tweets related to traffic. These hash tags also helped in studying the traffic related issues and its used as training dataset. Test dataset can be collected using hash tag related road traffic.

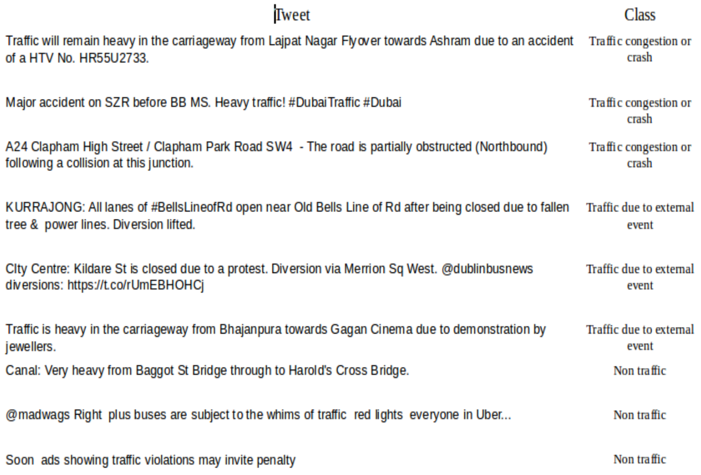
Most of the contents in the social media are often ambiguous. The previous studies showed that there may occur faulty assumptions because automatic algorithms were used without considering the quality of data. Latent dirichlet allocation (lda) is a popular topic modeling algorithm that can detect general topics from very large scale data [21], [22]. In order to identify traffic event in the tweets, an inductive content analysis which is the qualitative research method for analyzing the text content manually, was carried out on #traffictweet dataset. Our dataset consist of 1800 tweet, each category has 600 tweets.

A short descriptions of the 3 categories considered is given below

Traffic due to external events : discriminate traffic based on whether it is caused by an external event (e.g., a foot- ball match, a concert, a flash-mob, a political demonstration,a fire) or not. Even though the current release of the system was not designed to identify the specific event, knowing that the traffic difficulty is caused by an external event could be useful to traffic and city administrations,for regulating traffic and vehicular mobility, or managing scheduled events in the city.

Traffic due to congestion or crash : tweets related to traffic congestion, crashes, break down and jams(traffic congestion or crash class).

Non-traffic : tweets not related to traffic.



Tweet	Class
Traffic will remain heavy in the carriageway from Lajpat Nagar Flyover towards Ashram due to an accident of a HTV No. HR55U2733.	Traffic congestion or crash
Major accident on SZR before BB MS. Heavy traffic! #DubaiTraffic #Dubai	Traffic congestion or crash
A24 Clapham High Street / Clapham Park Road SW4 - The road is partially obstructed (Northbound) following a collision at this junction.	Traffic congestion or crash
KURRAJONG: All lanes of #BellsLineofRd open near Old Bells Line of Rd after being closed due to fallen tree & power lines. Diversion lifted.	Traffic due to external event
CIty Centre: Kildare St is closed due to a protest. Diversion via Merrion Sq West. @dublinbusnews diversions: https://t.co/rUmEBHOHCj	Traffic due to external event
Traffic is heavy in the carriageway from Bhajanpura towards Gagan Cinema due to demonstration by jewellers.	Traffic due to external event
Canal: Very heavy from Baggot St Bridge through to Harold's Cross Bridge.	Non traffic
@madwags Right plus buses are subject to the whims of traffic red lights everyone in Uber...	Non traffic
Soon ads showing traffic violations may invite penalty	Non traffic

B. *Text pre-processing*

Many symbols are being used by twitter to convey special meaning. For example, # is used to indicate a hashtag, @ is used to indicate a user account, and rt is used to indicate a re-tweet. Stop words “a, an, and, of, he, she, it”, non letter symbols, and punctuations also bring noise to the text. Thus the text needs to be per-processed before training the classifier:

1. The #traffic hash tags were removed and for other hash tags only # sign is removed and the hash tag texts were kept as such.
2. The tweets in languages other than english were also removed(chinese, italian, etc).
3. Non-letter symbols and punctuation contained words are removed which includes the removal of @ and http links and rts.
4. If we detected more than two identical letters repeating, we replaced them with one letter. Therefore, slooowdown and sooo were corrected to slowdown and so.
5. The common stop words were removed by using information retrievaltoolkit. We kept words like much, more, all, always, still, only, because the tweets frequently use these words to express extent.

C. *Combined SVM and DESICION tree classifiers*

Transformation of the multi-class classification problem into multiple single-class classification problems is one of the popular ways to implement the multi-class classifier. One-versus-all or binary relevance is one of the transformation methods which consist of assuming the independence among categories, and training a binary classifier for each category. All kinds of binary classifiers can be transformed to multi-class classifier using the one-versus-all heuristic.

Support vector machines are among the most robust and successful classification algorithms. They are based upon the idea of maximizing the margin maximizing the minimum distance from the separating hyperplane to the nearest example. The basic SVM supports only binary classification, but extensions [24][25] have been proposed to handle the multi class classification case as well. In these extensions, additional parameters and constraints are added to the optimization problem to handle the separation of the different classes. The formulation of [24] [25] can result in a large optimization problem, which may be impractical for a large number of classes. On the other hand, [26] is a better formulation with a more efficient implementation.

Decision trees are a powerful classification technique. Two widely known algorithms for building decision trees are classification and regression trees [28]and id3/c4.5 [29]. The tree tries to infer a split of the training data based on the values of the available features to produce a good generalization. The split at each node is based on the feature that gives the maximum information gain. Each leaf node corresponds to a class label. A new example is classified by following a path from the root node to a leaf node, where at each node a test is performed on some feature of that example. The leaf node reached is considered the class label for that example. The algorithm can naturally handle binary or multi class classification problems. The leaf nodes can refer to either of the k classes concerned.

The combination of SVM and decision tree may result in good performance since there is a big difference between them in both theory and technique. In certain cases in multi-class classification, decision tree classifier provides higher accuracy over SVM, and in certain other cases SVM provides higher accuracy. However, in the same classification problem, decision tree classifies some classes better than SVM. Thus a combined approach is used for provide higher accuracy. There are three methods for combining SVM and dtl, namely, minimum misclassification, maximum accuracy, and dominant class; here we use minimum misclassification method[10].

In this paper we use minimum misclassification method .the first step performed in this method is to construct contingency matrix for SVM and decision tree from the training data set. Then the test data is classified by using both SVM and decision tree. Any kernel function for SVM and any level of pruning for decision tree can be used in this method. From contingency matrix, probability of misclassification of each label is calculated. Then each tweet is labeled with label contain less probability[10].with all the above studies, it can be concluded that the combined SVM and decision tree classification algorithm can provide a better and effective performance than other classifiers such as naive BAYES multi-label classifier, decision tree classifier and linear multi-class SVM using the LIBSVM library [39] with the one-versus-all heuristic.

IV. IMPLEMENTATION AND RESULT

The main tools we have used for developing the system are: 1) twitter’s api, which provides direct access to the public stream of tweets; 2) tweepy,an easy-to-use python library for accessing the twitter api. 3) hunspell stemmer,a dictionary based stemmer which provide better per-processing than porter stemmer, word net stemmer and snowball stemmer. 4)scikit-learn,a machine learning library for python.

For feature extraction we have used combination of tf-idf and bags of words method. Irrelevant stems are removed before feature extraction using a dictionary consist of most frequently used words on twitter [30]

	Recall	Precision	F1-score
Non traffic	0.84	0.22	0.35
Traffic congestion or crash	0.49	0.94	0.65
Traffic due to external event	0.54	0.27	0.36

Table 1 decision tree

	Recall	Precision	F1-score
Non traffic	0.63	0.67	0.65
Traffic congestion or crash	0.56	0.62	0.59
Traffic due to external event	0.35	0.26	0.30

TABLE 2 SVM

TABLE 3 COMBINED SVM AND DECISION TREE

	Recall	Precision	F1-score
Non traffic	0.64	0.73	0.68
Traffic congestion or crash	0.58	0.77	0.66
Traffic due to external event	0.64	0.14	0.23

Tables 1,2 and 3 show recall,precision and f1-score of decision tree,SVM and minimum misclassification method respectively

TABLE 4.ACCURACY

	Accuracy
SVM	0.5300
Decision Tree	0.5526
Combined SVM & DTL	0.6071

Tables 4 shows accuracy of decision tree,SVM and minimum misclassification method

V. CONCLUSION

There are many limitations for the manual large scale computational analysis of user generated textual content. Machine learning based classifiers help the researchers in learning analytics, educational data mining, and learning technologies effectively. Social media data provides substantial details regarding traffic condition details in a particular geographical area. This data can be extracted and analysed using machine learning classifiers. This technique can also help to identify the traffic status about a particular area like traffic crash, traffic jam, etc., and help the drivers and authorities know the current situation to take necessary action.

The existing system detects the traffic related events by analyzing tweets in real-time. This system is basically based on italian tweets, i.e., it classifies only italian tweets. Tweet analysis in italian language is much more easier than english tweets. System classifies the tweets using various machine language techniques like multi class SVM, decision tree and naive bayes classifier. Multi class decision tree provides higher accuracy ,precision and recall over all other classifiers; classification based on SVM gave 55.26% accuracy. By using multi class SVM, it is impossible to achieve such higher accuracy levels in english tweets.

The proposed system will improve the accuracy of traffic event detection by combining machine language technique multi-class SVM and decision tree classifier. Minimum misclassification techniques used in proposed system. This methods will work well with english tweets. By using these method, we can achieve upto 60.71% accuracy over all other machine language techniques.

REFERENCES

[1] D Andrea,Ducange,Lazzerini,Marcellon. Real-time detection of traffic from twitter stream analysis,intelligent transportation systems, IEEE transactions on, 16(4), 2269-2283.(2015).

[2] Zhang, Shen. ”using twitter to enhance traffic incident awareness.”,intelligent transportation systems (itsc), 2015 ieee 18th international conference on. Ieee, 2015.

[3] Wanichayapong, Napong, et al., social-based traffic information extraction and classification., .its telecommunications (itst), 2011 11th international conference on. Ieee, 2011.

[4] Sakaki, Takeshi, et al., real-time event extraction for driving information from social sensors.,, cyber technology in automation, control, and intelligent systems (cyber), 2012 ieee international conference on. Ieee,2012.

[5] Hasby, Muhammad, And Masayu Leylia Khodra, optimal path finding based on traffic information extraction from twitter.,,ict for smart society (iciss), 2013 international conference on. Ieee, 2013.

[6] Wang,Al-Rubaie,Davies,Clarke, ”real time road traffic monitoring alert based on incremental learning from tweets”,.in evolving and autonomous learning systems (eals), 2014 ieee symposium on (pp. 50-57).ieee.(2014, december)

[7] GUTIERREZ, C., FIGUERIAS, P., OLIVEIRA, P., COSTA, R.,JARDIM-goncalves,twitter mining for traffic events detection, . In science and informationconference (sai), 2015 (pp. 371-378). Ieee.(2015, july).

[8] PARIKH, RUCHI, AND KAMALAKAR KARLAPALEM., ET: events from tweets., ,proceedings of the 22nd international conference on world wide web companion. International world wide web conferences steering committee, 2013.

[9] MADANI, AMINA, OMAR BOUSSAID, AND DJAMEL EDDINE ZEGOUR., WHATS HAPPENING: A SURVEY OF TWEETS EVENT DETECTION., INNOV 2014 (2014): 3rd.

[10] Nguyen, Thao, ”investigation of combining SVM and decision tree for emotion classification, multimedia, seventh IEEE INTERNATIONAL SYMPOSIUM ON. IEEE, 2005.

[11] Sakaki, Takeshi, Masahide Okazaki, and yoshikazu matsuo., tweet analysis for real-time event detection and earthquake reporting system development, knowledge and data engineering, IEEE transactions on 25.4 (2013):919-931.

[12] mathur, a., and g. M. Foody., multiclass and binary SVM classification:implications for training and classification users., geoscience and remote sensing letters, ieee 5.2 (2008): 241-245.

[13] duan, kai-bo, and s. Sathiya keerthi.,which is the best multiclass SVM method? An empirical study., multiple classifier systems. Springer berlin heidelberg, 2005. 278-285.

[14] takahashi, fumitake, and shigeo abe., d. ”decision-tree-based multiclass support vector machines.” Neural information processing, 2002.iconip’02. Proceedings of the 9th international conference on. Vol. 3.ieee, 2002.

[15] perera, kushani, and dileeka dias., a. An intelligent driver guidance tool using location based services., spatial data mining and geographical knowledge services (icsdm), 2011 ieee international conference on.ieee, 2011.

[16] using the twitter search api — twitter developers. [online]. Available: https://dev.twitter.com/docs/using-search. [accessed: 25-aug-2015].

[17] lee, k., palsetia, d., narayanan, r., patwary, m. M. A., agrawal, a., & choudhary, a. (2011, december). Twitter trending topic classification. In *data mining workshops (icdmw), 2011 ieee 11th international conference on* (pp. 251-258). Ieee.

[18] n. Cristianini and j. Shawe-taylor, *an introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[19] c.-c. Chang and c.-j. Lin, “libSVM -- a library for support vector machines.” [online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libSVM/. [accessed: 22-sep-2015].

[20] safavian, s. Rasoul, and david landgrebe. "a survey of decision tree classifier methodology."
(1990).

[21] d. M. Blei and j. D. Lafferty, “a correlated topic model of science,” *the annals of applied statistics*, vol. 1, no. 1, pp. 17–35, 2007.

[22] y.-c. Wang, m. Burke, and r. E. Kraut, “gender, topic, and audience response: an analysis of user-generated content on facebook,” in *proceedings of the sigchi conference on human factors in computing systems*, 2013, pp. 31–34.

[23] quinlan, j. Ross. "induction of decision trees." *machine learning* 1.1 (1986): 81-106.
[24] j. Weston and c. Watkins. Multi-class support vector machines. Technical report csd-tr-98-04, department of computer science, royal holloway,university of london, 1998.