Incident Detection Using Data from Social Media

Salas, Angelica
Sustainable Transport Research Group
University of Wolverhampton
Wolverhampton, United Kingdom
a.m.salasjones@wlv.ac.uk

Georgakis, Panagiotis; Petalas, Yannis
Sustainable Transport Research Group
University of Wolverhampton
Wolverhampton, United Kingdom
p.georgakis@wlv.ac.uk, i.petalas@wlv.ac.uk

Abstract — **Due to the rapid growth of population in the last 20 years, an increased number of instances of heavy recurrent traffic congestion has been observed in cities around the world. This rise in traffic has led to greater numbers of traffic incidents and subsequent growth of non-recurrent congestion. Existing incident detection techniques are limited to the use of sensors in the transportation network. In this paper, we analyze the potential of Twitter for supporting real-time incident detection in the United Kingdom (UK). We present a methodology for retrieving, processing, and classifying public tweets by combining Natural Language Processing (NLP) techniques with a Support Vector Machine algorithm (SVM) for text classification. Our approach can detect traffic related tweets with an accuracy of 88.27%.**

Keywords— Intelligent Transport Systems, Traffic Incident Detection, Social media analysis, Machine Learning.

I. INTRODUCTION

The increasing level of traffic congestion is one of the main challenges of Intelligent Transportation Systems (ITS). Traffic congestion constitutes a social, economic and environmental issue to modern cities as it can negatively impact travel times, fuel consumption and carbon emissions. In fact, the annual cost resulting from traffic congestion in the UK is set to rise to £21.4 billion by 2030 [1]. Traffic incidents are one of the major causes of traffic congestion. An incident is a non-recurrent event such as accidents, roadworks, disabled vehicles, and other events that disrupt normal traffic. During an incident, the capacity of the network is restricted leading to bottlenecks and delays [2]. It has been indicated that for every minute a traffic incident remains uncleared, it takes around four minutes for traffic to go back to normal [3]. The early detection of these incidents aid traffic management agencies to reduce fatalities, plan appropriate alternatives and alert road users accordingly.

A considerable amount of literature has been published on Automatic Incident Detection (AID) techniques. Most of AID systems rely on data sent from loop detectors and cameras installed on the transportation network [4]. However, it is expensive to install and maintain these sensors, as well as to cover broad areas of transport networks. On the other hand, traffic data may also be collected from mobile sensors such as vehicles equipped with Global Positioning Systems (GPS). Nevertheless, existing algorithms need a large volume of traffic data which will require a proportionally large number of users who are willing to offer GPS data in real time [5].

In recent years, social media has been exploited as a source of real-time data. Twitter is a popular microblogging site with 313 million users posting around 500 million tweets per day, mostly using mobile phones. But then, what makes Twitter unique it's the limitation of no more than 140 characters on tweets, as short information is easier to consume and faster to spread [6]. People use Twitter as a way of expressing their views in different matters, or to comment about real-time events happening around them. In particular, users tend to tweet about the traffic conditions they are currently facing, or to complain about the current state of the transportation network. This information has proven to be very useful for the detection of traffic incidents in real-time.

However, there are some challenges involved with using Twitter data for incident detection. Due to the short nature of tweets, they usually contain informal language, grammatical errors or misspellings [7]. In addition, at least 40% of Twitter content can be considered as useless [8]. As a result, a wide range of Natural Language Processing (NLP) techniques have been implemented to overcome these issues. On the other hand, although Twitter offers free access to their data, it has a limitation on the amount that can be obtained in real-time.

This paper proposes a framework for using data from Twitter to support incident detection in transport networks. The paper is structured as follows. Section II gives a brief overview of different studies for event detection using Twitter data. The methodology used for collecting, preprocessing, and classifying tweets is illustrated in section III. Next, in section IV, results from the experimental stage are being presented. Lastly, conclusions and recommendations for future work are proposed in section V.

II. BACKGROUND

Up to now, several studies have analysed the use of social media for event detection purposes. Abdelhaq et al. [9] presented a system capable of localising events and tracking their evolution over time. In [10], Twitcident was introduced as a framework for filtering, searching and analysing information about real-time events by monitoring information from emergency broadcasting services. Krstajic et al. [11] developed a methodology for the real-time detection of real-world events by identifying keywords whose frequency becomes significantly higher than expected. The

ReDites system [12] performed event detection, tracking and monitoring changes in emotions in discovered events. Weng et al. [13] suggested a detection technique that uses wavelet analysis to build signals for individual words and then cluster them to form events. Sakaki et al. [14] achieved a 96% accuracy in detecting earthquakes by using Support Vector Machine (SVM) as a classifier and Kalman filtering for location estimation. Schulz et al. [15] proposed a machine learning algorithm that combined text classification and semantic enrichment to identify small incidents with an accuracy of 89%.

Concerning traffic incident detection, different methodologies have been developed with the purpose of using Twitter as a sensor [7, 16-18]. D'Andrea et al. [7] presented a real-time traffic event monitoring system from Twitter stream analysis. Their system employed a SVM to classify tweets into Traffic event or not, and it achieved an accuracy of 95.75%. Although they obtained promising results, they only tested the accuracy of the algorithm on the training dataset. In this paper, we aim to test the classification algorithm in a different dataset to show that overfitting of the algorithms is not taking place. In [16], once a relevant tweet was detected using a SVM classifier, a clustering algorithm was used to trace the traffic event evolution. In their approach, they integrated tweets from specific traffic agencies, while we focus on real-time user generated tweets. Tweetple [17] provided traffic reports in real time by classifying traffic tweets into point and link categories using Natural Language Processing (NLP) and syntactic analysis. Unlike our approach, crawled tweets must contain a traffic related word, a verb and a location. In contrast, Gu et al. [18] compared historical Twitter data with existing incidents reports and were able to determine that a small sample of tweets cover most of the incidents reported. They used a Semi-Naïve-Bayes classifier and obtained an accuracy of 90.5% detecting traffic informing tweets. However, both studies fetched tweets using the Twitter REST API, while we propose crawling them through the Twitter Streaming API.

An important aspect to take into consideration is the Twitter statistics in the country of implementation. Studies in the literature explored the influence of Twitter for event detection in countries with high Twitter traffic, such as the United States, Japan, Italy, Portugal and Thailand. According to [19], Twitter has over 15 million active users in the UK, with more than 80% of these users accessing Twitter from their mobile device. This study provides insights into the potential of Twitter for traffic incident detection in the United Kingdom.

III. METHODOLOGY

In this section, we present the approach used to retrieve, process, and classify traffic related tweets. First, we acquired the tweets using the Twitter Streaming API with a geolocation filter. Then, tweets were filtered by road names and traffic related keywords. NLP was used to remove special characters and stop words. Finally, SVM was the algorithm used to classify them into 'traffic' or 'non-traffic' related. The following sections explain in more detail each step of the methodology. Figure 1, shows a tweet going through the different preprocessing stages.

A. Twitter data acquisition

Twitter provides free access to their data through two different tools. With the REST API users can query by keyword and location, and obtain recent or popular tweets. REST API queries are limited to 350 every 15 minutes. On the other hand, the Streaming API gives you access to the real-time stream of public tweets. However, it does not support queries with location and keywords simultaneously, and it only gives access to 1% of the volume of tweets per second at that moment.

For the purpose of working with a real-time stream of data, we selected the Streaming API. The connection was made through Python with a geolocation filter around the West Midlands area. Crawled tweets were filtered by road names and traffic related keywords (e.g. M6, accident, roadworks) using regular expressions.

B. Preprocessing

Tweets are written in a very informal way that usually contains emoticons, special characters, hashtags and so on. Before feeding tweets into the classifier, it is necessary to apply some text mining techniques to remove all these characters. The following text mining techniques were applied to our dataset:

   a) Tokenization: This step consists of transforming the text into a set of tokens (words). It includes the removal of non-alphanumeric characters (e.g. emoticons, hashtags, punctuation), so that the outcome being a set of words. This was accomplished by using regular expressions in Python.
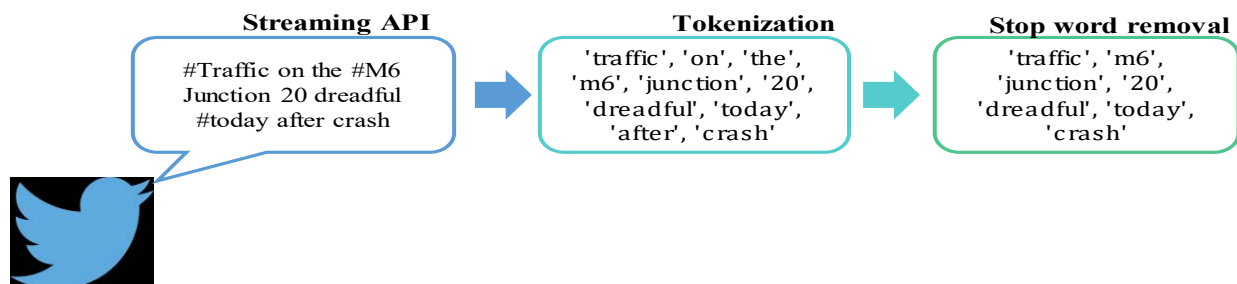

Streaming API

#Traffic on the #M6 Junction 20 dreadful #today after crash

Tokenization

'traffic', 'on', 'the', 'm6', 'junction', '20', 'dreadful', 'today', 'after', 'crash'

Stop word removal

'traffic', 'm6', 'junction', '20', 'dreadful', 'today', 'crash'

Fig. 1: Example of tweet preprocessing

b) Stop word removal: This involves the elimination of those words that do not help to characterize a text (e.g. conjunctions, prepositions, articles). To achieve this, the full list of English stop words from the Natural Language Toolkit (NLTK) library in Python was used.

C. Classification

The last stage of the methodology classifies the preprocessed tweets into 'traffic' or 'non-traffic' related. For this task, a wide range of machine learning algorithms have proven to be very effective. Many researchers in the literature have found SVMs to be the most efficient text classification algorithms for this type of data [7, 14, 15, 20]. For this reason, our classifier was based on a Support Vector Machine (SVM) and it was implemented through the machine learning library Scikit-Learn.

IV. IMPLEMENTATION

In this section, we discuss the implementation of the methodology. First, we present the different datasets used for the experiment. Then, the metrics used to evaluate the performance of the classifier are outlines, followed by the discussion of the results.

A. Datasets

We collected 3,956,871 tweets using the Twitter Streaming API from March 1st, 2017 to May 31st, 2017. From these data, the regular expressions filter was able to extract 13,410 tweets containing traffic related keywords. These tweets were manually labelled and divided into training and test datasets. It is important to state that even though this methodology does not identify the location of the incident, tweets that were traffic related but did not possess any information about the location of the incident, were labelled as non-traffic related tweets. Table I contains some examples of tweets and their corresponding label.

TABLE I.   EXAMPLES OF TWEETS IN THE DATASET

Tweet	Label
40 minute delays on the a5 at Grendon because of temporary lights that #noone seems to be working on roadworks #midlands	Traffic
temporary lights top of manor road wednesbury aren't working creating mayhem for traffic	Traffic
if you are making the trip to #cheltenhamfestival …. #accident on m6 southbound between junctions 8 and 7	Traffic
Success is no accident.. it is hard work, perservance, learning, studying sacrifice, and most of all love of what you are doing	Non-traffic
Absolutely horrendous crash at the f4 championship at donnington today! hope both drivers recover well from this!	Non-traffic

1) Training Dataset

The portion of filtered tweets used to train the algorithm were those from March 1st, 2017 to April 30th, 2017. These tweets were manually labelled into Traffic ('Good') or non-traffic ('Bad'). The final training dataset consisted of 871 traffic related and 871 non-

traffic related tweets. We performed a 10-fold cross validation on the training dataset, using different n-gram ranges.

2) Test Dataset

With the purpose of testing the algorithm with a different dataset, we labelled the remaining filtered tweets from May 1st, 2017 to 31st May 2017 as a testing dataset. We selected 290 traffic related and 290 non-traffic related tweets. The test dataset was used to evaluate the model trained on the training data, using the most accurate n-gram feature from the data validation stage.

B. Evaluation Metrics

In order to evaluate the performance of the classifier, the following four indicators were measured: True Positives (TP), False Positives (TP) True Negatives (TN) and False Negatives (FN). True negative and true positive are non-traffic and traffic related tweets, which were classified correctly as non-traffic and traffic related, respectively. False negative tweets are those traffic tweets that were misclassified as non-traffic, whereas false positive tweets are those non-traffic tweets that were misclassified as traffic tweets. From these values, we can then calculate the following statistical metrics:

a)   Accuracy is the fraction of the classification that is correct. It's calculated by dividing the correctly classified tweets by the total number of tweets (1).

$$accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)} \qquad (1)$$

b)   Precision of a class is the fraction of correctly classified tweets out of all tweets classified to that class (2).

$$Precision = \frac{TP}{TP+FP} \qquad (2)$$

c)   Recall of a class is the fraction of correctly classified tweets out of all tweets that actually belong to that class (3).

$$Recall = \frac{TP}{TP+FN} \qquad (3)$$

d)   F1-score is the harmonic mean of precision and recall (4).

$$F1\ Score = \frac{2 \times P \times R}{P+R} \qquad (4)$$

C. Results

From 13,410 tweets containing traffic related keywords, we identified 1161 traffic tweets. Perhaps one of the more important observations is all the sentiment that can be transmitted with a single tweet. Users tend to express their emotions when describing the current state of transportation services. Hateful language was common amongst users complaining about roadworks, traffic or delays. Similarly, stress could be perceived from users informing about a traffic accident. In contrast, users tend to express their comfort when an incident has been cleared, or when road conditions are good. This information can be very useful for traffic agencies to have an insight of the user's perception of the network.

TABLE II. TRAINING DATA VALIDATION WITH DIFFERENT N-GRAM FEATURES

N-gram	Class	Class precision	Class recall	Class F1	Accuracy
Unigrams	Positive	**88.61%**	**93.5%**	**90.97%**	**90.71%**
	Negative	**93.06%**	**87.93%**	**90.43%**	
Bigrams	Positive	89.24%	77.47%	82.94%	84.02%
	Negative	80.01%	90.61%	84.98%	
Trigrams	Positive	64.52%	46.04%	52.76%	60.12%
	Negative	57.8%	74.27%	64.96%	
Unigrams and Bigrams	Positive	85.47%	95.70%	90.4%	89.7%
	Negative	95.10%	83.66%	89.01%	
Unigrams, Bigrams and Trigrams	Positive	83.25%	96.07%	89.21%	88.35%

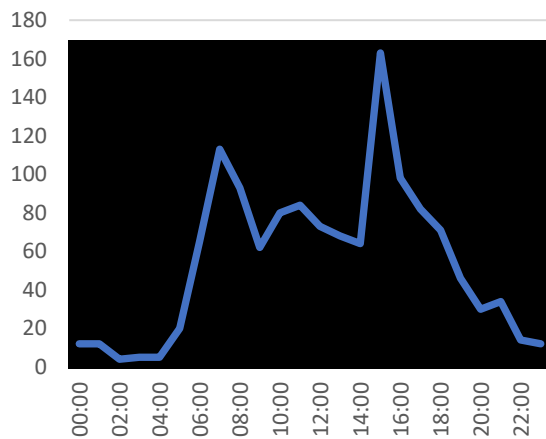
Fig 2: Tweets by time of the day

Fig. 3: Top 100 word-cloud inspections

TABLE III. EXPERIMENT RESULTS

Test data using Unigrams				
Class	Class precision	Class recall	Class F1	Accuracy
Positive	85.81	91.72	88.67	88.28%
Negative	91.11	84.83	87.86	

The period were more traffic tweets were received, was during peak times of the day. These are usually during the times that people are commuting to and from work. In figure 2, we plot the number of traffic related tweets by the time sent.

Table II, contains the results from the validation of the data using different N-gram ranges. It can be observed that unigrams had the highest accuracy (90.71%), while trigrams had the worst performance (60.12%). Based on these results, we evaluated the classifier on the testing dataset using unigrams.

As shown in Table III, classification using the test dataset resulted to an accuracy of 88.28%. These results are similar to those obtained on studies in the literature. It can be observed that the negative class had more *Precision*, but less *Recall* than the positive. This shows that the model identified a higher number of traffic related tweets, but had more precision identifying non-traffic ones.

The top 100 word-cloud inspections for the traffic related tweets are shown in figure 3. Words with larger font sizes in the word cloud are more weighted features than the ones with smaller font sizes. Traffic, accident, and delays being the most used traffic related keywords, while major motorways (e.g.: M6, M1, M5) and arterial roads (e.g.: A38, A27, A259) were the most mentioned locations.

V. CONCLUSIONS

This paper has shown the potential of Twitter for detecting traffic related events in the UK. We implemented a methodology for crawling, pre-processing, and classifying tweets using NLP. and a SVM. The algorithm achieved an overall accuracy of 88.27% on our test dataset.

We were able to identify valuable information regarding traffic events through Twitter. This information can be incorporated into existing incident detection techniques to improve their accuracy. On the other hand, it has been perceived that user emotions play a

strong role in traffic related tweets. By analyzing the sentiment within a tweet, traffic management agencies can have a better understanding of the user's perception of the network.

This is part of an on-going work for an automatic real-time traffic incident detection pipeline using Twitter streams. Future work will include using other NLP techniques to improve the accuracy of the classifier and exploring sentiment analysis within a tweet.

ACKNOWLEDGMENTS

REFERENCES

[1] Inrix (2012). Cost of congestion in the UK. Available: http://inrix.com/press-releases/traffic-congestion-to-cost-the-uk-economy-more-than-300-billion-over-the-next-16-years/.

[2] D. J. Fonseca, T. Brumback, G. Moynihan and H. Fernandes, "Development of a multi-algorithmic platform for traffic incident detection," in 2009, pp. 42.

[3] H. Dia and K. Thomas, "Development and evaluation of arterial incident detection models using fusion of simulated probe vehicle and loop detector data," Information Fusion, vol. 12, (1), pp. 20-27, 2011.

[4] Kinoshita, A. Takasu and J. Adachi, "Real-time traffic incident detection using a probabilistic topic model," Inf Syst, 2015.

[5] Siripanpornchana, S. Panichpapiboon and P. Chaovalit, "Incidents detection through mobile sensing," in 2016, pp. 1-6. DOI: 10.1109/ECTICon.2016.7561238.

[6] F. Atefeh and W. Khreich, "A Survey of Techniques for Event Detection in Twitter," Computational Intelligence, vol. 31, (1), pp. 132-164, Feb, 2015. . DOI: 10.1111/coin.12017.

[7] E. D'Andrea, P. Ducange, B. Lazzerini and F. Marcelloni, "Real-Time Detection of Traffic From Twitter Stream Analysis," Tits, vol. 16, (4), pp. 2269-2283, 2015.

[8] Analytics (2009). Twitter study. Available: www.pearanalytics.com/blog/wp-content/uploads/2010/05/Twitter-Study-August-2009. pdf.

[9] H. Abdelhaq, C. Sengstock and M. Gertz, "Eventweet: Online localized event detection from twitter," Proceedings of the VLDB Endowment, vol. 6, (12), pp. 1326-1329, 2013.

[10] F. Abel, C. Hauff, G. Houben, R. Stronkman and K. Tao, "Twitcident: Fighting fire with information from social web streams," in Proceedings of the 21st International Conference on World Wide Web, 2012, pp. 305-308.

[11] M. Krstajic, C. Rohrdantz, M. Hund and A. Weiler, "Getting there first: Real-time detection of real-world incidents on twitter," 2012.

[12] M. Osborne, S. Moran, R. McCreadie, A. Von Lunen, M. D. Sykora, E. Cano, N. Ireson, C. Macdonald, I. Ounis and Y. He, "Real-time detection, tracking, and monitoring of automatically discovered events in social media," 2014.

[13] J. Weng and B. Lee, "Event Detection in Twitter." Icwsm, vol. 11, pp. 401-408, 2011.

[14] T. Sakaki, M. Okazaki and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in Proceedings of the 19th International Conference on World Wide Web, 2010, pp. 851-860.

[15] Schulz, P. Ristoski, H. Paulheim, T. U. Darmstadt and T. Lab, "I See a Car Crash: Real-time Detection of Small Scale Incidents in Microblogs,". Extended Semantic Web Conference.

[16] Carlos Gutierrez, Paulo Figuerias, Pedro Oliveira, Ruben Costa and Ricardo Jardim-Goncalves, "Twitter mining for traffic events detection," in Jul 1, 2015, pp. 371.

[17] N. Wanichayapong, W. Pruthipunyaskul, W. Pattara-Atikom and P. Chaovalit, "Social-based traffic information extraction and classification," in ITS Telecommunications (ITST), 2011 11th International Conference On, 2011, pp. 107-112. DOI: 10.1109/ITST.2011.6060036.

[18] Y. Gu, Z. Qian and F. Chen, "From Twitter to detector: Real-time traffic incident detection using social media data," Transportation Research Part C: Emerging Technologies, vol. 67, pp. 321 342, 2016.

[19] "The Demographics of Social Media Users in 2016" , 2016. Available: http://www.thinkdigitalfirst.com/2016/01/04/the-demographics-of-social-media-users-in-2016/.

[20] S. Bhosale and S. Kokate, "Traffic Detection Using Tweets on Twitter Social Network" International Journal, Volume 4 (7).