Traffic event detection framework using Social Media

Salas, A., Georgakis, P., Nwagboso, C., Ammari, A. and Petalas, I.
Faculty of Science and Engineering
University of Wolverhampton
Wolverhampton, United Kingdom
a.m.salasjones@wlv.ac.uk, p.georgakis@wlc.ac.uk, c.nwagboso@wlv.ac.uk, a.ammari@wlv.ac.uk, i.petalas@wlv.ac.uk

***Abstract*— Traffic incidents are one of the leading causes of non-recurrent traffic congestions. By detecting these incidents on time, traffic management agencies can activate strategies to ease congestion and travelers can plan their trip by taking into consideration these factors. In recent years, there has been an increasing interest in Twitter because of the real-time nature of its data. Twitter has been used as a way of predicting revenues, accidents, natural disasters, and traffic. This paper proposes a framework for the real-time detection of traffic events using Twitter data. The methodology consists of a text classification algorithm to identify traffic related tweets. These traffic messages are then geolocated and further classified into positive, negative, or neutral class using sentiment analysis. In addition, stress and relaxation strength detection is performed, with the purpose of further analyzing user emotions within the tweet. Future work will be carried out to implement the proposed framework in the West Midlands area, United Kingdom.**

Keywords- Incident Detection, Twitter, Intelligent Transport Systems, Sentiment Analysis, Smart Cities.

I. INTRODUCTION

It is becoming extremely difficult to ignore the increasing volume of road users in modern cities. This has led to instances of recurrent traffic congestions, as people commute daily to/from work. In contrast, social events, adverse weather conditions, accidents and other unexpected incidents are responsible for causing non-recurrent traffic congestions. Up to now, many studies have developed algorithms to detect incidents using real-time data from sensors throughout the transportation network. However, this is not a cost-effective alternative, and it does not work well in conditions where traffic flows can be influenced by random factors [1].

Social network sites are an inexpensive source of human travel information [2]. Mainly, because people post events that are quickly spread by thousands of others in social media. Twitter is a popular microblogging site where users post short messages (called 'tweets') on a real-time basis. Twitter has 313 million active users that are currently producing 500 million tweets per day [3]. Many of these users tweet about specific events as they happen, or shortly after. This makes Twitter data a valuable source of information regarding incidents that differ significantly by type, location, and time [4]. As a result, several authors have studied the influence of Twitter at predicting real-world outcomes such as revenues, accidents, natural disasters and traffic. Unlike road sensors, there is no cost involved in using Twitter as the company offers free access to a subset of their data. Moreover, while algorithmic incident detection only considers changes in frequency counts, Twitter users can describe a wide variety of traffic events [5].

There are several challenges faced when identifying events using social media. First, there is the massive and unpredictable volume of data. Then, much of the content of Twitter may not be related to any event, or could be referring to an event that is not necessarily happening at the current instant. Also, as Twitter messages cannot exceed 140 characters, they usually have typos or grammatical errors [1]. In fact, Analytics [6] concluded that 40% of the tweets could be considered as 'pointless babble'. Finally, although Twitter data is free to access, there is a limit on the number of tweets that can be obtained in real-time.

This paper proposes an architectural framework for the identification of traffic events using Twitter data. The first section contains a brief overview on the impact of Twitter in predicting real-world outcomes. The paper then describes methods and techniques for traffic incident detection using Twitter. In addition, the potential of integrating user emotions into incident detection is discussed. Next, the proposed methodology for traffic event detection using Twitter-based content is described. Finally, conclusions are drawn, and future research is proposed.

II. BACKGROUND AND RELATED WORK

A. Twitter as a sensor

In recent years, there has been an increasing amount of literature on the use of social media data for various purposes. For instance, several studies have used it as a way of forecasting sales, polls, ratings and even diseases. Asur and Huberman [7] developed a model about the rate at which tweets can help make better market predictions. As an example, they used Twitter to forecast box office revenues for movies and discovered that such predictions are truly better than those produced by information markets. In contrast, Tumasjan et al. [8] successfully evaluated the ability of Twitter to predict the German election polls of 2009. Moreover, in [9] it was proposed a Twitter-based influenza detection method, which results outperformed Google Flu trends.

Some authors have exploited Twitter information to detect breaking news and events. Different machine learning algorithms have been proposed as a way to accomplish this.

Agarwal et al. [10] suggested an architecture based on Named Entity recognition and part-of-speech tagger to identify local news on Twitter. They used their proposed system to detect fire-in-factory and Labor-Strike events and were able to achieve an 80% of success rate using a Naïve Bayes classifier.

Li et al. [11] designed a system for the detection of Crime and Disaster related Events (CDE) from real-time incoming tweets. They created an interface where users could input a location and a time period, and the system will return CDE tweets ranked by their importance. Users could also input a keyword, and the system would find the spatial and temporal patterns for the query. The system accomplished an accuracy of 80% for detecting incident related tweets.

Sakaki et al. [12] proposed real-time event detection using social sensors. They integrated semantic analysis using a Support Vector Machine (SVM) to classify tweets, and applied Kalman and particle filtering for location estimation, creating an algorithm to detect a target event, in this case for earthquake prediction. They could identify 96% of earthquakes recorded by the Meteorological Agency in Japan.

Abel et al. [13] presented Twitcident, a system for searching, tracking, analyzing and visualizing information about real-time incidents. When an incident was detected from emergency services, Twitcident initiated a query for profiling the incident and collecting Twitter messages. These messages were processed using Named Entity Recognition (NER) to link them to a location and then classified into different event categories.

B. Twitter in transportation

Intelligent Transport Systems (ITS) is another area where a considerable amount of literature has grown around the influence of user-based content into traffic information. For example, in [14] it was suggested the incorporation of social media in traffic forecasting. They found a correlation between traffic measurement and tweet counts and incorporated tweets in traffic prediction via linear regression. By comparing linear regression using only traffic flow against the one that included social media, they were able to demonstrate the effectiveness of integrating Twitter data on traffic prediction. Another way of incorporating social media was by creating a system architecture based on social media information. To that extent, Wibisono et al. [15] developed an intelligent system capable of obtaining traffic information data from the Twitter accounts of official channels and disseminating this information to the users in their mobile devices, using neural networks. However, while this is a useful way of displaying traffic information, tweets from official accounts (e.g. police department, highway agencies) are not relevant to proactively warn users about real time incidents, since these are robots that broadcast data that has already been analyzed.

Few researchers have been focusing on the detection of real-time traffic events based on Twitter data. Wanichayapong et al. [16] implemented a unique traffic word dictionary-tokenizer of four categories: place, verb, ban and preposition. After crawling, tokenizing and filtering the tweets, it then geolocated the tweets using latitude and longitude from the place dictionary and Google geocoding. Similarly, in [1] it was developed an adaptive data acquisition following an iterative process. It consisted of queries to an API with a dictionary of initial keywords that iteratively expands the dictionary until the acquired data converges. A geocoder identified the location using road names. Although these studies showed promising results, they only dealt with historical Twitter data. In addition, one of the main challenges faced was the need for improvement in the geolocation techniques. Due to the short and informal nature of tweets, it's hard to identify locations from abbreviations or places with the same name.

1) User emotion analysis for incident detection

Sentiment analysis is the area of Natural Language Processing that studies opinions, sentiments, and emotions from the written language [17]. It analyses a text and estimates the polarity of it as positive, negative, or neutral. Microblogging websites are valuable sources for sentiment analysis as users tend to share their opinions on different subjects [18]. In recent years, people use social media to comment and complaint about events and problems of their social life [5]. This information can be beneficial to support traffic management agencies for two reasons. First, by assigning polarity to traffic tweets, it will help to identify if it's a negative occurrence (accident, roadworks, delays) or a positive one (lanes re-opened, roadworks finished). On the other hand, there is the stress factor. When people are commuting, factors such as congestions, the state of the road and any other unexpected incidents are likely to cause stress. By detecting the level of stress within a traffic related tweet, agencies can identify the factors causing non-recurrent congestions. Apart from [5], there is a general lack of research in incorporating user emotions for traffic related purposes.

III. PROPOSED ARCHITECTURAL FRAMEWORK

In this section, the methodology for a Twitter-based traffic event detection system is discussed. The proposed framework has the potential to retrieve and analyze real-time traffic incidents using Twitter data. The following processing pipeline is defined (See Fig. 1):

a) Gather real-time tweets through the Twitter Streaming API using a geolocation filter.
b) Text tokenization to process tweets for further analysis.
c) Identify traffic related tweets and classify into different event categories.
d) Sentiment analysis to allocate a positive or negative polarity.
e) Strength and relaxation strength detection.
f) Extract tweet location using name entity recognition and entity disambiguation.

While current techniques follow a similar approach, the proposed system aims to enhance existing methodologies by: 1) Using a real-time twitter feed, 2) Improving geolocation techniques, 3) Integrating user emotion analysis.

The proposed system will be implemented using two data processing pipelines using big-data storage and processing tools from the Hadoop ecosystem. In particular, Kafka and Flume will be used as the messaging platform, HBase as a permanent storage solution and Spark as the distributed processing tool. The first envisaged pipeline, will be offline, and will employ Spark and python machine learning algorithms for processing the tweets in bulk and training the different algorithms. Messages from Twitter will reach the HBase through Kafka and Flume. Through the use of connectors, data will be consumed by the machine learning libraries stated above. The second pipeline, will be implemented using Spark Streaming and will consume real-time tweets received in Kafka. This approach will allow real time tweets to be stored and processed through the pipeline. The following sections will describe the process at each step of the framework.

A. *Twitter API*

Twitter offers free access to its data through two different types of API. The REST API allows you to post Tweets, follow people, create lists, performing searches and more. On the other hand, the Streaming API is used to receive a real-time stream of the public tweets. Thus, it requires having an uninterrupted connection. However, these API's have a limit. On the REST API, applications can make 350 queries every 15 minutes. On the other hand, every Twitter account can obtain up to 1% of the volume of tweets per streaming second, using the Streaming API.

Unlike the Streaming API, the REST API allows to filter by location and keyword, being this the main reason that has led researchers in the literature to use the REST API for their studies. However, the Streaming API has the advantage of receiving tweets as they are posted. For the purpose of providing a real-time event detection, the Streaming API is the most efficient source of data. For the experimental stage of this study, the connection to the streaming API will be made through python, with a location filter using the West Midlands coordinates. Data received from Twitter will be pushed to a topic in Kafka for further processing.

B. *Text mining*

As it has been mentioned before, a tweet can be about anything. Since tweets will only be filtered by location, the data received will contain a wide amount of noise tweets (e.g. non-traffic related, spams). The objective of this stage is to process tweets using Natural Language Processing (NLP) techniques before they are fed into the classifier. This process includes tokenizing the tweets, removing stop words and special characters. It will also filter the tweets by specific traffic related keywords using regular expressions. This is with the purpose of removing as many noise tweets as possible before the next stage.

C. *Tweet classifier*

After obtaining a sample of tweets, the next step is to filter this information to get only traffic related data. There are a broad range of text classification algorithms, which given a training data set can successfully classify into the correct category. Support Vector Machine (SVM) has been selected as the classification algorithm for this stage. SVM has gained popularity as one of the most efficient machine learning algorithms. In fact, several authors in the literature has successfully used SVM as their text classification algorithm [5, 9, 19-21].

Official accounts (e.g. 'HighwaysWMIDS') contain validated traffic tweets that could serve as training data for our algorithm. Also, traffic related tweets generated by human users will also be part of the positive sample. In contrast, it is also needed a set of negative tweets. Due to the high demand of irrelevant messages on Twitter, it is easy to select a random sample of negative tweets to feed the algorithm.

Tweets will be first classified into traffic and non-traffic related tweets. Then, another classification algorithm will take place to sort them into the following categories:

a) Roadworks
b) Accidents
c) Weather
d) Social events

As an example of the process that is going to follow our methodology, we have extracted a small set of tweets from the United Kingdom. Table I has an example of the classification process.

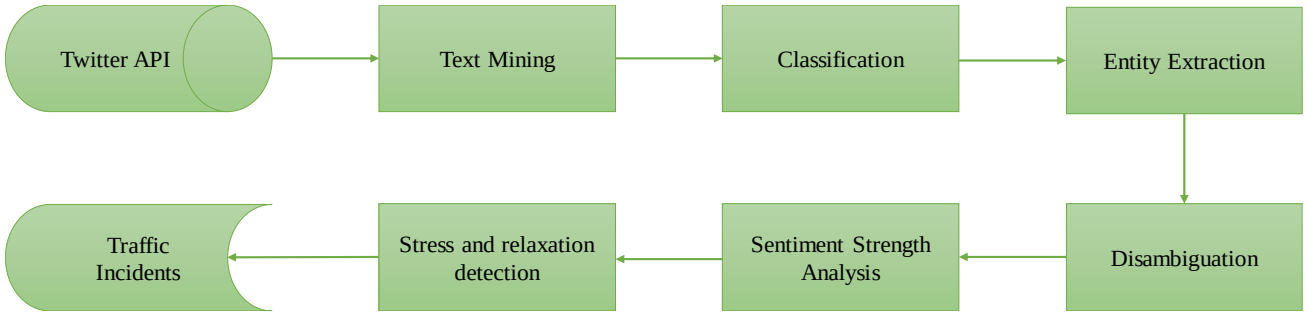

Figure 1. Proposed framework

TABLE I: EXAMPLE OF CLASSIFICATION

Tweet	Classifier
Highlight of the day, Catthorpe Interchange. Really good now, roadworks finished, 10/10 would use again.	**Traffic related** Roadworks
Starbucks crying because they have too much traffic. People sitting not buying. Are they just waiting for the bar across the street to open?	**Non-Traffic related**
#traffic on the #M6 junction 20 dreadful #today after crash	**Traffic Related** Accidents
Anybody have any super secret high-converting, mass traffic method they would like to share with the class?	**Non-Traffic Related** Accidents
The M6 northbound before and after Junction 18 in the roadworks is awful - lanes 1 and 2 are full of massive pot holes!	**Traffic related** Roadworks

D. *Geolocation*

On this phase, the already classified tweet sample will be allocated to a particular location. Some users associate a location within the tweet, but it does not necessarily represent the actual location of the event. Also, most of the time this information is not provided. Name Entity Recognition (NER) is used to label words in a text into entities (organization, person, and location). However, when using short and informal messages, one concept can refer to more than one place. To tackle this issue, a wide variety of studies have proposed Entity Disambiguation to link a concept to a unique location, through a knowledge base. Wikipedia has been a popular choice as a knowledge base, in which each page is treated as a named entity. In fact, [5] and [13], employed Wikipedia as the knowledge base in their studies.

The proposed system will use a NER to identify the location entity within tweets, and will then use Entity Disambiguation to link the reference to a unique location. In order to find the most accurate combination, different NER (e.g. Stanford NER) and knowledge base (e.g. Wikipedia) will be tested. Table II has an example of the process of geolocation using the previous set of classified tweets.

TABLE II: EXAMPLE OF GEOLOCATION

NER Tagger	Entity Link
Highlight of the day, Catthorpe Interchange. Really good now, roadworks finished, 10/10 would use again.	Catthorpe_Interchange
#traffic on the #M6 junction 20 dreadful #today after crash	M6_J20
The M6 northbound before and after Junction 18 in the roadworks is awful - lanes 1 and 2 are full of massive pot holes!	M6_J18

E. *Sentiment analysis*

This step of the methodology consists of assigning a polarity to the tweets. To achieve this, Sentiment strength detection will be performed. Sentiment strength detection predicts the strength of positive or negative sentiment within a text [22]. Thelwall et al. [23] presented SentiStrength, a classifier that uses additional linguistic information to detect sentiment strength in short informal text. For each text, the SentiStrength will output two integers: one for positive sentiment strength and another for negative sentiment strength. The scores range from 1 to 5 for positive sentiment and -1 to -5 for the negative one. As an example, Table III depicts some traffic related, and their sentiment strength analysis performed with the online version of SentiStrenght 2. While the first tweet is a compliment thus it has a positive polarity, the other two have a negative one due to complaints about traffic and roadworks.

F. *Stress and relaxation analysis*

Finally, for the purpose of analysing the level of stress or relaxation within the tweet, TensiStrength will be used. TensiStrength uses a lexical approach to detect indicators of stress and relaxation expressed in short text messages [24]. Similarly to SentiStrength, TensiStrength will output two values on a scale from 1 to 5 for relaxation and -1 to -5 for stress. Table III shows the results of the stress and relaxation analysis made with the online version of TensiStrength. It can be perceived that in the first tweet, the user is obviously happy and relaxed over the state of the network, thus the

TABLE III: SENTIMENT AND STRESS ANALYSIS USING SENTISTRENGTH AND TENSISTRENGTH

Tweet	SentiStrength		TensiStrength	
	+	-	+	-
Highlight of the day, Catthorpe Interchange. Really good now, roadworks finished, 10/10 would use again.	4	-1	4	-1
#traffic on the #M6 junction 20 dreadful #today after crash	1	-4	1	-4
The M6 northbound before and after Junction 18 in the roadworks is awful - lanes 1 and 2 are full of massive pot holes!	1	-4	1	-4

relaxation score is high (3) and shows not to be stressed (-1). As for the other tweets, they are obviously complaining, which makes their stress level high.

IV. CONCLUSIONS

This paper has proposed a framework for the identification of real-time traffic related events using Twitter data. We presented a methodology that will obtain real time tweets through the Streaming API with a geolocation filter. Acquired tweets will be processed using NLP. Next, it will proceed to classify them into traffic and non-traffic tweets, as well into different event categories. The classifier will be trained using tweets from official accounts (robots). Then, traffic tweets will be geolocated using NER and entity disambiguation. The geolocated tweets will then be assigned a polarity. Finally, stress and relaxation will be detected.

This method aims to enhance existing methodologies by using real-time tweets, instead of historical data. In addition, geolocation techniques are expected to be improved using NER and disambiguation. Finally, by incorporating sentiment analysis and stress and relaxation strength detection, it could be identified the nature of the event and the user's perspective.

A further experimental study to test the proposed framework will be carried out in the West Midlands area, United Kingdom.

ACKNOWLEDGMENTS

This research was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 636160-2, the Optimum project www.optimumproject.eu.

This paper is part of a PhD sponsored by the Dominican Republic's Ministry of Education (MESCyT).

REFERENCES

[1] 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 67*pp. 321 342. 2016. . DOI: 10.1016/j.trc.2016.02.011.

[2] D. Pathania and K. Karlapalem. Social network driven traffic decongestion using near time forecasting. Presented at Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems. 2015.

[3] Twitter, 2016. [online] Available at: <https://business.twitter.com/basics/what-is-twitter/>.

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

[5] Z. Kokkinogenis, J. Filguieras, S. Carvalho, L. Sarmento and R. J. Rossetti. Mobility network evaluation in the user perspective: Real-time sensing of traffic information in twitter messages. 2015.

[6] Analytics, 2009. Twitter study. [online] Available at: <www.pearanalytics.com/blog/wp-content/uploads/2010/05/Twitter-Study-August-2009. pdf>.

[7] S. Asur and B. Huberman. Predicting the future with social media. Presented at Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM

[8] A. Tumasjan, T. O. Sprenger, P. G. Sandner and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *Icwsm 10*pp. 178-185. 2010.

[9] E. Aramaki, S. Maskawa and M. Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. Presented at Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2011.

[10] P. Agarwal, R. Vaithiyanathan, S. Sharma and G. Shroff. Catching the long-tail: Extracting local news events from twitter. Presented at Icwsm. 2012.

[11] R. Li, K. H. Lei, R. Khadiwala and K. C. C. Chang. TEDAS: A twitter-based event detection and analysis system. Presented at 2012 IEEE 28th International Conference on Data Engineering. 2012. DOI: 10.1109/ICDE.2012.125.

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

[13] F. Abel, C. Hauff, G. Houben, R. Stronkman and K. Tao. Twitcident: Fighting fire with information from social web streams. Presented at Proceedings of the 21st International Conference on World Wide Web. 2012.

[14] J. He, W. Shen, P. Divakaruni, L. Wynter and R. Lawrence. Improving traffic prediction with tweet semantics. Presented at Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. 2013.

[15] A. Wibisono, I. Sina, M. A. Ihsannuddin, A. Hafizh, B. Hardjono, A. Nurhadiyatna and W. Jatmiko. Traffic intelligent system architecture based on social media information. Presented at Advanced Computer Science and Information Systems (ICACSIS), 2012.

[16] N. Wanichayapong, W. Pruthipunyaskul, W. Pattara-Atikom and P. Chaovalit. Social-based traffic information extraction and classification. Presented at ITS Telecommunications (ITST), 2011. DOI: 10.1109/ITST.2011.6060036.

[17] B. Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies 5(1)*, pp. 1-167. 2012.

[18] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. Presented at LREc. 2010.

[19] F. Yuan and R. L. Cheu. Incident detection using support vector machines. *Transportation Research Part C: Emerging Technologies 11(3–4)*, pp. 309-328. 2003. DOI: https://doi.org/10.1016/S0968-090X(03)00020-2.

[20] S. Bhosale and S. Kokate. Traffic detection using tweets on twitter social network.

[21] A. 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.

[22] M. Thelwall. TensiStrength: Stress and relaxation magnitude detection for social media texts. *Information Processing & Management 53(1),* pp. 106-121. 2017.