

## Real-time traffic event detection using Twitter data

Item Type	Journal article
Authors	Jones, Angelica Salas;Georgakis, Panagiotis;Petalas, Yannis;Suresh, Renukappa
Citation	Jones, A. S., Georgakis, P., Petalas, Y. & Suresh, R.(2018) 'Real-Time Traffic Event Detection Using Twitter Data', Infrastructure Asset Management, 5 (3) pp. 77-84
DOI	<a href="https://doi.org/10.1680/jinam.17.00022">10.1680/jinam.17.00022</a>
Publisher	ICE Publishing
Journal	Infrastructure Asset Management
Download date	2026-03-16 17:46:51
License	<a href="https://creativecommons.org/licenses/by-nc-nd/4.0/">https://creativecommons.org/licenses/by-nc-nd/4.0/</a>
Link to Item	<a href="http://hdl.handle.net/2436/621294">http://hdl.handle.net/2436/621294</a>

# Infrastructure Asset Management

## Real-time traffic event detection using Twitter: A case study

--Manuscript Draft--

<b>Manuscript Number:</b>	IAsMa-D-17-00022
<b>Full Title:</b>	Real-time traffic event detection using Twitter: A case study
<b>Article Type:</b>	Themed Issue: Highway infrastructure
<b>Corresponding Author:</b>	Angelica Milagros Salas Jones, MSc University of Wolverhampton Faculty of Science and Engineering Wolverhampton, West Midlands UNITED KINGDOM
<b>Corresponding Author Secondary Information:</b>	
<b>Corresponding Author's Institution:</b>	University of Wolverhampton Faculty of Science and Engineering
<b>Corresponding Author's Secondary Institution:</b>	
<b>First Author:</b>	Angelica Milagros Salas Jones, MSc
<b>First Author Secondary Information:</b>	
<b>Order of Authors:</b>	Angelica Milagros Salas Jones, MSc Panagiotis Georgakis, PhD Ioannis Petalas, PhD Renukappa Suresh, PhD
<b>Order of Authors Secondary Information:</b>	
<b>Abstract:</b>	Incident detection is an important component of Intelligent Transport Systems (ITS) and plays a key role in urban traffic management and provision of traveller information services. Due to its importance, a wide number of researchers have developed different algorithms for real-time incident detection. However, the main limitation with existing techniques is that they do not work well in conditions where random factors could influence traffic flows. Twitter is a valuable source of information as its users post events as they happen or shortly after. Therefore, Twitter data has been used to predict a wide variety of real-time outcomes. This paper aims to present a methodology for a real-time traffic event detection using Twitter. Tweets are obtained through the Twitter Streaming Application Programming Interface (API) in real-time with a geolocation filter. Then, we used Natural Language Processing (NLP) techniques to process the tweets before they are fed into a text classification algorithm that identifies if its traffic related or not. We implemented our methodology in the West Midlands region in the UK, and obtained an overall accuracy of 92.86%.
<b>Additional Information:</b>	
<b>Question</b>	<b>Response</b>
Please enter the total number of words in your main text.	3341
Please enter the number of figures, tables and photographs in your submission.	Tables: 7 Figures: 2

1  
2 **Title: Real-time traffic event detection using Twitter: A case study**  
3  
4  
5

6 **Author 1**

- 7  
8
  - Angelica Salas Jones, PhD student
  - Faculty of Science and Engineering, University of Wolverhampton, United Kingdom

9  
10

11 **Author 2**

- 12
  - Dr. Panagiotis Georgakis
  - Faculty of Science and Engineering, University of Wolverhampton, United Kingdom

13  
14

15 **Author 3**

- 16
  - Dr. Ioannis Petalas
  - Faculty of Science and Engineering, University of Wolverhampton, United Kingdom

17  
18  
19

20 **Author 2**

- 21
  - Dr. Renukappa Suresh
  - Faculty of Science and Engineering, University of Wolverhampton, United Kingdom

22  
23  
24  
25

26 **Full contact details of corresponding author.**

27 Email address: [a.m.salasjones@wlv.ac.uk](mailto:a.m.salasjones@wlv.ac.uk)

28  
29 Mobile number: 447481475793  
30

31 **Number of words in main text:** 3341

32  
33 **Number of tables:** 7

34  
35 **Number of figures:** 2  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63

64 This is an example created from parts of other articles, it is not designed to be read for sense.  
65

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

**Abstract (150 words)**

Incident detection is an important component of Intelligent Transport Systems (ITS) and plays a key role in urban traffic management and provision of traveller information services. Due to its importance, a wide number of researchers have developed different algorithms for real-time incident detection. However, the main limitation with existing techniques is that they do not work well in conditions where random factors could influence traffic flows. Twitter is a valuable source of information as its users post events as they happen or shortly after. Therefore, Twitter data has been used to predict a wide variety of real-time outcomes. This paper aims to present a methodology for a real-time traffic event detection using Twitter. Tweets are obtained through the Twitter Streaming Application Programming Interface (API) in real-time with a geolocation filter. Then, we used Natural Language Processing (NLP) techniques to process the tweets before they are fed into a text classification algorithm that identifies if its traffic related or not. We implemented our methodology in the West Midlands region in the UK, and obtained an overall accuracy of 92.86%.

**Keywords**

Transport management, Transport planning, Information Technology, Infrastructure planning.

1     **1     Introduction**

2     2     With 84% of people travelling by car at least once or twice a week (DfT, 2017), the need for  
3     3     more efficient traffic monitoring systems has become essential. Increases in traffic leads to  
4     4     more interaction between road users, and therefore, heightened likelihood of traffic incidents.  
5     5     Traffic incidents are non-recurrent events such as accidents, broken down vehicles, road  
6     6     maintenance, social activities and other unexpected events that affect the normal traffic flow.  
7     7     These incidents contribute to delays and have serious effects on safety, air pollution and the  
8     8     cost of travel. In order to reduce these adverse effects, incidents need to be detected and  
9     9     cleared as promptly as possible. For these reasons, Automatic Incident Detection (AID) has  
10    10    been widely studied in the last decades. AID is an important part of Intelligent Transportation  
11    11    Systems (ITS), and is designed to automatically detect incidents, or unexpected situations  
12    12    causing congestions in the transport network (D'Andrea, Marcelloni, 2017).

13  
14    14    Traditional AID systems exploit data collected from loop detectors and surveillance cameras on  
15    15    the transport network. These devices measure traffic data such as flow, speed, and occupancy  
16    16    for a given period of time. AID algorithms can then detect traffic incidents from anomalies found  
17    17    on these data. However, it is quite expensive to cover broad areas due to the high cost of  
18    18    installing and maintaining these types of devices. In contrast, this approach has poor  
19    19    performance on arterial roads, where traffic flows can be influenced by random factors. Recently,  
20    20    there has been a trend towards considering other data sources technologies, such as GPS and  
21    21    cellular geolocation systems (Parkany, Xie ,2005). Nevertheless, these approaches are limited  
22    22    by low sampling rate and high measurement errors (Siripanpornchana, Panichpapiboon &  
23    23    Chaovalit, 2016).

24  
25    25    It would be ideal if users could report incidents in real-time, as they are the ones that can  
26    26    provide more accurate information about the incident. In fact, virtually any person witnessing or  
27    27    involved in any event is able to disseminate it in real-time through microblogs (Atefeh, Khreich  
28    28    ,2015). Microblogging sites, particularly Twitter, have become a popular source of all kinds of  
29    29    information. Twitter is an online social network with over 300 million users posting short  
30    30    messages (tweets) on a real-time basis. Many of these tweets are about real-time events as

1 31 they happen, or shortly after. For instance, users turn to Twitter to report traffic incidents or to  
2 32 describe the traffic situation they are currently in, making Twitter a real-time source of human  
3  
4 33 travel information. For this reason, Twitter data has proven to be very useful for detecting traffic  
5  
6 34 events. In addition, people use Twitter to express their opinion and emotions on a certain  
7  
8 35 subject. Particularly, traffic related tweets tend to be filled with emotions as users usually  
9  
10 36 complain about the state of the network, or are stressed about a traffic incident. It is important to  
11  
12 37 include this subjective data into traffic incident detection, as it can give a better understanding of  
13  
14 38 the user perception of the transport network (Kokkinogenis et al., 2015).  
15

16 39  
17  
18 40 Using Twitter based data input for traffic incident detection overcome some of the issues faced  
19  
20 41 with conventional devices sensors. First, there is no cost involved as Twitter grants free access  
21  
22 42 to a subset of their data. Second, while traditional sensors only detect changes in traffic  
23  
24 43 measures, a tweet usually contains more detailed information about the traffic event taking  
25  
26 44 place. Third, users can tweet from any location, covering broader areas of the transport  
27  
28 45 network. Lastly, traditional approaches fail to provide an insight into the user's perception of the  
29  
30 46 flaws of the transport network. Nevertheless, there are some challenges involved with using  
31  
32 47 Twitter for incident detection. Traditional text mining techniques do not work well on tweets, as  
33  
34 48 they often contain emoticons, typos, and grammatical errors. Hence, with more than 500 million  
35  
36 49 tweets per day, it is difficult to detect useful information from noise (e.g.: non-traffic related,  
37  
38 50 spams). Finally, although Twitter data is free to access, there is a limitation on the amount that  
39  
40 51 can be obtained in real-time.  
41

42 52  
43  
44 53 This paper presents a methodology for traffic event detection by fetching, filtering and  
45  
46 54 processing public tweets in real-time. The procedure uses Natural Language Processing (NLP)  
47  
48 55 techniques to process the tweets before they are fed into a machine learning classifier. This is  
49  
50 56 an initial attempt to examine the accuracy and potential of incident detection through Twitter.  
51  
52 57 For this reason, although the methodology can be applied in real-time, we implemented it using  
53  
54 58 historical twitter data. The remaining part of the paper proceeds as follows. We first give an  
55  
56 59 overview of different implementations of Twitter for incident detection. The methodology for  
57  
58 60 crawling, processing and classifying tweets is described in section 3. In section 4, results and  
59  
60  
61  
62  
63  
64  
65

1 61 findings from the experimental implementation are presented. Finally, conclusions and  
2 62 recommendations are drawn.  
3

4 63

5  
6 64 **2. Related work**

7  
8 65 To date, several studies have analysed the use of Twitter for event detection. (Sakaki, Okazaki  
9 66 & Matsuo, 2010) were amongst the first to propose a methodology to detect events using  
10 67 Twitter. They were able to detect earthquakes with a 96% probability by using a Support Vector  
11 68 Machine (SVM) for classification, and a Kalman filtering for location estimation. (Abel et al.  
12 69 ,2012) developed a framework for filtering, searching, and analysing real-time world incidents  
13 70 from social web streams. Their system could collect Twitter messages, related pictures, and  
14 71 videos to the specific incident. In contrast, (Krstajic et al., 2012) detected potential events by  
15 72 monitoring the frequency of individual keywords and for those with unexpected high frequency  
16 73 values, it calculated additional scores that could help on describing the event. (R. Li et al., 2012)  
17 74 presented TEDAS, a system for detecting, ranking and locating crime and disaster related  
18 75 events by exploring information from Twitter. Similarly, Eventweet focused on detecting events  
19 76 by adopting a continuous analysis of the most recent tweets within a time frame (Abdelhaq,  
20 77 Sengstock & Gert, 2013). Lastly, (Osborne et al., 2014) introduced a system for monitoring  
21 78 security relevant events, and tracking changes in emotions over time.  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38

39 80 Concerning traffic incident detection, a number of researchers have presented different  
40 81 methodologies to exploit twitter data as a sensor. For instance, (Gutierrez et al., 2015)  
41 82 described an approach for integrating tweets from different traffic agencies in the UK, with the  
42 83 purpose of notifying drivers about the status of the network in real-time. Our approach  
43 84 concentrates on user generated tweets, rather than official traffic agencies tweets. (Schulz,  
44 85 Ristoski & Paulheim, 2013) presented a methodology for the identification of small scale  
45 86 incidents by combining text classification techniques with a machine learning algorithm. Their  
46 87 outcome was to identify car crashes, while we aim to detect any event that can influence the  
47 88 traffic condition. (D'Andrea et al., 2015) and (Gu, Qian & Chen, 2016) filtered tweets by traffic  
48 89 related keywords, and used a machine learning algorithm to classify them into traffic related or  
49 90 not. (D'Andrea et al., 2015) obtained promising results on the accuracy of the classifier, but they  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 91 tested it only on the training dataset. In this paper, we test the accuracy of the algorithm on a  
2 92 different dataset, with the purpose of showing that the model is not overfitted to the training  
3  
4 93 data. In addition, these studies used the Twitter REST API to crawl tweets, while we propose  
5  
6 94 fetching them through the Twitter Streaming API. Lastly, existing research for mining user  
7  
8 95 generated tweets for traffic incident has been applied in the United States, Italy, and Germany.  
9  
10 96 In this study, we employed the methodology in the West Midlands region, in the United  
11  
12 97 Kingdom.

### 14 98 15 16 99 **3. Methodology**

17 100  
18  
19 101 In this section, we describe the methodology used to identify traffic incident information from  
20  
21 102 twitter data. Figure 1 shows the system architecture and the different tools used on each phase.  
22  
23 103 We fetched tweets using the Twitter Streaming API with a geolocation filter. Road names and  
24  
25 104 traffic related words were used as keywords as an additional filter. Next, we trained five  
26  
27 105 machine learning algorithms with different word n-grams and tested their classification accuracy.  
28  
29 106 Finally, we selected the most accurate n-gram features, and evaluated each classifier on the  
30  
31 107 test dataset.

#### 33 108 34 35 109 **3.1 Fetching Twitter data**

36  
37 110 The first step entails the extraction of raw tweets using the Twitter Streaming API. One of the  
38  
39 111 limitations of using the Streaming API is that it does not allow to filter by location and keyword.  
40  
41 112 This is the main reason why authors in the literature have used the Twitter Search API for their  
42  
43 113 studies. However, the Search API searches against a sample of recent tweets focusing on their  
44  
45 114 relevance, while the streaming API gives real-time access to the streams of public data flowing  
46  
47 115 through Twitter (Twitter, 2017). For this reason, we selected the Streaming API for this stage.  
48  
49 116 Twitter API's are supported in many programming languages through a wide variety of libraries.  
50  
51 117 In our approach, we made an uninterrupted connection to the Streaming API with a geolocation  
52  
53 118 filter, using the Tweepy library in Python.

#### 54 119 55 56 120 **3.2 Traffic keywords filtering**

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

121 On this stage, we perform an additional filter to obtain the tweets mentioning traffic related  
122 words. To this end, we created a dictionary of highways, arterials, roads, and incident related  
123 words. We used regular expressions to filter the acquired tweets using the keyword dictionary.  
124 Table 1 shows an example of some of the keywords and road names used for filtration. In  
125 addition, we used this stage to remove all retweets, as they do only contain repeated  
126 information.

Traffic keywords	Road names
Accident	M6
Congestion	A449
Roadworks	M42
Traffic delays	A41
Stuck traffic	M5

**Table 1: Keywords for filtration**

127  
128

129 **3.3 Pre-processing**

130 Due to their informal nature, tweets usually contain mentions, hashtags, links, special  
131 characters and emoticons. This information needs to be removed before tweets are fed into the  
132 classifier. In the following sections, the text mining techniques applied to the dataset are  
133 described in detail.

134 *3.3.1 Tokenisation*

135 Tokenisation is the task of transforming a character sequence into pieces, called tokens, and at  
136 the same time removing certain characters. There are a wide range of tokenization tools,  
137 however they fail to recognise special tweet features such as @mentions, emoticons, URLs and  
138 hashtags as individual tokens. For this reason, we employed a pre-processing chain based on  
139 regular expressions that considers all these aspects. During this step, the tokeniser removes  
140 mentions, hashtags, URLs, punctuation and emoticons, and splits each tweet into a set of words  
141 ('tokens').

142 *3.3.2 Stop word removal*

143 Stop words are those common words that have little value in helping characterise a text, such  
144 as articles, conjunctions, and prepositions. These words are not very meaningful when deciding

1 145 if a tweet is traffic related or not, thus not valuable to be fed into a machine learning classifier. In  
2 146 our approach, the full list of English stop words from the Natural Language Toolkit (NLTK)  
3  
4 147 library was used to remove stop words from the set of tokens.  
5

6 148

### 8 149 **3.4 Classification**

10 150 Once tweets have been pre-processed, they were classified into traffic related or not. To  
11  
12 151 achieve this, a machine learning algorithm was employed. Studies in the literature have  
13  
14 152 employed and compared a wide range of text classification algorithms for incident detection  
15  
16 153 using Twitter data (Schulz, Ristoski & Paulheim, 2013, D'Andrea et al., 2015, Wanichayapong et  
17  
18 154 al., 2011, Gu, Qian & Chen, 2016). For this study, we compared a Ridge Classifier (RC), Naïve  
19  
20 155 Bayes (NB), k-Nearest Neighbour (kNN), Multilayer Perceptron (MLP) and a Support Vector  
21  
22 156 Machine (SVM). We combined and evaluated the classifiers with different word n-gram features  
23  
24 157 on the training dataset, and selected the most accurate parameters on each algorithm for the  
25  
26 158 test data. For this step, the machine learning library ScikitLearn was used.  
27

28 159

## 30 160 **4. Case study: West midlands region, England**

31 161

32  
33  
34 162 We evaluated our methodology using tweets from the West Midlands area in the United  
35  
36 163 Kingdom. Firstly, we measured the performance of the classifiers using different features on the  
37  
38 164 training dataset. Then, we selected the most effective feature amongst each classifier for the  
39  
40 165 test dataset. Lastly, we compared our work to similar studies in the literature.  
41

42 166

### 44 167 **4.1 Twitter data acquisition**

45  
46 168 We collected 4 million tweets using an uninterrupted connection to the Twitter Streaming API  
47  
48 169 from March 1st, 2017 to May 31st, 2017, with the coordinates to the West Midlands region as a  
49  
50 170 geolocation filter. From these data, the regular expressions filter extracted 13,410 tweets, using  
51  
52 171 a dictionary of 265 road names and traffic related keywords. Tweets were then manually  
53  
54 172 labelled into traffic and non-traffic related, and divided into the following datasets:

- 55 173 • Training: This is the portion of tweets used to train and validate the text classification  
56  
57 174 algorithms. It consisted of 785 traffic related tweets and 785 non-traffic related.

58

59

60

61

- Test: To show the effectiveness of the classifiers on a different dataset than the training, we built a test dataset of 196 traffic tweets and 196 non-traffic related tweets.

From the three-month period, May obtained the highest amount of traffic related Tweets (see figure 1). This was influenced by a high traffic of tweets on the 15<sup>th</sup> and 16<sup>th</sup> of May, due to the identification of an undetonated WWII bomb in the city centre of Birmingham. Table 2 has some examples of traffic and non-traffic related tweets from May 2017. It is important to mention that even though our methodology does not include geolocation, we only took into consideration as traffic tweets those that specify the location of the incident.

Tweet	Label
Brum traffic chaos all entry and exit slip roads to m6 at spaghetti junction and the whole a38m are closed due to a bomb being found #ww2	Traffic
massive car crash on pedmore road by merry hill going towards halesowen road all shut off so avoid it	Traffic
traffic chaos bingo big delays in #birmingham #ww2bomb #aston	Traffic
just heard... interview car crash is an understatement	Non-traffic
after a few rough days following my crash im working hard staying positive to get fixed for	Non-Traffic

**Table 2: Examples of tweets and their label**

## 4.2 Experimental results

With the purpose of identifying which feature works best with the different machine learning algorithms, we tested each classifier with different n-gram values on the training dataset. For this step, we used a k-fold cross validation methodology. K-fold crossvalidation randomly partitions the dataset into k equal sized folds. From these folds, one is retained for testing the model, while the remaining k-1 are used as training data. This process is repeated k times,

193 using each of the k folds exactly once as test data. We performed the k-fold crossvalidation with  
 194 n = 10 on the training dataset for each classifier/n-gram.

195  
 196 To evaluate the performance of the classifiers, we calculated the statistical metrics shown in  
 197 table 3. True negative (TN) and true positive (TP) correspond to the tweets that were classified  
 198 correctly as non-traffic and traffic related, respectively; while False negative (FN) and False  
 199 positive (FP) tweets are those that were misclassified as non-traffic and traffic tweets. Accuracy  
 200 is the overall efficiency of the classifier and corresponds to the fraction of correctly classified  
 201 tweets by the total number of tweets. Precision of a class represents the fraction of correctly  
 202 classified tweets within that class. Recall of a class is the number of correctly classified tweets  
 203 over the total number of tweets that belong to that class. F1-score is the weighted mean of  
 204 precision and recall.

<b>Metric</b>	<b>Formula</b>
<i>Accuracy</i>	$acc = \frac{(TP+TN)}{(TP+FP+TN+FN)}$
<i>Precision</i>	$Prec = \frac{TP}{TP + FP}$
<i>Recall</i>	$Rec = \frac{TP}{TP + FN}$
<i>F1 score</i>	$F1 = \frac{2 \times P \times R}{P + R}$

205 **Table 3: Evaluation metrics**

206  
 207 Table 4 shows the results from the cross validation of the training data using different n-gram  
 208 ranges. For each classifier, we performed the 10-fold cross validation using unigrams, bigrams,  
 209 unigrams and bigrams, and unigrams, bigrams and trigrams. We calculated the average of the  
 210 10 values of accuracy obtained in the cross validation. It can be perceived that most of the  
 211 classifiers have higher performance using unigrams or the combination of the three features,  
 212 while the worst performance amongst all is observed on the trigrams.

213

Model	Unigrams	Bigrams	Trigrams	Unigrams and Bigrams	Unigrams, Bigrams and Trigrams
RC	<b>90.19%</b>	84.92%	64.20%	90.04%	88.87%
KNN	86.88%	77.78%	50.38%	87.83%	<b>87.90%</b>
NB	87.96%	78.30%	56.18%	88.66%	<b>88.98%</b>
MLP	89.49%	84.87%	64.59%	<b>90.32%</b>	89.87%
SVM	<b>90.32%</b>	84.82%	64.20%	89.77%	88.64%

**Table 4: Classifiers vs word n-gram features**

We selected the feature with the highest accuracy for each classifier, and proceeded to evaluate them on the test dataset. Table 5 depicts the classification results for each classifier on the test dataset. The classifier with the highest accuracy was the Ridge classifier (RC) with a 92.86%. MLP and SVM had similar performance to the Ridge classifier both with 92.6%, while the NB was the one with the lowest accuracy with an 89.54%. These results show that the classifiers are not overfitted to the events in the training data. The classifiers had more precision predicting non-traffic related tweets, but less recall. This shows that while the model identified a higher number of traffic related tweets, they had more precision identifying non-traffic related ones.

Model	Traffic			Non-Traffic			Accuracy
	Prec	Rec	F1	Prec	Rec	F1	
RC	90.38%	95.92%	93.07%	95.65%	89.80%	92.63%	<b>92.86%</b>
KNN	86.18%	95.41%	90.56%	94.86%	84.69%	89.49%	90.05%
NB	84.14%	97.45%	90.31%	96.97%	81.63%	88.64%	89.54%
MLP	89.57%	96.43%	92.87%	96.13%	88.78%	92.31%	<b>92.60%</b>
SVM	89.57%	96.43%	92.87%	96.13%	88.78%	92.31%	<b>92.60%</b>

**Table 5: Results on the test dataset**

Results from the test dataset showed that a RC, MLP or a SVM would obtain high accuracy on classifying tweets into traffic related or not. However, there are other aspects that need to be taken into consideration, such as the training and prediction time. Table 6 contains the training and prediction time of each algorithm on the test dataset in seconds. RC and SVM are the fastest in both training and prediction both with 0.04s and 0.008s respectively. However, although MLP obtained one of the highest accuracy scores, it needed 43.53s to train. This is more than 1000 times more of what was needed by the RC and the SVM. Contrary to RC and

234 SVM, MLP obtained more accuracy using unigrams and bigrams, instead of only unigrams,  
 235 which increases the computing time. However, MLP always obtained the highest computing  
 236 time amongst all the n-gram variations.

237

Algorithm	Training time	Prediction time
RC	<b>0.044</b>	<b>0.008</b>
KNN	0.149	0.048
NB	0.176	0.039
MLP	43.53	0.02
SVM	<b>0.04</b>	<b>0.008</b>

238 **Table 6: Training and prediction time (sec)**

239

240 As seen in table 7, results from our RC outperformed studies in the literature. We only took into  
 241 consideration studies that tested their classifiers on a dataset different than the training one.  
 242 (Gu, Qian & Chen ,2016) obtained an accuracy of 90.5% on their test dataset, using a Naïve  
 243 Bayes classifier identifying traffic related tweets. On the other hand, (Schulz, Ristoski &  
 244 Paulheim ,2013) compared SVM, RIPPER and NB for the identification of car accidents, with  
 245 accuracies of 89.06%, 84.21% and 79.21%, respectively. In this paper, we used a split of  
 246 75%/25% of the train and test data, which was similar to the ones used by these studies. Both  
 247 studies employed the REST API for crawling tweets, while we used the Streaming API.

Author	Algorithm	Train/Test split	Accuracy
(Gu, Qian & Chen ,2016)	Naives Bayes	77.5%/22.5%	90.5%
(Schulz, Ristoski & Paulheim, 2013)	Support Vector Machine	75%/25%	89.06%
	RIPPER		84.21%
	Naïves Bayes		79.21%

248 **Table 7: Results from the literature**

249

250 **4. Conclusions and future work**

1 251 We have developed a methodology for crawling, processing, and classifying traffic related  
2 252 tweets in real-time. We fetched tweets using an uninterrupted connection to the Streaming API.  
3  
4 253 Then, we used natural language processing techniques to remove special characters and stop-  
5  
6 254 words. We compared five different machine learning algorithms, and obtained an overall highest  
7  
8 255 accuracy of 92.86% with a Ridge Classifier on our test data. Our results outperformed similar  
9  
10 256 studies in the literature.

11  
12 257  
13  
14 258 Our experimental results show the ability of the system in detecting traffic incidents on real-time.  
15  
16 259 This information can be incorporated on AID systems to improve their accuracy to wider areas  
17  
18 260 of the network. Social media data can also be used to detect the feedback of the users in  
19  
20 261 specific parts of the network.

21  
22 262  
23  
24 263 This paper is part of an on-going work for a real-time pipeline for incident detection using  
25  
26 264 Twitter. Future work includes the use of additional NLP techniques to improve the accuracy of  
27  
28 265 the classifier and to detect the location of the incident. Finally, sentiment and stress analysis will  
29  
30 266 be performed to obtain the user's perspective of the network.

31 267  
32  
33 268 **Acknowledgements**  
34  
35 269 This research was supported by the European Union's Horizon 2020 research and innovation  
36  
37 270 programme under grant agreement No 636160-2, the Optimum project  
38  
39 271 [www.optimumproject.eu](http://www.optimumproject.eu).

40  
41 272  
42  
43 273 This paper is part of a PhD sponsored by the Dominican Republic's Ministry of Education  
44  
45 274 (MESCyT).

46  
47 275  
48  
49 276 **References**

50  
51 277 Abdelhaq, H., Sengstock, C. & Gertz, M. 2013, "Eventweet: Online localized event detection  
52 278 from twitter", *Proceedings of the VLDB Endowment*, vol. 6, no. 12, pp. 1326-1329.  
53  
54 279 Abel, F., Hauff, C., Houben, G., Stronkman, R. & Tao, K. 2012, "Twitcident: fighting fire with  
55 280 information from social web streams", *Proceedings of the 21st International Conference on*  
56 281 *World Wide Web* ACM, , pp. 305.

1 282 Atefeh, F. & Khreich, W. 2015, "A Survey of Techniques for Event Detection in Twitter",  
2 283 *Computational Intelligence*, vol. 31, no. 1, pp. 132-164.

3  
4 284 D'Andrea, E., Ducange, P., Lazzerini, B. & Marcelloni, F. 2015, "Real-Time Detection of Traffic  
5 285 From Twitter Stream Analysis", *IEEE Transactions on Intelligent Transportation Systems*,  
6 286 vol. 16, no. 4, pp. 2269-2283.

7  
8 287 D'Andrea, E. & Marcelloni, F. 2017, "Detection of traffic congestion and incidents from GPS  
9 288 trace analysis", *Expert Systems with Applications*, vol. 73, pp. 43-56.

10  
11 289 DfT 2017, *Road Traffic Estimates: Great Britain 2016*. Available:  
12 290 <https://www.gov.uk/government/statistics/road-traffic-estimates-in-great-britain-2016>

13  
14 291 Gu, Y., Qian, Z.(. & Chen, F. 2016, "From Twitter to detector: Real-time traffic incident detection  
15 292 using social media data", *Transportation Research Part C: Emerging Technologies*, vol.  
16 293 67, pp. 321 342.

17  
18 294 Gutierrez, C., Figuerias, P., Oliveira, P., Costa, R. & Jardim-Goncalves, R. 2015, "Twitter mining  
19 295 for traffic events detection", *Proceedings of the 2015 Science and Information Conference*,  
20 296 *SAI 2015*, , pp. 371-378.

21  
22  
23 297 Krstajic, M., Rohrdantz, C., Hund, M. & Weiler, A. 2012, "Getting there first: Real-time detection  
24 298 of real-world incidents on twitter", .

25  
26 299 N. Wanichayapong, W. Pruthipunyaskul, W. Pattara-Atikom & P. Chaovalit 2011, "Social-based  
27 300 traffic information extraction and classification", *ITS Telecommunications (ITST), 2011 11th*  
28 301 *International Conference on*, pp. 107.

29  
30 302 Osborne, M., Moran, S., McCreddie, R., Von Lunen, A., Sykora, M.D., Cano, E., Ireson, N.,  
31 303 Macdonald, C., Ounis, I. & He, Y. 2014, "Real-time detection, tracking, and monitoring of  
32 304 automatically discovered events in social media", .

33  
34 305 Parkany, E. & Xie, C. 2005, *A complete review of incident detection algorithms & their*  
35 306 *deployment: what works and what doesn't*.

36  
37  
38 307 R. Li, K. H. Lei, R. Khadiwala & K. C. C. Chang 2012, "TEDAS: A Twitter-based Event Detection  
39 308 and Analysis System", *2012 IEEE 28th International Conference on Data Engineering*, pp.  
40 309 1273.

41  
42 310 Sakaki, T., Okazaki, M. & Matsuo, Y. 2010, "Earthquake shakes Twitter users: real-time event  
43 311 detection by social sensors", *Proceedings of the 19th international conference on World*  
44 312 *wide webACM*, , pp. 851.

45  
46 313 Schulz, A., Ristoski, P. & Paulheim, H. 2013, "I See a Car Crash: Real-Time Detection of Small  
47 314 Scale Incidents in Microblogs", Springer, Berlin, Heidelberg, , pp. 22.

48  
49 315 Siripanpornchana, C., Panichpapiboon, S. & Chaovalit, P. 2016, "Incidents detection through  
50 316 mobile sensing", *IEEE*, , pp. 1.

51  
52  
53 317 Twitter 2017. *The Search API*. Available: <https://dev.twitter.com/rest/public/search>.

54  
55 318

56  
57  
58  
59  
60  
61  
62  
63  
64  
65

Figure 1. Tweets per month

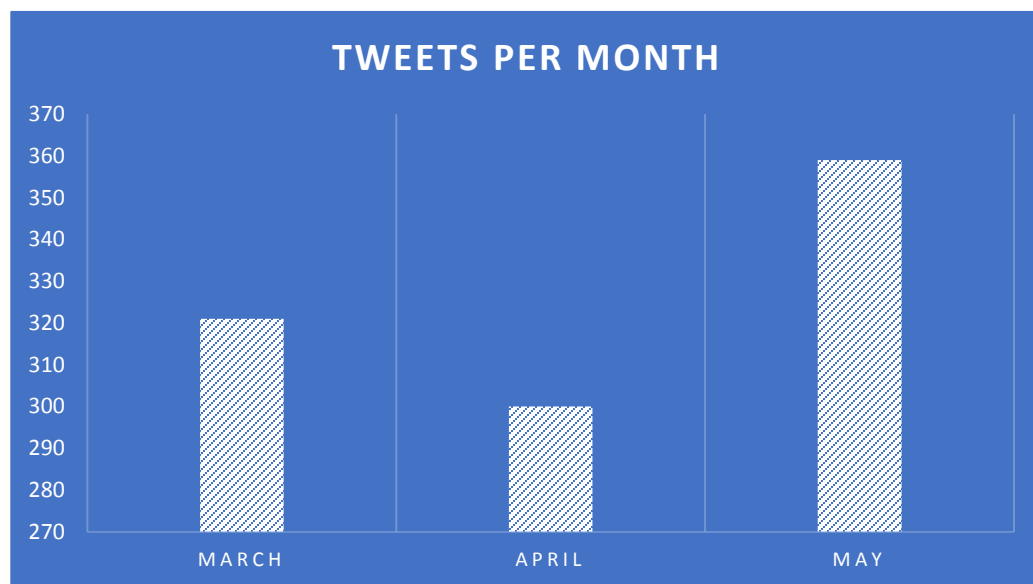


Figure 2. System architecture

