Public Twitter Data and Transport Network Status

Adel Almohammad
Faculty of Science and Engineering
University of Wolverhampton
Wolverhampton, UK
a.almohammad@wlv.ac.uk

Panagiotis Georgakis
Faculty of Science and Engineering
University of Wolverhampton
Wolverhampton, UK
p.georgakis@wlv.ac.uk

***Abstract*— Twitter data can be collected and analysed to be used for predicting the status of a transport network at a given time and geographic location (e.g. forecasting disruptions, congestions, or road closures). However, this requires geolocating the tweets to define the parts of the transport network which may be related to these tweets. This paper investigates the relationship between the actual transport network status, with that being synthesised using public Twitter data in the Greater Manchester conurbation. Therefore, it answers the following question: are the sentiments of tweets around the incidents and accidents areas (or bounding boxes) different from the sentiments of tweets in the seamless traffic areas?. According to the used research methodology, analysis techniques, and sentiment detection APIs, it has been concluded that there is no significant difference between the sentiments in the tweets regardless the prevailing traffic conditions of the locations the tweets refer to.**

Keywords—Twitter, transport network, traffic, incidents and accidents

I. Introduction

There are various events that affect the normal traffic flow (i.e. accidents and road work), and unexpectedly cause congestions in the transport network and therefore delay passengers for reaching their destinations on time. Social media websites may be considered as a common source of traffic and transport information such as incidents and accidents. Users may use such platforms to describe the current traffic situation around them as well as to express or share their feelings towards such events. Twitter, which is one of these social media platforms, is commonly used for sharing information and expressing own opinion and emotions on a certain subject. It contains different data (e.g. traffic related tweets) that can be used for detecting traffic events. For example, users may express their impression (e.g. angry, stressed, or exhausted) about a traffic jam, or a slow mobility service.

Social media data can be collected or streamed and then analysed for predicting the status of the transport network at a given time and geographic location. This requires the determination of the locations that the tweets refer to, as to identify the relevant parts of the transport network. One of the methods which may be used to define the tweets' location is getting the GPS traces of the producers of these tweets. However, this feature needs to be enabled by the users and thus is not available by default. Graham, Hale, and Gaffney (2014) reported that as few as 0.7% of 19.6 million tweets contained geo coordinates [1]. Another method of locating Twitter users may depend on the location information included in the tweets themselves. However, this may represent a very small proportion of the tweets collected since the majority of tweets does not involve any location information in their texts. Therefore, an alternative methodology should be used to identify the location of tweets or the place from which these tweets have been produced (i.e. location of Twitter users). The Twitter includes a geolocation feature that automatically adds the user's neighbourhood or town and state information to each Tweet. If this geolocation service is selected, geolocation information on public tweets from Twitter users can be obtained. This means that the same location is being used for every tweet produced by a specific user. Therefore, it is more beneficial to collect Twitter data produced by as many users as possible.

The traditional methods for collecting traffic information using physical sensors are expensive. They are unable to cover every road on the network and requires regular maintenance. Additionally, the deployment of such physical sensors is an inefficient solution for widespread tracking of traffic flows. Therefore, social media information regarding road and traffic conditions can be used as an alternative method for collecting traffic information [2]. This paper investigates the relationship between public Twitter data and transport network status, by comparing Twitter data geolocated near the location of a road incident or accident with those geolocated in parts of the network not close to such incident or accident locations. Moreover, this relationship determines whether public Twitter data which has been analysed in such a way can be used for traffic forecasting or not.

The remainder of this paper is organized as follows. In Section 2 we discuss related work and in Section 3 we present the proposed methodology. In Section 4 we discuss data collection and in Section 5 we present the different data preparation procedures. In Section 6 we discuss the results of the experiments and in Section 7 we conclude with a summary of the current status regarding the implementation of the proposed methodology.

II. Related work

Twitter represents a promising source of information as most posted messages are about daily experiences and opinions of people. Therefore, tweets can be used as sensors or detectors for a range of issues. Twitter offers GPS-enabled messaging and therefore such enabled tweets can be geolocated (tweets represent a source of geo information). However, if no GPS coordinates are sent with tweets, other methods can be used to geolocate these tweets, such as the geographical description held in the tweets.

Many studies have investigated the idea of using Twitter as a valid source of information in general, including the use of Twitter as a source of transport and traffic information. [3] proposed an open big-data architecture for road traffic prediction in large metropolitan areas. They investigated the functional characteristics of this proposed architecture which allows processing of data from various sources including Twitter. Additionally, this study relies on the contents of the tweets themselves (available geographical description such as road, junction, or station name) to geolocate the tweets. [4]

created a concept for (specific domain) information extraction that is based on machine learning and largely independent from the grammatical correctness of the analysed messages. This concept can be used to analyse public available messages on communication platforms like Twitter to get processable data for a specific domain e. g. traffic information.

Since existing incident detection techniques are limited to the use of sensors in the transportation network, [5] investigated the use of Twitter for supporting real-time incident detection in the United Kingdom (UK). They present a methodology for streaming, processing, and classifying public tweets by combining Natural Language Processing (NLP) techniques with a Support Vector Machine algorithm (SVM) for text classification. Furthermore, [2] used real-time Twitter data to analyse traffic congestion in Los Angeles, USA. The proposed model extracts traffic-related tweets and classifies the extracted information in order to estimate the traffic status on roads. In this model, real-time tweets containing the word 'traffic' was collected and a machine learning classifier was used to classify traffic information.

III. Methodology

In this paper, the study area of Greater Manchester ([SW: (Lat=53.361, Long=-2.484), NE: (Lat= 53.593, Long= -2.007]) has been divided into thousands of bounding boxes (55448 bounding boxes, each H=111m x W=133m). We consider each of these bounding boxes as the essential independent location unit and all the assumptions and outcomes will rely on this consideration. For example, analysing an incident or accident depends on all the events inside the bounding box which contains the location of this incident or accident.

The methodology for defining the relationship between the Twitter data and the transport network status includes two separate data processing parts which connect to each other to form a unified data processing pipeline (Figure 1). The first data processing part is for public Twitter data streaming and analysis (Figure 1, blue colour part) and the second is for incidents and accidents data collection and analysis (Figure 1, green colour part). However, this methodology can be summarised as follows:

1- Data Collection: actual incidents and accidents information in Greater Manchester is being collected using the Transport for Greater Manchester (TfGM) - Open Data Service API. In addition, the public tweets are being collected using the Twitter Streaming API and a geolocation filter.

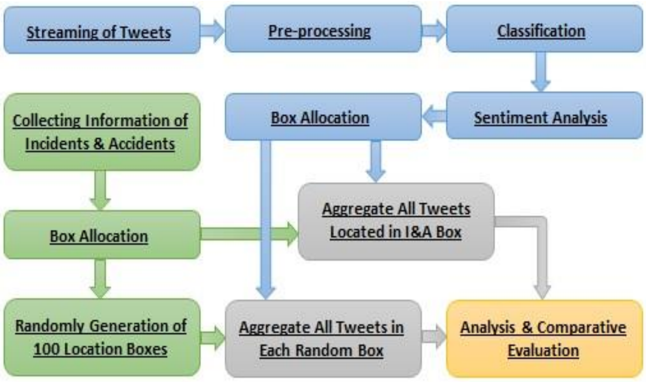

Streaming of Tweets
Pre-processing
Classification
Collecting Information of Incidents & Accidents
Box Allocation
Sentiment Analysis
Aggregate All Tweets Located in I&A Box
Box Allocation
Randomly Generation of 100 Location Boxes
Aggregate All Tweets in Each Random Box
Analysis & Comparative Evaluation

Fig. 1. Data Processing Pipeline.

2- Data Preparation: represents all pre-processing actions applied to the collected data (step 1) to be ready for the next stage of data analysis. These actions include Twitter data cleaning, classification, box allocation, and sentiment detection. Moreover, they include incidents and accidents boxes allocation, and random boxes generation for each incident and accident.

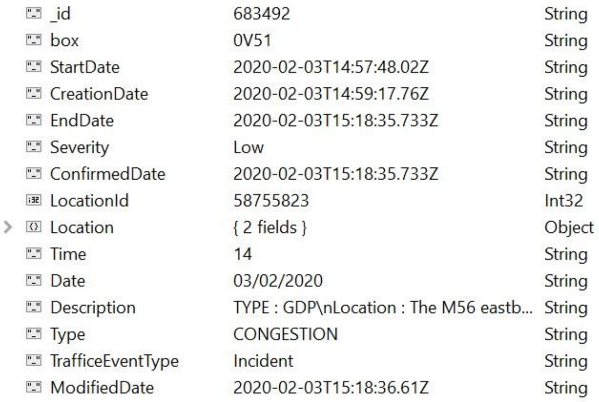
3- Data Analysis and Comparative Evaluation: represents the process of considering the sentiments of all aggregated tweets within each incident and accident box to be compared with the sentiments of all aggregated tweets within randomly selected non-incident and non-accident boxes. Such comparison will show if the sentiments of the tweets around the incidents and accidents areas are different from the sentiments of the tweets in the seamless traffic areas.

IV. Data Collection

Twitter data as well as incidents and accidents data are required in the study area (Greater Manchester). Therefore, this data must be located within Greater Manchester [SW: (Lat=53.361, Long=-2.484), NE: (Lat= 53.593, Long= -2.007].

A. *Incidents and Accidents Information*

Open Data Service API (ODS API), an Open Data RESTful API, provides real-time data about the transport network in Greater Manchester [6]. This data contains incidents and accidents information from Highways England for Greater Manchester and the surrounding area (motorways and sliproads only). Incidents include things such as congestion, broken down vehicle, animals on the road and obstructions while accidents include things such as collision. Figure 2 shows an example of one incident data record collected using this API. Each incident or accident data contains various fields or features, including: ID, start date and time, and location. Since incidents and accidents will be considered the same in the rest of this paper, the term of ‘incident’ will be used to refer to the ‘incident or/and accident’ term.



_id	683492	String
box	0V51	String
StartDate	2020-02-03T14:57:48.02Z	String
CreationDate	2020-02-03T14:59:17.76Z	String
EndDate	2020-02-03T15:18:35.733Z	String
Severity	Low	String
ConfirmedDate	2020-02-03T15:18:35.733Z	String
LocationId	58755823	Int32
Location	{ 2 fields }	Object
Time	14	String
Date	03/02/2020	String
Description	TYPE : GDP\nLocation : The M56 eastb...	String
Type	CONGESTION	String
TrafficeEventType	Incident	String
ModifiedDate	2020-02-03T15:18:36.61Z	String

Fig. 2. An incident data collected by the ODS API after preparation.

B. *Twitter Data*

Twitter provides a REST API and Streaming API for real-time access to their data. The REST API can be used for applications that need to use Twitter functionalities such as finding the posts that contain a given keyword (request and response). However, the Streaming API works differently as a streaming connection can be established, tweets are streamed as they occur to be processed and the results can be stored. The limitation of this API is that it can only request public tweets [7].

Twitter data can be accessed for free and this data provides a rich source of real time information about traffic and transport status. However, one of the limitations of the Twitter streaming API is that less than 1% of the total tweets collected may contain geo coordinates [1]. Twitter is widely used to express the opinion and emotions of users on a certain subject such as traffic. Therefore, Twitter data has been included as a source of information to be analysed in order to understand the relationship between the users' perception of the transport network and the actual status of the transport network (incidents information provided). Figure 3 shows an example of a tweet from the Twitter Streaming API and using a geolocation filter of Greater Manchester area.

V. DATA PREPARATION

Twitter and incidents data that has been collected needs further processing and preparation to be used in the analysis and evaluation stage. Therefore, there are many procedures needs to be applied to each collected data set.

A. *Preparing the Incidents Information*

For each incident record collected, the location of this incident is presented as coordinates (latitude and longitude). This incident location needs to be identified by one bounding box of the 55448 boxes designed for the Greater Manchester study area. Therefore, this preparation finds out the bounding box, the date, and the time of each incident. Figure 2 shows an incident after preparation and it shows that this incident is located in the bounding box 0V51 and occurred on Date: 03/02/2020 at Time 14.

B. *Preparing the Twitter Data*

As mentioned before, the collected Twitter data belongs to the users who mentioned in their profiles that their locations are located within the Greater Manchester area (in one of the 55448 bounding boxes). This includes all public tweets posted by all users within this predefined area. In order to use these tweets for sentiment analysis and detection, the texts need to be clean, readable and compatible with the requirements of sentiment analysis tools or APIs. Additionally, it would be more beneficial and practical to classify these tweets and get those related to the study question. Thus, in order to get more accurate results in terms of users' sentiment toward traffic and transport, only "traffic" and "transport" related tweets are considered. Therefore, the preparation steps of Twitter data can be described in the following subsections.

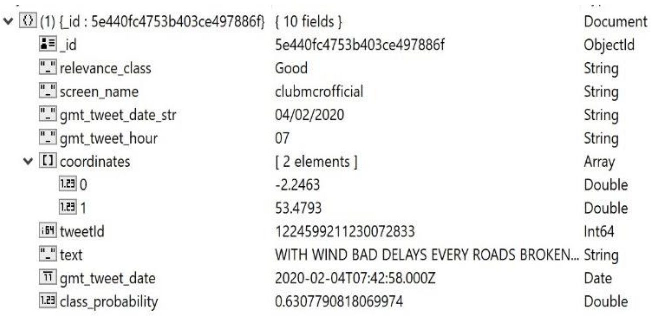
1- Tweets Cleaning

The goal of this process is making the texts or tweets ready for next processing steps. Tweets Cleaning includes removing some characters such as hashtags, emotions, mentions, and punctuations. In the next step, the Text Classification process may rely on the weight or the meaning of the words in the text. However, the text contains many words which has almost no effect on this classification process (i.e. prepositions and conjunctions). Therefore, these words will be deleted from the text.

2- Tweets Classification

The goal of this process is classifying the tweets into traffic (and transport) related tweets and non-traffic related tweets. As a result of Tweet Classification, only the tweets that are related to traffic and transport will be used and analysed in this study.

There are various algorithms for text classification which require training data. One of the applicable machine learning algorithms for such classification is Linear Support Vector Classifier (Linear SVC). The objective of the Linear SVC is to categorise or classify the data according to some training data. Therefore, by having a training data set, some data can be fed to this classifier and this classifier will be able to predict the correct class of this data (or tweet). This algorithm represents a suitable algorithm for tweet classification and can be used for this study. Therefore, the classification model and training data which have been developed and used by [3] and [5] will be used to classify the collected tweets. As a result, each tweet will be classified and have a relevance class (either Good or Bad). Figure 3 represents a tweet after it has been cleaned and classified according to its relevance to the traffic and transport concept. This example shows that the value of 'revelance_class' field is 'GOOD' which means that this tweet is traffic related.



(1) {_id : 5e440fc4753b403ce497886f}	{ 10 fields }	Document
_id	5e440fc4753b403ce497886f	ObjectId
relevance_class	Good	String
screen_name	clubmcrofficial	String
gmt_tweet_date_str	04/02/2020	String
gmt_tweet_hour	07	String
coordinates	[2 elements]	Array
0	-2.2463	Double
1	53.4793	Double
tweetId	1224599211230072833	Int64
text	WITH WIND BAD DELAYS EVERY ROADS BROKEN...	String
gmt_tweet_date	2020-02-04T07:42:58.000Z	Date
class_probability	0.6307790818069974	Double

Fig. 3. A tweet after cleaning and classification.

3- Defining the Bounding Box(s) of Tweets

Streaming tweets according to a geolocation filter provides tweets with different formats of location (i.e. geo, coordinates, and place fields). According to the values of these three fields, three different types for a tweet location can be described.

1) Regardless of the "place" field value, the "coordinates" and "geo" fields have only two double values which means that these values represent the latitude and the longitude of the tweet location (point).

2) The "coordinates" and "geo" fields have "null" values and the "place" field contains a polygon type location (place.bounding_box.coordinates). However, these four points of the location are similar to each other. This means that the location of the tweet is a point rather than a polygon.

3) The "coordinates" and "geo" fields have "null" values and the "place" field contains a polygon type location (place.bounding_box.coordinates) but the four points of the location are different. This means that the location of the tweet is a wide geographic area or a polygon (general address) not a point.

Figure 4 shows an example for the relationship between the location types of tweets and the number of bounding boxes which are covered by these tweets. Tweet-1, tweet-5, and tweet-8 have locations of type-1 or type-2 since the location of each tweet represents a point within a bounding box (e.g. the tweet-1 is located in the Box-7). However, the rest of the tweets have locations of type-3 as the location of each tweet represents a polygon (e.g. the tweet-3 covers all the 10 Boxes).

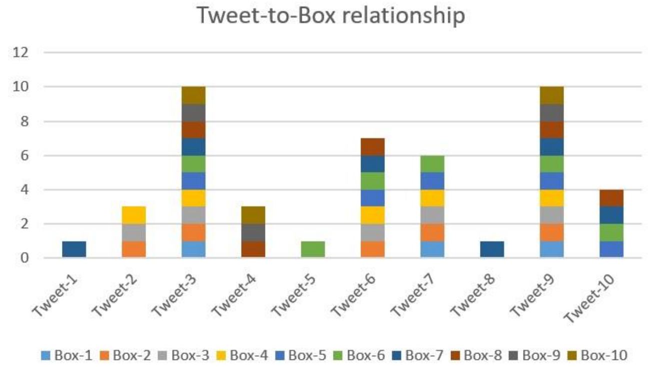


Fig. 4. The relationship between the Tweets' location-types and the bounding boxes.

The tweets that have locations of type-1 or type-2 belongs to one bounding box and the tweets that have locations of type-3 may cover a geographic area equal to thousands of bounding boxes. Experimentally, the number of tweets that have locations of type-1 and type-2 represents less than 6% of the tweets collected. Therefore, considering only these location types of tweets means that the majority of the bounding boxes will have zero tweets almost over all the time. So, tweets with all types of location (type-1, type-2, and type-3) will be considered and each tweet with a location type-3 will be considered in each bounding box covered by this tweet's location. For example, tweet-3 covers all the 10 bounding boxes in Figure 4 and therefore this tweet will be considered in each of these ten boxes.

Accordingly, the tweets that have locations of type-1 or type-2 will belong to only one bounding box of the Greater Manchester study area. However, the tweets that have locations of type-3 will belong to many bounding boxes of the Greater Manchester study area. Therefore, this preparation finds out the bounding box(s) of each tweet ('Box' field) in addition to the date ('gmt_tweet_date_str' field) and time ('gmt_tweet_hour' field) of this tweet (Figure 5 & Figure 6).

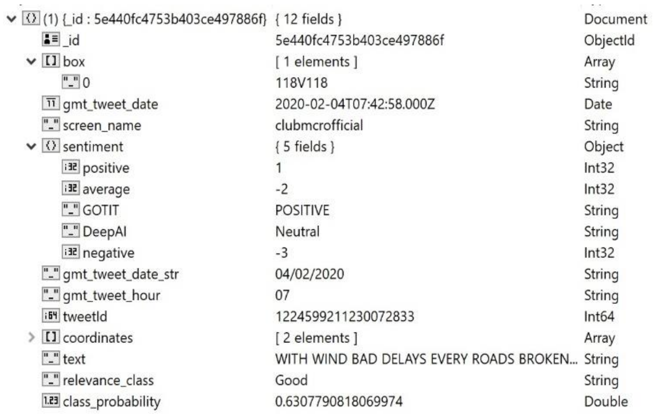

```
(1) {_id : 5e440fc4753b403ce497886f}   { 12 fields }                              Document
    _id                                5e440fc4753b403ce497886f                   ObjectId
    box                                [ 1 elements ]                             Array
        0                              118V118                                    String
    gmt_tweet_date                     2020-02-04T07:42:58.000Z                   Date
    screen_name                        clubmcrofficial                            String
    sentiment                          { 5 fields }                               Object
        positive                       1                                          Int32
        average                        -2                                         Int32
        GOTIT                          POSITIVE                                   String
        DeepAI                         Neutral                                    String
        negative                       -3                                         Int32
    gmt_tweet_date_str                 04/02/2020                                 String
    gmt_tweet_hour                     07                                         String
    tweetId                            1224599211230072833                        Int64
    coordinates                        [ 2 elements ]                             Array
    text                               WITH WIND BAD DELAYS EVERY ROADS BROKEN...  String
    relevance_class                    Good                                       String
    class_probability                  0.6307790818069974                         Double
```

Fig. 5. Same tweet in Figure 3 after sentiment analysis.

4- Sentiment Analysis of Tweets

The purpose of this process is analysing the tweets and calculating their sentiments' features or values. This calculation will be done by performing three different methods or APIs: SentiStrength [8], DeepAI [9], and GotIt [10].

SentiStrength classifier has been developed by [11]. It detects sentiment strength in short informal texts and predicts the strength of positive or negative sentiment within a text. Therefore, this classifier or sentiment detector will be used to calculate both positive and negative sentiment scores of tweets. For each tweet, the SentiStrength outputs two integers: one for positive sentiment strength and another for negative sentiment strength. The scores range from 1 to 5 for positive sentiment and from −1 to −5 for the negative sentiment. Therefore, the average sentiment of a tweet can be calculated by adding up the positive sentiment value of this tweet to its negative sentiment value (Positive=1, Negative=-3, Average=-2). The sentiment of a tweet will be considered positive if the average sentiment value is positive and will be considered negative if the average sentiment value is negative (e.g. Average=-2 means 'Negative' sentiment while Average=1 means 'Positive' sentiment).

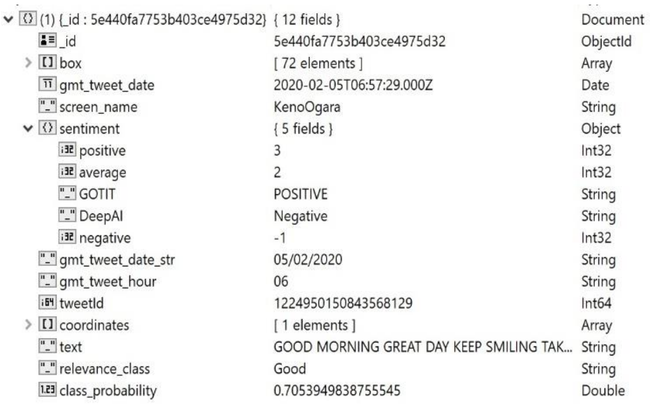

```
(1) {_id : 5e440fa7753b403ce4975d32}   { 12 fields }                              Document
    _id                                5e440fa7753b403ce4975d32                   ObjectId
    box                                [ 72 elements ]                            Array
    gmt_tweet_date                     2020-02-05T06:57:29.000Z                   Date
    screen_name                        KenoOgara                                  String
    sentiment                          { 5 fields }                               Object
        positive                       3                                          Int32
        average                        2                                          Int32
        GOTIT                          POSITIVE                                   String
        DeepAI                         Negative                                   String
        negative                       -1                                         Int32
    gmt_tweet_date_str                 05/02/2020                                 String
    gmt_tweet_hour                     06                                         String
    tweetId                            1224950150843568129                        Int64
    coordinates                        [ 1 elements ]                             Array
    text                               GOOD MORNING GREAT DAY KEEP SMILING TAK...  String
    relevance_class                    Good                                       String
    class_probability                  0.7053949838755545                         Double
```

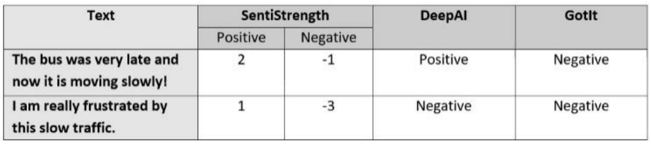
Fig. 6. An example of a tweet with a location type-3.

DeepAI sentiment analysis API uses text analysis, natural language processing, and computational linguistics to identify subjective information from the input text. This API classifies each text into one of the following categories: very negative, negative, neutral, positive, or very positive. In this study, all these sentiment categories will be turned into three main categories only: positive, negative, and neutral.

GotIt sentiment analysis API identifies, extracts and quantifies emotions, feelings, and subjective information involved in a text. This API examines the relationship between the words and considers the emotion of each word depend on the context in which it is used. It classifies the feeling of a sentence, whether it is POSITIVE, NEGATIVE, NEUTRAL or have CONFLICTING feeling (which will be considered as neutral).

Figure 7 shows two examples for traffic related tweets and their sentiment analysed by the online versions of the SentiStrength, DeepAI, and GotIt APIs.

Figure 5 shows the same tweet in Figure 3 after defining its location box (one point within one bounding box: 118V118) and after calculating its text sentiment values. The SentiStrength value for this tweet is 'Negative' (positive:1, negative: -3, average: -2), DeepAI value is 'Neutral', and GotIt value is 'POSITIVE'. However, Figure 6 shows an example for another tweet but with a location type-3. It is clear that this tweet covers 72 bounding boxes.



Text	SentiStrength		DeepAI	GotIt
	Positive	Negative		
The bus was very late and now it is moving slowly!	2	-1	Positive	Negative
I am really frustrated by this slow traffic.	1	-3	Negative	Negative

Fig. 7. Examples of text sentiment analysis and detection APIs

VI. Data Analysis and a Comparative Evaluation

In practice, the total number of incidents collected (reported in the Greater Manchester study area) from 03 Feb 2020, 13:00H till 05 Feb 2020, 14:00H (49 hours in total) was 151. Additionally, the total number of streamed tweets (during the same period) was 27878 which contain only 7059 traffic-related tweets. Since the non-traffic tweets are irrelevant and may lead to incorrect results, only the tweets which have 'Good' relevance class (traffic-related tweets) will be considered in the data analysis.

The analysis starts from the bounding box of each incident and analyse the Twitter date which has been collected and prepared. For each incident, the bounding box, the date, and the time of this incident (e.g. 03/02/2020 between 14:00-15:00) will be considered to select the appropriate Twitter data to be aggregated and analysed. Therefore, the data analysis steps can be described for each incident record as follows:

1- Determine the bounding box (location), date, and time of this incident (e.g. box: 0V51, Date: 03/02/2020, Time: 14).

2- Aggregate all the tweets (only traffic-related tweets) which have the same date and time of this incident (step 1) as well as the location of these tweets ('box' field) involves the 'box' of this incident (step 1).

3- Calculate the overall sentiments of this incident box (three values represent the three different sentiment APIs). The sentiment API/value (e.g. SentiStrength/POSITIVE) of an incident box is calculated by counting all the tweets which are included in the aggregated tweets (step 2) and have this sentiment API/value (e.g. SentiStrength: Positive).

4- Randomly select 100 bounding boxes (from the 55448 boxes) but they do not include this incident box (e.g. 0V51 is not included in the randomly selected boxes).

5- For each randomly selected box (e.g. 12V34), aggregate all the tweets (only traffic-related tweets) which have the same date and time of this incident as well as the location of these tweets ('box' field) involve this randomly selected box (e.g. 12V34).

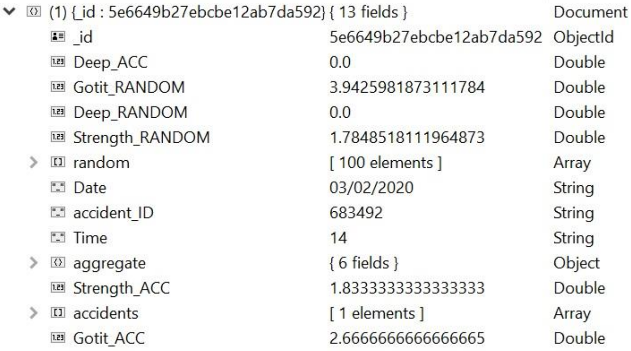
6- For each randomly selected box (e.g. 12V34), calculate the overall sentiments of this box. The sentiment API/value (e.g. SentiStrength/POSITIVE) of a random box is calculated by counting all the tweets which are included in the aggregated tweets (step 5) and have this sentiment API/value (e.g. SentiStrength: Positive).

7- Calculate the overall sentiments of the 100 randomly selected boxes (which are simply selected for this incident - 0V51). The average value of a specific sentiment (e.g. SentiStrength POSITIVE) for the 100 random boxes is calculated by adding up all the values of this sentiment (e.g. SentiStrength POSITIVE) in all the random boxes (step 6).

8- For this incident box (output of step 3), the positive sentiment value related to a specific API (e.g. SentiStrength POSITIVE) is divided by its corresponding negative value (e.g. SentiStrength NEGATIVE).

9- For all the random boxes (overall) selected for this incident (output of step 7), the positive sentiment value related to a specific API (e.g. SentiStrength POSITIVE) is divided by its corresponding negative value (e.g. SentiStrength NEGATIVE).

10- The output of step 8 (data1) represents Twitter data within incidents boxes while the output of step 9 (data2) represents Twitter data out of incidents boxes (randomly selected boxes). Figure 8 shows an example for one incident box (box: 0V51 mentioned in step 1) with the calculations described above (step 8 and step 9). For example, 'SentiStrength_ACC', 'Deep_ACC' and 'Gotit_ACC' are the outputs of step 8 while 'SentiStrength_RANDOM', 'Deep_RANDOM' and 'Gotit_RANDOM' are the outputs of step 9.



```
(1) {_id : 5e6649b27ebcbe12ab7da592} { 13 fields }   Document
    _id               5e6649b27ebcbe12ab7da592       ObjectId
    Deep_ACC          0.0                            Double
    Gotit_RANDOM      3.9425981873111784             Double
    Deep_RANDOM       0.0                            Double
    Strength_RANDOM   1.7848518111964873             Double
    random            [ 100 elements ]               Array
    Date              03/02/2020                     String
    accident_ID       683492                         String
    Time              14                             String
    aggregate         { 6 fields }                   Object
    Strength_ACC      1.8333333333333333             Double
    accidents         [ 1 elements ]                 Array
    Gotit_ACC         2.6666666666666665             Double
```

Fig. 8. Same Incident in Figure 2 after calculating the sentiments of this incident box and the sentiments of the related random boxes.

As we mentioned before, each value of the dataset 'data1' represents the total number of 'positive-sentiment' tweets divided by the total number of 'negative-sentiment' tweets within one incident box. However, the value of the dataset 'data2' represents the total number of 'positive-sentiment' tweets divided by the total number of 'negative-sentiment' tweets within all the 100 random boxes (which are selected for this incident mentioned above). Therefore, the value of the following division will be used in the data analysis:

$$\frac{\boldsymbol{Number\ of\ Positive\_Sentiment\ Tweets}\ (\boldsymbol{NPST})}{\boldsymbol{Number\ of\ Negative\_Sentiment\ Tweets}\ (\boldsymbol{NNST})}$$

For any box, if this value (regarding a specific sentiment API) is larger than one (>1), it means that the number of positive-sentiment tweets inside this box is larger than the number of negative-sentiment tweets. For example, Figure 8 shows that Strength_ACC=1.83 and Strength_RANDOM=1.78.

These two values can be compared with each other in order to investigate the difference between the tweets inside incident boxes and those out of incident boxes. The result of this comparison for the previous example in Figure 8 is: Strength_ACC (1.83) > Strength_RANDOM (1.78). For instance, if the majority of incidents achieves the following equation:

$$\frac{\boldsymbol{NPST}}{\boldsymbol{NNST}}_{\boldsymbol{incident\ box}} < \frac{\boldsymbol{NPST}}{\boldsymbol{NNST}}_{\boldsymbol{random\ boxes}}$$

This means that the ratio of the negative-sentiment tweets to the positive-sentiment tweets in the incident box is larger than that of random boxes. For example, the values of this equation can be 0.25 < 0.9 if an incident box has 5 positive-sentiment tweets and 20 negative ones while overall random boxes have 900 positive-sentiment tweets and 1000 negative ones (in 100 random boxes).

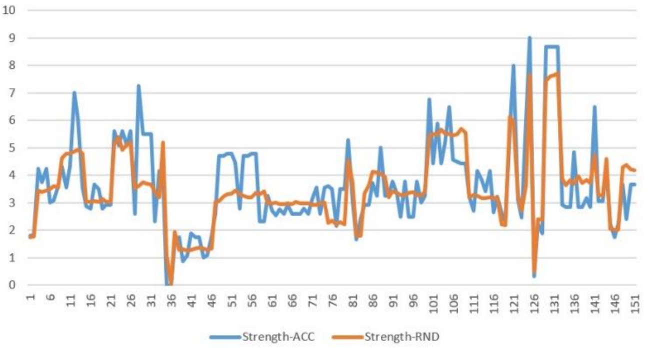


Fig. 9. SentiStrength of Incidents boxes vs SentiStrength of Random boxes.

Figure 9 shows data1 and data2 values for SentiStrength sentiment API. It is quite obvious that both values are moving beside each other without any noticeable difference between them (e.g. blue line of Strength-ACC and orange line of Strength-RND). Therefore, the statistics of each incident box look like those of related overall random box. This means that the existence of an incident within a specific bounding box is not reflected in the contents (and therefore in the sentiments) of the tweets aggregated for this bounding box. Also, this is true for the other two sentiment APIs, DeepAI and GotIt. Figure 10 shows data1 and data2 values for the three sentiment APIs and it reveals the same behaviour discussed above.

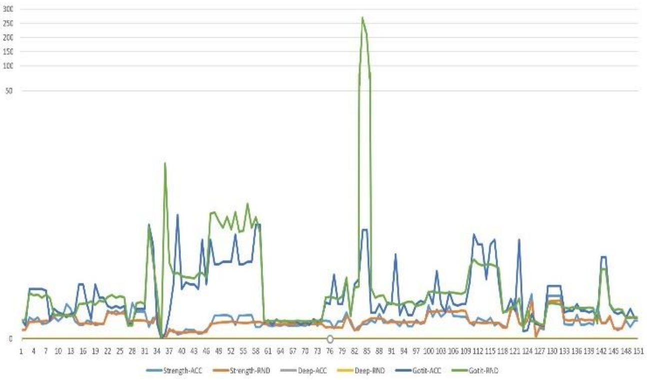


Fig. 10. Sentiments of Incidents boxes vs sentiments of Random boxes.

VII. Conclusion

This study investigated the possibility of using Twitter feed as a source of information to assess the status of transport network and may be to forecast short-term traffic. Therefore, all tweets have been streamed according to a geolocation filter for Greater Manchester area. The collected tweets have been classified, analysed, and grouped into small bounding boxes according to their locations.

The experiments showed that the number of tweets with positive-sentiment values to the number of tweets with negative-sentiment values of incidents' boxes and those of the overall random box are quite close to each other. So, both values of incidents' boxes and random boxes have no explicit difference from each other.

Therefore, Twitter data analysed and utilized in such a method is unlikely to be used neither as an indication for transport network status nor for traffic forecasting. This is due to the similarity of the results related to the incidents' boxes compared to those related to the non-incident boxes.

In conclusion, Twitter data usage in such a way described in this paper is unlikely to be beneficial neither for examining the status of transport network nor for traffic forecasting. The reasons behind this may include the following:

1- The location of the tweets (derived from the profiles of their users) may be inaccurate.

2- As the majority of tweets cover a wide geographic area (ten thousands of bounding boxes), the sentiment of a tweet may be considered in all the boxes covered by this tweet.

3- As the methodology classifies tweets into traffic-related or non-traffic related tweets, the classification algorithm or method may require further improvement.

4- Reasons related to the tweet sentiment detection tools or APIs. There were many cases in which the sentiment values are not the same for a given tweet when the three sentiment APIs were used (the sentiment of a tweet can be positive with one API, negative with another API, and neutral with the third API). Finally, the size of the data used in this study (151 incident and 7059 traffic-related tweets) is unlikely to have impact on the results since all Twitter data that cover all incidents locations, date, and time has been considered.

Acknowledgment

Research reported in this paper has received funding by the H2020 EC project MaaS4EU (GA no. 723176).

References

[1] S. Haustein, R. Costas, "Determining Twitter audiences: Geolocation and number of followers", The 2015 Altmetrics Workshop, Amsterdam, 9 October 2015.

[2] M. A. A. Al-qaness, M. A. Elaziz, A. Hawbani, A. A. Abbasi, L. Zhao and S. Kim, "Real-Time Traffic Congestion Analysis Based on Collected Tweets," 2019 IEEE International Conferences on Ubiquitous Computing & Communications (IUCC) and Data Science and Computational Intelligence (DSCI) and Smart Computing, Networking and Services (SmartCNS), Shenyang, China, 2019, pp. 1-8.

[3] Y. G. Petalas, A. Ammari, P. Georgakis, C. Nwagboso C, "A Big Data Architecture for Traffic Forecasting Using Multi-Source Information,". In: Sellis T., Oikonomou K. (eds) Algorithmic Aspects of Cloud Computing. ALGOCLOUD 2016. Lecture Notes in Computer Science, vol 10230. Springer, Cham.

[4] C. Engel, S. Magnus, J. Krause, "Using Twitter as source of traffic data - Concepts and experiences of the project AUSWEG," Procedia Computer Science, vol. 110, pp. 479-485, 2017.

[5] A. Salas, P. Georgakis, Y. Petalas, "Incident detection using data from social media", 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), pp. 751-755, 2017.

[6] Travel for Greater Manchester (TfGM), https://developer.tfgm.com/.

[7] Twitter Developer Website, https://developer.twitter.com/.

[8] http://sentistrength.wlv.ac.uk/.

[9] https://deepai.org/machine-learning-model/sentiment-analysis.

[10] https://www.gotit.ai/en-us/Home/Sentiment.

[11] T. Mike, B. Kevan, P. Georgios, C. Di, K. Arvid, "Sentiment Strength Detection in Short Informal Text." Journal of the American Society for Information Science and Technology. 61. 2544-2558. 10.1002/asi.214.