A Near Real-Time Traffic Congestion Monitoring System Using Sentiment Analysis On Twitter Data

Goboitshepo Ororiseng Leroke, Manoj Lall

Abstract: Traffic congestion is a major challenge facing urban areas around the globe today. A common approach adopted by government agencies for the monitoring of traffic conditions is by making use of CCTV cameras or electronic sensors. These approaches requires the maintenance of a large network of sensors and cameras to monitor every street in the city. This is impractical and very costly. However, with the advancement of social media in all its forms, including blogs, online forums, Facebook, and Twitter, it is possible to treat social media as a human sensor network. In this article, an alternative traffic monitoring approach that is inexpensive and provides traffic information in near real-time is developed. The proposed approach makes use of Twitter data analytics to report on the prevailing traffic conditions in a particular locality. In addition, the reason behind the traffic congestion is also highlighted. Knowing the cause of the traffic congestion is important as it gives an indication of the severity of the problem. For the modelling of the proposed near real time Twitter-based model, 5 000 tweets collected over a period of six months were collected for a particular geographical location. The relevant Twitters were pre-processing to obtain the applicable features such as the location of the origin of a particular post, the time when the tweet was posted. Random Forest, Naïve Bayes, Support Vector Machine and K-Nearest Neighbour were used in the construction of the classification model. The best performing model (Naïve Bayes) was selected for real-time tweet classification. Python's Natural Language Toolkit (NLTK) and associated libraries, was applied to enhance the suitability of tweets for conducting sentiment analysis and topic modelling. The emotions expressed in the tweets were captured by sentiment analysis and the reason behind traffic congestions were determined by topic modelling. The location, the sentiment and the reasons for the traffic congestions were visualized using street map. It is envisaged that such a model will assist commuters in making an informed decision on route selection.

Index Terms: Social Media; Twitter; Traffic Congestion; Sentiment Analysis; Natural Language Processing; Machine Learning; Topic Modelling.

1 INTRODUCTION

Since the emergence of road transportation network, which is conclusively one of the biggest revolutions of this century, the impact it has on the lives of people around the globe is immense (Elsafoury, 2013; Musakwa, 2014; Sukhai, Jones & Haynes). In spite of the benefits of this revolution, the problems associated with its emergence, such as traffic congestion has the potential to impact severely on a country's economy. In an effort to monitor the traffic flow, various communities, especially in urban areas, have adopted numerous approaches. The commonly utilized mechanisms utilize hardware sensors, cameras and radar instruments (Gong, Deng & Sinnott, 2015). However, these tools have certain constraints, such as, being expensive to maintain and can only focus on limited areas of the transport network (Elsafoury, 2013). With the incorporation of Web 2.0 technologies, the social media platforms such as Facebook, Twitter, and LinkedIn are able to provide location based information, and have successfully shifted social media platform from just being a cyberspace to real places (Kurashima et al., 2010; Somwanshi, 2015; Mahmood et al., 2017)

These social media platforms have become an integrated part of people's daily lives and could be viewed as a virtual sensor network full of human-agents with high mobility, quick response time, and greater flexibility. It is estimated that Twitter has more than a billion registered accounts, with over 317 million active users worldwide (Khatri, 2018). Twitter data consists of the following information or features: tweet text, geolocation (if available), date and time the tweet was posted, and user ID. Twitter allows its users to posts messages (tweets) limited to a maximum of 140 characters per tweet. The limitations imposed on the number of characters allowed in a tweet enables users to update posts quickly and easily (Grosenick, 2012). Additionally, when users posts, they express their views, thoughts and their emotional state (Musakwa, 2014). Besides the text message, the tweets frequently contain geospatial information such as the location from where the tweets originated. Combining these three characteristics provides possibilities to mine patterns and phenomenon associated with what is trending at any given time and place, or in establishing the positive or negative sentiment on a particular topics across the population (Gong, Deng & Sinnott, 2015; Kharche & Bijole, 2015; Khatri, 2018; Mahmood et al., 2017). This article makes use of these three characteristic of Twitter for the proposed near real-time traffic congestion monitoring system. The remainder of the paper is organized as follows. In section 2, we review related work on transport issues and focus on the requirements/challenges of our research. Section 3 covers the architecture and methodology used to harvest and cluster tweets. Section 4 shows implementation, results and discussion of model. Finally, in Section 5 conclusions on the work is presented and future enhancements are suggested.

2 BACKGROUND AND RELATED WORK

The commonly used approaches to traffic congestion monitoring and traffic information dissemination systems are organized in a centralized approach, where information are collected by using devices such as inductive loops, infrared sensors, or video cameras. Sensor-based traffic monitoring systems are deployed directly on the side of the road to collect information about the current traffic density as illustrated in Fig. 1. The measured data are transferred to a central Traffic Information Centre (TIC), where the data packets from all

- *Goboitshepo Ororiseng Leroke, Tshwane university of Technology, Pretoria, South Africa. E-mail: lerokego@tut.ac.za.*
- *Manoj Lall, Tshwane University of Technology, Pretoria, South Africa, E-mail: LallM@tut.ac.za*

sensors are received and the current traffic situation is analysed (LEROKE). The result of this situation analysis is packed into messages for the Traffic Message Channel2 (TMC), forwarded to the FM radio broadcast station and/or TV channels, and transmitted via Radio Data System (RDS). Alternatively, the traffic messages can be transferred on demand via a cellular mobile phone network or internet.

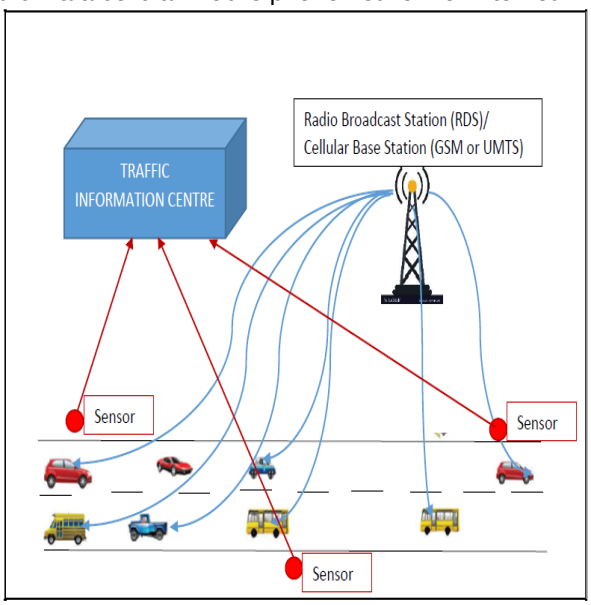

Fig. 1. Conventional, centralized TTI analysis and distribution.

A current centralized service for distributing traffic information, such as the traffic news or the TMC broadcast by a radio station, has several technical disadvantages, namely:

- A large number of sensors are needed to be deployed in order to monitor the traffic situation.
- The traffic information service is limited to streets where sensors are integrated.
- The recorded traffic density data are transmitted to a central unit and evaluated, before it is broadcasted to the drivers.
- Traffic information is distributed with a relatively high delay. The time delay before receiving an update of the current traffic situation (especially for the local area) is the most crucial point in all TTI systems.
- An extremely large investment for the communication infrastructure (sensors, central unit, wired and wireless connections) is necessary.

Based on these reasons, an alternative approach for monitoring the traffic situation and distributing the traffic messages to interested parties is presented in this article. It is important for commuters to know about traffic incidents and traffic congestions in real time, so that appropriate action can be taken to avoid problems resulting from the congestions. As mentioned earlier, using traditional approaches, such as using sensors and cameras, a limited area can be monitored. In this article, social media is used to overcome these limitations. These limitations may be overcome by use of social media (crowdsourcing) platforms. Crowdsourcing data from mobile applications have become an emerging data source for transportation systems due to its ubiquitous influence (Khatri, 2018; Silva, De Melo, Viana, Almeida, Salles & Loureiro, 2013 Almeida, Salles & Loureiro, 2013; Tsou, 2015). The application of crowdsourcing for real time traffic information has been used in the creation of applications such as Waze, Inrix, and Autonavi (Khatri, 2018; Pandhare & Shah, 2017; Silva et al., 2013). The kind of information these real time applications provide includes construction related details on various streets, traffic congestion due to events and games, safety and visibility, accidents, objects in the street, and congestion and traffic jams on the streets. These applications provide an interface for people to report and update the real time maps (Mahmood et al., 2017). However, the limitations of such systems are twofold. Firstly, reporting incidents requires users to log into the particular application and create reports that fall under one of the pre-determined incident categories. Secondly, these data are treated as proprietary and owned by private companies. Hence, it may be difficult for public agencies to obtain this data (Hasan & Ukkusuri, 2014; Li, Zhang, Wang & Ran, 2018; Yu, 2016; Zheng, Chen, Wang, Shen, Chen, Wang, Zhang & Yang, 2015 Shen, Chen, Wang, Zhang & Yang, 2015).

As the Twitter is used by a lot of people from all walks of life, enormous quantities of data are produced and used (Gong, Deng & Sinnott, 2015; Nguyen, Liu, Rivera & Chen, 2016 2016).

The use of Twitter comes with numerous advantages in comparison to the other social media platforms, for example:

- Tweets (posts on Twitter) are short, resulting in Twitter information being easy and quick to retrieve, store, and disseminate.
- A large number of tweets are geotagged, enabling users to identify the location of the tweet's origin. The geographical context can be used for spatial related analysis.
- The date and time features embedded in tweets make it useful for temporal related analysis.
- An exceptionally large coverage and user base across South Africa, makes the data available from areas that are much wider than those covered by cameras and sensors.

In the context of this research Twitter has used as a data source for collecting traffic related data. In a study by Kosala & Adi (2012) real time traffic data in the form of geotagged tweets from the city of Jakarta were retrieved and used for public purposes. In their model, tweets were filtered using some statistical properties before plotting the results on real time maps. Additionally, the researchers applied spatial and temporal clustering (Kosala & Adi, 2012). In another study, Chen and Krishnan (Chen & Krishnan) developed a real time Twitter monitoring system that automatically extracts tweets that are related to transportation safety. They calculated public sentiment and presented a visualized map by using OpenStreetMap. This model did not focus on incorporated machine learning functions for monitoring traffic congestions and their causes. Mai and Hranac (2013) made use of data from public social interactions on Twitter as a potential complement to traffic incident data. The researchers compared incident records from the California Highway Patrol with Twitter messages related to roadway events over the same time period. The aim of their study was on real time sentiment analysis. However, they focused only on incidents related to accidents and not on any other cause of traffic congestion. In another study, Elsafoury (2013), the author proposed a system

that uses traffic information shared on Twitter for real time analysis. Additionally, it uses a customized part of speech (POS) tagging technique that enables the extraction of the tweets and extraction of the tweets' geolocation, using various custom developed location libraries. Another tweet geolocation technique used was Google Geocode application programming interface (Google Geocode API). Not only regular users' information was extracted, but tweets were also used from official Twitter accounts that report traffic. The results are displayed on the map by highlighting the route, which is the road mentioned in the tweet. Wang et al.(2017) implemented a model to determine the "talking points" of people when facing traffic jams and to provide support to relevant authorities in making successful and effective decisions for real time traffic jam response and management. They applied the NLP and data mining techniques to extract traffic jam related information from Tianya.cn database. In another study Kumar, Jiang and Fang (2014) explored the use of Twitter for road hazard detection in California by aggregating hazard-related information posted by Twitter users. Their study focuses only on the detection of hazardous roads conditions using Twitter as a primary source of information. An architectural and novel harvesting and analytics approach that exploits tweets to identify near real time transport congestion, was presented by (Gong, Deng & Sinnott, 2015). Their research specifically presents an algorithm for targeted harvesting of tweets solely on a road network, using the definitive road network data for Australia. Spatial-temporal clustering algorithms are developed to identify spatial-temporal clusters of tweets on roads to identify potential traffic congestion. In another research, Lin and Li (Lin & Li, 2020) investigated the methods to predict the complicated behaviour of traffic flow evolution after traffic accidents using crowdsourcing data. They divided the traffic congestion conditions into four levels, namely severely congested, congested, slow moving and uncongested conditions. A hierarchical scheme was designed for identifying the most congested level and sequentially predicting duration of each level. Their model was validated using traffic accident data from an anonymous source in Beijing, China by embedding three machine learning algorithms, random forest (RF), support vector machine (SVM) and neural network (NN) in the scheme. Lucic, Wan, Ghazzai and Massoud (2020) developed an automated traffic alert system based on Natural Language Processing (NLP) which filters and extracts important traffic-related bullets. They made use of fine-tuning Bidirectional Encoder Representations from Transformers (BERT) language embedding model to filter the related traffic information from social media. They then applied a question-answering model to extract necessary information characterizing the report event such as its exact location, occurrence time, and nature of the events. In research conducted by Yu (2016) the author proposed a methodology to crawl, process and filter tweets that are posted by generic users and shared with the public. The tweets are processed to generate a dictionary of important keywords related to traffic incidences and geolocations of the tweets. The tweets are then classified into one of predetermined incidence categories. The model proposed in this article complements this work by conducting sentimental analysis and topic modelling to determine the reason for the traffic congestion. In addition, clustering is achieved by making use of SVM, RF, NB and KNN. The clustering algorithms were used to determine the most effective and accurate model.

3 METHODOLOGY

For the modelling of the proposed near real time Twitter-based model, Twitter feeds related to traffic conditions were obtained for a particular geographical location (in this study, the city of Pretoria was used). The relevant Twitters were subjected to data cleaning and pre-processing to obtain the applicable features such as the location of the origin of a particular post, the time when the tweet was posted, and the textual content of the post. The cleaning and pre-processing of Twitter feeds are essential as these texts are unstructured and usually contain informal, abbreviated words, misspellings and grammatical errors. The pre-processed data was subsequently used to build the model. Machine learning algorithms such as Random Forest and K-Nearest Neighbour (KNN) were used in the construction of the model. Natural language processing (NLP), using Python's Natural Language Toolkit (NLTK) and associated libraries, was conducted to enhance the suitability of tweets for carrying our sentiment analysis and topic modelling. The reason for the sentiment analysis was to link the text to the emotions (such as happiness, sadness or anger) of the people posting the tweets. Topic modelling was performed to help in identifying trends and the reason behind traffic congestions. Having created the model and evaluated it to determine the best performing one, new tweets (as they are posted) are then fed as input to the model to be classified, its sentiments, and the causes of the congestions are obtained in near real time. Fig. 2 depicts the major steps followed to achieve the objectives of this research.

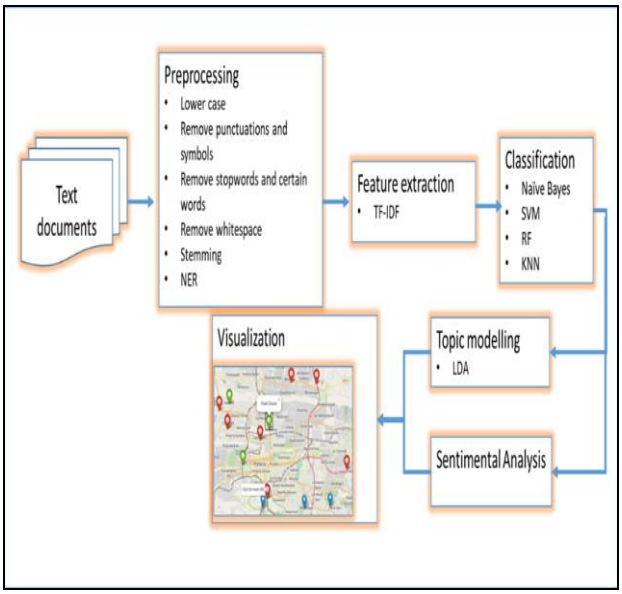

Fig. 2. A block diagram of the methods applied.

3.1 TWITTER DATA EXTRACTION

Twitter has numerous application program interfaces (APIs) that enable access to public tweets (Kosala & Adi, 2012; McHugh, 2015). Twitter Streaming APIs were used to extract real time tweets using geolocation and keywords filtering (Chen & Krishnan, 2013; Elsafoury, 2013; Mahmood et al., 2017; Mai & Hranac, 2013). Fig. 3 shows the Python code snippet used to extract the various features from each downloaded tweet. The extracted tweets from Tweeter were filtered for the word "traffic". While these tweets contained the

word “traffic” in its text, it does not guarantee that they are related to road vehicle traffic information. It could, for example, refer to network traffic or human traffic. For this research 5000 tweets were collected over a period of six months. Out of the 5000 tweets, there were 2,000 tweets that were found to be traffic related.

```
all_data=json.loads(data)
tweet=all_data["text"].encode("utf-8")
tweet=" ".join(re.findall("[a-zA-Z]+", tweet.decode('utf-8')))
blob=TextBlob(tweet.strip())
location=all_data["coordinates"]
geo=all_data["geo"]
place=all_data["place"]
```

Fig. 3. Snapshot of coding to extract relevant information from a tweet

Using geolocation, south-west and north-east latitude-longitude coordinates were obtained to create a boundary box (see Fig. 3). Only the tweets that have been geotagged and fall within the boundary box were harvested. Filtering by keywords, tweets with specific keywords were obtained.

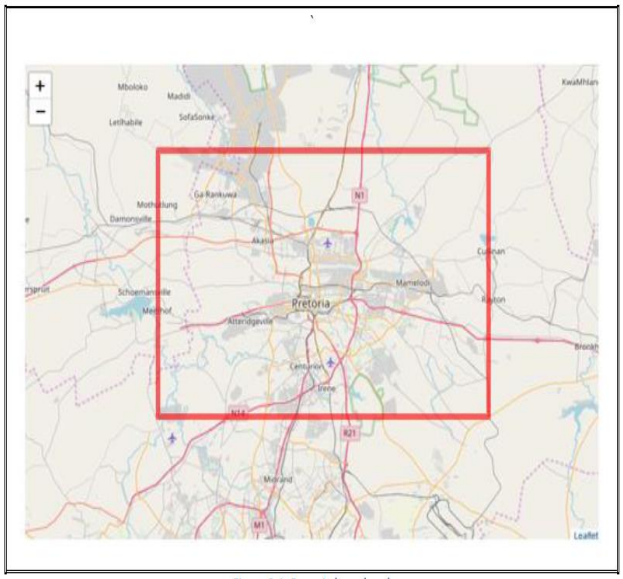

Fig. 4. Pretoria boundary-box.

3.2 CREATION OF DICTIONARIES

The traffic related dictionaries are stored as Python objects using a Pickle (a Python library) (Abrahams, Grosse-Kunstleve & Overloading, 2003). For this research, two dictionaries were created - the traffic-incidence dictionary and the non-traffic dictionary. These dictionaries are essential for creating the classification model and topic modelling. The traffic-incidence dictionary contains all words related to road vehicle traffic or synonyms of traffic related words together with the causes of the congestion. Table 1 shows the words that form the traffic-incidence dictionary. These words were obtained from the tweets collected.

TABLE 1.
WORDS IN THE TRAFFIC-AND-INCIDENCE DICTIONARY

CarCrash	Protests	Construction	Traffic lights down
Blocked lane	Strike	Flooding	Accident
Bus stuck	March	Veld fire	Stationary vehicle
Loadshedding	Fog	Congestion	Overturned

Similarly, the words in the non-traffic dictionary are shown in Table 2.

TABLE 2
WORDS IN THE NON-TRAFFIC DICTIONARY.

sextrafficking	antihumantrafficking
networktraffic	humantrafficking
trafficked	trafficking
trafficker	endhumantrafficking
humantrafficker	airtraffic

Using the words from the two dictionaries, the tweets are labelled as “traffic” and “non-traffic” and used in the creation of the classification models.

3.3 DATA PRE-PROCESSING

Once the Twitter data are collected, it goes through some pre-processing steps. The pre-processing of data is essential as the tweets are unstructured and contain irregular texts and some irrelevant information such as URLs and usernames, misspelt words, slang words, repeated characters (for example OOMMMGG). Commonly applied techniques in the pre-processing pipeline include tokenization, changing text to lower case, removal of stop word and stemming (Elsafoury, 2013; Gong, Deng & Sinnott, 2015, LALL). Raw text documents are stripped into individual tokens and the words of the documents are converted into lower case allowing for a case insensitive comparisons. This is often followed by the removal of all non-alphabetical characters like white space, punctuation marks, accent marks and other diacritics. Numbers are removed or changed into words. Stop words – words that commonly appear in documents but do not add any value to the task of text mining (such as a, an, the) - are removed from the dataset. Subsequently, stemming of the tokenized words is carried out to reduce it to a minimized form by extracting the root of the words. For instance, the root word of “going” is “go”. Stemming is an important pre-processing step as it helps in dimension reduction of the document representation. In addition to the steps mentioned above for pre-processing, Named entity recognition is used to extract the named entities in the tweet text. The named entities are pre-defined categories such as names of persons, organizations and locations. For example, a tweet such as “RT @netstartraffic: #PTATraffic @TrafficFreeflow Traffic lights out: In Pretoria CBD, corner of Stanza Bopape St and Du Toit St. Pointsmen” would pick up Stanza Bopape St and Du Toit St as location entities. Furthermore, locations are also collected by

means of the geotagged tweets and filtered according to Pretoria's boundary-box. This information assists in plotting the origin of the tweets. OpenStreetmap is used for the visualization of the location of the tweets.

3.4 FEATURE EXTRACTION

To perform clustering on text documents, an important step is to transform these documents into document representations. Various document representation techniques have been proposed to represent the texts, such as bag-of-words, term frequency-inverse document frequency (TF-IDF) (Lucic et al., 2020), one-hot vector and distributed representation of words. In this article TF-IDF is applied. TF-IDF is a statistical measure that reflects the importance of a word to a document in a collection or corpus (Christian, Agus & Suhartono, 2016). It is a commonly used document representation technique in which each document is represented by a vector d, in the term space, such that d = {tf1, tf2, …, tfn}, where tfi , i = 1, … ,n is the term frequency of the term ti in the document. A term frequency matrix is created to represent the document vectors. The document frequency dfi is the number of documents in the collection of N documents in which the term ti occurs. The inverse document frequency (idf) factor helps diminish the weights of terms that occur very frequently in the corpus and increases the weight of terms that occur rarely. A typical idf factor of this type is given by log(N/dfi). The weight of the term ti in a document is given by w = tfi × log(N/dfi). For example, tfidf for the word "traffic" in the corpus D is ["traffic", "is", "heavy"], ["there", "bad", "taxi", "accident"] for document/tweet number 1 is:

$$\text{tfidf(t="traffic", d=1,D)} = \text{tf("traffic", d=1)} \times \text{idf("traffic",D))} \quad (1)$$

$$\text{tfidf(t="traffic", d=1,D)} = 1 * \log 2 / |1| \approx 0.3010 = 0.3020 \rangle\!\rangle \quad (2)$$

Equations 1 and 2 show the term document matrix after applying tf-idf to a few of the tweets collected. In such a document representation, each row represents a term, each column a document, and each element is the term influence in the respective document (Lucic et al., 2020).

3.5 CLASSIFICATION MODEL

Although the tweets for this research was collected and filtered based on the word "traffic" appearing in the tweet text, not all tweets having the word "traffic" are related to road/vehicle traffic. Hence a classification model is built using machine learning to classify any tweet as either traffic related or non-traffic related. The two dictionaries - traffic-and-incidence dictionary and the non-traffic dictionary - are used for labelling the tweets as either traffic related or non-traffic related. This is achieved by looking up the words in these dictionaries and then labelling them appropriately. For example, the following tweet would be labelled as "traffic" related (Fig. 5) and tweet in Fig. 6 would be labelled as non-traffic related.

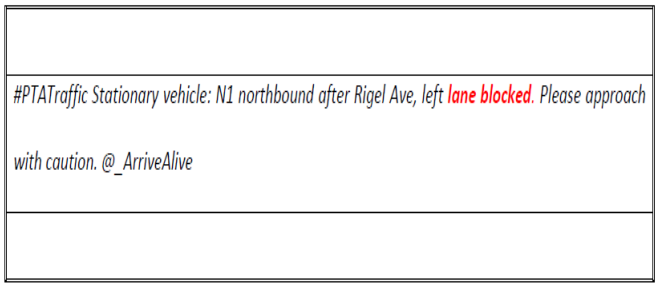

Fig. 5. Tweet labelled as traffic related.

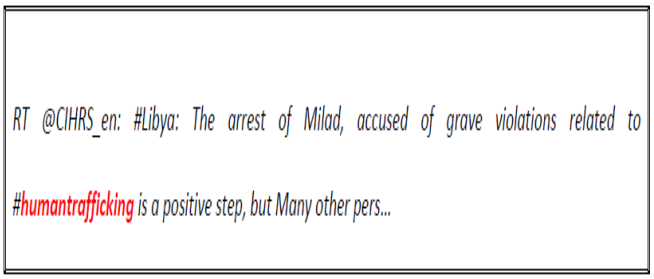

Fig. 6. Tweet labelled as not traffic related

The classification techniques and the performance evaluation matrix are presented below. The most accurate classification technique was selected for classifying real-time tweets. Naïve Bayes: The Naïve Bayes is a probabilistic classifier (Khatri, 2018) and is based on applying Bayes theorem with strong independence assumption between the features. It is one of the most prevalent classifiers for text classification (Chen & Krishnan, 2013; Kumar, Jiang & Fang, 2014; Pang, Lee & Vaithyanathan, 2002). The algorithm of Naïve Bayes Classifier entails the following: Given a set of variables, X = {x1, x2, ..., xn}, the objective is to construct the posterior probability for the event Cj among a set of possible outcomes C = {c1, c2, …., cn} like class "traffic" or "non-traffic". In a more familiar language, X is the predictors and C is the set of categorical levels present in the dependent variable. Using Bayes' rule:

$$Cj|x1,2,\ldots,xn) \propto (x1,2,\ldots,xn \mid Cj)p(Cj) \quad (3)$$

where $p(Cj \mid x1,x2,\ldots,xn)$ is the posterior probability of class membership, that is the probability that X belongs to Cj. Using Equation 3, a new case X from training data is labelled with a class level Cj that achieves the highest posterior probability. An advantage of Naïve Bayes is that it only requires a small amount of training data to estimate the parameters necessary for classification. Support Vector Machine: Support vector machines (SVMs) were initially established by concentrating on binomial classification. It makes use of hyperplane or a decision boundary to separate multidimensional data into classes (Chen, Patel, & Vasques, 2019; Chen & Krishnan, 2013; Pang, Lee & Vaithyanathan, 2002). The formula below is used to linearly separate data if the ordered w and b exists.

$$w^T x_1 + b \geq 1 \text{ for all } x_1 \in C_+ \text{ or } w^T x_1 + b \leq -1, \text{ for all } x_1 \in C\_ \quad (4)$$

The data that are retrieved from real life generally come in nonlinear format, meaning that it is not easy to draw a line to separate data into two groups. SVM achieves this nonlinearity challenge by coming up with the notion of a "kernel-induced feature space", which transfers the data into higher

dimensional space to convert it as separable data (Chen, Patel, & Vasques, 2019). Random Forest: Random forests can be considered as an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the classes (for classification problems) or mean prediction values (for regression problems) of the individual trees (SANDHU, 2021). Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. Random forests is considered as a highly accurate and robust method because of the number of decision trees participating in the process. Furthermore, Random forests do not suffer from the problem of overfitting as they take the average of all predictions hence cancelling out the biases. However, Random forest is reported to be show in generating predictions because it makes use of multiple decision trees, and the generated models are more difficult to interpret than the decision trees (Khatri, 2018). k-Nearest-Neighbours (kNN): where “k” represents the number of nearest neighbors, uses proximity in parameter space as a substitution for similarity. kNN is nonparametric, making no prior assumptions about the probability distribution of the observed data, and is arguably the simplest, yet effective, machine learning algorithm. A major limitation of kNN is that it relies on the selection of a “good value” for k and being a lazy learning method excludes it from many applications such as dynamic web mining.

3.6 PERFORMANCE EVALUATION MATRIX

For calculating the performance of machine learning, the dataset are split into training and testing sets. A 10-fold cross-validation technique was used. From the confusion matrix, number of true negatives (TN), false negatives (FN), true positives (TP) and false positives (FP) are used to compute the performance measure of the classifier. The Equations 13 - 16 show the performance measures computed

$$\text{Sensitivity} = TP/(TP+FN)\times 100\% \quad (5)$$

$$\text{Specificity} = TN/(TN+FP)\times 100\% \quad (6)$$

$$\text{Accuracy} = (TP+TN)/(TP+FP+TN+FN)\times 100\% \quad (7)$$

$$\text{F-score} = 2TP/(2TP+FP+FN) \quad (8)$$

In addition to the above mentioned performance measures, the Receiver Operating Characteristic curve (ROC curve) was used to compare the performances of the various classifiers.

3.7 SENTIMENTAL ANALYSIS

The goal of sentiment analysis is to determine the attitude or feeling a speaker or writer has towards a particular topic, or the overall contextual polarity of a document (Elsafoury, 2013). The feeling may be the writer or speaker’s judgement or evaluation of the topic concerned. In this study, the sentiment of the traffic related tweets is analysed. Sentiment analysis is the automated process of analysing text data and sorting it into sentiments, positive, negative, or neutral. For the purpose of conducting the sentiment analysis, this article makes use of dictionary based approach in conjunction with NPL parsing. The tools used are WordNet and Natural Language Toolkit (NLTK) libraries with Python 3, and TextBlob 0.7. TextBlob integrates NLTK’s WordNet interface to make interaction with WordNet easier. Every word in this English dictionary has a score assigned to it and by using these scores the syntactic representation of the sentence or document log probabilities are calculated. These probabilities indicate how positive or negative the sentiment in the tweet/document is. Fig. 7 shows the sentiment scores of randomly selected tweets.

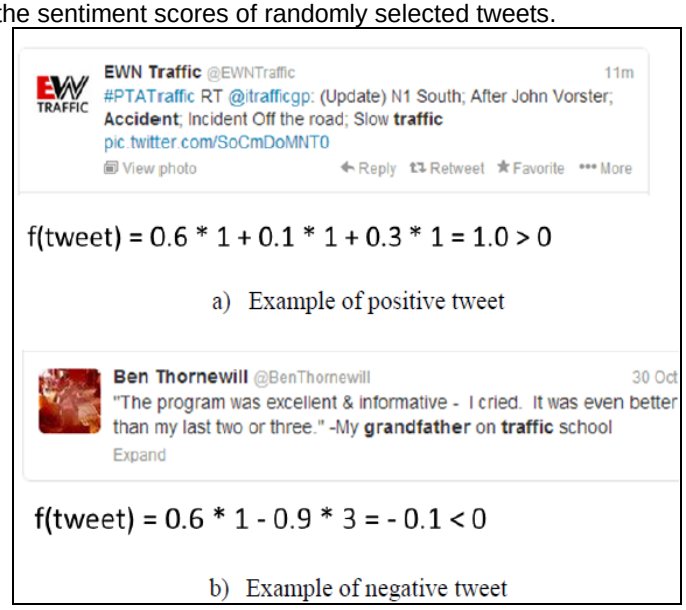

Fig. 7. Sentiment of tweet.

3.8 TOPIC MODELLING

Topic modelling may be regarded as a type of Bayesian statistical model for discovering the latent "topics" that occur in a collection of documents through machine learning (Muchaene 2020). Topic modelling may be considered as a technique to reveal, discover and annotate thematic structure in a collection of documents (tweets, in this case) (Blei, Ng & Jordan, 2003). Although the data used in topic modelling could be in the form of image, video, genetic and biochemical sequence, it is predominantly used with text data (Vayansky & Kumar, 2020). In topic modelling, the topics are defined as a distribution of words, with documents modelled as combinations of topics. The aim of topic modelling is to discover patterns of words used within documents. Topic models may also be considered as methods that group documents, that share similar words, and words that occur in a similar set of documents (Curiskis, Drake, Osborn & Kennedy, 2020). Topic modelling, like document clustering, may be used for clustering of documents by giving a probability distribution over a range of topics for each document (Erman, Arlitt & Mahanti, 2006). This form of clustering can be viewed as soft partition clusters. A Commonly used topic modelling technique is the LDA (Blei et al., 2003). LDA assumes that a discrete distribution of phrases known as topics is used in text generation, which then forms a document that contains a probabilistic distribution of topics (Blei, 2012). LDA seeks to define the underlying structure of the latent topic based on the information observed. LDA is considered as a mechanism to attach topical content to text documents and each document is seen as a mix of multiple distinct topics. LDA is a “bag of words” model in which the words do not have any particular order (Chen, Patel & Vasques, 2019).

3.9 VISUALIZATION

The geographical position, the sentiment expressed in the tweet and the reasons related to traffic conditions, are displayed on a map. Presenting this information in a visual manner will assist the road users in making an informed decision when planning their road journey (Wang et al., 2014). For the visualization of the road conditions, this study makes use of the Folium Library. Folium is a tool for visualizing data that has been manipulated in Python on an interactive leaflet map (Middleton, Middleton & Modafferi, 2014; Tsou, 2015). It enables both the binding of data to a map for choropleth visualizations as well as passing rich vector/raster/HTML visualizations as markers on the map.

4 IMPLEMENTATION, RESULTS AND DISCUSSION

As mentioned in Section 3, it is of paramount importance to pre-process the tweets. This step requires changing all text to lower case, removing symbols such as @, #, and punctuations. Additionally, stop words are removed and the words are stemmed. For creating and testing of the classification model, the tweets collected were labelled as “traffic” related or “non-traffic” related using the traffic-and-incidence dictionary and the non-traffic dictionary and divided into 80:20 ratio for training and testing of the model. The following classification models were created and their performance evaluated using accuracy, precision, recall and F-score. Table 3 shows the performance of the classifiers and Fig. 8 displays the ROC curve.

TABLE 3
SUMMARY OF THE PERFORMANCE OF THE CLASSIFICATION MODELS.

	Accuracy	Precision	Recall	F-score
Random Forest	0.775	0.76	0.77	0.77
Naïve Bayes	0.967	0.99	0.94	0.97
Support Vector Machine	0.870	0.89	0.83	0.86
K-Nearest Neighbour	0.775	0.76	0.77	0.77

It is observed that, for the given dataset, Naïve Bayes model performed the best. Hence, it is used for the classification of real time tweets.

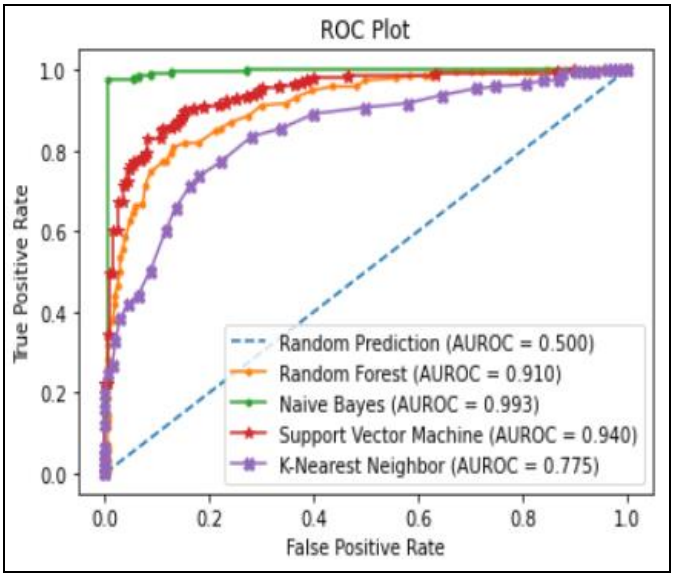

Fig. 8. ROC curve of the classifiers

For the current research, the sentiments expressed in the tweets were found to be mostly negative in nature (see Fig. 9). The red, green and blue points in Fig. 9 indicates negative, positive and neutral sentiments respectively.

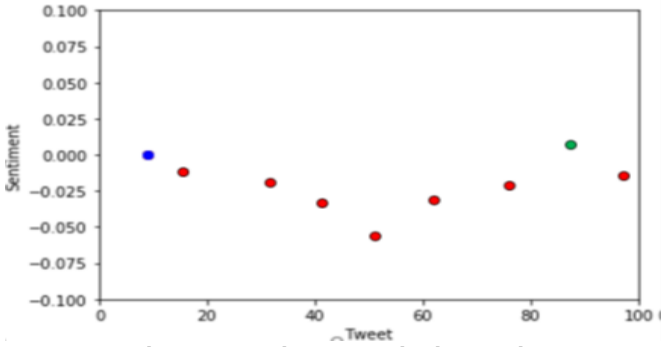

Fig. 9. Sentiment analysis graph.

In this article, LDA is applied for topic modelling. Fig. 10 depicts the ten most often appearing topics in the dataset of tweets. Since a topic is defined as a probability distribution over all words in the corpus that captures the salient themes that run through the corpus, these topics provide characterization of the traffic incidents identified by the token frequency and association analysis.

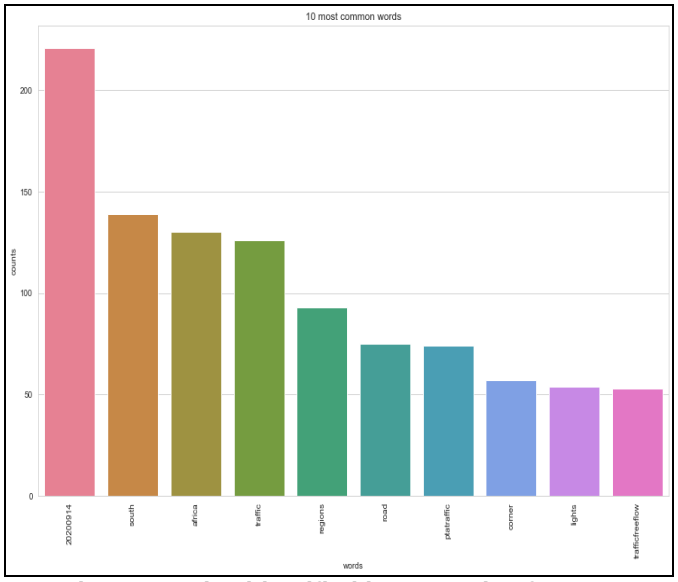

Fig. 10. Topics identified in a sample of tweets

Words like “queuing”, “blocked” and “freeflow” are the some of the words appearing in the top 10 word stems embodying information about incident response, while some secondary terms like “South Africa”, “regions”, and “please” further clarify the topic. A word cloud of the topics appearing in a sample of the dataset is shown in Fig. 11 below.

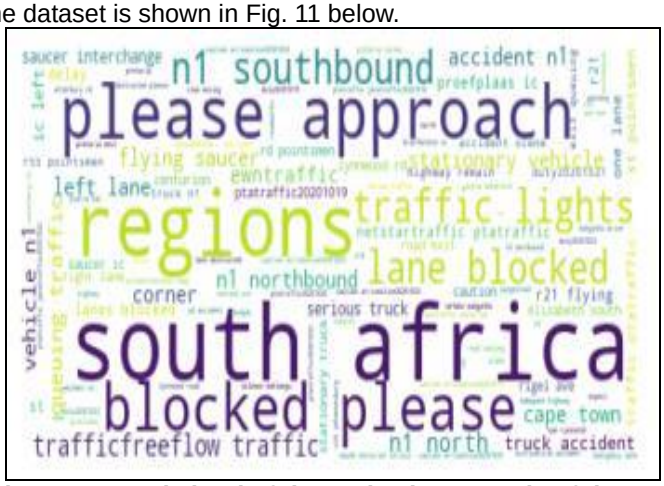

Fig. 11. A word cloud of the topics in a sample of dataset.

The geographical position, the sentiment expressed in the tweet and the reasons related to traffic conditions, are displayed on a map. Presenting this information in a visual manner will assist the road users in making an informed decision when planning their road journey. Figure 14 displays the sentiments of tweets and possible causes of traffic congestion in the central business district of Pretoria.

Fig. 12. Sentiments of tweets and possible causes of traffic congestion

For the analysis of a real-time tweet, assuming that the tweet shown below in Fig. 13 has just been received.

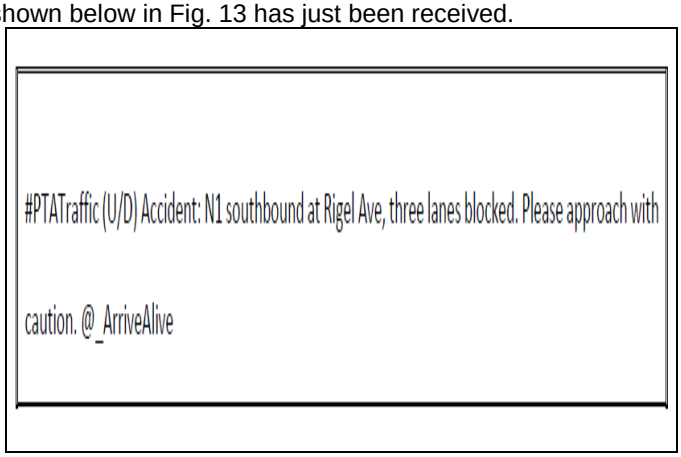

Fig. 13. Example of a real-time tweet received

After subjecting this tweet to the pre-processing pipeline and applying the selected classification model to it, sentiment analysis and topic modelling of the tweet is then plotted on the street map as shown in Fig. 14.

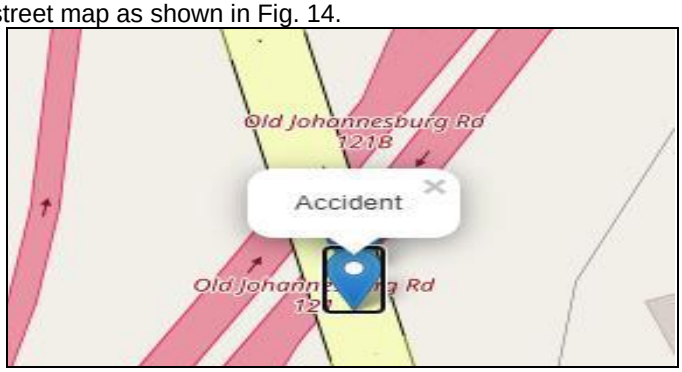

Fig. 14. Plotting of the received tweet on street map.

5 CONCLUSION AND RECOMMENDATIONS

In this research, it was established that using traffic related tweets could be a cost effective way to monitor traffic congestion in near real time. An important benefit of such a system is that it can be easily generalizable to other geographical areas. Another important contribution of this study is that it not only displayed a sentiment-aware visual traffic condition, but also shows the reasons for the congestions. This is an important contribution of this research. An additional contribution of this study is the machine learning

based classification model for near-real time traffic congestion monitoring. It is shown that Naïve Bayes was the most suitable algorithm for the given dataset. In this study, experiments are conducted using Random Forest, Naïve Bayes, Support Vector Machine and K-nearest neighbours. It is recommended that further research be conducted by implementing other feature extracting and feature engineering techniques to boost the accuracy of the algorithms. Using a larger dataset, other machine learning algorithms like the Artificial Neural Network (ANN) can be used to build the classification model. Additionally, a more comprehensive traffic related vocabulary may be used and developed to build the traffic-related dictionary.

REFERENCES

[1] F.A. Elsafoury, "Monitoring urban traffic status using Twitter messages", pp. 1–46, 2013.

[2] W. Musakwa, "The use of social media in public transit systems: the case of the Gautrain, Gauteng province, South Africa: analysis and lessons learnt", Proceedings of 19th International Conference on Urban Planning, Regional Development and Information Society, pp. 721-727, 2014.

[3] Y. Gong, F. Deng, and R.O. Sinnott, "Identification of (near) Real-time Traffic Congestion in the Cities of Australia through Twitter", Proceedings of the ACM First International Workshop on Understanding the City with Urban Informatics, pp. 7-12, 2015.

[4] T. Kurashima, T. Iwata, G. Irie, and K. Fujimura, "Travel route recommendation using geotags in photo sharing sites", Proceedings of the 19th ACM international conference on Information and knowledge management, pp. 579-588, 2010.

[5] Y. Somwanshi, V. Salegaonkar, and S. Sharma, "Understanding Social Media Phenomenon, Diversity and Research", International Journal of Computer Applications, 129(9):5-8., 2015.

[6] T. Mahmood, G. Mujtaba, L. Shuib, N.Z. Ali, A. Bawa, and S. Karim, "Public bus commuter assistance through the named entity recognition of twitter feeds and intelligent route finding", IET Intelligent Transport Systems, 11(8):521-529, 2017.

[7] C. Khatri, "Real-time road traffic information detection through social media", arXiv preprint arXiv:1801.05088, 2018.

[8] S. Grosenick, "Real-time traffic prediction improvement through semantic mining of social networks", Thesis (Master's), University of Washington. URI available at http://www.hdl.handle.net/1773/20911, 2012.

[9] S. Kharche, and L. Bijole, "Review on sentiment analysis of twitter data", International Journal of Computer Science and Applications, 2015.

[10] G.O. Leroke, and M. Lall, "A (near) real-time traffic monitoring system using social media analytics", Journal of Engineering and Applied Sciences, Vol 14, No. 21, pp 8055 – 8060, 2019.

[11] T.H. Silva, P.O.V. De Melo, A.C. Viana, J.M. Almeida, J. Salles, and A.A. Loureiro, "Traffic condition is more than colored lines on a map: characterization of waze alerts", International Conference on Social Informatics. Springer:309-318, 2013.

[12] M.H. Tsou, "Research challenges and opportunities in mapping social media and Big Data", Cartography and Geographic Information Science, 42(sup1):70-74, 2015.

[13] K. R. Pandhare, M.A. Shah, "Real time road traffic event detection using Twitter and spark", 2017 International conference on inventive communication and computational technologies (ICICCT). IEEE:445-449, 2017.

[14] S. Hasan, S.V. Ukkusuri, "Urban activity pattern classification using topic models from online geo-location data", Transportation Research Part C: Emerging Technologies, 44:363-381, 2014.

[15] L. Li, J. Zhang, Y. Wang, and B. Ran, "Missing value imputation for traffic-related time series data based on a multi-view learning method", IEEE Transactions on Intelligent Transportation Systems, 20(8):2933-2943, 2018.

[16] G. Yu, "Strategies of newsroom convergence: comparing UK and Chinese newspaper groups", Doctoral Thesis, University of Westminster, UK, 2016.

[17] X. Zheng, W. Chen, P. Wang, D. Shen, S. Chen, X. Wang, Q. Zhang, and L. Yang, "Big data for social transportation", IEEE Transactions on Intelligent Transportation Systems, 17(3):620-630, 2015.

[18] H. Nguyen, W. Liu, P. Rivera, and F. Chen, "Trafficwatch: real-time traffic incident detection and monitoring using social media", Pacific-asia conference on knowledge discovery and data mining pp. 540-551, 2016.

[19] R. Kosala, and E. Adi, "Harvesting real time traffic information from Twitter", Procedia Engineering, 50:1-11, 2012.

[20] F. Chen, and R. Krishnan, "Transportation sentiment analysis for safety enhancement", Technologies for Safe and Efficient Transportation, Carnegie Mellon University, 2013.

[21] E. Mai, and R. Hranac, "Twitter interactions as a data source for transportation incidents" 92th Annual Meeting on TRB, 2013.

[22] S. Wang, H. Dong, Y. Zhou, L. Jia, and Y. Qin, "Exploring traffic accident locations from natural language based on spatial information retrieval", 2017 29th Chinese Control And Decision Conference (CCDC), pp. 3490-3495, 2017.

[23] A. Kumar, M. Jiang, and Y. Fang, "Where not to go?: detecting road hazards using twitter", Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, pp. 1223-1226, 2014.

[24] Y. Lin, and R. Li, "Real-time traffic accidents post-impact prediction: Based on crowdsourcing data", Accident Analysis & Prevention, 145:105696, 2020.

[25] M.C. Lucic, X. Wan, H. Ghazzai, and Y. Massoud, "Leveraging Intelligent Transportation Systems and Smart Vehicles Using Crowdsourcing: An Overview", Smart Cities, 3(2):341-361, 2020.

[26] D. McHugh, "Traffic prediction and analysis using a big data and visualisation approach", Department of Computer Science, Institute of Technology Blanchardstown, 2015. Retrieved from http://leeds.gisruk.org/abstracts/GISRUK2015_submission_20.pdf

[27] D. Abrahams, R.W. Grosse-Kunstleve, and O. Overloading, "Building hybrid systems with Boost", Python. CC Plus Plus Users Journal, 21(7):29-36, 2003.

[28] M. Lall, "Exploring Interdisciplinary Nature of Postgraduate Research in the Field of Computing Using Text Mining: A Case Study", IEEE 15th International Conference on Industrial and Information Systems (ICIIS-2020), 500-505, 26th – 28 November, 2020. Ropar, India.

[29] H. Christian, M.P. Agus, and D. Suhartono, D. "Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF)", ComTech: Computer, Mathematics and Engineering Applications, 7(4):285-294, 2016.

[30] A.K. Sandhu, and R.S. Batth. "Software reuse analytics using integrated random forest and gradient boosting machine learning algorithm." Software: Practice and Experience 51, no. 4: 735-747, 2021.

[31] K. Shah, H. Patel, D. Sanghvi, et al. "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the

Text Classification", Augment Hum Res, pp. 5-12, 2020.
[32] L. Muchene, W. Safari, "Two-stage topic modelling of scientific publications: A case study of University of Nairobi, Kenya", PLoS ONE 16(1): e0243208. https://doi.org/10.1371/journal.pone.0243208, 2021.
[33] S.A. Curiskis, B. Drake, T.R. Osborn, and P.J. Kennedy, "An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit", Information Processing & Management, 57(2), 102034, 2020.
[34] J. Erman, M. Arlitt, and A. Mahanti, "Traffic classification using clustering algorithms", Proceedings of the 2006 SIGCOMM workshop on Mining network data, pp. 281-286, 2006.
[35] D. M. Blei, A. Y. Ng, anfd M. I. Jordan, "Latent dirichlet allocation", The Journal of machine Learning research, 3, pp. 993-1022, 2003.
[36] D. M. Blei, "Probabilistic topic models", Communications of the ACM, 55(4), pp. 77-84, 2012.
[37] D. M. El-Din, "Enhancement bag-of-words model for solving the challenges of sentiment analysis", International Journal of Advanced Computer Science and Applications, 7(1), 2016.
[38] X. Wang, K. Zeng, X. L. Zhao, and F. Y. Wang, "Using web data to enhance traffic situation awareness", 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 195-199, 2014.
[39] S. E. Middleton, L. Middleton, and S. Modafferi, "Real-time crisis mapping of natural disasters using social media", IEEE Intelligent Systems, 29(2):9-17, 2014.