International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 3133-3141 www.ijaem.net ISSN: 2395-5252

Road Traffic Event Detection Using Twitter Data,Machine Learning, and Apache Spark

Ms.Soumya Santhosha, Aishwarya Naik, Nisha D Costa, Shetty Kshama Umesh, Shrilekha

Computer science Department, Assistant Professor in Department of Computer Science and Engineering,Visvesvaraya Technological University
Computer Science Department, Students of Department of Computer Science and Engineering, Visvesvaraya Technological University

Submitted: 10-07-2021 Revised: 23-07-2021 Accepted: 26-07-2021

ABSTRACT—Road transportation is the backbone of modern societies, yet it costs annually over a million deaths and trillions of dollars all over the global economy. Social media such as Twitter have increasingly become an important source of information in many aspects of smart societies. The detection of road traffic events using Twitter data is one such area of a great many applications and great potential, facing major challenges concerning the management Various approaches were done on the subject have been proposed in recent years, but the methods and outcomes are in their infancy. This project proposes a method for the detection of road traffic-related events from tweets in India using machine learning and big data technologies. Firstly, we build and train a classifier using two machine learning algorithms, Naïve Bayes, and Rando forest algorithms, to filter tweets into relevant and irrelevant. Subsequently, we train the classifiers to detect multiple types of events including traffic conditions and social events. The results will analyze the traffic condition. To the best of our knowledge, this is the work on traffic event detection from Indian tweets using machine learning and Apache Spark.

I. INTRODUCTION

Road transportation is the backbone of modern cities and societies, yet it costs, annually, 1.25 million deaths and 20-50 million people injured across the world. Moreover, road traffic congestion is the most significant problem in modern cities. The annual cost of congestion to the US economy alone exceeds $305 billion[3]. The increasing number of vehicles, social events, lane closures, road works, adverse weather, and other unexpected incidents have a negative impact on traffic flow and cause traffic congestions. Therefore, those causes, namely events, should be detected in an efficient and timely manner in order to support decision-making and set management strategies to reduce or eliminate congestion. Smart cities provide "State of the art approaches for urbanization, having evolved from the knowledge-based economy, digital economy, and intelligent economy". In smart cities and societies, a large amount of diverse information is produced daily by heterogeneous sources including GPS, cameras smartphones as well as user-generated content from social media[5]. Such data offer the potential for developing novel solutions that will support decision-making for smart transportation. In recent years, several approaches related to transportation in smart cities have been proposed, for example, autonomic transportation systems and intelligent disaster management. Social media such as Twitter and Facebook are relatively inexpensive and conveniently available sources of information comparing to physical sensors that cost greatly to install at a large scale to monitor the traffic flow. Twitter is one of the best popular microblogging media used for communication and sharing personal status, events, news, etc. Twitter allows users to post text messages called tweets. A huge amount of data is posted by millions of users on various topics including transportation and road traffic.

II. LITERATURE SURVEY

In this section, we review some literature related to social media-based event detection. Subsequently, we discuss the existing works about the detection of various events (not necessarily traffic events) from Indian social data. These sections do not include any works that use big data. We analyzed the works on traffic event detection that use big data technologies.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 3133-3141 www.ijaem.net ISSN: 2395-5252

A.Traffic event detection using social media:

In recent years, researchers had proposed different approaches in online event detection from social media. Agarwal et al.[11] identified the complaints reported in road irregularities and worst road conditions. After extracting the information; such as the problem and the location, they applied a rule-based classifier and categorized them, nearly-useful, and irrelevant complaint reports. Sakaki et al.[11] focus on detecting huge traffic information and weather information. They classified Japanese tweets into positive and negative classes using the Support Vector Machine classifier. Additionally, Klaithin and Haruechaiyasak extracted information with respect to traffic using lexicon-based and rule-based techniques. They applied ML classifiers based on the Naive Bayes to classify tweets about traffic into six categories include accident, announcement, question, orientation, request, and sentiment.

B.General event detection from Arabic Tweets:

The amount of research about analyzing Indian social information for event detection is considerably limited compared to what is done in other languages. In this, the tweets are classified into event and non-events tweets using a Naïve Bayes model. Also, they applied an online clustering algorithm to classify the topic of an event. Moreover, they extended the work and used the clustering algorithm to detect the events of the riots. Other researchers trained classification algorithms by using the training matrix that contains the selected terms and their corresponding TF-IDF (Term Frequency-Inverse Document Frequency) weights. They tested several algorithms. The results show that RF was promising in terms of accuracy. However, the model was trained on a small dataset of about 1000 Indian tweets to detect one type of event which is a high-risk flood. In addition, none of the above-discussed on event detection from the Arabic language used big data platforms. Furthermore, their main focus was not on traffic events such as traffic jams.

C.Traffic Events detection using big data technologies

Alomari and Mehmood applied SAP HANA, which is a memory processing platform to identify the Indian tweets related to traffic conditions. In addition, they extracted the causes of traffic congestion. Furthermore, they extended the proposed sentiment analysis approach for traffic events. However, their approach was dictionary-based. They did not use machine learning techniques. Salas et al.[10] proposed a framework for the detection of traffic events from tweets in the English language using Apache Spark and Python ML algorithms. Additionally. They used the RF classification algorithm and classified the tweets into traffic and non-traffic-related tweets. Suma et al.[10] have analyzed tweets to detect events related to road traffic. They proposed a classification model to identify the tweets into traffic-related and non-traffic-related by using logistic regression with stochastic gradient descent. To detect events, they identify the most frequent among the traffic-related tweets. They improved the methodological and event detection aspects of their work. All of these approaches used supervised classification algorithms on the Apache Spark platform. They analyzed tweets in the English language. Lau used the Latent Dirichlet Allocation (LDA) topic modeling module for unsupervised topic mining.

III.METHODOLOGY

Fig.1 shows the proposed architecture for traffic event detection from tweets in the dataset using supervised Naive Bayes and Random Forest algorithms and Apache Spark. It consists of six main components: (1) Data collection and storage component, (2) Data pre-processing component, (3) Feature extractor component, (4) Tweet filtering component, (5) Event detection component, and (6) Validation and results in visualization component. First, the data are collected using Twitter API, and the fetched JSON objects are stored in MongoDB. After removing the duplicates, we split the tweets into a labeled and unlabeled dataset.

A. Data collection

Tweets are collected via Twitter REST API using geolocation filtering to obtain tweets posted in India. In addition, we collected tweets in hashtags that are usually used to post about events in cities such as. We collected all Indian tweets in the period, since our data required scalable and flexible schemas-based storage, we selected NoSQL databases instead of relational databases. The collected tweets are stored in MongoDB, which is a document-oriented database suitable for storing and managing Big Data-sized collections of documents like text. The fetched JSON objects from Twitter API are inserted into the database.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 3133-3141 www.ijaem.net ISSN: 2395-5252

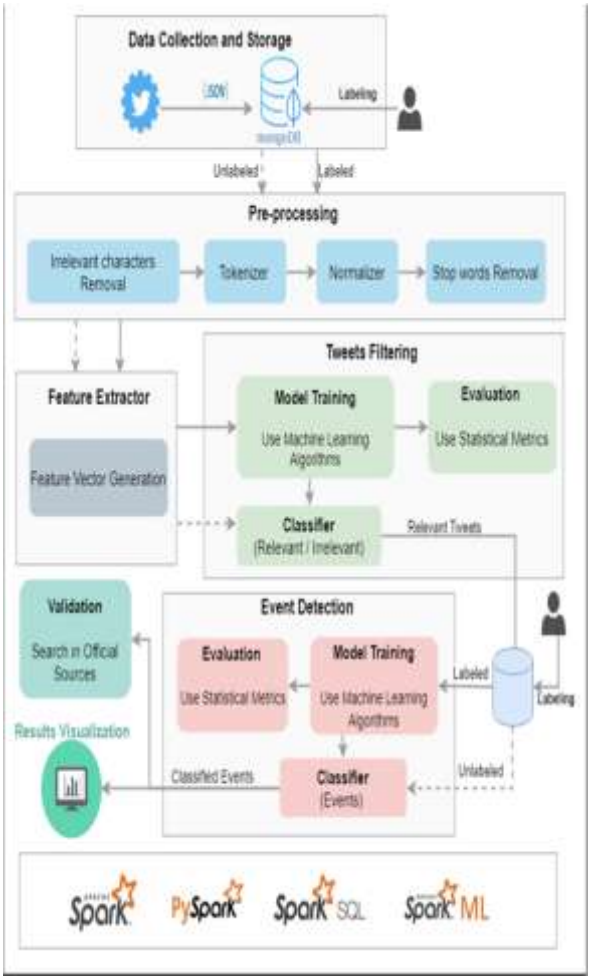

Fig. 1. Architecture of the proposed event detection system using Twitter data, machine learning, and Apache Spark

Additionally, the Tweets contain attributes including (i) 'created_at', which represents the time when the tweet was posted, and (ii) 'full_text' contains the message content. After that, we checked the redundancy and removed duplicate tweets (retweets). The total number of tweets after removing the duplicates is about 1 million.

B.Pre-processing

Pre-processing the text is an essential task since the Indian morphology is rich and the text usually has typos or grammatical mistakes. Also, it is a critical step to reduce the amount of noise before classification because performing analysis directly on text may lead to poor results.

Algorithm1 summarizes the main pre-processing steps. First, SparkConnector is used to connect MongoDB. Then, the tweets are loaded and saved in Spark DataFrame. The next step is iterating over the tweets to remove all numbers, English alphabets, and punctuations

such as commas (,), period (.), semi-colons (;), colons (:), question marks (?), etc... Removing punctuations helps to

reduce the size of the feature set since users rarely use formal language. Therefore, most of the punctuation marks are not used properly, and keeping them will not give any valuably the hash (#) and underscore (_) symbols and keep the keywords because it almost includes useful information like the place/ event name.

C.Feature Extraction

We use Feature Extractors algorithms provided in the Spark ML package. We apply TF-IDF (Term Frequency-Inverse Document Frequency), which is a measure of how important a word is to a document (tweet). The TF-IDF is merely the product of TF and IDF. The TF(t, d) is the frequency of the appearance of term t in document d while the IDF is a numerical measure

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 3133-3141 www.ijaem.net ISSN: 2395-5252

of how much information a term provides. The IDF is calculated using the following equation:

$$IDF(t,D) = \log \quad (1)$$

where |D| is the total number of documents present in the collection D.The Document Frequency DF(t, D) is the number of documents where the term t appears.

$$TFIDF(t, d, D) = TF(t, d) \cdot IDF(t, D) \quad (2)$$

To generate the term frequency (TF) vectors, we used the CountVectorizer algorithm. The algorithm gets the list of tokens in the 'Tokens' column as input and then converting them into token counts vectors. Then, the resultant term frequency vectors are passed to the IDF algorithms. After that, the IDFModel will rescale the feature vectors, and the output will be stored in a new column named 'Features'. This column is passed as input for classification algorithms.

D.Classification (Tweet Filtering)

Since not all the collected tweets are relevant to traffic, we filter the tweets before detecting events. So, we build a classifier to filter out the irrelevant tweets to traffic. We used machine learning algorithms in the Spark ML package.

We split the manually labeled data into training sets (80%) and testing sets (20%). After that, we build and train the model using Naïve Bayes and Random Forest algorithms. The models in this are trained on the training set. To find the best algorithm, we evaluate them over the testing set. The common statistical metrics, such as precision, accuracy, recall, and F-score are used to evaluate the trained classifier. To clarify the meaning of these metrics, we refer to traffic-related tweets as positive class and none related as negative class.

E.Event Detection

For event detection, we build and train the classifier using the Naïve Bayes and Random Forest algorithms. To train the events classifier, the authors manually label part of the filtered data from the previous step into eight event categories, which are Fire, Weather, Social Events, Traffic Condition, Roadwork, Road Damage, Accident, and Road Closures. The traffic condition category includes negative and positive tweets about the traffic condition. For Fire events, all tweets about fires are included under this category even though it is not a vehicle fire because it mayaffect negatively thetraffic and cause congestion. Furthermore, for the social event, we focus only on the events that could affect the traffic. affect negatively thetraffic and cause congestion. Furthermore, for the social event, we focus only on the events that could affect the traffic.

During our analysis, we notice that some event types have a large number of tweets compared to the others. So, we divided them into small-scale and large-scale events based on the number of tweets. The small-scale events are Traffic Condition, Roadwork, Road Damage, Accident, and Road Closures. The number of tweets for these events is small compared to Fire, Weather, and Social Events. So, we consider them as large-scale events. Furthermore, we have a multi-label classification problem, since the classes (event types) are not mutually exclusive and the same tweet can belong to more than one class. For example, the tweet can be about Traffic conditions and Accidents at the same time. To address this problem, we treat each label as a separate binary classification problem. Thus, we trined eight binary classifiers. For each event type, we consider the tweets about the event as positive while all the remaining tweets about the other types of events as negative. However, this will lead to imbalance sampling where the number of negative is larger than the positive. To adjust the class distribution and eliminate the effect on evaluation results, we perform undersampling for the negative (majority) class using the random undersampling method to make the data set balanced before evaluation. We prefer undersampling by removing samples from the majority class instead of oversampling by taking repeated samples from the minority class. Since the number of the negative labels is very large compared to the positive where it contains all the tweets about the other event types. Even though undersampling leads to loss of information, in our case, correctly classifying the negative labels is less important than the positive labels. Moreover, after detecting the events, we extract the time of occurrence using the time, and date information from the 'created_at' attribute in the tweets object. Furthermore, we extract information about each event including location information using the top frequent terms since people usually refer to the event place using the hashtag. For model evaluation, we use the same evolution method explained in section 4.C. To validate the effectiveness of our event detection approach, we extract the top vocabularies from the tweets of each detected event. Then, we use these vocabularies to search in the official news/newspapers websites to confirm the occurrence of the events. After that, we

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 3133-3141 www.ijaem.net ISSN: 2395-5252

compare the extracted information by our method including time and location with the real information in the official sources.

IV. RESULTS AND DISCUSSION

A. Login Page and Analysis

This section involves the opening of the login page. Fig.2 shows the login page which has username and password once the user enter the correct username and password the another page will open that is Analysis page

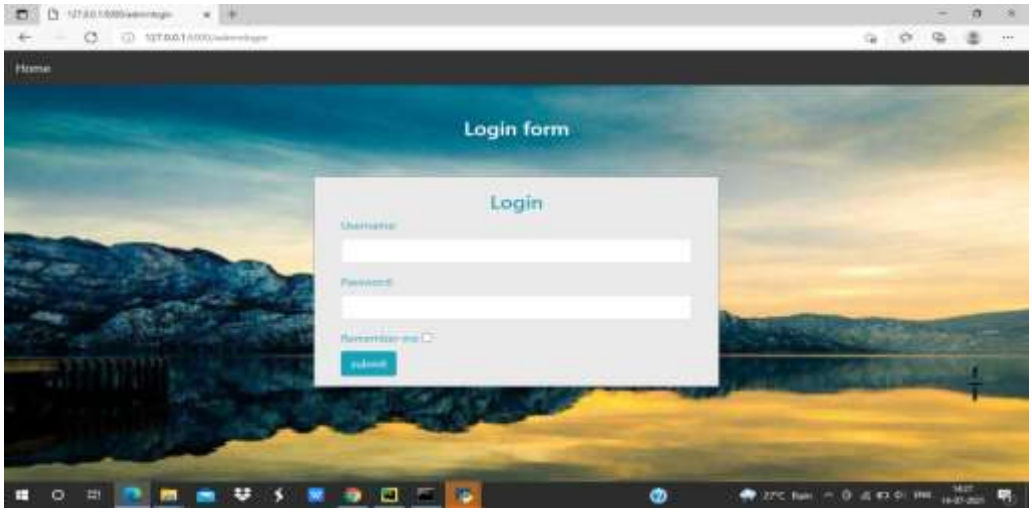

Fig.2 Login Page

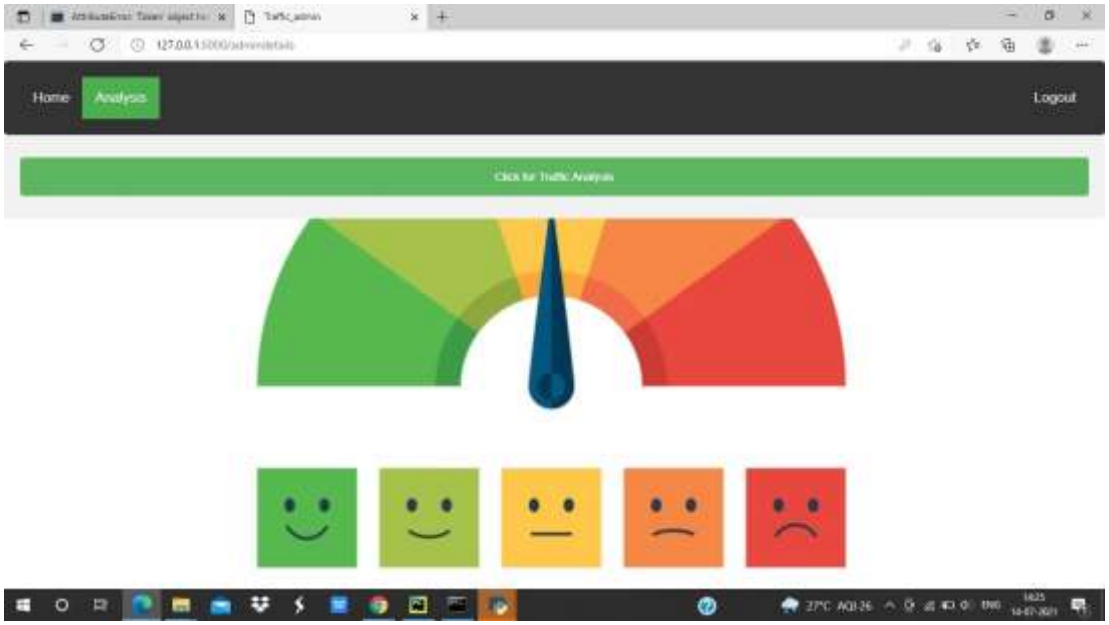

Fig.3 Click for Analysis

Fig.3 shows the Analysis page where on clicking the click for traffic analysis at the background it will analyse the static dataset using the two algorithms(Naïve Bayes and Random Forest).

B. Results for event detection

This section involves the discussion on the classification model and outcomes. First, we checked the performance of different machine learning algorithms such as Multinomial Naive

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 3133-3141 www.ijaem.net ISSN: 2395-5252

Bayes and Random Forest algorithm based on traffic dataset.

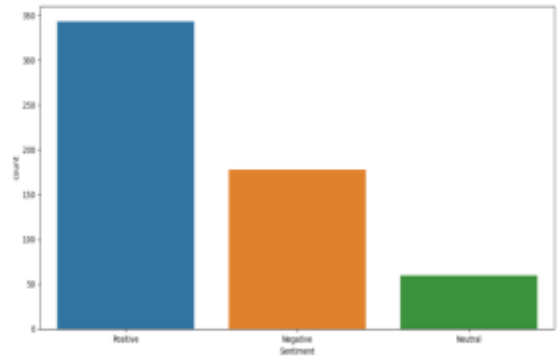

Fig.4 Dataset analysis

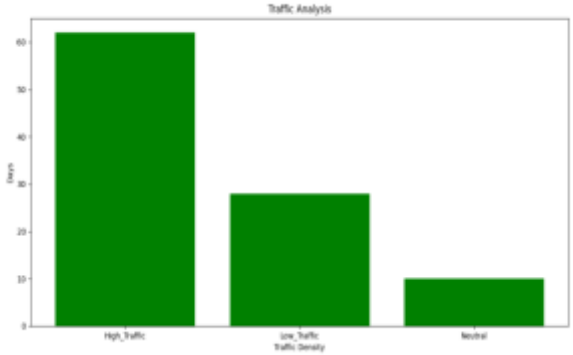

Fig.5 Traffic Analysis

Fig.4 shows the analysis of the static dataset with the three-parameter which includes positive, negative, and neutral sentiments with respect to the count of the dataset. As the positive parameter indicates the high traffic, in this we found there is more traffic.Fig.5 shows the traffic analysis of the tweets in the dataset. The traffic Analysis was carried on the basis of the day with respect to the three parameters,i.e, High_traffic, low_traffic and Neutral Density. In this analysis, it indicates High_traffic is more count of traffic compared to other parameters.

The performance of the two classification algorithms (Naïve Bayes and Random Forest algorithms) for tweet filtering is measured using the evaluation metrics. Fig.6 shows that Random Forest is better than Naïve Bayes algorithms in terms of Accuracy.Fig.7 The above figure shows the confusion matrix for the datasets. Confusion Matrix is a matrix that is used for the description of the performance of a classification model on a dataset for which the true values are known to us.Fig.8 shows the final accuracy result of both MultinomialNB and Random Forest algorithms. The MultinomialNB algorithm gives 79% and the Random Forest algorithms have an accuracy of 85%, which is better thanMultinomialNB.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 3133-3141 www.ijaem.net ISSN: 2395-5252

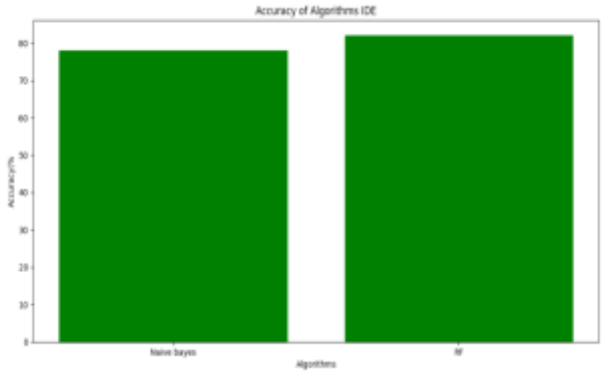

Fig.6 Accuracy of Algorithm IDE

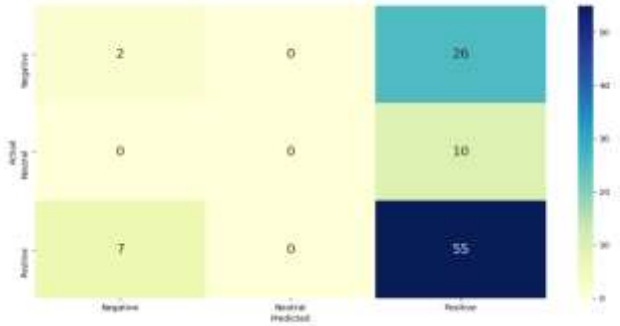

Fig.7 Confusion Matrix

Fig.8 Results

V. CONCLUSION

We focused on detecting road traffic related events to enable smarter transportation. We proposed a method for automatic detection of traffic events from tweets in dialect using machine learning algorithms and Apache Spark platform. Since the raw text is not suitable as direct input to classification, the text was divided into tokens and normalized after removing numbers, punctuation, diacritic, and non-Arabic words. TF-IDF was selected as weighting schemes and the tokens are converted into a vector of terms.

Furthermore, we trained a classifier to filter tweets intorelevant (to traffic) and irrelevant. We used three machine learning algorithms, Naïve Bayes, SVM, and logistic regression. Subsequently, we trained the other classifiers to detect the occurrence of multiple traffic-related events in Saudi Arabia. We extracted information about each event including location information using the top frequent terms. Then, we searched in the official sources such as the newspaper websites to validate our approach. The results showed that our method is able to detect the traffic-related events, as well as

their location and time, automatically, without any prior knowledge of the events.

REFERENCES

[1]. G. Cookson, "World Health Organization: Road traffic injuries."[Online]. Available: https://www.who.int/news-room/factsheets/detail/road-traffic-injuries. [Accessed: 18-Feb-2019].

[2]. "INRIX Global Traffic Scorecard." [Online]. Available: http://inrix.com/scorecard/. [Accessed: 18-Feb-2019].

[3]. R. Mehmood, B. Bhaduri, I. Katib, and I. Chlamtac, Eds., Smart Societies, Infrastructure, Technologies and Applications, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering (LNICST), Volume 224, vol. 224. Cham: Springer International Publishing, 2018.

[4]. J. Schlingensiepen, F. Nemtanu, R. Mehmood, and L. McCluskey,"Autonomic Transport Management Systems—Enabler for Smart Cities, Personalized Medicine, Participation and Industry Grid/Industry 4.0," in Intelligent Transportation Systems – Problems and Perspectives, Volume 32 of the series Studies in Systems, Decision and Control, Springer International Publishing, 2016, pp. 3– 35.

[5]. Z. Alazawi, O. Alani, M. B. Abdljabar, S. Altowaijri, and R. Mehmood, "A Smart Disaster Management System for Future Cities," WiMobCity '14. Int. Work. Wirel. Mob. Technol. Smart Cities, pp. 1– 10, 2014.

[6]. D. Wang, A. Al-Rubaie, J. Davies, and S. S. Clarke, "Real time road traffic monitoring alert based on incremental learning from tweets," in In 2014 IEEE Symposium on Evolving and Autonomous Learning Systems (EALS), 2014, pp. 50–57.

[7]. M. Ni, Q. He, and J. Gao, "Forecasting the Subway Passenger Flow under Event Occurrences with Social Media," IEEE Trans. Intell. Transp. Syst., vol. 18, no. 6, pp. 1623–1632, 2017.

[8]. S. Wang, L. He, L. Stenneth, P. S. Yu, and Z. Li, "Citywide traffic congestion estimation with social media," in Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '15, 2015,pp.1–10.

[9]. S. Agarwal, N. Mittal, and A. Sureka, "Potholes and Bad Road Conditions- Mining Twitter to Extract Information on Killer Roads," ACM India Jt. Int. Conf. Data Sci. Manag. Data CoDS-COMAD 2018, 2018.

[10]. A. Oussous, F.-Z. Benjelloun, A. A. Lahcen, and S. Belfkih, "Big Data technologies: A survey," J. King Saud Univ. - Comput. Inf. Sci., 2017.

[11]. T. Sakaki, Y. Matsuo, T. Yanagihara, N. P. Chandrasiri,and K. Nawa, "Real-time event extraction for driving information from social sensors," in Proceedings - 2012 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems, CYBER 2012, 2012, pp. 221–226.

[12]. E. D'Andrea, P. Ducange, B. Lazzerini, and F. Marcelloni, "RealTime Detection of Traffic from Twitter Stream Analysis," IEEE Trans. Intell. Transp. Syst., vol. 16, no. 4, pp. 2269–2283, 2015.

[13]. R. Y. K. Lau, "Toward a social sensor based framework for intelligent transportation," in 2017 IEEE 18th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2017, pp. 1–6.

[14]. N. Pavlopoulou, A. Abushwashi, F. Stahl, and V. Scibetta, "A text mining framework for Big Data," Expert Updat., vol. 17, no. 1, 2017.

[15]. S. Klaithin and C. Haruechaiyasak, "Traffic Information Extraction and Classification from Thai Twitter," Comput. Sci. Softw. Eng. (JCSSE), 2016 13th Int. Jt. Conf., pp. 1–6, 2016.

[16]. A. Kumar, M. Jiang, and Y. Fang, "Where not to go?: detecting road hazards using twitter," in Proceedings of the 37th international ACM …, 2014, vol. 2609550, pp. 1223–1226.

[17]. D. A. Kurniawan, S. Wibirama, and N. A. Setiawan, "Real-time Traffic Classification with Twitter Data Mining," in In 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE), 2016, pp. 1–5.

[18]. D. Semwal, S. Patil, S. Galhotra, A. Arora, and N. Unny, "STAR: Real-time Spatio-Temporal Analysis and Prediction of Traffic Insights using Social Media," in In Proceedings of the 2nd IKDD Conference on Data Sciences, 2015, p. 7.

[19]. M. R. Alifi and S. H. Supangkat, "InformationExtraction for Traffic Congestion in Social Network,"

[20]. R. Hanifah, S. H. Supangkat, and A. Purwarianti, "Twitter information extraction

for smart city," Proc. - 2014 Int. Conf. ICT Smart Soc. "Smart Syst. Platf. Dev. City Soc. GoeSmart 2014", ICISS 2014, pp. 295–299, 2014.

[21]. P. Tejaswin, R. Kumar, and S. Gupta, "Tweeting Traffic: Analyzing Twitter for generating real-time city traffic insights and predictions," Proc. 2nd IKDD Conf. Data Sci. - CODS-IKDD '15, pp. 1–4, 2015.

[22]. N. Dhavase and A. M. Bagade, "Location identification for crime & disaster events by geoparsing Twitter," 2014 Int. Conf. Converg. Technol. I2CT 2014, pp. 2–4, 2014.