IEEE Open Journal of
Intelligent Transportation
Systems

Received 14 April 2020; revised 19 June 2020 and 8 September 2020; accepted 10 September 2020. Date of publication 15 September 2020; date of current version 5 October 2020.

Digital Object Identifier 10.1109/OJITS.2020.3024245

Empowering Real-Time Traffic Reporting Systems With NLP-Processed Social Media Data

XIANGPENG WAN (Graduate Student Member, IEEE), MICHAEL C. LUCIC (Graduate Student Member, IEEE), HAKIM GHAZZAI (Senior Member, IEEE), AND YEHIA MASSOUD (Fellow, IEEE)

School of System and Enterprise, Stevens Institute of Technology, Hoboken, NJ 07030, USA

CORRESPONDING AUTHOR: H. GHAZZAI (e-mail: hghazzai@stevens.edu)

A part of this work has been accepted for publication in IEEE Technology and Engineering Management Conference (TEMSCON), Novi, MI, USA, 2020 [1].

ABSTRACT Current urbanization trends are leading to heightened demand of smarter technologies to facilitate a variety of applications in intelligent transportation systems. Automated crowdsensing constitutes a strong base for ITS applications by providing novel and rich data streams regarding congestion tracking and real-time navigation. Along with these well-leveraged data streams, drivers and passengers tend to report traffic information to social media platforms. Despite their abundance, the use of social media data in ITS has gained more and more attention as of now. In this article, we develop an automated Natural Language Processing (NLP)-based framework to empower and complement traffic reporting solutions by text mining social media, extracting desired information, and generating alerts and warning for drivers. We employ the fine-tuned Bidirectional Encoder Representations from Transformers classification model to filer and classify data. Then, we apply the Question-Answering model to extract necessary information characterizing the reported incident such as its location, occurrence time, and nature of the incidents. Afterwards, we convert the collected information into alerts to be integrated into personal navigation assistants. Finally, we compare the recently posted incident reports from both official authorities and social media in order to provide more complete incident pictures and suggest some open research directions.

INDEX TERMS Intelligent transportation system, natural language processing, social media, BERT, question-answering, name-entity recognition.

I. INTRODUCTION

AS THE world's urban population rapidly grows, cities all over the world are experiencing severe traffic congestion. According to the 2019 mobility report issued by the New York State Department of Transportation (DOT), the average midtown core speed is consistently 30% slower than that of one of the central business districts in Manhattan and, in general, the average speed has decreased by 20% from 2010 to 2019 [2]. Traffic congestion causes widespread socioeconomic and environmental issues for cities. According to a study of the Partnership for New York City, the New York City economy is expected to incur losses of over $20 billion over the next five years because of traffic congestion that leads to unforeseen business expenses, wasted fuel consumption from increased idling, and increased vehicle maintenance costs [3]. In order to address the severe congestion problems, New York City has begun to charge fees for bringing vehicles into the city starting in 2019 [4]. In addition to the rapid growth of the number of vehicles on the road, unexpected roadway conditions caused by roadside construction/maintenance, car accidents, asynchronous traffic signals, changes in driver behavior, among other factors also cause the traffic congestion in urban areas [5]. In fact, the authors of [6] have show that every minute of unexpected roadway conditions leads to an average of four additional minutes of traffic delay. The propagation of traffic disturbances along with a lack of rapid information sharing between drivers can easily lead to widespread delays in an entire city's traffic network. To address the root problem of urban traffic congestion,

The review of this article was arranged by Associate Editor Jia Hu.

many technical solutions have been proposed to provide real-time traffic information to drivers. This information can lead them to reroute their trip to avoid congested traffic links in advance [7].

Novel technologies have been implemented by municipal governments across the United States to provide the basis for such information-based systems. In Columbus, OH, USA, the city utilizes vehicle-to-infrastructure (V2I) communications systems to collect traffic flow data, which is utilized to time the traffic signal cycles.[1] In Atlanta, GA, USA, a 2.3-mile smart corridor was opened in 2017. This smart corridor incorporates Internet of Things (IoT) technologies such as sensors, cameras, and wireless communications technology to collect traffic data. This data powers various interconnected subsystems to improve traffic flow efficiency and safety in the corridor [8]. In San Francisco, CA, USA, local authorities installed sensors alongside the roads to identify open street parking, which reduces the amount of time and energy wasted by drivers hunting parking spots [9].

Researchers have developed systems independently of municipal governments as well. The authors of [10] introduced a framework to collect accident evidence from vehicular sensors to provide a new data source related to traffic incidents. The authors of [11] developed a support vector machine (SVM) model that classifies short-term traffic conditions as: smooth, basically smooth, mild congestion, moderate congestion, and serious congestion. These classifications are based on the data collected in the area of interest, and are then used to predict future traffic conditions. These innovations contribute to a full stack of technologies that can be leveraged to create an automated sensing pipeline for ITS. As the technologies improve, collecting such data in urban locations will become easier [12], opening access for additional information for drivers. This increased information can lead to overall improvement in the ITS operational efficiency.

Research related to vehicular social networks (VSN) has emerged in parallel with the implementation and exploration of IoT technology in ITS [13], [14]. VSNs represent a solution that will improve the information sharing among drivers through vehicular communication technologies [15]. Data collection and propagation techniques have improved in recent years. In the past, most relied exclusively on a single entity, while most of them now involve the collaboration from multiple participants, including road users, road side units, and Web-based information to combat traffic congestion. For example, Waze built a navigation application that collects vehicle speed, while offering user input functionality to report traffic conditions for other drivers. One drawback to this application is that incident reporting requires manual input from the user, which is dependent on the participation and veracity of the user. The manual input of data by a driver into a mobile device during operation of a vehicle is also inherently unsafe. Input synchronization with local authorities would also improve the performance of the application. In Fig. 1, we illustrate a collaborative crowdsourcing framework with multiple input streams. In the framework, the area of interest is managed by a central data warehouse server with multiple functionalities, including acceptance, processing, and preparation of data from multiple sources in order to monitor urban traffic conditions in real-time.

Applying Natural Language Processing (NLP) in social media for the purpose of leveraging underutilized transportation-related posts for ITS applications is an active area in research. In this article, we propose to support and complement traffic reporting systems and navigation assistants by exploiting the abundance of data in social media and use it as an additional source of traffic information. To this end, we develop an automated framework that automatically processes social media data from the Web, classifies it, and extracts traffic-related reports to convert them into navigation alerts. One situation where this framework has the ability to effectively work is the case when drivers are stuck in heavily-congested roadways. There is a tendency that users could post about the situation in social media, which is meaningful for other drivers. However, this is not the only unique possibility. The framework also uses inputs from specialized agencies and from regular people who are not necessarily driving such as cyclists, pedestrians, or passengers in vehicles, ride-sharing taxis, or buses, etc. The application can also employ speech-to-text technology so drivers could safely share information in real-time to social media platforms while driving. Another advantage of this framework is that it does not require a specific application to share the information, e.g., Waze, where only the app users could share this type of information together. On the contrary, this solution does not require registration in any specific application, and allows the extraction of traffic information from any social media platform and share to any navigation or traffic-related app. For example, regular social media users can be sharing a publication/tweet about a damaged car on the road in front of their buildings without necessarily having the intention to report incidents to other drivers like in Waze. The platform-agnostic nature of this solution allows for it to act as a complementary source of information that enhances the information provided in dedicated vehicular social media platforms (e.g., Waze). The framework automatically browses social media platforms, distinguishes traffic-related messages, extracts/understands the incidents information, and converts them into alerts in the navigation apps.

A joint text processing framework based on fine-tuning BERT classification models for filtering the traffic related information, and either a Question-answering (QA) model or a name-entity recognition (NER) model for extracting real-time traffic details data from social media are developed. In this two-phase NLP pipeline, we first develop a filter that classifies numerous collected text inputs into different groups. We conduct either binary classification or multi-classification to filter the social media data. With binary classification, the social media data would be classified into

1. https://smart.columbus.gov/projects/connected-vehicle-environment

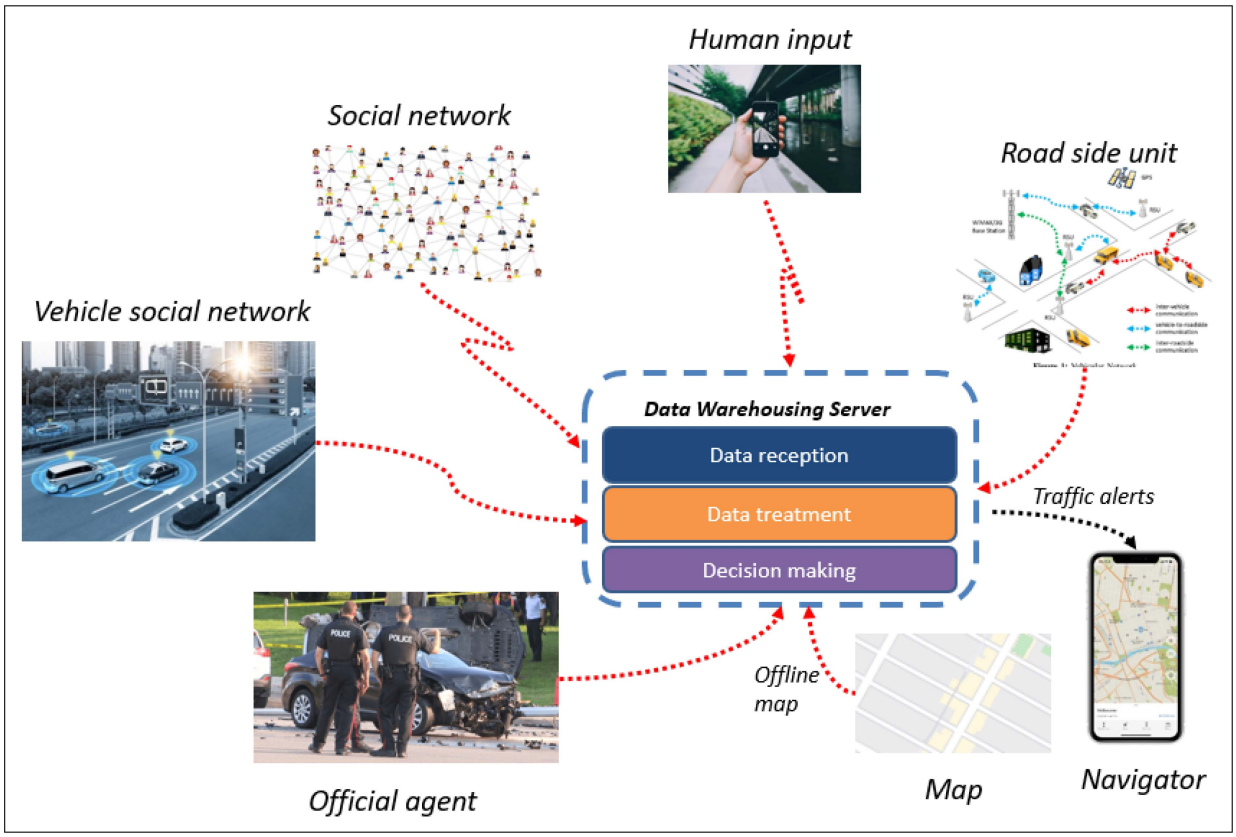

FIGURE 1. A high-level system architecture illustrating the different crowdsourcing data sources and the main components of the traffic reporting system.

two groups: traffic related and traffic unrelated. With multi-classification, the social media data would be classified into several groups such as incident related, congestion related, road construction related, public transportation, etc. The accuracy of binary classification is higher but requires lower computational resources. The multi-classification allows more categorized data and eases the operation of the next phase, which aims to extract necessary information from these filtered groups. Its objective is to automatically understand and characterize the reported traffic event from social media by determining its location, its occurrence time, and its nature, e.g., blocked road, accident, etc. To do so, we implement the QA model [16]. In fact, recently published QA models including BERT and XLNet [17] are shown to achieve human-like performance when tested on the Stanford Question Answering data set [18]. Once the detailed traffic information is obtained, we implement an automated system that converts collected posts from social media into real-time traffic information that are used to update navigation maps and warn drivers about any traffic events. Our analysis is based on experiments conducted on Twitter, where real-world traffic events-related data are studied. Our results investigate the efficiency of the proposed automated framework and show its ability to outperform traditional NLP techniques.

The main contribution of this article is to propose an automated pipeline that uses public-facing traditional social media as a mobile crowdsourcing data source to augment current ITS technologies such as vehicular social networks in reporting traffic-related incidents and information along roadways. This is achieved by applying state-of-the-art natural language processing (NLP) models to the application of interest and using them in a structured pipeline. To the best of our knowledge, we are the first to apply these algorithms and adapt them to the context of mining and processing traditional social media for ITS incident reporting. In other words, the trained models are specifically adapted to the context of transportation via a curated data set harvested with the twitter API, and then structured to convert the outputs of these models into an automated incident reporting system that attaches the incidents to a geographical location. In addition, the pipeline which combines the classification and question-answering models together in a single framework is novel and has not been explored in previous studies. This enhanced data stream will provide additional information for drivers to navigate more efficiently and safely.

The rest of the paper is organized as follows. Section II provides an overview of related work. Section III lists the classification results. The question-answering and name-entity recognition models are presented in Section IV. In Section V, we present the comparison between car incidents posted in official report and social media in NY. We then consider future research directions in Section VI, and we conclude in Section VII.

II. RELATED WORK

In this section, we explore the application of data sourced from traditional social media sites and official transportation agencies. Traditional social media platforms such as Twitter provide a rich data environment because a multitude of users post information about everything, including transportation. On Twitter alone, there are over 330 million monthly active users [19], who generate more than 600 million micro-blog posts per day. As a result of the sheer volume of data produced on these platforms, there is a large potential data source for additional traffic data that may be leveraged to complement existing technologies that gather real-time traffic data. Currently, huge volumes of traffic related information are posted in social media by specialized traffic agencies and regular users, but many of them are not checked by road users, especially drivers who cannot read them while driving.

On the other hand, expenses associated with installation of roadside sensors and relying on authorities to report information are high. These costs are motivation for work on optimizing roadside sensor unit placement in order to stay within an installation and operational budget [20]. Mining social media data may provide a cost effective alternative or complement existing sensors-based solutions, as drivers or the passengers accompanying them tend to post about the traffic conditions they are currently facing by texting when they are caught in traffic jams and free of hand. Rapid information sharing on social media platforms can lead to faster alerts, more widespread coverage, and also more precise information about the incident compared to existing methods [21].

Due to the large volume of social media posts, there exist challenges in extracting useful information. First, many of the messages posted on social media are usually short, written in informal language with multiple grammatical errors/misspellings/abbreviations, and include peculiar Unicode characters such as emojis. Second, only a small fraction of the posts are relevant to traffic status. Third, even with a successful extraction of relevant information, additional analysis would be required to link the metadata of the post (location, time, etc.) with the actual incident, and also to develop a plan of action. Therefore, it is worthwhile to develop a framework to collect, filter, and extract essential information from the social media posts.

There are some existing solutions that address the challenges of data mining with social media in ITS, and several social media-powered ITS applications are represented in existing literature. The authors of [22] developed an unsupervised learning model for clustering areas based on *placeness* - the cultural and semantic characteristics of a location. In [23], the authors built an artificial neural network (ANN) that classifies traffic congestion conditions based on data sourced from Twitter and Waze. The authors of [24] explored applying data mining methods to predict traffic congestion and movement patterns from social media posts.

Natural Language Processing (NLP) is the use of computers to analyze human language and thus can be an efficient tool to analyze and understand social network data. The raw computing gains made since the inception of the discipline as enabled many applications, including but not limited to sentiment analysis and prediction [25] or detecting rumors on social networks [26].

Latent semantic indexing (LSI) is one of the first of many algorithms that helped search engines identify related keywords and processing synonyms to deliver accurate related results [27]. Many newer NLP models are built on LSI, such as latent semantic analysis (PLSA) [28] and latent Dirichlet allocation (LDA) [29]. However, it is hard to apply classic NLP approaches to unstructured social media data, as the text contains short messages with a low frequency of meaningful words, and a high degree of overlap.

On the other hand, topic modeling is a branch of NLP that aims at discovering the abstract "topic" that occurs in a collection of documents [27]–[29]. One use-case of this approach is to extract the "topics" from product or service reviews, and to utilize these "topics" to distill user reviews into the main issues or feelings that reviewers are trying to state. In a similar approach, text summarizing methods are employed to extract the main essence of information contained in large documents [30]–[35].

Recent advances have focused on combining word embedding models with machine learning algorithms on very large data sets [36]. Most state-of-the-art NLP models are usually designed as the combination of a large, pre-trained word-embedding layered with a transformer model structure. In [37], the authors proposed an embedding vector representation for finding phrases in text called Word2vec, while the authors of [38] proposed a paragraph vector named word2doc to overcome the weakness of the classic bag of words approach. Another well-known word embedding called Glove was developed by the authors of [39]. However, in recent years, after the introduction of the transformer model [40], there is a new trend in designing NLP models based on the attention mechanism. A successful example is the bidirectional encoder representations from transformers (BERT), developed by Google in 2018, which redefined the start-of-the-art for eleven NLP tasks [41]. Similar pre-trained models that have been proposed include: embeddings from language model (ELMo) [42], generative pre-training language understanding model [43], and a lite BERT (ALBERT) [44].

The growth of interest in NLP techniques and applications has opened up new avenues of exploiting unstructured text data. As a result, there has been a dramatic increase of potential applications for social media data. As NLP models have improved, the classification accuracy and flexibility of models that leverage social media data in ITS have improved in lockstep. The authors of [45] presented a natural language interface for trip planning in complex multi-modal urban transportation networks, which provides robust understanding of complex requests with the aim of giving users flexibility in their language. The model's main shortcomings, however, stemmed from a lack of flexibility and limited training data. The authors of [46] processed a SVM algorithm based on word vector features that can

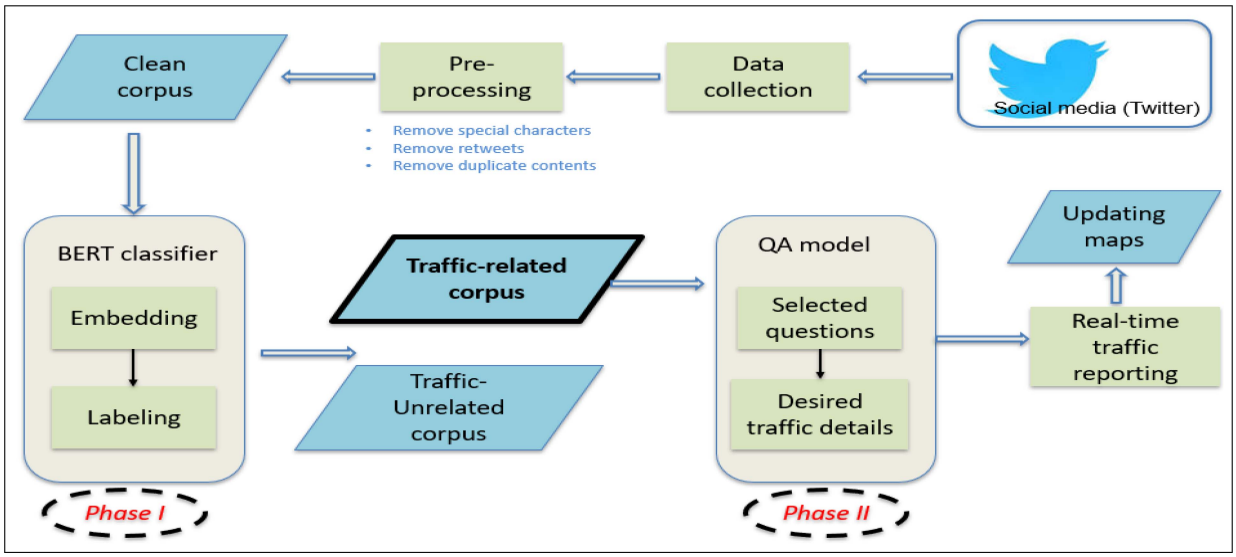

FIGURE 2. The proposed NLP-based traffic reporting system using social media input.

detect traffic-related tweets with an accuracy of 88.28%. The authors of [47] proposed a series of convolutional neural network (CNN), long short-term memory neural network (LSTM), and hybrid LSTM-CNN models that are trained to classify micro blog posts related to traffic. They found their approaches achieved higher than 91% accuracy. In [48], the authors explored the BERT to verify that its context-aware representation can achieve similar performance improvement in sentiment analysis. They indicated that the accuracy of classifying traffic-related and traffic-unrelated messages could also be improved by applying BERT.

Social media can provide supplemental information for official agent reports. In [49], the authors explored applied crowdsourcing in ITS, including tools, technologies, and methods. Scraped social media data that powers a traffic alert system may be considered a form of mobile and/or spatial crowdsorcing application in the context of ITS. The authors of [50] studied the user satisfaction towards the public transportation in Chile through Twitter via sentiment analysis and other NLP methods, instead of sending traditional surveys, which are far more expensive. They found that while surveys provided more detailed information, their twitter mining methods allowed for faster and more widespread analysis of trends, leading to more rapid solutions to problems. Opinion mining techniques on reviews towards public services in the Pujiang and Changxingdao suburb parks were leveraged to generate meaningful feedback for government concerns by the authors of [51]. The authors of [52] developed a methodology for identifying the most influential individual in a social media network in the context of transportation. This technique can be leveraged for many purposes including, but not limited to limiting the spread of misinformation and encouraging acceptance of a new transportation product or service. Finally, the authors of [53] compared the delay/incident report from both social media and official agents (MTA) regarding to a Metro North Railroad system in NY through resilience analysis and data visualization. This non-exhaustive collection of studies show the great potential and interest in leveraging social media as a complementary information source in public transportation services.

From our review, we can conclude that the accuracy of classification models related to traffic prediction can be improved by including innovative NLP models. Furthermore, unlike previous studies which mainly classify the information into traffic related and traffic unrelated groups, we propose to employ a multi-classification approach to categorize the social media input into several groups like public transportation updates, incident reports, road construction alerts, road closure alerts, special event notifications, road delay alerts, etc. While this problem can be framed as a multi-class classification problem, there exists the real possibility of a large imbalance of data with real-world data in this problem context, as a much higher proportion of traditional social media may contain unrelated information. To adjust for this, the problem may also be broken into two sub-tasks, where the first problem is a binary classification that filters out unrelated data. The related data is then fed through a multi-class classifier in the second task that categorizes each post based on how it relates to transportation. More importantly and to the best of our knowledge, previous studies did not develop a systematic method that could automatically extract necessary information from related social posts and build a data pipeline that can be used to update mapping applications in real-time. In this article, we develop this pipeline and implement the necessary mechanisms to deliver automated and reliable traffic alerts extracted from social media. The proposed framework is presented in Fig. 2.

III. INPUT TRAFFIC CLASSIFICATION

In this section, we present the detailed approach to filter the input social media data into two or multiple traffic related/unrelated groups.

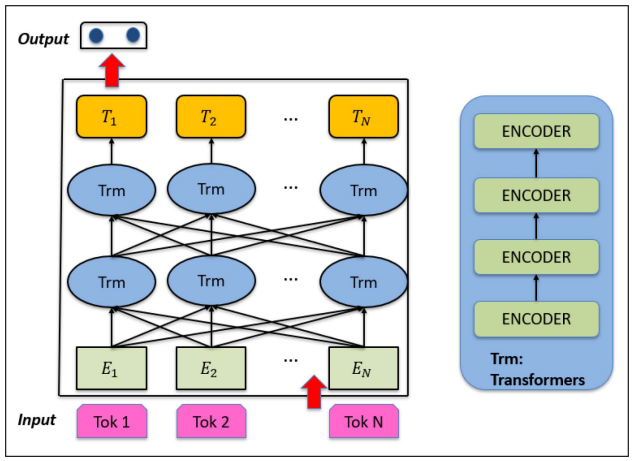

FIGURE 3. Architecture of the BERT binary classification model.

TABLE 1. Examples of labeled tweets in the data set.

Tweet	Label
Crash investigation and Crash with Injuries on Garden State Parkway southbound North of Exit 63B - NJ 72 West (Stafford Twp) right lane blocked.	Related
Incident on I295 NB at Exit 61 - Arena Dr.	Related
Cleared: Construction on West 11Th Street from 7th Avenue to 6th Avenue	Related
I will visit the transportation office tomorrow at Garden Parkway but my car is crashed.	Unrelated
Can confirm former @lakers star Rick Fox WAS NOT among the passengers on the helicopter with Kobe Bryant. Source: his daughter.	Unrelated
Can confirm former We had car troubles along the jos-Abuja road this morning, the army officers at the military check point close to where the incident happened went beyond the call of duty to help, one literally went under the vehicle to try and fix it.	Unrelated

A. BINARY CLASSIFICATION

In order to build traffic-related corpus from the social input data, we utilize a fine-tuned BERT classifier model, which is a pre-trained deep bi-directional transformer model for language understanding tasks. The BERT model is pre-trained using two unsupervised tasks, i.e., Masked LM and Next Sentence Prediction. For the Masked LM task, they mask 15% percentage of the input tokens at random, and then predict those masked tokens. Note that they sometimes replace "masked" words with other random words to create cross entropy loss. For the Next Sentence Prediction (NSP) task, they choose sentences A and B for each training example where in half of the training examples, B is the sentence directly succeeding sentence A, and in the other half of the examples, B is a random sentence. In each training example, the input features are pairs of sentences (A, B) that follow the aforementioned rules, with the first case labeled "`IsNext`" or 1, and the other case labeled "`NotNext`" or 0. From this set, the model is trained to predict whether or not a sentence B directly succeeds a sentence A. For the pre-training data, they use the BooksCorpus (800 million words) and the English-language Wikipedia (2.5 billion words). It has been shown that the BERT model is achieving extraordinary performance in many tasks such as named entity recognition and question answering via transfer learning. BERT is a pre-trained model that can be extended to other NLP problems by adding an output layer at the end of the network [41]. Indeed, BERT relies on a transformer architecture which contains several self-attention encoders to read the text input; the model architecture is shown in Fig. 3.

The input vectors E_n where $n = 1, \ldots, N$ represent the sequence of tokens converted by the input text using WordPiece embedding [54], which would be processed in the neural network later. The output vector T_n represents the final hidden vector. The first token of every sequence is always a special classification token, E_1 in this case. The final hidden state corresponding to this token is used for the classification tasks, T_1 in this case. The model is pre-trained on two separate tasks: predicting masked tokens and predicting the next sentence. As a result, the model with the pre-trained parameters could be used for a wide variety of fine-tuning tasks that beat previous proposed NLP models. On the other hand, pre-training and fine-tuning procedures of the model are using the same architecture but different output layers. All parameters are fine-tuned during the fine-tuning process, where there are total 340 million parameters in $BERT_{LARGE}$ models. In fact, fine-tuning is relatively inexpensive compared to the pre-training process. It usually takes few hours on a General-Purpose Graphics Processing Unit (GPGPU).

We then illustrate and evaluate our classification model by analyzing recently published tweets data from Twitter that are written in English. We start by feeding a large amount of labeled tweets into the classification model to train the parameters, i.e., the weights of the BERT classification model. Our model must be able to identify the context of the tweets to be considered accurate; i.e., it must be able to learn if tweets are related to traffic incidents if no traffic-related terms are used, and it must be able to learn if tweets are unrelated to traffic even if they do contain traffic-related terms. Some examples are given in Table 1. We collect 106437 tweets via the Twitter API. Some are written by traffic management agencies, and others by regular Twitter users. We then categorize them into three major groups, with two response labels (i.e., we manually label the response variable of the tweets). A label of one corresponds to relevant traffic-related tweets (group 1). A label of zero corresponds unrelated tweets that either contain a high (group 2) or low (group 3) frequency of transportation-related terms. In order to facilitate the training of an accurate model, we curate our data set to contain a roughly equal proportion of data points with zero and one labels. We verify the accuracy of the tweets with a label of one by manually parsing scraped tweets from the official government twitter accounts. One example from an official twitter account @511NY is shown in Fig. 4. Those kind of twitter accounts exist for larger cities which are responsible for updating people about traffic and transit situations. In order to generalize our model's ability to distinguish context, we insert random sentences unrelated to transportation into our traffic-related tweets.

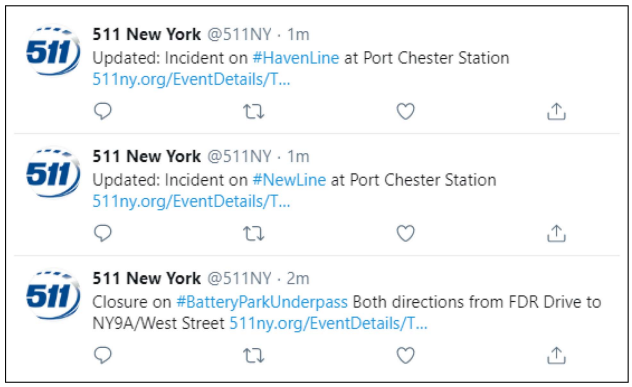

FIGURE 4. Social media input example: 511 New York's traffic account on Twitter.

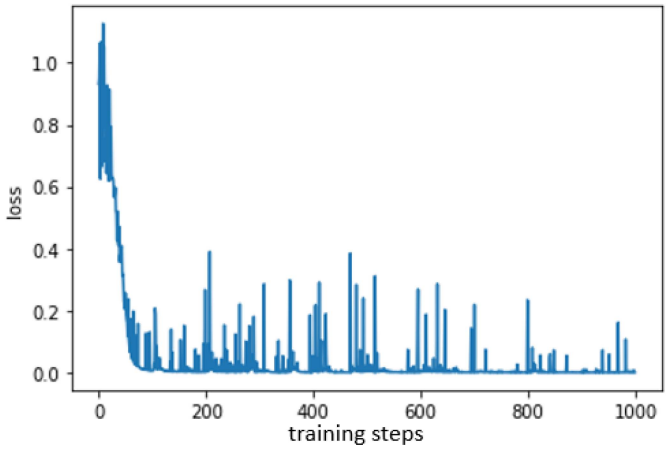

FIGURE 5. Training phase of the binary classification model: achieved loss vs. training steps.

TABLE 2. Testing phase: Binary classification results.

Class	Precision	Recall	Accuracy	MCC
Naive Bayes	95.6	92.7	94.45	88.9
SVM	99.0	94.3	96.8	93.7
Fine-tuning BERT	99.6	99.3	99.45	98.9

To obtain the second group of tweets, we parse the data set and search for high-frequency traffic words such as "incident", "delay", "construction", "crash", "lanes", "road", etc. Those high-frequency traffic words are manually selected after collecting the traffic-related information, counting the word frequency, and analyzing the obtained word map. Then we select those tweets that contain one or more high-frequency traffic words but not belong to traffic-related, and we label them as zero. For the last group, we simply select the tweets that do not contain any of the high-frequency traffic words, and assign a label of zero. In our study, there are 50 high-frequency traffic words that are considered.

The standard evaluation process of a classification model involves construction of a confusion matrix, which contains the true positive (TP), false negative (FN), true negative (TN), and false positive (FP) rates. Our first objective is to build an accurate BERT classifier for our data set, so we may utilize it in a pipeline to later on extract desired traffic-related information from new tweets. After the construction of the confusion matrix, we compute the precision, recall, accuracy, and Matthews correlation coefficient (MCC) metrics to evaluate the model performance. Those are defined as follows:

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

To effectively train the BERT classifier, we setup our training environment as follows. First, we split the 106437 tweets into a training group (80% of the data), and a test group (20%). The learning rate for our optimizer is set to 2×10^{-5}, with a batch size of 24, maximum iteration cap of 4018, a dropout rate of 0.1, and utilize the Gaussian error linear unit (GELU) activation function in the hidden layers. Our model is trained in one epoch. Fig. 5 illustrates the descent of the loss function over the iterations that the optimizer operates over. We can visualize that the loss function descends rapidly within the first 100 iterations, then stabilizes slowly afterwards.

In order to define the effectiveness of our BERT model in classifying the types of messages collected, we compare its performance in the accuracy, precision, recall, and MCC metrics with the performance of Naive Bayes and SVM classifiers. When calculating these metrics, we define a label of one (transportation-related) to be a positive result, and a label of zero (non-transportation-related) a negative result. For example, the classification accuracy of our BERT model is $\frac{4775+5221}{4775+5221+204+280} = 98.9\%$. We notice in Table 2 that our BERT classifier has the best performance in all four metrics compared to the other two methods.

Although its training computational complexity is low, the act of classifying the data into two categories is the main limitation of the binary classification. It is usually insufficient to just distinguish traffic-related information from the traffic-unrelated one from social media. For instance, people driving in personal automobiles would likely not be interested about the on-time status of a municipal bus. Similarly, people commuting via heavy rail public transportation may not need to know about an accident that happened in a distant location. Nevertheless, binary classification can be a valuable pre-step before conducting multi-classification.

B. MULTI-CLASS CLASSIFICATION

It is a very worthwhile endeavour to classify the social media data beyond the aforementioned simple binary classification. To pursue this, the BERT model can be restructured into a multi-class classifier by changing the output layer schema, as shown in Fig. 6. In order to provide a more useful set

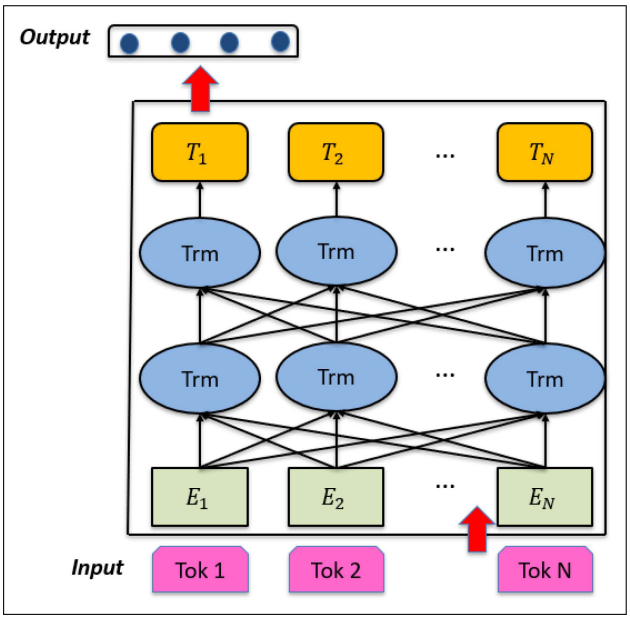

FIGURE 6. Architecture of the BERT multi-classification model.

TABLE 3. Examples of labeled tweets in the data set.

Tweet	Label
Albuquerque, NM – Car Crash at Charleston St and Trumbull Ave.	Incident
Update from the Reconstruction of Kawo Flyover Bridge with Three Ground Rotaries and Access Road, Construction of temporary road as alternative route to be completed soon for traffic diversions.	Construction
The right lane of I-43 remains closed at this time. Please use an alternate route if possible.	Closure
Delays on George Washington Bridge westbound from New York Side/Upper Level (Manhattan). Travel time of 25 minutes from the Cross Bronx Expressway at Jerome Avenue to Interstate 95 at Interstate 80	Delay
MRT Line Update: We have a train not moving between Phileo Damansara to Pusat Bandar Damansara. Rescue is in progress. Please expect slight delay.	Public transportation
American school buses are yellow because humans see yellow faster than any other color, which is important for avoiding accidents.	Unrelated

of information for drivers, we aim to partition the data into six classes: information related to incidents, road construction, road closures, traffic delays, public transportation, and unrelated information. Examples are provided in Table 3.

To assess the viability of the proposed multi-class classifier, we train our model on 29418 tweets, and test on 7357 tweets, where the tweets in both groups were labeled with one of the six aforementioned classes. We set the learning rate to 2×10^{-5}, batch size to 24, the iteration cap to 1225, and the dropout rate to 0.1. Again, we utilize the GELU activation function, with one training epoch. Fig. 7 illustrates the descent of the loss function over the iterations that the optimizer operates over. We notice that the loss function descends rapidly within the first 200 iterations, then stabilizes

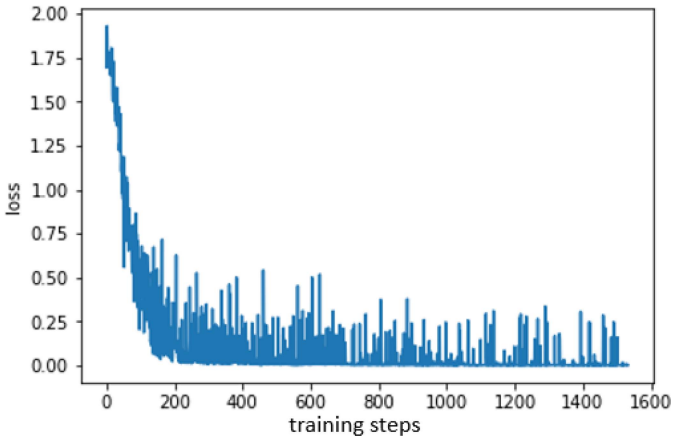

FIGURE 7. Training phase of the Multi-classification model: Achieved loss vs. training steps.

slowly afterwards, the training time are longer compared to binary class as more training steps needed.

We summarize our training results as the confusion matrix presented in Table 4, along with the performance metrics in Table 5, and it appears that our BERT classifier performs better than Naive Bayes, Decision Trees, and SVM. The BERT classifier has a $\frac{1170+1045+1083+1041+1035+1937}{7357} = 99.37\%$ accuracy. We conclude that the BERT classifier achieves significantly better performance than other traditional methods.

C. BINARY CLASSIFICATION VS MULTI-CLASS CLASSIFICATION

In the previous section, we have proposed two approaches for social media data classification. The first is the binary classification that divides the social media inputs into traffic-related and traffic-unrelated corpuses. The second approach is more elaborate and aims to divide the social media data into multiple corpuses, e.g., based on the ITS applications such as public transportation and traffic incident reports. The binary classification requires less training time and data. Hence, it is easy to classify the input data into traffic-related and traffic-unrelated. This quick and rapid classification could be used with a small dataset and less complicated applications where the knowledge of the concerned ITS applications is not required beforehand. However, for complex ITS applications, it has the disadvantage of making the next step to extract essential information from the traffic-related corpus less efficient than the case with multi-class classification. Nevertheless, the binary classification could be used as a filter for the data in a multi-step pipeline, where unrelated data are filtered out in the first step via the binary classifier, and then the traffic-related data may be further classified with the multi-class classifier. Hence, it can help improve the multi-class classification results if the training data set is not sufficient or unbalanced (e.g., the size of traffic-related corpus is much smaller than traffic-unrelated corpus). We test both of these approaches on an $N = 5000$ sample of our data set. Table 6 represents the confusion matrix for the multi-class classification only approach. We

TABLE 4. Multi-class classification confusion matrices for naive Bayes, decision tree, SVM, and BERT classifiers.

	-	Incident	Construction	Closure	Delay	Public	Unrelated
Naive Bayes:	Incident	838	64	13	50	15	192
	Construction	26	867	59	29	29	83
	Closure	90	94	763	59	1	94
	Delay	240	12	30	584	6	121
	Public	65	45	29	14	615	296
	Unrelated	4	1	3	7	4	1913
	-	Incident	Construction	Closure	Delay	Public	Unrelated
Decision Tree:	Incident	929	29	27	61	33	93
	Construction	15	911	28	10	42	87
	Closure	7	29	845	17	19	184
	Delay	53	20	19	704	25	172
	Public	28	50	12	24	620	330
	Unrelated	16	4	9	37	177	1689
	-	Incident	Construction	Closure	Delay	Public	Unrelated
SVM:	Incident	967	20	21	70	44	8
	Construction	35	974	8	7	44	4
	Closure	23	20	989	28	17	4
	Delay	77	23	40	820	17	14
	Public	27	32	4	11	793	23
	Unrelated	43	24	39	57	149	1879
	-	Incident	Construction	Closure	Delay	Public	Unrelated
BERT:	Incident	1170	0	0	4	7	4
	Construction	0	1045	0	1	2	3
	Closure	0	0	1083	0	4	1
	Delay	0	0	2	1041	0	4
	Public	0	2	1	0	1035	1
	Unrelated	1	3	0	2	4	1937

TABLE 5. Testing phase: Multi-classification results.

Class	Precision	Recall	Accuracy	MCC
Naive Bayes	78.79	72.63	75.87	69.40
Decision tree	80.37	75.99	77.47	72.57
SVM	85.97	87.59	87.31	83.90
Fine-tuning BERT	99.37	99.37	99.37	99.24

see an accuracy of 95.97% in this case. Table 7 represents the confusion matrix for the binary to multi-class pipeline approach. In this case, we see a lower rate of unrelated miss-classifications, and a higher accuracy of 96.60%. In our case, we curated a balanced dataset, so we just directly employed a multi-class classifier.

IV. INCIDENT IDENTIFICATION AND EXTRACTION

The development of the BERT model is the first step in our pipeline, which allows us to filter and classify transportation-related information. The next step in our pipeline development is to extract the desired information include type of the event, time of the event, locations of event, etc, from each category that we classified with our BERT model. To this end, we consider two information extraction models in parallel: the Question-Answering (QA) and Named Entity Recognition (NER) models, and apply the most effective of the two to transform landmark information into geospatial coordinates.

A. QUESTION-ANSWERING MODEL

The QA model is a potential candidate to complete the information extraction step. Fig. 8 illustrates the general architecture of the QA model. It is similar to the BERT model's architecture, with one key difference: the output layer. The data is characterized by three main parts: the question, the data containing the answer (i.e., a tweet), and the separator character *SEP*, which is between the question and answer parts. The output layer consists of three neurons: "start", "end", and "span". The first two neurons indicate the locations of the answer tokens in the tweet and the last neuron represents the total length of the answer, e.g., number of words. With properly selected questions, we are able to automatically extract the desired traffic information from the collected tweets and use them for navigation in real-time, for example, by automatically updating the navigation maps.

Our aim is to extract the desired traffic information from these tweets by using an appropriate QA model. To extract the details, we apply the "BERT-For-Question-Answering", which was trained on the Stanford Question Answering Data set (SQuAD) [55] using three classes of questions: "What", "When", and "Where" to identify the type of incident, its occurrence time, and its location, respectively. SQuAD is a reading comprehension data set, consisting of a set of Wikipedia articles with regarding questions/answers provided by crowd workers. We adopt the model presented in [56], the open sourced library that gathers state-of-the-art general-purpose pre-trained models including BERT-For-Question-Answering model. We train it for 2 epochs with a batch size of 8. After training, the model is shown to achieve a matching score of 82.1% and a F1-score 85.0% in the testing data set from SQUAD, where human performance was shown to be 86.8% and 89.5% for the matching and F1 scores, respectively [55]. With a fast training time, the adopted QA model was shown to achieve close performance

TABLE 6. Multi-class classification only - confusion matrix for validation of the trained BERT classifier.

		Actual					
	-	Incident	Construction	Closure	Delay	Public	Unrelated
Predicted	Incident	1123	0	3	11	44	24
	Construction	0	999	0	1	30	5
	Closure	4	1	1077	4	23	6
	Delay	39	14	3	1027	15	9
	Public	1	29	1	2	935	6
	Unrelated	4	7	2	3	5	1900

TABLE 7. Two step classifier - Binary Classification to filter unrelated data, then multi-class classification to further classify related data - confusion matrix for validation of the trained BERT classifier pipeline. Red values correspond to the binary classifier, black values correspond to the multi-class classifier.

		Actual						
			Related					
		-	Incident	Construction	Closure	Delay	Public	Unrelated
Predicted	Related	Incident	1097	1	14	15	23	
		Construction	1	1017	1	0	15	
		Closure	18	8	1053	7	27	15
		Delay	52	13	17	1026	8	
		Public	3	11	1	0	979	
		Unrelated	0					1935

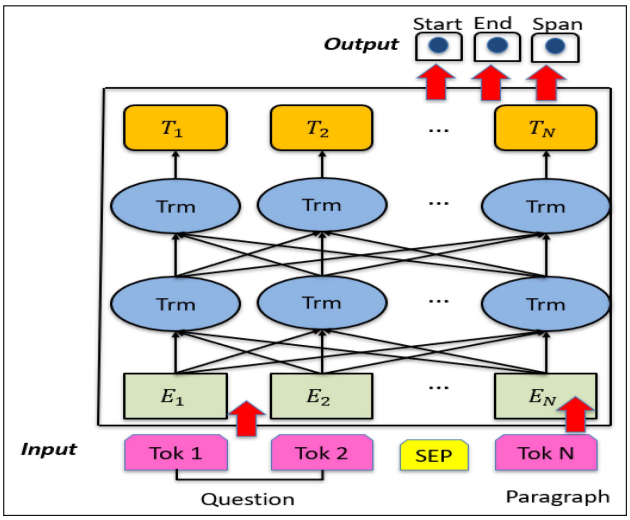

FIGURE 8. Architecture of the BERT question-answering model.

to the one achieved by humans [56]. The validation results on the test data set are a matching score of 80.3% and a F1-score 83.7%, where the test data is the filtered 1468 tweets related to transportation happened in NYC. Examples showing how the question-answering model extracts desired traffic details are illustrated in Fig. 9.

B. NAME ENTITY RECOGNITION

Another potential solution for extracting specific locations from tweets is through NER, *aka* entity extraction, which is presented as a token classification model in BERT. The method locates and classifies named entities from given text including names, locations, etc. The tags include the action, time and location, which we manually label. One example is shown in Fig. 10, which can infer the "What", "Where", and "When" tags by classifying the relevant information in the tweets. We then utilize the labelled data to train the BERT Tokenization Classification model. The model's objective is to predict the tags based on the relevant tokens in the tweets. In our study, we train it on 1468 samples and validate our results on 468 samples, and we set the learning rate to 3×10^{-5}. The model achieves a 85.9% matching score and 87.7% for F1-score in the validation step. NER achieves slightly better results compared to the QA approach, partly because we trained our model on our own curated data set rather than the more broad SQuAD set. However, the limitation of NER is that the output answer contains short phrases only while the answers provided by the QA model may contain a complete sentence which can disclose more information such as the relationships between two streets where an event is happening.

C. TRANSFORMING STREET NAMES INTO GEOGRAPHICAL COORDINATES

We explored the application of NER and QA models in parallel in order to determine which one was more appropriate to apply to our problem when extracting incident spatio-temporal information from the tweets. While NER performs slightly better than QA (due to our training of the NER model on our tailor-fit data set), the QA model is more generalized and required a smaller amount of training resources. Therefore, we proceed with the QA model for the location information extraction. While the desired traffic details could be extracted, it is not straightforward to convert extracted text information to the exact coordinates. For instance, the exact location at which the incident occurred could be a short phrase with several road names, e.g., "Main Rd between B3335 Highbridge Rd and B2177 Portsmouth Rd." Our solution is to create a comprehensive word document that includes terms such as "and," "between," "of," etc. This would allow for us to split the short sentences into different phrases, where each phrase represents a specific location. In other words, we could split "Main Rd between B3335 Highbridge Rd and B2177 Portsmouth Rd" to "Main Rd," "B3335 Highbridge Rd," and "B2177 Portsmouth Rd."

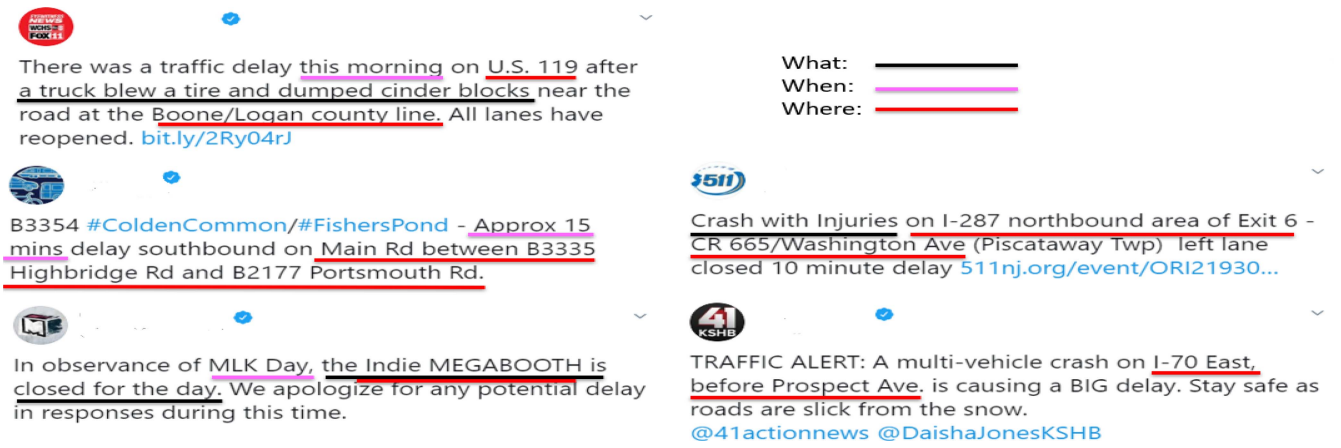

FIGURE 9. Examples of the desired traffic information extracted using the question-answering model.

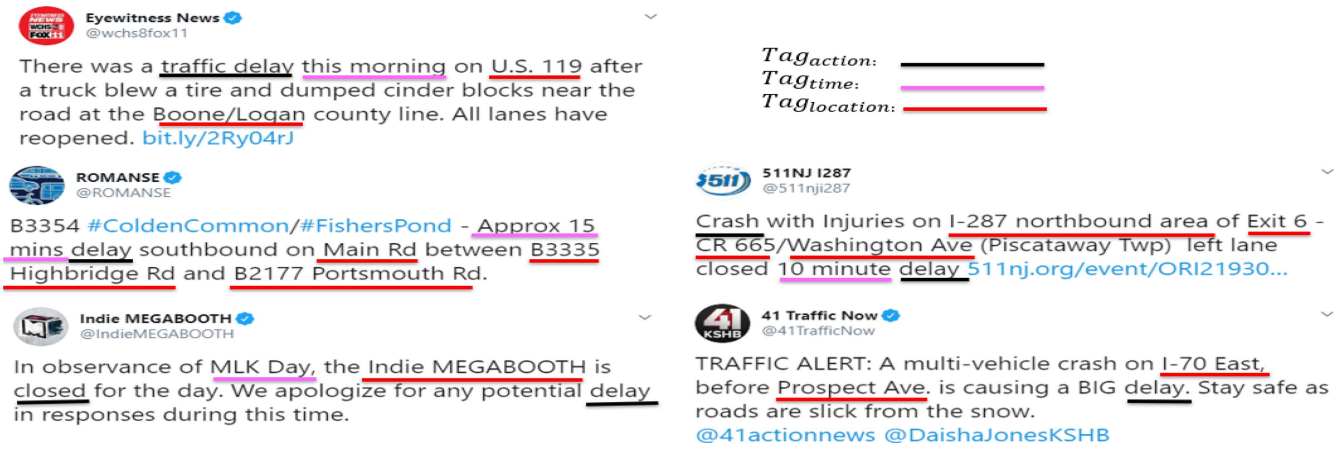

FIGURE 10. Examples of the desired traffic information extracted using the name-entity recognition model.

Note that we split the sentences due to the way Google Maps API queries work. If we were to search "Main Rd between B3335 Highbridge Rd and B2177 Portsmouth Rd" via the API, it would not return the desired coordinates. If we break the API call into two separate ones, "Main Rd & B3335 Highbridge Rd" and "Main Rd & B2177 Portsmouth Rd," the API returns the coordinates for those respective intersections.

We must also take into account the importance of the information contained in the relations between those specific locations. For example, if the relation is "between," then we must update the delayed routes between the two locations and potentially other affected routes that are connected to that location.

In the previous example, the "Main Rd" between "B3335 Highbridge Rd" and "B2177 Portsmouth Rd" is the affected route. If the relation is "of", like in "I-287 northbound area of Exit 6 - CR 665/Washington Ave" then we need to identify the delayed routes as "I-287 northbound area" in the region of "Exit 6 - CR 665/Washington Ave". In other words, our designed automated system needs to handle these issues and convert them into geographic information system (GIS) data which can be mapped in standard mapping applications. We illustrate examples of how we aim to convert the QA tags into GIS data in Fig. 11. The red underline represents the critical preposition words that split the sentence into different streets, and the black underline represents the redundant information to be removed since the navigation map API, e.g., Google Maps API can not recognize it. Finally, through the Maps API, we are able to transform the intersection (represented by two street names with sign in between) into GIS coordinates. We aim to create a complete document that contains all the combinations of critical preposition words like ("at"), ("from", "to"), ("from", "to"), ("from"), etc, so that we could properly split the sentence into different locations and combine them together as intersections with filtering process.

It is possible for some distortions in the temporal information extracted by the models. For example, the aforementioned example tweet, the model picked up "approx. 15 mins" as the "when" token extracted by the model. In this case, another layer of verification may be considered by comparing the time of the post and the time information extracted from its content. Hence, the time of the post can be used to improve the accuracy of the extracted temporal information. If needed, the timestamp on the post could be considered as the closest time instant, and may be used as the basis for the

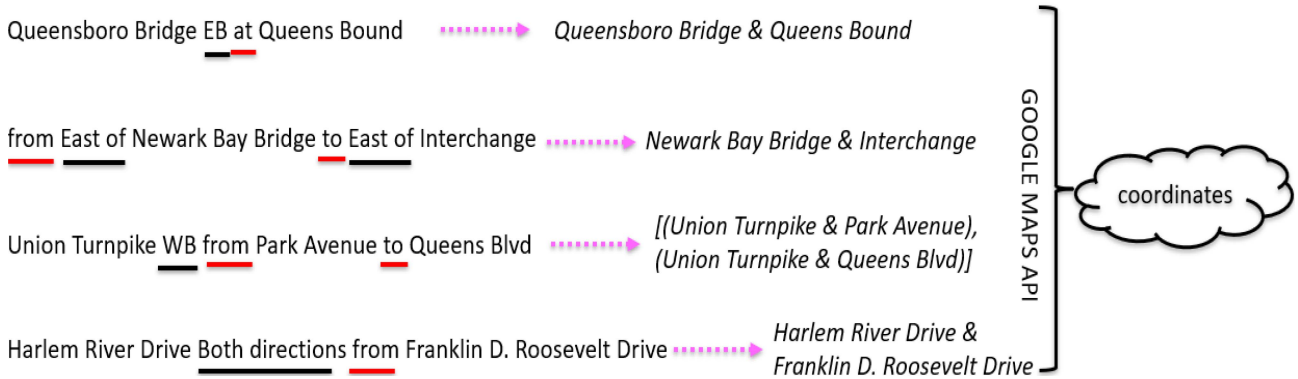

FIGURE 11. Automatic system that transforms street addresses into coordinates from extracted answers.

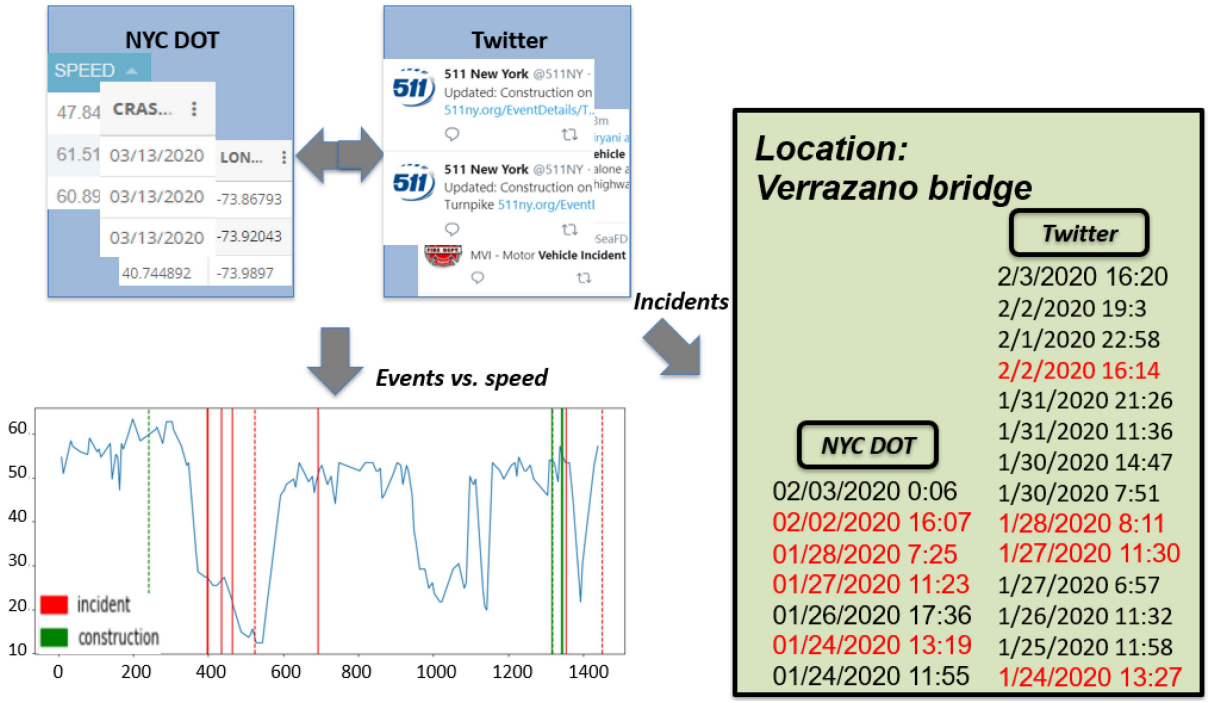

FIGURE 12. Comparison of traffic information from the NYC DOT and from Twitter using the QA BERT model.

"when" information extraction. On the other hand, in practice, some cross-validation approaches can be added to ensure that the framework provides accurate information to users. For example, an incident may be announced by a heterogeneous mix of data sources within a short interval of time especially when the announcers are regular social media users.

V. COMPLEMENTARY VEHICLE INCIDENT REPORTS FROM SOCIAL MEDIA

Social media information can also serve as a cost-effective complementary source for the official transportation reports. In our study, the official transportation reports are published by NYC DOT, which includes the real-time speed data captured by cameras in some streets and the incident reports generated by DOT agents. We present a comparison between the DOT report-generated information and the information generated by our information extraction method in Fig. 12. First, we explore the relationship between unexpected traffic events with the real-time speed in the Event vs. speed figure at the bottom left of Fig. 12, where the x-axis represents the minutes in a day, and the y-axis represents the speed at that moment. A red line represents an incident and a green line represents construction, while a dashed line represents a previous situation getting resolved. We notice that some incidents do cause the decline of the traffic speed while the situation is resolved. However, since the speed data provided by NYC DOT is sampled every five minutes, there is a chance that patterns may be lost.

We compare the incidents extracted from social media with the official reports generated by the NYC DOT; an example of this is shown in the Incidents box, right-hand side of Fig. 12. The location of interest is the Verrazano-Narrows bridge and the time of interest spans 1/24/2020 to 2/3/2020. The timestamps are red if the event is represented in both

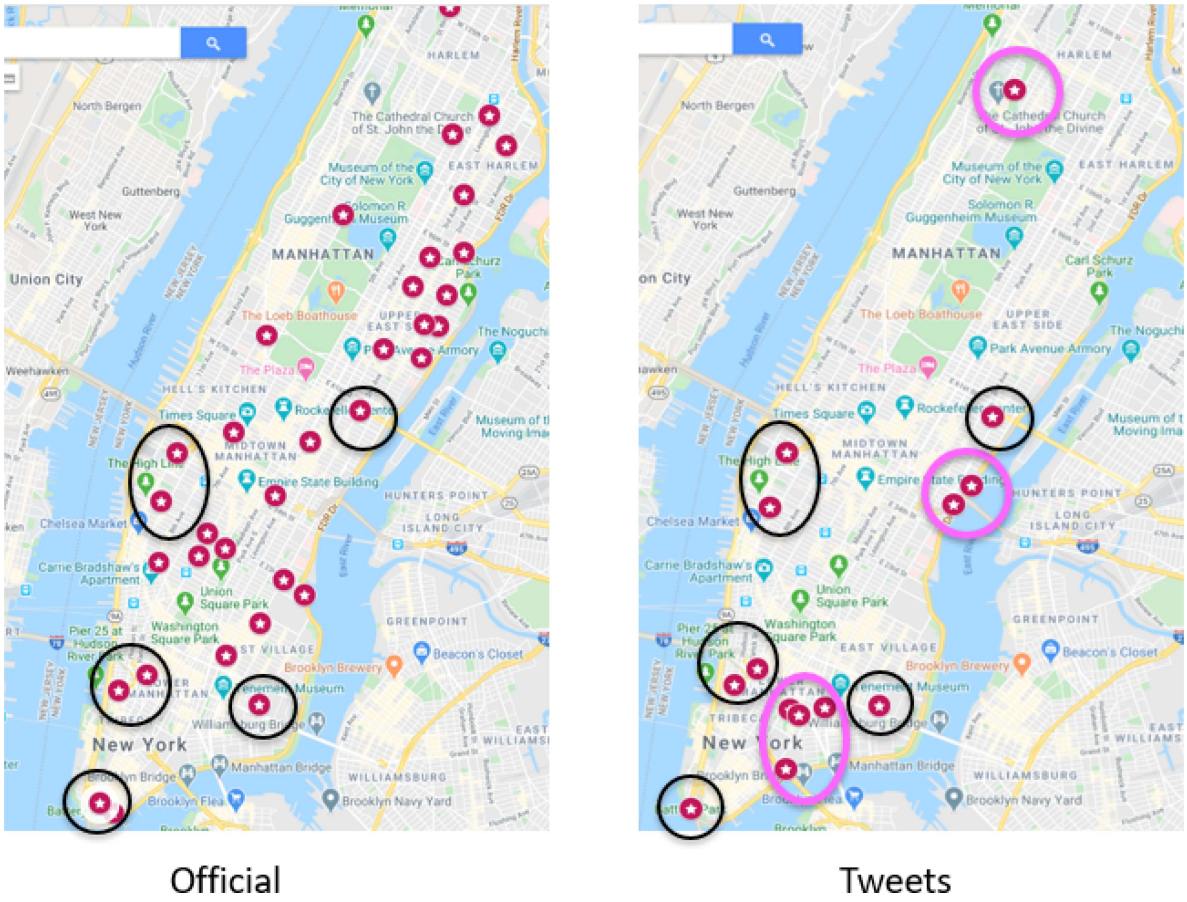

FIGURE 13. Accidents reported by NYC DOT and by social media (tweets) in 01/27/2020, Manhattan.

the DOT reports and the extracted tweets. There is a gap between the two, as the DOT officials are likely to report and respond to incidents faster than motorists can report them. We also notice that the multiplicity of posts for the same event is usually larger than the same for the official reports in areas that tend to have a high density of motorists, such as bridges and tunnels, showing the effectiveness of tweets as a complimentary source of data. We can speculate from this pattern that the multiplicity of tweets may positively correlate with the severity of the incident (i.e., a larger impact to service).

Fig. 13 and Fig. 14 illustrate the reported accidents by two different sources in 01/27/2020 and 01/28/2020, respectively. We plot the geographical coordinates of events reported by the NYC DOT (left-hand side plot), *a posteriori* non-social media source, and the data gathered, in real-time, from Twitter (right-hand side plot) on 01/27/2020. The incidents reported by both sources are outlined with black ovals, while the incidents that are exclusively reported by social media are highlighted with pink ovals. Note that although DOT reports cover more locations than social media posts, it still misses some incidents, for example, incidents that are not covered by official agents or unreported incidents due to human error. On the other hand, the social media posts tend to be concentrated around the main external access points to the city such as bridges and tunnels where severe congestion usually occur. At these instants, drivers or passengers opt to tweet about the traffic situation they are facing. Our proposed framework may serve as a complementary source of information in the under-served coverage ares, such as local residential roadways.

VI. FUTURE RESEARCH DIRECTIONS AND PERSPECTIVES

In this section, we consider logical extensions of our work, along with other novel methods that may arise though future directions related to our study.

A. TASK RECRUITMENT/REWARDING SYSTEMS

From previous sections, we show that there is potential for the use of crowd-sourced social media posts for improving the reporting of traffic-related incidents. One main obstacle to a more widespread adoption of such a system is that there exists a dearth of posts related to transportation-related incidents and information. This can be addressed through effective recruitment of users to post transportation-related information.

There exists a multitude of research related to crowdsourcing recruitment and reward methodologies, but this area is still new and in need of additional work, especially related to the issue presented in this manuscript. Potential recruitment and reward policies can be to improve the standing of the

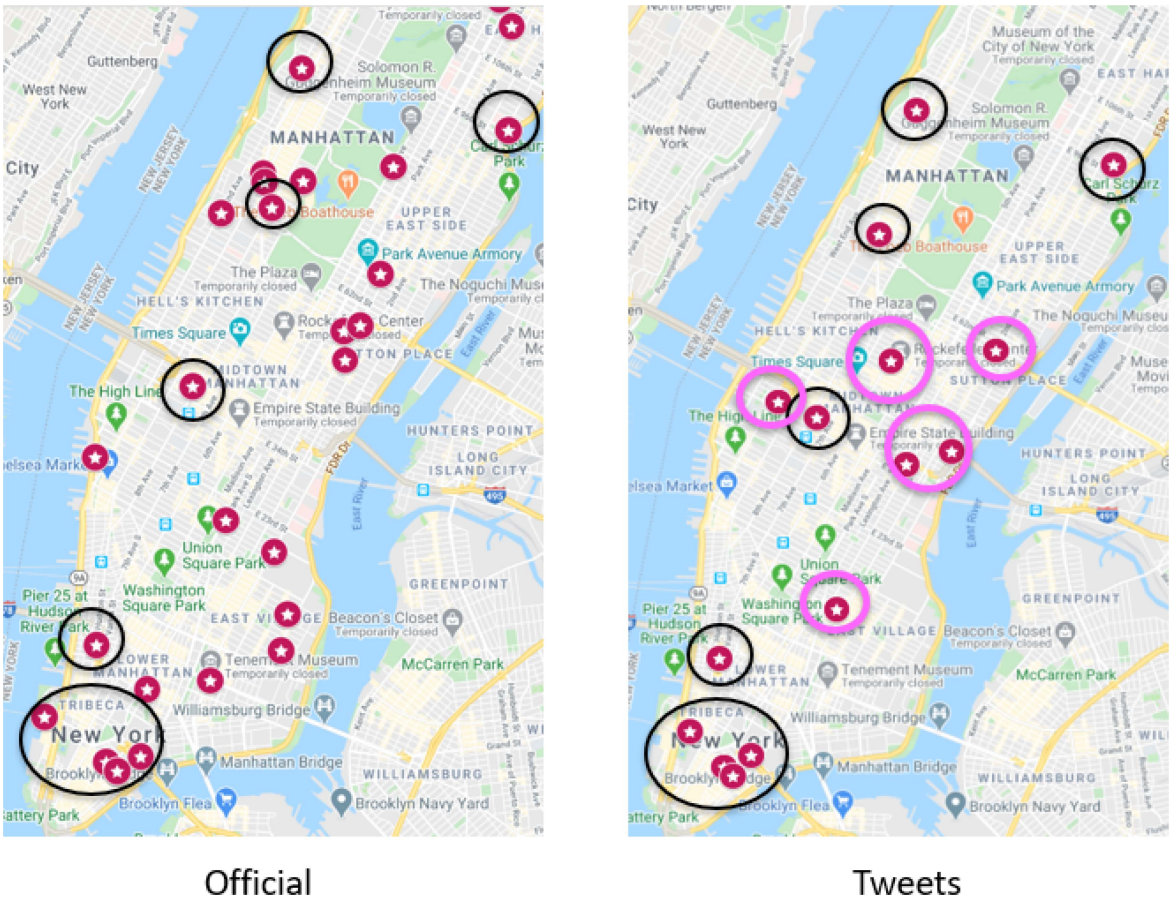

FIGURE 14. Accidents reported by NYC DOT and by social media (tweets) in 01/28/2020, Manhattan.

poster by granting them titles such as "contributor", "senior contributor", etc., based on their performance (i.e., how useful their posts are). This title can upgrade their posts on the feed, so their posts are more likely to lead to the generation of incident reports. Also, exploration of rewards can revolve around financial incentives, so work may focus on how to price information based on its usefulness and the influence of the poster.

Continuing along the rewarding systems; another interesting (and necessary) research direction is fact checking. We are currently working on implementing such solutions to cross-validate the extract information either by only disclosing an incident after validation from different sources, comparing social media information with other data, e.g., from installed sensors along the roadway, or assigning an accuracy score to different sources that evaluate their reliability. This accuracy score could correspond with the aforementioned contributor titles, where people that post information with a high degree of veracity are rewarded, and malicious or uninformed users are ranked lower, or even blocked if necessary.

B. REDUCED DATA ENTRY BARRIERS

Another approach may revolve around reducing the barrier for drivers to post in the first place. Effective Speech to Text (STT) systems would improve hands-free interaction between the driver and their mobile devices. For example, if a driver passes by an area of road construction while in operation of a motor vehicle, it would be much easier (and safer!) for them to speak about the incident rather than manually input it to twitter.

In addition, many newer models of vehicles in developed countries are equipped with Apple CarPlay and Android Auto functionality, which allow for mobile devices to directly interface with on-board systems. As vehicles are equipped with sophisticated arrays of sensors such as LiDARs, Radars, cameras, etc, there is potential for systems that automatically collect and process information related to the vehicle's surroundings. Coupled with other machine learning methods, systems that detect incidents can automatically generate social media posts related to the detected incident. This could lead to the sharing of incident information by a passing driver to be reduced to a simple yes/no response, which would further reduce the barrier to post about information for drivers.

C. QUANTIFYING INCIDENT SEVERITY

As mentioned *a priori*, we speculated that the multiplicity of social media posts may correlate positively with the severity of the incident (i.e., the more people that post about a traffic

jam, the worse it likely is). The use of information to quantify the severity of an incident is useful as an information filter, so that users of the traffic monitoring system are not overwhelmed with unnecessary alerts. On the other hand, machine learning can be utilized to quantify the severity of incidents.

D. STANDARD TEXT GENERATION OF GIS COORDINATES AND PRIVACY ISSUES

One of the main challenges is to transform the street name to GIS coordinates, as our method has some limitations. First, the city of origin of the post must be narrowed down, as different cities may have overlapping street names (for example, hundreds if not thousands of streets in the United States may be called "Main Street"). Second, there is a possibility for the Google Maps API to fail to generate coordinates from highway information, leading to unreported incidents.

Collection of GIS coordinates from the driver's post would simplify the process, but this would require robust privacy and security considerations, such as end-to-end encryption, to protect users from malicious actors. Applying end-to-end encryption and location obfuscation methods is a current field of study. Future work can explore how to incorporate these into VSN traffic reporting systems, while minimizing sacrifices to performance.

VII. CONCLUSION

In this article, we discussed the related efforts of researchers to develop data-driven intelligent transportation systems with the based on crowdsourcing. In addition, we proposed an automated traffic information extractor framework based on state-of-art NLP techniques. After, we utilized our BERT model to filter out unrelated transportation information. We then built and compared QA and NER models that classified reported events, and extracted relevant information from them. Our methods were then used to build a system that can rapidly integrate generated reports into platforms that can help drivers assess the traffic conditions and plan their trip accordingly. The performance of the proposed framework has been experimented on the area of Manhattan, New York City, where real-time traffic maps are automatically updated using information extracted by scraping Twitter. We then compared different data sources and show that social media sources could be a useful complementary data stream for the official transportation agencies. Finally, we discussed potential research directions to consider future improvements of the performance and viability of the applying NLP and crowdsourced social media data in ITS. Our study focused on the use-case of assisting the navigation of drivers, but NLP has great potential for other ITS applications such as public transportation operation/alert systems, monitoring malicious/dangerous/inattentive driver behavior, etc. at low cost, quickly, with widespread coverage, and from a very large pool of potential information.

REFERENCES

[1] X. Wan, H. Ghazzai, and Y. Massoud, "Leveraging personal navigation assistant systems using automated social media traffic reporting," in *Proc. IEEE Technol. Eng. Manag. (TEMSCON)*, Novi, MI, USA, Jun. 2020, pp. 1–6.

[2] "New York City mobility report 2019," NYC Dept. Transp., New York, NY, USA, Rep., Aug. 2019. [Online]. Available: https://www1.nyc.gov/html/dot/downloads/pdf/mobility-report-singlepage-2019.pdf

[3] A. Noto, "NYC economy May be losing 20 billion a year due to traffic congestion," New York Bus. J., New York, NY, USA, Rep., Jan. 2018. [Online]. Available: https://www.bizjournals.com/newyork/news/2018/01/17/nyc-economy-may-lose-due-to-traffic-congestion.html

[4] J. R. Short, "New York gets serious about traffic with the first city-wide us congestion pricing plan," Convers. Media Group, Parkville, VIC, Australia, Rep., Apr. 2019. [Online]. Available: https://www.salon.com/2019/04/05/new-york-gets-serious-about-traffic-with-the-first-citywide-us-congestion-pricing-plan_partner/

[5] A. Rosen, "What really causes traffic congestion?" Bklyner, Brooklyn, NYC, USA, Rep., Jul. 2013. [Online]. Available: https://bklyner.com/what-really-causes-traffic-congestion-sheepshead-bay/

[6] H. Dia and K. Thomas, "Development and evaluation of arterial incident detection models using fusion of simulated probe vehicle and loop detector data," *Inf. Fusion*, vol. 12, no. 1, pp. 20–27, 2011.

[7] X. Wan, H. Ghazzai, and Y. Massoud, "Mobile crowdsourcing for intelligent transportation systems: Real-time navigation in urban areas," *IEEE Access*, vol. 7, pp. 136995–137009, 2019.

[8] K. Lord, "North avenue smart corridor: Intelligent mobility innovations in atlanta improving safety and efficiency," SNC-Lavalin, Montreal, QC, Canada, Rep., Jul. 2018. [Online]. Available: https://www.snclavalin.com/en/beyond-engineering/north-avenue-smart-corridor#:~:text=The%20North%20Avenue%20Smart%20Corridor%20integrates%20several%20smart,to%20improve%20safety%20and%20efficiency.&text=The%20City%20of%20Atlanta%20and,and%20efficiently%20into%20the%20future

[9] "SFPart: Putting theory into practice," SFMTA, San Francisco, CA, USA, Rep., Jun. 2014. [Online]. Available: https://www.sfmta.com/sites/default/files/reports-and-documents/2018/08/sfpark_pilot_overview.pdf

[10] R. Hussain, F. Abbas, J. Son, D. Kim, S. Kim, and H. Oh, "Vehicle witnesses as a service: Leveraging vehicles as witnesses on the road in VANET clouds," in *Proc. IEEE Int. Conf. Cloud Comput. Technol. Sci. (CloudCom)*, Bristol, U.K., Mar. 2013, pp. 439–444.

[11] H. Yan and D. Yu, "Short-term traffic condition prediction of urban road network based on improved SVM," in *Proc. IEEE Int. Smart Cities Conf. (ISC)*, Wuxi, China, Sep. 2017, pp. 1–2.

[12] N. Caceres, L. M. Romero, F. G. Benitez, and J. M. del Castillo, "Traffic flow estimation models using cellular phone data," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 3, pp. 1430–1441, Sep. 2012.

[13] X. Chen and L. Wang, "A cloud-based trust management framework for vehicular social networks," *IEEE Access*, vol. 5, pp. 2967–2980, 2017.

[14] X. Wan, H. Ghazzai, and Y. Massoud, "A generic data-driven recommendation system for large-scale regular and ride-hailing taxi services," *Electronics*, vol. 9, no. 4, p. 648, 2020.

[15] Z. Ning, F. Xia, N. Ullah, X. Kong, and X. Hu, "Vehicular social networks: Enabling smart mobility," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 16–55, May 2017.

[16] A. Kumar *et al.*, "Ask me anything: Dynamic memory networks for natural language processing," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1378–1387.

[17] Z. Zhang, Y. Wu, J. Zhou, S. Duan, and H. Zhao, "SG-Net: Syntax-guided machine reading comprehension," 2019. [Online]. Available: arXiv:1908.05147.

[18] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," 2016. [Online]. Available: arXiv:1606.05250.

[19] J. Clement, "Twitter: Number of monthly active U.S. users 2010–2019," Statista, Inc., New York, NY, USA, Rep., Aug. 2019. [Online]. Available: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[20] M. C. Lucic, H. Ghazzai, and Y. Massoud, "A generalized dynamic planning framework for green UAV-assisted intelligent transportation system infrastructure," *IEEE Syst. J.*, early access, Feb. 17, 2020, doi: 10.1109/JSYST.2020.2969372.

[21] X. Zheng *et al.*, "Big data for social transportation," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 3, pp. 620–630, Mar. 2016.

[22] G. Kalra, H. M. Nguyen, W. Yoon, D. Lee, and D. Kim, "Location digest: A placeness service to discover community experience using social media," in *Proc. IEEE Int. Smart Cities Conf. (ISC2)*, Kansas City, MO, USA, Sep. 2018, pp. 1–8.

[23] A. Sidauruk and Ikmah, "Congestion correlation and classification from twitter and waze map using artificial neural network," in *Proc. 3rd Int. Conf. Inf. Technol. Inf. Syst. Elect. Eng. (ICITISEE)*, Yogyakarta, Indonesia, Nov. 2018, pp. 224–229. [Online]. Available: https://ieeexplore.ieee.org/document/8720995/authors#authors

[24] H. Shekhar, S. Setty, and U. Mudenagudi, "Vehicular traffic analysis from social media data," in *Proc. Int. Conf. Adv. Comput. Commun. Informat. (ICACCI)*, Jaipur, India, Sep. 2016, pp. 1628–1634.

[25] B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge, U.K.: Cambridge Univ. Press, 2015.

[26] Y. Zhang, W. Chen, C. K. Yeo, C. T. Lau, and B. S. Lee, "Detecting rumors on online social networks using multi-layer autoencoder," in *Proc. IEEE Technol. Eng. Manag. Conf. (TEMSCON)*, Santa Clara, CA, USA, Jun. 2017, pp. 437–441.

[27] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.

[28] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, nos. 1–2, pp. 177–196, 2001.

[29] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.

[30] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the Web," Stanford InfoLab, Stanford, CA, USA, Rep. 1999-66, 1999.

[31] F. Barrios, F. López, L. Argerich, and R. Wachenchauzer, "Variations of the similarity function of textrank for automated summarization," 2016. [Online]. Available: arXiv:1602.03606.

[32] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Found. Trends Inf. Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[33] A. R. Pal and D. Saha, "An approach to automatic text summarization using WordNet," in *Proc. IEEE Int. Adv. Comput. Conf. (IACC)*, Gurgaon, India, Feb. 2014, pp. 1169–1173.

[34] T. Vodolazova, E. Lloret, R. Muñoz, and M. Palomar, *The Role of Statistical and Semantic Features in Single-Document Extractive Summarization*, vol. 2. Richmond Hill, ON, Canada: Sciedu Press, Apr. 2013, pp. 35–44.

[35] X. Wan, H. Ghazzai, and Y. Massoud, "Word embedding-based text processing for comprehensive summarization and distinct information extraction," in *Proc. IEEE Technol. Eng. Manag. (TEMSCON)*, Novi, MI, USA, Jun. 2020, pp. 1–5.

[36] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013. [Online]. Available: arXiv:1301.3781.

[37] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran, 2013, pp. 3111–3119.

[38] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. 31st Int. Conf. Mach. Learn.*, Beijing, China, Jan. 2014, pp. 1188–1196.

[39] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*, Doha, Qatar, Oct. 2014. pp. 1532–1543.

[40] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Assoc. Inc., 2017, pp. 5998–6008.

[41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018. [Online]. Available: arXiv:1810.04805.

[42] M. E. Peters *et al.*, "Deep contextualized word representations," 2018. [Online]. Available: arXiv:1802.05365.

[43] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," White Paper, 2018. [Online]. Available: https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf

[44] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," 2019. [Online]. Available: arXiv:1909.11942.

[45] J. Booth, B. Di Eugenio, I. F. Cruz, and O. Wolfson, "Robust natural language processing for urban trip planning," *Appl. Artif. Intell.*, vol. 29, no. 9, pp. 859–903, 2015.

[46] A. Salas, P. Georgakis, and Y. Petalas, "Incident detection using data from social media," in *Proc. IEEE 20th Int. Conf. Intell. Transp. Syst. (ITSC)*, Yokohama, Japan, Oct. 2017, pp. 751–755.

[47] Y. Chen, Y. Lv, X. Wang, L. Li, and F. Wang, "Detecting traffic information from social media texts with deep learning approaches," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 8, pp. 3049–3058, Aug. 2019.

[48] Z. Gao, A. Feng, X. Song, and X. Wu, "Target-dependent sentiment classification with BERT," *IEEE Access*, vol. 7, pp. 154290–154299, 2019.

[49] M. C. Lucic, X. Wan, H. Ghazzai, and Y. Massoud, "Leveraging intelligent transportation systems and smart vehicles using crowdsourcing: An overview," *Smart Cities*, vol. 3, no. 2, pp. 341–361, 2020.

[50] J. T. Méndez, H. Lobel, D. Parra, and J. C. Herrera, "Using twitter to infer user satisfaction with public transport: The case of santiago, chile," *IEEE Access*, vol. 7, pp. 60255–60263, 2019.

[51] Y. Chen, J. Wang, and G. Lai, "Research on improving the government service quality by public comments monitoring: Take suburb park an example," in *Proc. 15th Int. Conf. Serv. Syst. Serv. Manag. (ICSSSM)*, Hangzhou, China, Jul. 2018, pp. 1–5.

[52] P. Pimpale, A. Panangadan, and L. V. Abellera, "Analyzing spread of influence in social networks for transportation applications," in *Proc. IEEE 8th Annu. Comput. Commun. Workshop Conf. (CCWC)*, Las Vegas, NV, USA, Feb. 2018, pp. 763–768.

[53] G. G. Svartzman and J. E. R. Marquez, "Listen to the people! comparing perceived and documented disruptions in public transportation, through quantitative quality of experience, the case study of NYC," in *Proc. IEEE Int. Conf. Syst. Man Cybern. (SMC)*, Bari, Italy, Oct. 2019, pp. 947–952.

[54] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016. [Online]. Available: arXiv:1609.08144.

[55] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for SQuAD," 2018. [Online]. Available: arXiv:1806.03822.

[56] T. Wolf *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," 2019. [Online]. Available: https://arxiv.org/abs/1910.03771.

XIANGPENG WAN (Graduate Student Member, IEEE) received the bachelor's degree in electrical engineering from the Harbin Institute of Technology, Harbin, China, in 2015, and the master's degree in applied mathematics from the University of Minnesota, Duluth, MN, USA, in 2017. He is currently pursuing the Ph.D. degree in engineering management with the Stevens Institute of Technology, Hoboken, NJ, USA. His research interests are mainly on smart city design, big data analysis, and applied machine learning.

MICHAEL C. LUCIC (Graduate Student Member, IEEE) received the Bachelor of Science degree in industrial and systems engineering with a minor in mathematics from the University of Florida, Gainesville, FL, USA, in 2018. He is currently pursuing the Ph.D. degree (third-year) in systems engineering with the School of Systems and Enterprises, Stevens Institute of Technology, Hoboken, NJ, USA. His research interests revolve around applied optimization and machine learning in intelligent transportation systems.

HAKIM GHAZZAI (Senior Member, IEEE) received the Diplome d'Ingenieur (with Highest Distinction) in telecommunication engineering and the master's degree in high-rate transmission systems from the Ecole Superieure des Communications de Tunis (SUP'COM), Tunis, Tunisia, in 2010 and 2011, respectively, and the Ph.D. degree in electrical engineering from KAUST, Saudi Arabia, in 2015. He is currently a Research Scientist with the Stevens Institute of Technology, Hoboken, NJ, USA. Before joining Stevens, he was a Visiting Researcher with Karlstad University, Sweden, and a Research Scientist with Qatar Mobility Innovations Center, Doha, Qatar, from 2015 to 2018. His general research interests are at the intersection of wireless networks, UAVs, Internet-of-Things, intelligent transportation systems, and optimization.

YEHIA MASSOUD (Fellow, IEEE) received the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA. He is currently the Dean of the School of Systems and Enterprises, Stevens Institute of Technology, Hoboken, NJ, USA. He has held several industrial and academic positions, including Member of the Technical Staff with the Advanced Technology Group, Synopsys, Inc., Mountain View, CA, USA, a Tenured Associate Professor with the Departments of Electrical and Computer Engineering and Computer Science, Rice University, Houston, TX, USA, the W. R. Bunn Head of the Department of Electrical and Computer Engineering, University of Alabama at Birmingham, Birmingham, AL, USA, and the Head of the Department of Electrical and Computer Engineering, Worcester Polytechnic Institute, Worcester, MA, USA. He has authored over 225 papers in peer-reviewed journals and conferences. He leads research efforts in various areas of electrical and computer engineering, computing, and systems and software engineering. He was a recipient of the Rising Star of Texas Medal in 2007. He was also a recipient of the National Science Foundation CAREER Award in 2005, the DAC Fellowship in 2005, the Synopsys Special Recognition Engineering Award in 2000, several best paper award nominations, and two best paper awards at the IEEE International Symposium on Quality Electronic Design in 2007, and the IEEE International Conference on Nanotechnology in 2011. He was selected as one of ten MIT Alumni Featured in the MIT EECS Newsletter in 2012. He was an elected member of the IEEE Nanotechnology Council from 2009 to 2011. He served as the Theme Leader of Novel Interconnects and Architectures in the SRC Southwest Academy of Nanoelectronics from 2006 to 2011. He was named as a Distinguished Lecturer by the IEEE Circuits and Systems Society from 2014 to 2015.