Detection and Prediction of Traffic Accidents using Deep Learning Techniques

Anique Azhar · Saddaf Rubab · Malik M Khan · Yawar Bangash · Mohammad Dahman Alshehri · Fizza Illahi · Ali Kashif Bashir

Abstract Road transportation is a statutory organ in a modern society; however it costs the global economy over a million lives and billions of dollars each year due to increase in road accidents. Researchers make use of machine learning to detect and predict road accidents by incorporating the social media which has an enormous corpus of geo-tagged data. Twitter, for example, has become an increasingly vital source of information in many aspects of smart societies. Twitter data mining for detection and prediction of road accidents is one such topic with several applications and immense promise, although there exist challenges related to huge data management. In recent years, various approaches to the issue have been offered, but the techniques and conclusions are still in their infancy. This paper proposes a deep learning accident prediction model that combines information extracted from tweet messages with extended features like sentiment analysis, emotions, weather, geo-coded locations, and time information. The results obtained show that the accuracy is increased by 8% for accident detection, making test accuracy reach 94%. In comparison with the existing state-of-the-art approaches, the proposed algorithm outperformed by achieving an increase in the accuracy by 2% and 3% respectively making the accuracy reach 97.5% and 90%. Our solution also resolved high-performance computing limitations induced by detector-based accident detection which involved huge data computation. The results achieved has further strengthened confidence that using advanced features aid in the better detection and prediction of traffic accidents.

Keywords Accident Detection · Accident Prediction · Deep Learning · Social Media Accident Detection, High Performance Computing

A. Azhar E-mail: aneeqathar@gmail.com · S.Rubab E-mail: saddaf@mcs.edu.pk · M.Khan E-mail: mmurtaza@mcs.edu.pk · Y.Bangash E-mail: yawar@mcs.edu.pk · F.Illahi E-mail: fizza.ncsael@mcs.edu.pk
National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan

A K Bashir E-mail: dr.alikashif.b@ieee.org
School of Electrical Engineering and Computer Science, National University of Sciences and Technology(NUST), Islamabad, Pakistan
Department of Computing and Mathematics, Manchester Metropolitan University, Manchester, United Kingdom

M D Alshehri E-mail: alshehri@tu.edu.sa
Department of Computer Science, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia

1 Introduction

Over the past few years, travelers and the public, in general, are witnessing one downside of the population increase in the form of traffic congestion and blockage. Traffic blockage sometimes depends on the area, so most researchers categorize it into recurrent and non-recurrent events [16]. The root cause of traffic congestion can be various reasons such as accidents, road construction, marathon events, and faulty and illegal parking, etc. The traffic congestion caused due to accidents holds extreme significance as it involves human life. The death rate in traffic accidents of young people is high [17]. According to the authority of the World Health Organization (WHO), the death rate due to accidents is increasing rapidly. Approximately 1.35 million deaths occur every year because of road accidents. This is the prime reason why researchers are feverishly corroborating from the past few decades to reduce the death rate caused by accidents.

Manually observing and reporting traffic jams is an arduous task that involves high error. [12][9]. In this regard, many researchers and practitioners are making use of social media platforms to extract the necessary information to understand the patterns of accidents, to make better decisions. The previous studies of the researcher's conventional detection-based strategies with only traffic information faced some challenges. One of the challenges is the authenticity of the data. Previous research conducted was mostly on the assumption that the data collected through detectors are reliable. However, traffic operations cannot solely rely on it as there are numerous detector failures and communication errors. For example, the Illinois Department of Transportation (IDOT) in Chicago revealed that around 5% of their detectors fail to work at the time of the accident. The failure of sensors leads to difficulties in the investigation and detection of the root cause of an accident. [24]. The traditional ways of detecting an accident include the installation of sensors like cameras and collision detectors. However, traditional ways are difficult to implement in long roads because of high deployment and maintenance costs. Gradually, there has been a paradigm shift in collecting the data through social media platforms and developing knowledge basis on the research studies done on the gathered data. [3]

Social media platforms are a source of a copious amount of data. For example, every day Twitter has 400 million tweets on average and 100 million active users all around the world [16]. People are relying on social media platforms for a wide array of tasks including businesses, advertisements, circulating important news, interacting with friends, etc. Users actively post the real-time events as well; this means that sometimes news and events are circulated through social media more quickly as compared to the news channel; for example, in April 2013 the death of the Boston Marathon and the death of the former British Prime Minister Thatcher was first announced on the social website. Researchers are now making use of these platforms to extract the data related to accidents. Whether it is about accident detection, traffic analysis, or election results, social media platforms played vital role by providing important data. This data then became the backbone of research studies conducted in past times. Hence, social media is considered a reliable source for research purposes. As compared to the traditional methods, Twitter-based method has many benefits which are listed below:

- Data extraction from Twitter is very cost-effective as compared to traditional data collection i.e. Hardware sensors.
- User's tweet in natural language and there are stable algorithms developed to process those tweets for information
- Tweets posted, entail important information like time at which it was posted, and the location as well; so the detection of the incident reported on Twitter is in a timely manner
- The system of tweets-based occasion acknowledgment can be stretched out and applied to various types of traffic data and issues [10]

Nowadays, information on social media platform spreads within seconds, very large community including Governments has started relying on social media news [27] On the other hand, different official resources also use social network services to update information on the internet which also includes traffic conditions, for example, Dubai police also have an account on Twitter where they post road updates frequently. In research [13], Indonesian election results were predicted using Twitter status message posts. According to the results yielded, the Mean absolute error they got from their architecture was only 0.6% which was very less as compared to other ways of prediction. This proves the importance of social media platforms especially when it is integrated with modern technologies like natural language processing and deep learning for purpose of detection and prediction.

These facts demonstrate that there are rich assets in social network services, which can be utilized to distinguish and foresee congested roads [1]. Using Twitter as a source of extracting useful information has drawn quite an attention in the recent past. Tweets are often succinct and to-the-point which makes them the most suitable candidate for spreading or circulating any news. Users often tweet about the situation of the traffic in case it is crowded, or it is jammed [19]. There has been a constant effort to deploy machine learning algorithms to achieve high accuracy and efficiency not only for detection but also prediction of road accidents. The purpose of this study is to include different factors present in the message updates (tweets) i.e., emojis which were neglected and removed as noise in previous researches. Emojis provide a lot of information about the context of message especially in accident news by classifying it into sad and happy news. Merging emojis with other information i.e, Weather info, location, Geo-coded coordinates, wind, and time attributes extracted from the text using different natural language processing techniques allowed us to detect and predict road traffic accidents by utilizing Twitter data.

To summarize, integrating social network data into a traffic-related study opens up a slew of new opportunities for transportation scientists. The findings of this study show that social network details may be noisy due to certain limitations such as use of slang terms, variation in sentence structure, and the limited capacity of characters in a tweet, etc. but when combined with modern Deep Learning techniques such as LSTM, CNN, etc. they contribute in the detection and prediction of traffic accidents. During the selection of algorithms of deep learning, various factors are considered which affect the accuracy of the model like class imbalance. The majority of real-world classification problems include some degree of class imbalance, which occurs when there aren't enough instances of the data that correspond to either of the class labels. As a result, it's critical to choose the right assessment measure for your model. Moreover, another factor affecting accuracy is the weight of the features selected [28] [29] [30] This study's model will be used to detect and predict traffic collisions in real time, potentially contributing to quicker emergency response and decision making. In the future, even more accurate models can be calibrated by creating datasets that can be commonly used for research.

The contribution of this paper is to capture the traffic accidents occurring via social platforms and then to use the extracted information to predict the trend of accidents. In the traditional techniques, detectors like sensors of collision and cameras are used which not only feed visual data but also numeral data continuously into the system. Accident detection mechanisms installed at Transportation Departments or Analysis Wings continuously process these feeds for the extraction of information. Compute process continuously takes place which leads to the problem of high-performance computing. Such systems need multiple compute nodes based MapReduce mechanisms.

Our Proposed solution is not dependant on reducing map architecture mechanism as data is only fed when an accident occurs. Deep learning enables computer models to learn representations of data with several degrees of abstraction, as opposed to traditional machine learning approaches, which were confined to processing natural data in its raw form. This research aims to address the major industry-leading problems of high-performance computing; and also saves other resources like power and network bandwidth. The manifold contributions of the paper are:

- This work includes emotions and sentiments present in tweet text to identify and classify traffic accident related posts. To the best of our knowledge, this is the first time such parameters are included, and the results clearly show the enhanced accuracy of our model
- We present additional models for detection and prediction which helps in better understanding and decision making related to transportation
- We merged rule matching classifier with other related modules i.e. Sentimental Analyzer, Emotion Classifier in a systematic hierarchy to develop an automatic classifier which aid in classifying tweets. It works with accuracy of $> 80\%$
- We also developed modules for Geo Coding, Historical weather, Impact analyser etc. which aids in extracting useful information for better prediction of accidents

Rest of the paper is distributed as: Existing work is presented in Section 2. Architecture of the proposed algorithm is presented in Section 3, whereas Section 4 presents the evaluation of the model. Section 5 discusses the validation and comparison of the proposed model with the other models, and finally Section 6 presents conclusion and future work.

2 Literature Review

The rapid surge in population is directly proportional to the rate of accidents and only a few systems have been built to address this problem. According to the previous studies, traffic mishaps can be controlled if the driver knows the areas where most of the incidents took place [2]. Only traffic control rooms can envision current conditions through customary or traditional sensors. But, there is also a need for a computerized and automatic monitoring of the traffic to present the real-time clarification of traffic state. Many methods have been proposed by different researchers to perceive incidents that happened on roads. These systems mostly used the information of the single data source whose information is in the form of the single language, but the suggested system has some issues, they may either fail to automatically identify the reason for the accidents or do not renovate the traffic condition simultaneously. To develop the exact automatic system, researchers face some challenges such as merging multiple data resources, accrediting the exact location of the post, and detecting and predicting traffic jams [1].

For decades, traffic jams are a major issue and researchers proposed many methodologies to overcome the traffic issues developed the Traffic Watch framework for slithering and handling open tweets by conveying propelled web automation and state-of-the-art machine learning algorithm. If properly gathered, processed, and analysed, social network platforms such as

Twitter and Facebook may enable fast transmission of unstructured traffic-related information through application programming interfaces (APIs)[25] . The authors of [10] used two strategies ("Twitter API and initial data crawling" and "Adaptive Data Acquisition") to build up the approach to interpret tweets into traffic occurrence data by leading the underlying information obtaining from Twitter servers, trailed by an iterative procedure. They utilized the tweet writings to extract the occurrence data on the two parkways and arterial as a proficient and financially savvy option in contrast to existing information sources. They additionally mapped the tweets into features or component space and characterized them by the Semi-Naive-Bayes (SNB) by identifying the tweets as either traffic incident occurrence tweets or non-Traffic incident tweets. All tweets were geo-coded and geo-parsed to distinguish the area. At that point, geo-coded tweets were named Supervised Latent Dirichlet Allocation, arranged to recognize the status of the accident. The limitation of this strategy is the absence of ground truth at the time when the mishap occurred (not revealed time), so they are not able to give an absolute solution with respect to time. [14]

A. Pereira recognized oddities as per drivers conducting manners on the metropolitan road network. The identified anomalies were conferred by a sub-diagram of a road network where driver routing behavior significantly contrasts from their original arrangement. They also mined GPS direction information to recognize significant steering changes, the sub-graph of the street arranges on which an oddity is found that is used to recover pertinent internet-based life to portray the anomaly. They visualized two use-cases for their framework, one is situated towards an independent end-user traveling around an anomaly and the other is patterned towards city organizers and traffic controllers to encourage checking and visual examination. This framework also provided the services for the individual clients i.e. real-time alert showing the anomaly area and also evaluated characteristics of the anomaly such as speed and steering changes and a semantic setting to offer significance to the irregularity, i.e. the social site terms that portray the occasion. The system consists of three parts: offline mining, abnormality discovery, and irregularity investigation [18].

The researchers of [24] applied and correlate some deep learning methods i.e. "Deep Neural Network (DNN)" and "Long Short-Term Memory (LSTM)". They proposed an efficient characteristic selection operation for removing both, the independent and matched token characteristics from social websites. They disclosed the language procedure of the tweet clients in depicting the mishap locations. They authenticated the potency of deep learning access in classifying web-based information. Their results concluded that the deep learning approach exceeded the other conquering data mining methods. The pros and cons of the mishap recognition are checked and totally examined by contrasting tweets and both mishap log from the state Department of Transportation and traffic information from a huge number of loop identifiers.

A. Parsa presented the Social network service Jam technique, a powerful and effective framework to distinguish and foresee street car influxes supporting multi-language (English and Arabic) information gathered from multiple sources, for example, Twitter and Instagram. This framework comprised six modules i.e. multi-source information gathering, data stream handling, post-characterization, area or location Recognition, traffic examination and visualization, and traffic forecast. The objective of the proposed framework was to gather tweets from multiple data sources of social media. Moreover, data was processed and filtered to expel irrelevant information regarding street traffic, the raw posts were then arranged according to their language, and afterwards, raw posts were tokenized into words. All the raw posts were checked demonstrating whether its information is related to traffic incidents or non-traffic incidents. The ordered traffic-revealing RPs are further sub-categorized into the reasons for traffic. The grouped RPs are geo-tagged to distinguish their physical areas. The raw posts, which were accumulated from various information sources, experienced the late combination procedure and then they were analysed to detect traffic incidents and their root causes. The distinguished traffic was then envisioned on a spatiotemporal map. The posted messages were nourished into a Prediction model to foresee future traffic congestion [17].

The authors of [8] developed a framework for zoomed-in grounding (below city level) for short messages (e.g., tweets). This framework joined diverse characteristics of language handling and AI methods to expand the quantity of geo grounded tweets, which is basic to numerous applications, such as fiasco reaction and constant traffic checking. To process this system, the researchers started by tuning in to the Twitter feed in order to discover significant social posts utilizing a high-quality rundown of 70 watchwords identified to the context of the system (e.g., road traffic). Returned posts were then pushed through a three-step pipeline where they double-check the relevance of the post using a binary classifier (the Content Filter), at that point, they extract the area names with the reference present in the posts. Afterward, they geo-locate the identified areas to their exact spot referenced on the map. This procedure allowed to filter undesirable posts, and increase the pertinent ones with exact geolocation arrangements which were finally uncovered for utilization through RESTful API. [20]

P. Chen developed a unified investigative substructure that consolidated the two models (modeling-based language model and community deduction model) dependent on pivot misfortune Markov irregular fields (HLMRFs). They formulated the new factual system (L-HLMRFs) that incorporates point models and pivots misfortune Markov arbitrary fields by breaking down the traffic blockage design from Twitter and related the aftereffects of the gathered information with genuine INRIX traffic speed information. As per the analysts, L-HLMRFs make it conceivable to cooperatively deduce blocked street connects over the whole street arrange dependent on pertinent tweet information, at that point develop a rough calculation for LHLMRFs MAP derivation. They assessed the L-HLMRFs by leading broad tests on genuine Twitter and INRIX traffic speed datasets [4].

The authors of [19] examined the capability of Twitter for supporting ongoing occurrence locations in the United Kingdom (UK). They presented a system for recovering, preparing, and characterizing open tweets by consolidating Natural Language Processing (NLP) procedures with a Support Vector Machine calculation (SVM) for content arrangement. After separating the tweets by street names and traffic-related catchphrases, NLP was utilized to evacuate uncommon characters and stop words. They utilized the support vector machine calculation to characterize them into 'traffic' or 'non-traffic' related. Tweets experienced the accompanying pre-processing stages which are Twitter information securing, Pre-processing and Classification. For the ongoing stream of information, they chose the Streaming API and the association was made through Python with a geolocation channel. Before bolstering the tweets into the classifier, the scientists applied some content mining systems to evacuate every one of the characters. They applied some content mining methods to their dataset for example Tokenization and Stop word expulsion. Finally, they predicted and categorized the traffic and no traffic accident posts using their SVM achieving 88.27% accuracy. Table 1 presents the summary of the related work. Although, there exists adequate research in general DM modelling, there is lack of investigation and comparative analysis of ML algorithms for the detection of traffic related events [26]

3 Architecture

The work presented in this manuscript develops an effective approach to convert the tweets into useful information for predicting future accidents. The proposed architecture comprises of two main phases i.e. Detection and Prediction. Social media posts are fetched, pre-processed, and labeled through our smart classifying techniques in the detection phase. In this phase different rule matcher, sentimental analysis, and emotion classification techniques are stacked in such a way that yields the best classification results. Further, manual labeling has been performed to correct errors and ensure training of deep learning model from noise free dataset. Flow diagram of the detection phase is shown in Fig. 1 . In the Prediction phase, a thorough analysis of the time and location information, collected in the detection phase, is performed to extract more precise data of the reported accident like exact lat, long, hour of the day, day of the week, and weather information of the accident area. We further used these features to train a deep learning model to predict the accident severity. Flow diagram of prediction phase is shown in Fig. 2

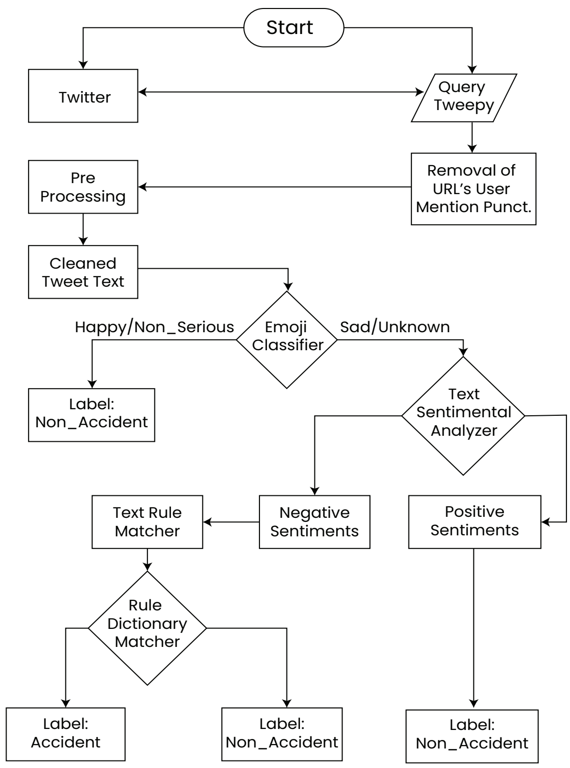

Fig. 1: Detection Phase Flow Diagram

Table 1: Literature Comparison

Reference	**Data Source**	**Accuracy/ Precision**	**Learning & Classification Technique**	**No. of Data Set**	**Limitations & Future Work**
Yiming Gu (2016) [10]	Twitter	SNB has 90.5% classification	Semi Naive Bayes	22,200	- To improve the categorization of the traffic incident tweets, more extensive NLP algorithms can be applied - Acquire more tweets to examine incident coverage - The geocoder can be updated by joining extra names of streets and focal concentrations with the ability to address incorrectly spelled names - Manual Labelling
Y. Chen (2018) [5]	Sina Weibo	89%	LSTM, CNN, LSTM-CNN	11000	- Word embedding is created using only one micro blog i.e.Sina Weibo. Multiple blog services can be combined to increase accuracy - Extraction of Features like location and time from micro blogs and combining them as feature to increase accuracy - Manual Labelling
Zhenhua Zhang (2018) [24]	Twitter	The accuracy of this system is 85%	DBN (Deep Belief Network) & LSTM (Long term short memory)	900	- The occurrence of traffic events may identify through non-geotagged tweets - Merging of data from different information sources. - Manual Labelling
Alexandra S.Pereira [18]	Twitter (Waze-TrafficSF)	88% & 93%	K-nn & SVM	3000	Specialize the suggested model by considering other aspects, rewrite the architecture, introduce new techniques and information sources and implement new tests - Manual Labelling
Balsam Alkouz (2020) [1]	Twitter and Instagram	The accuracy for English tweets is 89.2% The accuracy for Arabic tweets is 92.2%	SVM	English: 940 & Arabic: 767	The fusion of Twitter and Instagram increased efficiency. More social sources can be fused to achieve better accuracy and dimensionality in the dataset - Both Manual and Automatic labelling to enhance speed
Eleonora D'Andrea (2015)[6]	Twitter, News & blogs	- 88% for 3 class - 95% for 2 class	SVM (Support vector machine)	1330	- In Future: Merging data with physical sensors for more accurate real time reporting - They used News Papers and other reported accidents to perform analysis instead of relying totally on social media feeds - Manual Labelling
A. Salas (2017)[19]	Twitter, Micro blogs	88% Accuracy	NLP & SVM	13410	- Sentiments and emotions play important role, so they can be included in future studies - Location information and extraction techniques can be enhanced to improve results - Manual Labelling
H. Nguyen (2016)[15]	Twitter, News & blogs	>90% for Precision and Recall, Accuracy not mentioned as it varied	NLP & SVM	5000	- Twitter accounts which frequently post accident news can be included with other feed to support automatic collection of more precise and accurate information. - Engaging feeds from multiple source can boost the early detection of accidents. - Labelled db provided by TMC (AUS Gov.)

3.1 Detection Phase

In the detection phase, it is determined whether the tweet would be labeled as an accident or a non-accident. It consists of the modules as follows: Data collection, pre-processing, emoji classifier, textual sentimental analysis, rule matching for classification of SMP (Social Media Post). A detailed explanation regarding implementation is discussed in the sections below

3.1.1 Data Collection

On a daily basis, approximately 500 million tweets are posted related to economic, social, political, and personal matters. Therefore, it is a hub of useful information related to accidents occurring throughout the world. The majority of researchers use this platform because of the availability and authenticity of the information. Accident related tweets are requested from Twitter using API requests. 'Tweepy' library [23] is used to gather information by providing multiple informational parameters like query, geocodes, and radius. However, the limitation of this library is that it provides data for a week only. Data is gathered by running a query for several weeks. There are certain keywords used for the efficient collection of accident related tweets like "traffic accident", "road accident", and "car accident". Twitter allows a maximum of 140 characters in a tweet, so people often find very limited space to report any accident related information. Hence, the reporting incidents are specific and entail the particular information required for analysis. Other attributes of tweet like time, date, geo-tagged location, and username are also provided as result of a request, hence these attributes aid in refining information like country of origin, city, and sometimes location. In different countries, there are same names for a road; people usually do not find enough space to write city or country name. This problem has been addressed by using meta data attributes which helped in the classification.

3.1.2 Pre-Processing

Often a tweet contains user mentions, links, hashtags, and punctuation which are irrelevant to the detection or prediction of traffic accident reporting. So, pre-processing is performed to extract the short text from the entire text. Scripting languages were used to enable extraction process. The raw text is structured, and transformation of short text takes place for following things:

1. LowerCase: All the text of a post is converted into lower case
2. User Mentions: In an SMP, other users of media platforms are tagged or mentioned. Format @"" is filtered from every SMP and removed from the text
3. URLs: Links of the websites, pdfs, images, videos or any other content are filtered and removed from the SMP
4. Punctuation: Special characters and punctuation has no information in it about the accident and has nothing to do with reporting of traffic accident, so such characters are also extracted and removed from the SMP

The processed data, along with other extracted information, is saved in structured databases so that steps like stemming, Dimension handling, stop word removal, and feature selection can be easily performed as required. Algorithm 1 shows functionality and flow of preprocessing phase.

Algorithm 1: Pre Processing SMP

```
input  : A String type tweettext
output: Filtered Processed Post
DictionaryUrls ← {http, https, www}
  DictionaryMentions ← {#, @}
  DictionaryPunctuations ← {:,' ,' ," ,!,' ,}
for RPi ∈ of Raw posts, i : 1....n do
    if RPi contain KW and KW ∈
      DictionaryUrls then
        remove mention
    if RPi contain KW and KW ∈
      ofDictionalyMentions then
        remove url
    if RPi contain KW and KW ∈
      ofDictionalyPunctuations then
        remove punctuation
return filtered
```

3.1.3 Emoji Emotion Classifier

Emotions play an integral role in the field of humor detection, information security, and social media researches. Many researchers are now deploying it in the classification problems because manual training datasets classification is a time consuming and intensive task [5]. Vast number of tweets contains emoticons in it. We make use of this feature to make our classifier smarter and reduce workload. Use of emoticon in a tweet reveals a lot about the context of text without reading and processing it. In this module, it looks for the emojis in the text and passes the emotion-related information to the emotion classifier stated as Algorithm 2 where emotions are matched to the groups of sad and happy emotions. A post reflecting the happy behavior is labeled as a non-accident post while the post which reflects sadness or those which do not contain such information is further processed towards the text classifier for in-depth review and analysis of text.

3.1.4 Text Sentimental Analyzer

For an in-depth and comprehensive analysis of text, we have used Natural Language processing techniques and methodologies. Context is still important at this stage; we have analysed that posts which reports accidents reflects grief and sorrow in them which explains presence of negativity in them. Hence tweets are filtered

Algorithm 2: Emoji Emotion Classifier

Result: Classified Emoji, Happy or Sad
Happy : ;
Sad : ;
for $PPi \in set\ of\ processed\ postsPPs, i : 1....n$ **do**
  if $PPi\ contain\ Emojis \in happy$ **then**
    $non\ traffic\ accident\ label$
  if $PPi\ contain\ Emojis\ \in sad\ emojis$ **then**
    $text\ classifier()$
return post;

through sentimental analysis techniques to further dissect into Positive and Negative reflection tweets. Positive tweets are labeled as 'non accident' tweets while negative tweets are still not classified and needs further in-depth analysis. In this module, polarity scores are measured using Vader sentiment analysis library of python.

3.1.5 Rule Matcher Classifier

Different rules are written which help in deriving information related to accident from the dissected part of speech (POS). A text is dissected in POS format and then analysis is conducted to look for matching rules which reflect the presence of accident information in the post. Using these, tweets are labeled as accident or non-accident reporting SMP. Rules generated for detection are shown in detail in Algorithm 3. Sample tweets fetched and processed are shown in Table 2

Algorithm 3: Rule Matcher

POS: Part of Speech, Adj.: Adjective, NUM: number
$R1 \leftarrow \{POS \in [NUM, ADJ, NOUN]\}$
$R2 \leftarrow \{POS \in NOUN\}\&\{LOWER \in [at, on]\}$
$R3 \leftarrow \{POS \in NOUN\}\&\{IS\_PUNCT == True\}$
$R4 \leftarrow \{POS \in ADP\}\&\{POS \in NOUN\}$
$R5 \leftarrow \{LOWER \in [accident, accidents, incident, injury, damage, death]\&POS \in NOUN\}$
$R6 \leftarrow \{LOWER \in [no, none, major, minor, severe]\&\ LOWER \in [injured, injury, injure, hurt, damage]\}$
$R7 \leftarrow \{POS \in [NUM, ADJ, NOUN]\&LOWER \in [injured, injury, injure, kill, die, damage, hit, bruise, crash]\&POS \in VERB\}$
$R8 \leftarrow \{POS \in NUM\&LOWER \in IN[vehicle, car, truck, semi, van, sedan, people, pedestrian, suv, person, bike, motorcycle, automobile, auto, bus, cyclist, buggy, cruiser]\}$

Table 2: Classified Tweets

Tweet	**Label**
She mad ash she had that black bag on accident, car damaged https://t.co/B64fgGqwe2 , 24/12/2019 0:09 Pittsburgh PA	not accident
ma andover major accident I-495 @ I-133 u/d cmnd rpts tools going to work att. UNK severity of injs. usa 03,25/12/2019 17:50, West Mifflin PA	accident
2019-12-29 20:08 est ct fairfield major accident I-95 sb btwn 22 amp; 21 pd o/s w/mva. tt vs veh. neg pin. unk injs att. r lns clsd. hvy delays uea07, 30/12/2019 1:12, West Mifflin pa	accident
@maddoxem666 It was an accident bro 30/12/2019 18:26,Oakdale pa	not accident

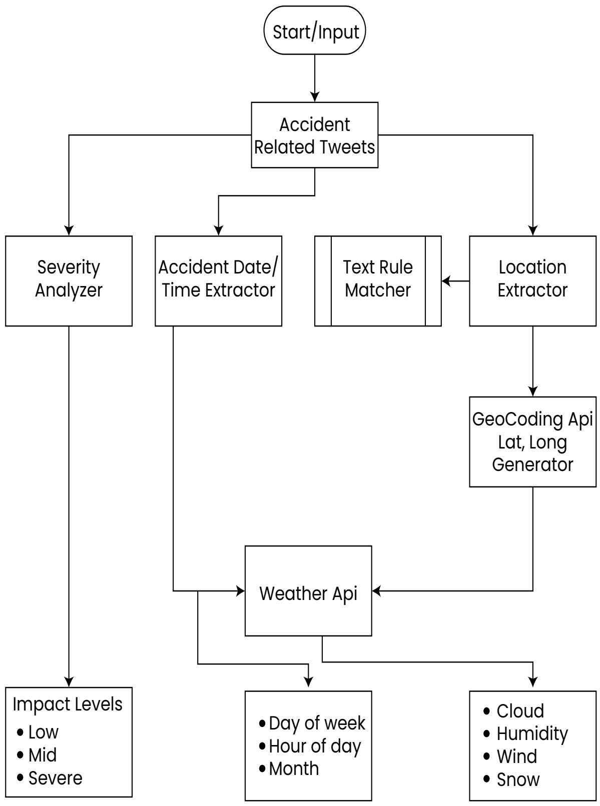

Fig. 2: Prediction Phase Flow Diagram

3.2 Prediction Phase

In prediction phase, features are extracted and selected for training of deep learning model. Features from previous section i.e. location, Date and Time are used to further process and extract more precise information to train the model. This phase consists of the modules as follows: Location Processor, Geo Coding, Date & Time Extractor, Impact Analyzer and weather.

3.2.1 Location Processor

Most of the time, tweets contain information related to the site of accident. There are copious ways to mention

such information and it gets difficult to extract such information due to lack of a uniform method. Moreover, use of limited characters in a tweet makes it more challenging as people often use abbreviations and short forms. However, we analyzed the gathered data set and then, gleaned common reporting styles of users and created bag of words. Information regarding location is more often seen before and after those words. A rule matcher technique is used along with text scrapper to extract most useful information about location from the provided SMP.
BOG: "accident before", "accident at", "accident on", "accident in", "between", "from"

3.2.2 Geo Coding

It is integral to convert the location of accident collected in the previous step to numerical data in order to train a model. Accident location needs to be reversed to latitude and longitude for multi purposes. First, it is used in fetching weather information along with date and time. Second, lat and long are directly used as feature set in model training. Google service for geo coding is used in this process but text location alone does not give exact information because there are multiple roads, corners, and streets that have same names. Another problem encountered is the usage of local area names, native residents give short names to the places for their ease and make use of it in their posts. These are not recognized by the Google Maps dictionary. We solved these problems by using the city origin of tweets provided by twitter and merging them with country and most possible location information gathered from our own location processor to make it more accurate. We took the nearest latitude and longitudes provided by Google API against the text location provided.

3.2.3 Date and Time Extractor

Information like time and date are considered most important for detection and prediction of accident. People reporting an accident via tweet mostly mention the time of the accident but does not include the date. So, time is extracted if provided in the text otherwise date and time of the post, provided by Twitter, is considered. This information is included as features for the training after transforming it into most comprehensive way. Moreover, day of week, hour of day, and month are extracted using different library function of scripting languages and included as feature set.

Table 3: Bag of word - Impact

Keyword	**Score**	**Keyword**	**Score**
traffic jam	0	blocked	0
slow traffic	0	stationary traffic	0
queueing traffic	0	delays	0
injured	1	damage	1
collision	1	wreck	1
crash	1	smash	1
bump	1	strike	1
killed	2	blood	2
dead	2	fire	2
fatal	2	lethal	2
tragic	2	mortal	2

3.2.4 Weather Extractor

Weather extremes like precipitation, high winds, and temperature fluctuation affects the driver's capabilities, vehicle's stability and maneuverability, and roadway infrastructure etc. Adverse weather and slick pavements accounts to increased traffic accidents. Hence, such factors are to be considered when it comes to accident prediction. In order to extract information of weather; location attributes such as latitude and longitude along with date and time information are sent to historical weather API. 'WeatherMeteostat' library is used to fetch information like snow, rain and other factors of weather. This API gives checks for nearest weather station against geo codes, in 90% of cases it found weather station in radius of $< 5km$.

3.2.5 Accident Severity Analyzer

Author of a post usually mentions the consequences of an accident like the fatality rate, situation of traffic or the severity of damage occurred. For the purpose of this research, the impact of accident is categorized into

- Delays or congestion on roads
- Financial loss
- Loss of life

First one is labeled as least severe impact, second as medium impact and third is considered most severe. Severity is extracted from text using a scrapper along with word matcher. Bag of word is created after analysis of gathered dataset which is shows in Table 3.

4 Evaluation

In evaluation phase, we evaluated our all three modules; first module is about our effective and unique way of labeling data using meta data and architectural approaches which is introduced for the first time. Second module is about the detection of accident tweets using deep learning models which includes LSTM, Simple RNN, and GRU. Third module is about the prediction of accidents using ANN. In all of our modules and simulations, we achieved accuracy ranging from minimum of 80% to 94%. Detailed discussion is done below.

4.1 Automated Tweet Classifier

Deep learning algorithms heavily rely on the quality of dataset. The more accurately data is labelled, the more accurate training will be, hence predictions through these models will also be accurate and precise. Quantity of data is required along with quality, hence achieving both at same time is very difficult task as it requires more effort and resources. We used our architecture to not only derive high quality data set but also adopted techniques to label our classes automatically. As shown in Fig 1 labelling is done provided on different parameters like Emotions and sentimental analysis of text data (tweets). Furthermore, applying sophisticated rule matcher which base on natural language techniques including identifying different parts of speech and mapping them against grammar rule to identify whether accident related information is present in text. Based on such artifacts we labelled all our dataset which is more than 14000 between two classes.. i.e., Accident or Non- Accident.

For evaluation of our automated classifier, we run our algorithm on all the collected tweets and labeled them accordingly. We ,then, labeled all the tweets manually for finding accuracy of our smart algorithms. The results showed that 11265 tweets were correctly labeled; this makes accuracy 80.46%. Although accuracy of 80% is not enough for the training of deep learning models but it helps a lot and reduces human error. We improved this accuracy to 94.2%, mechanism and results are discussed in next section.

4.2 Accident Tweet Detection Model

In this module, deep learning algorithms are used to train model for detection of accidents. High quality dataset which is built in the previous phase is used for the training of artificial neural networks. Data is almost evenly distributed between two classes i.e, Accident and Non Accident. For training and testing purpose, splitting is done as 80% is reserved for training purpose while 20% is reserved for testing purpose. As per our previous discussions in architecture, we have derived very important artifacts from SMPs like impact of accident, sentiment of a tweet, and emotion representational attributes. Emotions and sentiments representation is included for the first time for detection of accident tweets which not only improved the efficiency in terms of training time but also improved the accuracy of model.

Our sample neural network model is shown in Fig 3. Model has three input layers, First one is of word embedding; words are converted into vector space to represent each word in numerical value. Word embedding like Glove is used to transform each word. In the embedding layer, values of max length, dimension size and word count used is 150, 200 and 250 respectively. Second input layer is of impact; numeric values which represents traffic pay-off value which is derived from text of SMP using bag of words. Third input layer is also numeric value which expresses sentimental in the text of post. All the three input layers are concatenated using Concatenate layer followed by three dense layers and single output layer which classifies the value into binary and represents the SMP as Accident or Non Accident related.

Activation function of 'sigmoid' is applied on last layer to get binary classification. The Function formula is shown in equation 1.

$$S(z) = \frac{1}{1 + e^{-z}} \tag{1}$$

In the dense layer, we have applied Relu function as per in Eq.2 and Eq. 3

$$R(z) = max(0, z) \tag{2}$$

$$R(z) = \{0 \; z < 01 \; z > 0\} \tag{3}$$

Multiple input layers are concatenated using concatenation layers followed by multiple dense layers using functions of 'relu' and 'sigmoid' for the binary classification at the end as shown in Fig 3 .

Output value is a binary value, which classifies tweet as an accident or not accident. For the evaluation of our architecture, we simulated training's of multiple models including GRU, RNN, and LSTM. All training simulations were performed for 50 epochs and with batch size of 128. We achieved high score on our test data with an accuracy of 93.7%, 91.6% and 94.2% respectively. An accuracy comparison chart of all three models is shown in Fig 5. Chart displaying the F1 score, precision, and recall of all the three models is shown in Fig 6

The test results revealed that false positives were very less which proves the soundness of the architecture. False positives values are 77, 100, and 140 for

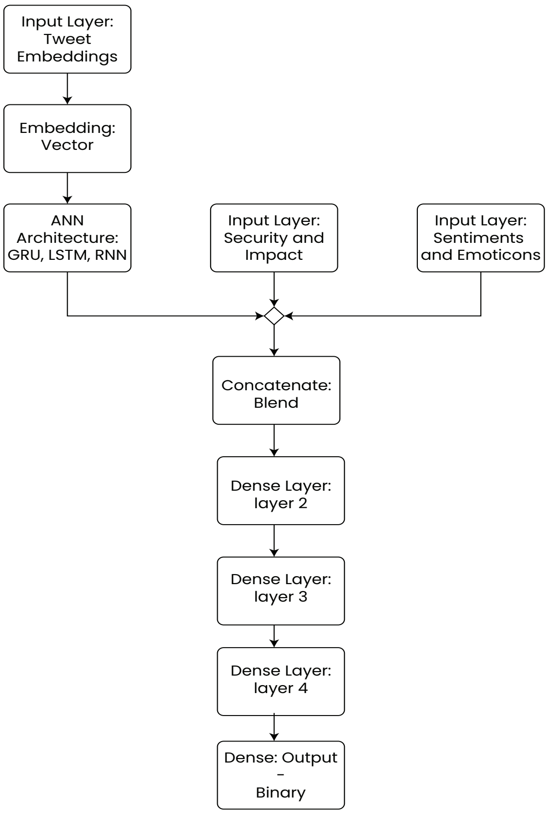

Fig. 3: Neural Network Model

GRU, LSTM, and RNN respectively. Confusion matrix showing True positives, True negative, False positive, and False negative has been created which are shown in Fig 4 for GRU, LSTM, and Simple RNN respectively.

In order to cross check our hypothesis and prove the importance of attributes like sentiment, severity impact, and emotion representative value, we trained additional deep learning model of LSTM, which did not have input layers for the attributes mentioned above , and contained only single input layer of word embedding. Same scenario of dense layers, number of epochs, and batch size is used to cross validate results on the same dataset. We found accuracy to be 83-87% with multiple simulations. Moreover, the false positive rate also got more than double, by increasing it to 339. All the experiments have been done with different algorithms like RNN, LSTM, and GRU. Average accuracy for all the models was approximately 85% which clearly shows 8% improvements in accuracy.

Above experiments clearly show that our new methodology is giving 8% higher accuracy as compared to using simple single input layer, while correlating word embedding with other useful feature, we extracted from

prediction outcome

actual value	**p**	**n**	**total**
p′	1474	104	P′
n′	77	1207	N′
total	P	N	

(a) GRU

prediction outcome

actual value	**p**	**n**	**total**
p′	1496	65	P′
n′	100	1201	N′
total	P	N	

(b) LSTM

prediction outcome

actual value	**p**	**n**	**total**
p′	1487	101	P′
n′	140	1134	N′
total	P	N	

(c) Simple RNN

Fig. 4: Confusion Matrices showing results for GRU, LSTM and Simple RNN

SMP text showed significant improvement in accuracy of deep learning models.

4.3 Accident Prediction Model

In this module, we focus on training of model which predicts accidents that can be used by different departments of transportation for controlling traffic and decision making. For such analysis, attributes are mentioned in Table 4. e.g., latitude, longitude, snowfall status, temperature, wind speed, wind direction, hour of day and day of week are very important to correlate.

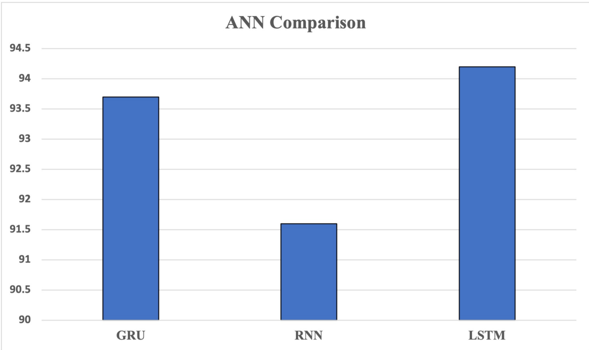

Fig. 5: ANN Accuracy Comparison Chart

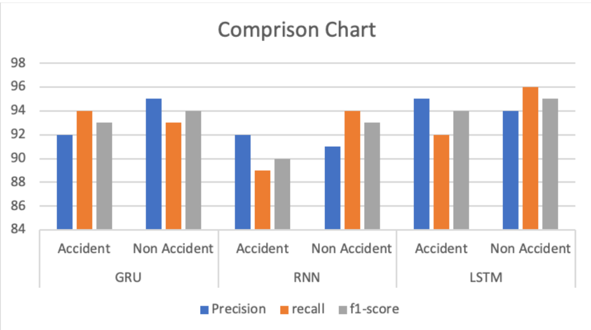

Fig. 6: ANN F1, Precision, Recall Comparison Chart

Table 4: Categorized Feature List

Category	**Attributes**
Location	Latitude, Longitude
Time	Hour of day, Month, Day of week
Weather	Snowfall, Max-Min Temperature, Wind speed, Precipitation, Direction of wind

Model is predicted for the impact of accident. Severeness of accident is predicted as output of model which is categorized into two levels. Above mentioned attributes are passed in input layers which are followed by four dense layers resulting in a binary value. Data set for prediction training and testing is little bit tweaked so that output results can be obtained. All tweets did not had all the value above mentioned so different techniques were adopted to fill the missing values. The values which were unable to be filled, such SMP's data was dropped. We used 'adam' optimizer for this training and 'relu' as activation function in all dense layers. Data is splitted in same ratio i.e., 80% and 20%. Training is completed with batch size of 32 and for 500 epochs. After multiple simulations best score of 75% accuracy is achieved. We got average recall value of 75% and f1-score of 81%.

5 Validation & Comparison

Detection of accidents from the analysis of Twitter social media posts is not new, many researchers have proposed different techniques to achieve better detection rates. However, very few researchers have made their dataset public. In order to cross validate our methodology, we have acquired their collections to apply our methodology on them. A dataset of Dabiri, Sina [7], is acquired which is compiled from Twitter using search query API in 2018. The labeled dataset consists of two types of classes. i.e., Traffic Incident and Non-Traffic Incident. Luckily, tweet text was also available with which enabled us to run our methodology on it.

For the sake of comparison, we first implemented their techniques by using the same deep learning models listed in the paper. We achieved F-score of 95.9% which is the same as mentioned by the authors. Then, we executed our mechanisms to calculate and extract features like sentiment score, emoji score, and impact factors. Such attributes contain important information about traffic incident and they play a very vital role in the decision. However, these scores and features were attached to the same dataset which we acquired and then the results were compiled again. We split the dataset of 50000 tweets by ratio of 80:20 in between training and testing sets. In multiple run scenarios, we achieved an increase in accuracy and F-score by 2%. For 100 epochs and batch size of 64, we got accuracy of 97.5%.

Precision for accident and not accident are 98.5% and 97.5% respectively. Confusion matrix of results is shown in Fig 7a. Comparison chart showing accuracy before and after including proposed metrics is shown in Fig 8

For purpose of extended evaluation and cross validation, we acquired another dataset which is compiled by C.Guckelsberger [21]. It contains more than 11000 SMP which are collected from Twitter. Posts have been labeled between traffic incident and non traffic incident which makes it 2 class dataset. Data had been compiled in 2015. The paper of this dataset describe different technique like semantic cross examination of keywords within different cities w.r.t reporting of incidents. However, we have executed our methodology on the provided data to cross check accuracy. Author of paper has used F-1 measure to examine his methodology, so, instead of accuracy we calculated F-1 measure to compare. We achieved 3% increase as avg score on F-1 measure. Results are shown in Table 5. Column with original marking shows results of author of paper while column without original tag represents our results. Besides this, accuracy is also computed for future references. On average 90% accuracy is achieved on this

prediction outcome

		p	n	total
actual value	p′	4953	142	P′
	n′	116	4932	N′
	total	P	N	

(a) 1st Dataset

prediction outcome

		p	n	total
actual value	p′	1929	107	P′
	n′	177	660	N′
	total	P	N	

(b) 2nd Dataset

Fig. 7: Confusion Matrix of comparisons with 1st and 2nd Dataset

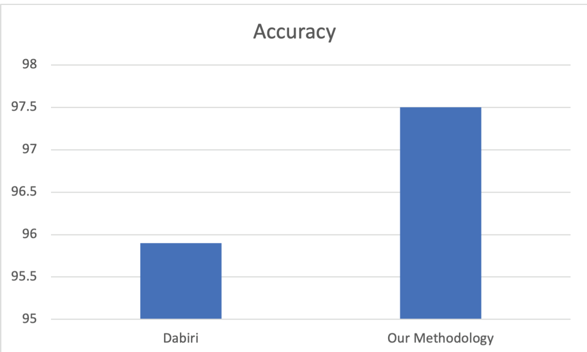

Fig. 8: Accuracy Comparison Chart

dataset. Another simulation is performed by combining data of all cities. Accuracy of 90% is achieved in that simulation with F-1 measure of 88%. For all the simulation dataset was divided in 80:20 ratio for train and test sets. Confusion matrix of the simulation is shown in Fig. 7b

We have analysed two datasets for the cross validation and evaluation of our methodology of detection phase, but unfortunately we were unable to find datasets for prediction phase. Prediction phase require different attributes which are missing in datasets. Missing attributes include dateOfTweet, Tags, and GeoLocations. These all attributes are further used to extract extended list of features which play important role in determining the reason of accident. e.g., Weather info using data and Geo location. In Conclusion, Our enhanced features aid in the detection decision of the deep learning model. This increase in accuracy truly represents the importance of sentiment score and emojis. This dataset was compiled in 2018, when the use of emojis was not that common as compared to nowadays. We are confident enough that if algorithm is executed on latest dataset, it will give more promising results.

Table 5: F-Measure Comparison Table

City	**F-1 Measure (Original)**	**F-1 Measure**	**Accuracy**
Boston2C	86%	91%	94%
Brisbane2C	86%	93%	94%
Chicago2C	91%	86%	91%
Dublin2C	93%	88%	91%
London2C	81%	88%	90%
Memphis2C	80%	85%	86%
NYC2C	85%	86%	89%
SanFrancisco	86%	85%	90%
Seattle	74%	86%	87%
Sydney2C	82%	85%	88%
Avg. %	84%	87%	90%

6 CONCLUSION, LIMITATIONS AND FUTURE WORK

In this paper, we make use of deep learning methods to not only detect traffic accidents from social media posts but also perform prediction of traffic accidents. Our major steps were a collection of SMP, classification, feature selection, model training, and results validation. To the best of our knowledge, this is first time emotions and sentiments reflected in texts are being used to aid the classifier, the results clearly show that it increases the accuracy of the classifier. In conclusion, using social media posts for the study of traffic accident patterns bring a vast variety of possibilities for transportation research especially for arterial roads where sensor-based detection of accidents is difficult and major cases remain undetected and unreported. We applied various deep learning models and compared the results between them, we achieved best accuracy of 94.2% for detection of accident related tweets and 75% for prediction of accidents.

This study was constrained by limitations that were encountered during this research. First, all SMPs do not include any kind of location information, we resolved this issue by using the location of a tweet, in case geo-tag with Twitter post is not available then that item is dropped. Second, there is non-availability of historical weather information corresponding to that specific Latitude and Longitude; in order to resolve this issue, we used weather information like snow, humidity, and clouds of near-by surrounding areas of city within 10km radius range. Third, frequent use of local names for roads instead of official names makes it difficult to find location or area of accident. Moreover, there was non availability of data of all fields, for example, if location, weather and hour of day is available, one of attribute like temperature or wind speed might be unavailable which makes it very difficult to run simulation on all the available dataset. In this study, simulations are done on the items whose all attributes are available.

Road conditions and traffic on roads (No. of cars on road in 1 km of area) are not in scope of this study because of non availability of data, however we believe that if we include this data for training of model, accuracy can be increased by a remarkable value. Detection of fake news and filtering of irrelevant inputs which affects the detection and prediction accuracy can also help in improving the technique. Multiple platform data can be fused together to merge information from different sources, like, multiple people are reporting same incident, one person may post more information as compared to the other, hence they can be merged together for better and enhanced featured data set.

Funding

The research is supported by project under University Faculty Start-Up Research Grant Program (SRGP) by the HEC, Grant No: HEC(FD)/2019 dt 12 Jun 2020. The research is also supported by Taif University Researchers Supporting Project number (TURSP-2020/126), Taif University, Taif, Saudi Arabia.

Consent Statement

Not Applicable

Conflict of Interests

Authors have no conflict of interest to declare

Availability of Data and Material

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Author Contribution Statement

Saddaf Rubab conceived of the presented idea and supervised the research. Anique Azhar has developed the theory and performed the computations. Yawar Abbas Bangash and Malik M Khan verified the analytical methods. Mohammad Dahman Alshehri assisted with the analysis of the results. Ali Kashif Bashir encouraged Anique Azhar to investigate the proposed method on the variants of datasets and analysed the findings of this work. Fizza Elahi interpreted the results and drafted the manuscript. All authors discussed the results and contributed to the final manuscript.

References

1. B. Alkouz and Z. Al Aghbari, "SNSJam: Road traffic analysis and prediction by fusing data from multiple social networks", Information Processing & Management, vol. 57, no. 1, p. 102139, 2020. Available: 10.1016/j.ipm.2019.102139
2. S. Alhumoud, "Twitter Analysis for Intelligent Transportation", The Computer Journal, vol. 62, no. 11, pp. 1547-1556, 2018. Available: 10.1093/comjnl/bxy129
3. M. Abdel-Aty, J. Keller and P. Brady, "Analysis of Types of Crashes at Signalized Intersections by Using Complete Crash Data and Tree-Based Regression", Transportation Research Record: Journal of the Transportation Research Board, vol. 1908, no. 1, pp. 37-45, 2005. Available: 10.1177/0361198105190800105
4. P. Chen, F. Chen and Z. Qian, "Road Traffic Congestion Monitoring in Social Media with Hinge-Loss Markov Random Fields", 2014 IEEE International Conference on Data Mining, 2014. Available: 10.1109/icdm.2014.139
5. Y. Chen, Y. Lv, X. Wang, L. Li and F. Wang, "Detecting Traffic Information From Social Media Texts With Deep Learning Approaches", IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 8, pp. 3049-3058, 2019. Available: 10.1109/tits.2018.2871269
6. E. D'Andrea, P. Ducange, B. Lazzerini and F. Marcelloni, "Real-Time Detection of Traffic From Twitter Stream Analysis", IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 4, pp. 2269-2283, 2015. Available: 10.1109/tits.2015.2404431
7. S. Dabiri (2018), "Tweets with traffic-related labels for developing a Twitter-based traffic information system.", Mendeley Data, V1, doi: 10.17632/c3xvj5snvv.1

8. N. Emadi, S. Abbar, J. Borge-Holthoefer, F. Guzman and F. Sebastiani, "QT2S: A System for Monitoring Road Traffic Via Fine Grounding of Tweets", Aaai.org, 2020.
9. "Fatality Facts 2018: Yearly snapshot", IIHS-HLDI crash testing and highway safety, 2020. .
10. Y. Gu, Z. Qian and F. Chen, "From Twitter to detector: Real-time traffic incident detection using social media data", Transportation Research Part C: Emerging Technologies, vol. 67, pp. 321-342, 2016. Available: 10.1016/j.trc.2016.02.011
11. P. Giridhar et al., "ClariSense+: An enhanced traffic anomaly explanation service using social network feeds", Pervasive and Mobile Computing, vol. 33, pp. 140-155, 2016. Available: 10.1016/j.pmcj.2016.03.005
12. "Global status report on road safety 2018", World Health Organization, 2020.
13. M. Ibrahim, O. Abdillah, A. Wicaksono and M. Adriani, "Buzzer Detection and Sentiment Analysis for Predicting Presidential Election Results in a Twitter Nation", 2015 IEEE International Conference on Data Mining Workshop (ICDMW), 2015. Available: 10.1109/icdmw.2015.113
14. H. Nguyen, W. Liu, P. Rivera and F. Chen, "TrafficWatch: Real-Time Traffic Incident Detection and Monitoring Using Social Media", Advances in Knowledge Discovery and Data Mining, pp. 540-551, 2016. Available: 10.1007/978-3-319-31753-3_43 [Accessed 4 July 2020].
15. H. Nguyen, W. Liu, P. Rivera and F. Chen, "TrafficWatch: Real-Time Traffic Incident Detection and Monitoring Using Social Media", Advances in Knowledge Discovery and Data Mining, pp. 540-551, 2016. Available: 10.1007/978-3-319-31753-3_43
16. K. Pandhare and M. Shah, "Real time road traffic event detection using Twitter and spark", 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), 2017. Available: 10.1109/icicct.2017.7975237
17. A. Parsa, A. Movahedi, H. Taghipour, S. Derrible and A. Mohammadian, "Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis", Accident Analysis & Prevention, vol. 136, p. 105405, 2020. Available: 10.1016/j.aap.2019.105405
18. A. Pereira, T. Silva, F. Silva and A. Loureiro, "Traffic Event Detection Using Online Social Networks", 2017 13th International Conference on Distributed Computing in Sensor Systems (DCOSS), 2017. Available: 10.1109/dcoss.2017.36 [Accessed 4 July 2020].
19. A. Salas, P. Georgakis and Y. Petalas, "Incident detection using data from social media", 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), 2017. Available: 10.1109/itsc.2017.8317967
20. R. Sinnott and S. Yin, "Accident Black Spot Identification and Verification through Social Media", 2015 IEEE International Conference on Data Science and Data Intensive Systems, 2015. Available: 10.1109/dsdis.2015.34
21. A. Schulz, C. Guckelsberger and F. Janssen, "Semantic Abstraction for generalization of tweet classification: An evaluation of incident-related tweets", Semantic Web, vol. 8, no. 3, pp. 353-372, 2016. Available: 10.3233/sw-150188
22. A. Salas, P. Georgakis, C. Nwagboso, A. Ammari and I. Petalas, "Traffic event detection framework using social media", 2017 IEEE International Conference on Smart Grid and Smart Cities (ICSGSC), 2017. Available: 10.1109/icsgsc.2017.8038595 [Accessed 4 July 2020].
23. "Python Tweepy", PyPI, 2021. [Online]. Available: https://pypi.org/project/tweepy/.
24. Z. Zhang, Q. He, J. Gao and M. Ni, "A deep learning approach for detecting traffic accidents from social media data", Transportation Research Part C: Emerging Technologies, vol. 86, pp. 580-596, 2018. Available: 10.1016/j.trc.2017.11.027
25. Essien, A., Petrounias, I., Sampaio, P. et al. A deep-learning model for urban traffic flow prediction with traffic events mined from twitter. World Wide Web (2020). Available: https://doi.org/10.1007/s11280-020-00800-3
26. Alotaibi, S.; Mehmood, R.; Katib, I. (2019) Sentiment Analysis of Arabic Tweets in Smart Cities: A Review of Saudi Dialect. In Proceedings of the 2019 Fourth International Conference on Fog and Mobile Edge Computing (FMEC), Rome, Italy, June 10–13, pp. 330–335
27. K. Fu, C. Lu, R. Nune and J. Tao, Steds: Social Media Based Transportation Event Detection with Text Summarization, 2015 IEEE 18th International Conference on Intelligent Transportation Systems, 2015.
28. M. Shafiq, X. Yu, A. Bashir, H. Chaudhry and D. Wang, "A machine learning approach for feature selection traffic classification using security analysis", The Journal of Supercomputing, vol. 74, no. 10, pp. 4867-4892, 2018. Available: 10.1007/s11227-018-2263-3
29. M. Shafiq, Z. Tian, A. Bashir, X. Du and M. Guizani, "CorrAUC: A Malicious Bot-IoT Traffic Detection Method in IoT Network Using Machine-

Learning Techniques", IEEE Internet of Things Journal, vol. 8, no. 5, pp. 3242-3254, 2021. Available: 10.1109/jiot.2020.3002255

30. M. Shafiq, Z. Tian, A. Bashir, X. Du and M. Guizani, "IoT malicious traffic identification using wrapper-based feature selection mechanisms", Computers & Security, vol. 94, p. 101863, 2020. Available: 10.1016/j.cose.2020.101863