International Journal of Digital Earth

ISSN: 1753-8947 (Print) 1753-8955 (Online) Journal homepage: www.tandfonline.com/journals/tjde20

A systematic review and comparative analysis of deep learning models for Twitter/X-based traffic event detection

Danya Qutaishat & Songnian Li

INTERNATIONAL JOURNAL OF DIGITAL EARTH
2026, VOL. 19, NO. 1, 2604977
https://doi.org/10.1080/17538947.2025.2604977

REVIEW ARTICLE

A systematic review and comparative analysis of deep learning models for Twitter/X-based traffic event detection

Danya Qutaishat and Songnian Li

Department of Civil Engineering, Toronto Metropolitan University, Toronto, Canada

ABSTRACT

Traffic anomalies caused by accidents, sports events, and lane closures are spatiotemporal events that reduce free-flow speed, increase vehicular queues, and impair human mobility. Early detection may provide better route planning before traffic gets worse. Recent and ongoing research, as well as a review of transportation literature, have revealed three essential topics: big data, data mining and representation, and Deep Learning (DL). Furthermore, traffic studies have adopted DL to extract hidden features that efficiently infer human activities and interactions and detect the underlying relationships to generate useful fine-grained information. This paper reviews current research that adopts state-of-the-art DL in detecting traffic events from big data, specifically Twitter/X data. In addition, it investigates the detailed pipeline for developing a DL-based model using data from Twitter/X for traffic event detection (TED). The review is a timely addition that clarifies the roadmap of detecting traffic events from big social media data, which benefits transportation and DL community researchers.

ARTICLE HISTORY
Received 24 February 2025
Accepted 12 December 2025

KEYWORDS
Deep learning; feature learning; word embedding; model selection; Twitter/X

1 Introduction

Traffic Event Detection (TED) holds significant promise in assisting road users. It helps choose optimal paths, reduce travel time, mitigate traffic congestion, minimise fuel consumption, and reduce environmental pollution (Nejjari, Benhlima, and Bah 2016; Kim et al. 2023; Lee et al. 2023; Gannina Kumar et al. 2024; Qutaishat and Li 2025a). Moreover, it help reduce traffic accidents that lead to injuries, fatalities, and property damage, which result in substantial social and economic costs (Es Swidi et al. 2023; Gannina Kumar et al. 2024). Additionally, generating multiple scenarios equips traffic management departments with a tool to suggest timely and practical plans to improve traffic conditions. This enables prompt responses by emergency services and supports traffic rerouting to enhance traffic management (Nejjari, Benhlima, and Bah 2016).

Transportation agencies' most significant challenge is acquiring real-time, large-scale, and up-to-date observational data (Anda, Erath, and Fourie 2017; Liu et al. 2020; Saeedi et al. 2020). Traditional data sources consist of structured data collected using physical condition monitoring devices deployed in the field or sensing devices installed in moving vehicles (Hall, Shi, and Atala 1993; Sethi et al. 1995; Samant and Adeli 2000). Although incorporating conventional data for detecting traffic events provides accurate information regarding their location and time, several challenges have been identified. First, studies are built on the assumption of data reliability; however, incident detection has proven to be difficult due to detector failure, sensor malfunctions, and communication errors in large-scale areas. Second, external factors impact traffic operations, reducing the effectiveness of traffic metrics in detecting traffic incidents. Third, physical sensors or detectors require regular maintenance and cover small-scale areas such as intersections or short road segments. This limitation restricts the capture of traffic patterns and non-recurring events across entire urban areas (Münz, Sa, and Georg 2007; Zhang et al. 2018).

CONTACT Songnian Li snli@torontomu.ca Department of Civil Engineering, Toronto Metropolitan University, Toronto, Canada
Supplemental data for this article can be accessed online at https://doi.org/10.1080/17538947.2025.2604977.
This article has been corrected with minor changes. These changes do not impact the academic content of the article.

Nowadays, social media data makes a significant contribution to traffic studies, including human activity patterns or travel behaviours (Hasan and Ukkusuri 2014; Hasnat and Hasan 2018), traffic flow forecasting (Lin et al. 2015; Ni, He, and Gao 2016; Cottrill et al. 2017), transportation management and planning (Cottrill et al. 2017), travel mode extraction (Maghrebi, Abbasi, and Waller 2016), and destination choice modelling (Huang, Gallegos, and Lerman 2017; Molloy and Moeckel 2017; Hasnat and Hasan 2018; Hasnat et al. 2019). However, concerns are rising due to the lack of tools and techniques to unlock the power of data and extract valuable knowledge from this kind of massive, complex, and diverse big data.

Traditional machine learning (ML) models have been recognised as cornerstones of Twitter/X-based TED. They laid the foundation for utilising structured textual data and feature engineering to classify and predict traffic-related events. The most commonly investigated models are Support Vector Machine (SVM) (Noori and Mehra 2020; Afyouni, Aghbari, and Razack 2022; Dinesh, Kuhaneswaran, and Ravikumar 2023), Random Forest (Alomari, Mehmood, and Katib 2019; Jiang and Deng 2020; ElSahly and Abdelfatah 2023), and Naive Bayes (Alomari, Mehmood, and Katib 2019; Nirbhaya and Suadaa 2023). However, several challenges arise from issues related to data quality, model performance, and computational efficiency. Selection bias arises from non-representative Twitter/X data, leading to skewed outcomes and reduced generalisability (Liu et al. 2024). Class imbalance negatively impacts detection accuracy due to the higher number of non-traffic tweets compared to relevant ones (Liu et al. 2024). Linguistic challenges, including informal language, expressions, and negative phrases, contribute to frequent detection errors (Dhiman and Toshniwal 2020; Liu et al. 2024). Additionally, short tweet lengths limit contextual understanding, reducing model effectiveness (Dhiman and Toshniwal 2020). High computational costs are another issue, as processing large-scale Twitter/X data with complex algorithms demands significant resources (Dhiman and Toshniwal 2020). Lastly, feature extraction issues arise from automated social accounts mimicking human behaviour, making traffic event detection more difficult (Sethurajan and K 2023).

In the context of TED using Tweets, Deep Learning (DL) models have overcome ML in multiple aspects. The first aspect is the ability of DL models to excel in automatically extracting relevant features from unstructured text data, eliminating the need for manual feature engineering that is labour-intensive and prone to human error (Hussain 2024; Qutaishat and Li 2025a). Second, techniques like transformers enable these models to understand the context of informal language, expressions, and abbreviations commonly found in tweets (Neruda and Winarko 2021). They also process sequential and multimodal data, such as text, time, and location, using architectures like Recurrent Neural Networks (RNN) and multi-input neural networks, which provide a more comprehensive analysis (Alifi and Supangkat 2018; Zhang et al. 2018). Third, DL scales efficiently and can handle large, real-time datasets while adapting real-time patterns and language changes on social media. Finally, its ability to detect complex, non-linear patterns leads to significantly higher accuracy in identifying traffic events, making it a superior choice for real-time, dynamic environments (Kisters and Bauer 2023 ; Li, Dou, and Zhou 2023; Nirbhaya and Suadaa 2023; Qutaishat and Li 2025a, Suat-Rojas, Gutierrez-Osorio, and Pedraza 2022; Yang 2022; Qutaishat and Li 2025b).

This paper reviews the literature on DL-based TED using social media data, focusing on Twitter/X. Twitter/X has emerged as a valuable platform for supporting the detection and modelling of traffic events and deserves close attention in this context. Several reviews have explored the topic of event detection using social media data, with some studies focusing on traditional machine learning techniques and statistical methods for analysing Twitter/X data to identify traffic-related events (Garg and Kumar 2016; Xu, Li, and Wen 2018; Liu et al. 2024). These studies lack emphasis on DL approaches, which have proven to be more effective in handling unstructured text, capturing complex relationships, and integrating multimodal data sources. Other studies have focused on investigating event detection on Twitter/X, without specifically addressing traffic events or detailing the techniques used, such as (Saeed et al. 2019; Atefeh and Khreich 2013). There has been a lack of systematic reviews that provide comprehensive insights into the research framework for traffic event detection using Twitter/X data, specifically in the context of DL techniques.

The main contributions of this paper are as follows:

1. A comprehensive systematic review that consolidates and critically analyses published studies related to Twitter/X-based TED using DL techniques.
2. A detailed description of the general Twitter/X-based TED workflow, including preprocessing steps and model development.

3. An analysis of feature learning and models, evaluating their strengths and weaknesses, and the influence of parameters on DL performance.
4. Identification of the challenges and future research directions, offering essential guidance in transportation engineering and related domains.

2 Review methodology

This research followed the PRISMA 2020 guidelines for systematic reviews (Page et al. 2021; Page and McKenzie, 2020). Figure 1 presents the PRISMA flow diagram, which shows the number of studies identified, screened, excluded, and ultimately included in the review.

2.1 Eligibility criteria

Studies were assessed based on the following inclusion and exclusion criteria:

2.1.1 Inclusion criteria

Studies were included if they:

- Applied DL models such as CNNs, RNNs, or LSTMs for TED.
- Used Twitter/X as a standalone source or integrated it with other data.
- Were published in English, from 2010 onward, reflecting DL advancements from that time.
- Provided detailed methodologies, including DL architecture, preprocessing, and evaluation metrics (e.g. accuracy, F1-score).
- Were published in peer-reviewed journals or reputable conferences in relevant fields, such as transportation engineering or computer science.

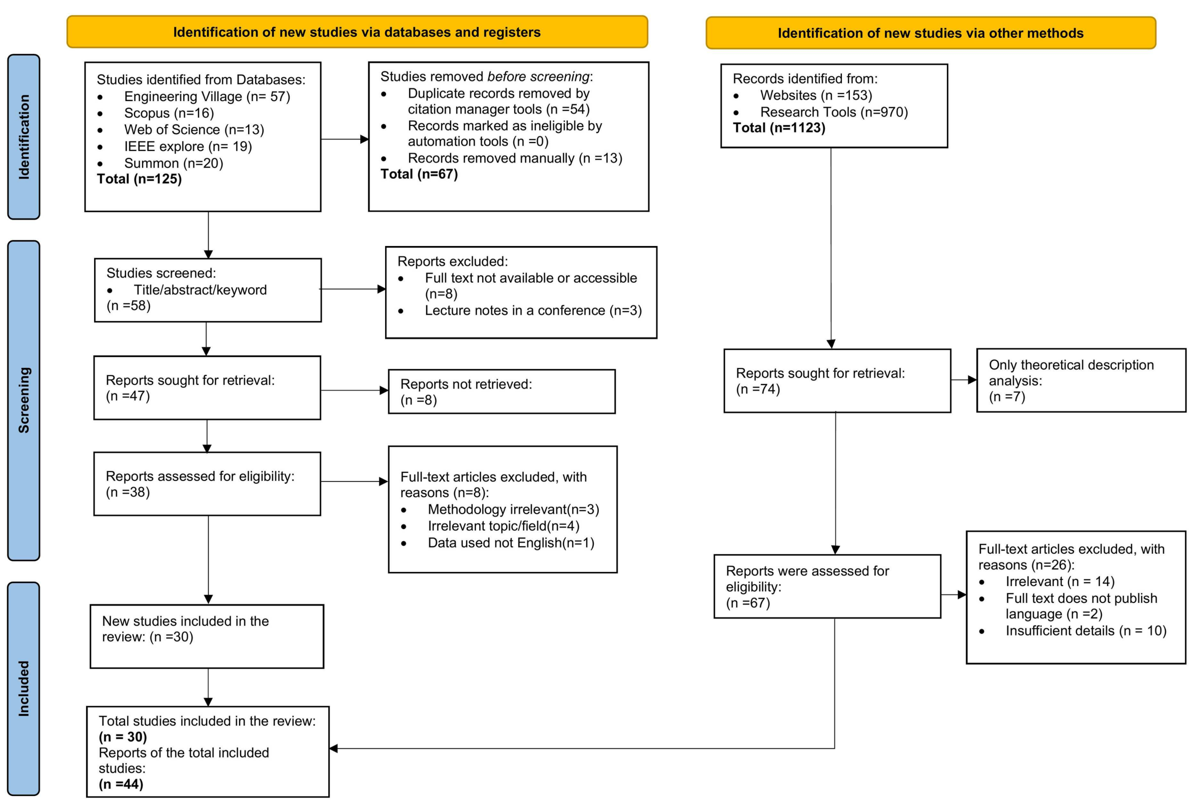


Figure 1. The PRISMA flow diagram of the article selection process. Note: Although the main flow shows 30 studies from database searches, additional studies were identified via citation/web methods, bringing the total to 44.

- Cover both recurring (e.g. rush hour) and non-recurring (e.g. accidents) traffic events.

2.1.2 Exclusion criteria

Studies were excluded if they:

- Used observational or non-social media data for TED.
- Depended on traditional ML instead of DL.
- Focused on events unrelated to traffic (e.g. natural disasters, crime, health)
- Lacked sufficient methodological detail or empirical evaluation.
- Were theoretical, non-English, or published before 2010.

2.2 Information sources and search strategy

For the scope of this review, the search was narrowed down to journal articles and conference proceedings. Five scholarly databases were searched: Engineering Village, Scopus, Web of Science, IEEE, and Summon 2.0. Citation searches help trace prior, derivative, or related works. ResearchRabbit, a citation-based literature mapping tool, facilitated this process in conjunction with Google Scholar. The timeframe, set from 2010 onward, reflects the rise of DL, driven by the revolution in computational power and the availability of large datasets. Earlier studies predominantly used ML or statistical techniques.

Boolean Keyword combinations used included:

(traffic event detection AND deep learning) AND (social media OR Twitter/X OR crowdsourcing).

Initial database search yielded:

- Engineering Village: 57
- IEEE Xplore: 19
- Scopus: 16
- Web of Science: 13
- Summon 2.0: 20

An additional **1,123** studies were identified via citation and web-based searching.

2.3 Screening and selection process

A double-screening strategy was applied (Nama et al. 2019) in which two reviewers independently screened records. First, EndNote X9 was used to import references, remove duplicates, and sort studies by publication year, title, and author to structure the screening sequence. Each reviewer then independently assessed titles and abstracts; a random subset was cross-checked to ensure consistency. Rayyan facilitated collaborative full-text review and inclusion/exclusion decisions.

To ensure inter-rater reliability, a random subset of studies was jointly reviewed, and reviewer agreement was monitored. Although Cohen's kappa coefficient was not formally calculated, consistency was verified through regular calibration and consensus discussions. Any discrepancies between reviewers were resolved through consensus discussion, and a third reviewer was not needed.

2.4 Final inclusion and data extraction

After removing duplicates and applying inclusion/exclusion criteria, a total of 44 studies were included, with 30 from database searches and 14 from other methods, as shown in Figure 1.

We extracted the following data from each included study:

- Title, year, publication type, and author list
- Country or region of focus

- DL models used
- Data representation methods
- Performance metrics (e.g. accuracy, precision, F1-score)

3 Results and discussion

3.1 Characteristics of reviewed studies

Figure 2 presents a word cloud of key terms. 'Deep Learning', 'Social Networking', and 'Twitter/X', which were clearly stated in the article search, appear with high prominence in the word cloud. The CNN and LSTM models are the most frequently used, alongside the terms related to semantic feature representation, such as word embeddings and information management, which are essential for Twitter/X data mining.

Studies focusing on keyword generation emphasise the term 'Traffic' in various contexts, such as 'Traffic Congestion', 'Traffic Incident', and 'Traffic information', indicating a primary focus on unexpected or non-recurring traffic events. For model evaluation, terms like 'accuracy' and 'Accuracy assessment' were commonly used, suggesting that the studies emphasise textual feature representation, model selection, and framework construction over evaluation measures. Given the novelty of social media data analysis in traffic research, diverse keywords like 'Detection', 'Event Detection', and 'TED' were employed.

Figure 3 is a world map showing the geographic distribution of study sites. India leads in publications, followed by the UK and the US, which together account for approximately 45% of the reviewed studies. Indonesia, China, and Pakistan contributed 25%. While India-centric data (8/44 studies) demonstrates feasibility in local contexts, reliance on region-specific corpora may restrict generalisability. Future research should expand validation efforts in underrepresented regions such as Africa and South America to ensure global applicability.

Figure 2. Word cloud of the prominent keywords in 44 studies.

Country	Number. Of Publications
India	8
United Kingdom	5
United States	5
Indonesia	4
China	3
Pakistan	3
Greece	2
Japan	2
Saudi Arabia	2
Belgium	1
Egypt	1
Iraq	1
Portugal	1
Qatar	1
Singapore	1
South Korea	1
Switzerland	1
Thailand	1
United Arab Emirates	1
	44

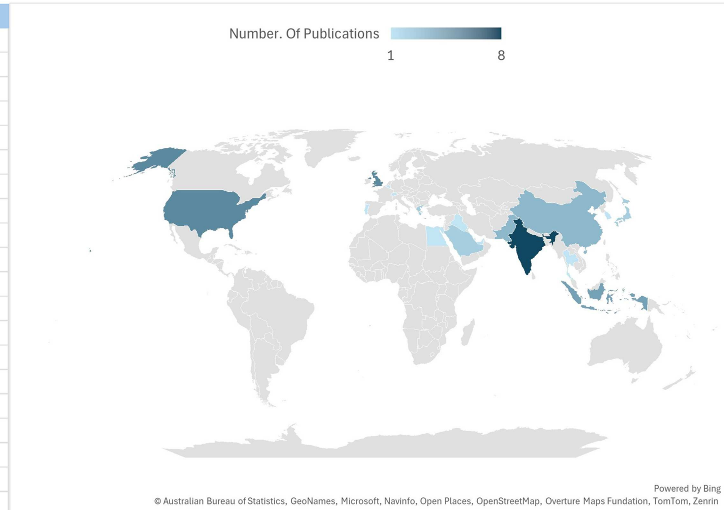


Figure 3. Map of the world based on No. of publications related to Twitter/X TED using DL.

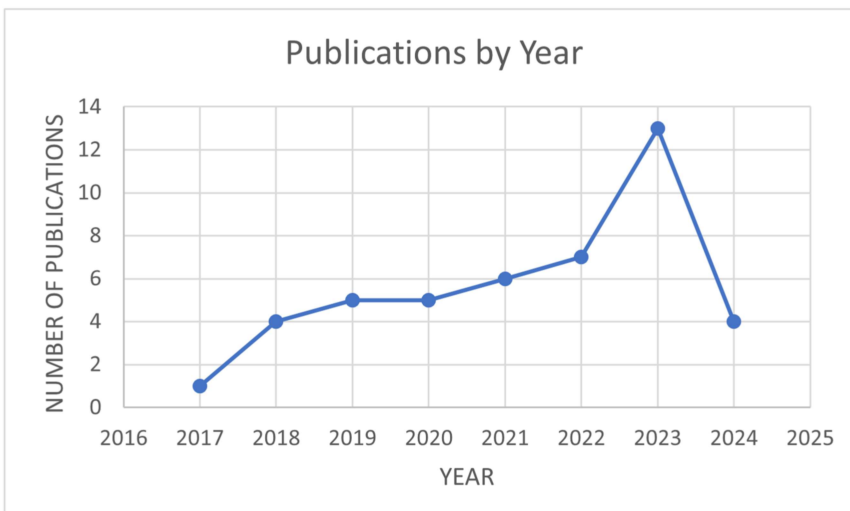


Figure 4. The trend of published studies on Twitter/X TED using DL (2017–2024) peaked in 2022–2023.

Based on Figure 4, 60.98% of studies were published between 2021 and 2023, with publications doubling in 2023, reflecting a surge in research interest during that period.

The dominant subject areas were computer science (40.4%), engineering (25.8%), and mathematics (12.4%). India's dominance is attributed to its large social media user base, extensive volume of Twitter/X data, and robust AI/ML capabilities, which are supported by collaborations among academia, industry, and government. The UK and US benefit from top-tier research institutions, advanced infrastructure, and robust funding for AI, ML, and DL studies in transportation research (Sahni and Raja 2018).

3.2 Deep-learning Twitter/X Data analysis for traffic events detection preprocessing workflow

A detailed workflow for deep-learning-based Twitter/X data analysis in TED, as shown in Figure 5, is summarised from the reviewed studies. However, variations exist in the techniques and algorithms as researchers explore different models, methods, and architectures. At its core is the data crawling stage, where tweets are collected via REST or Streaming APIs using filters like keywords and geolocation. Data

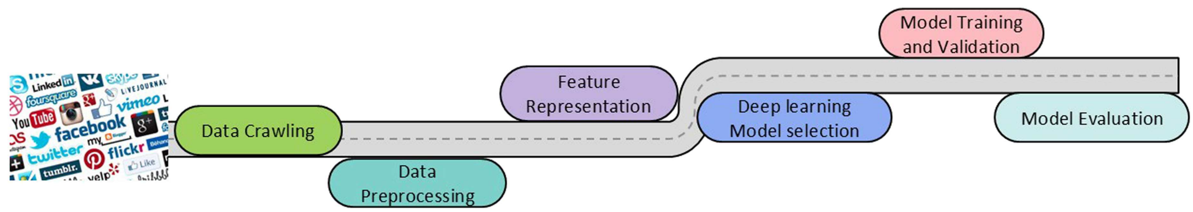


Figure 5. Generalised workflow for DL-based social media TED. It includes data crawling from social media platforms (e.g. Twitter/X, Facebook), preprocessing, feature representation, DL model selection, training and validation, and final model evaluation for accurate event detection.

quality is enhanced during preprocessing, while feature representation prepares the data for the DL model. A suitable model is selected, followed by training and evaluation to improve and assess performance. This framework is adaptable to various social media platforms, including Facebook, Instagram, Twitter/X, and Sina Weibo.

3.2.1 Traffic tweets crawling on twitter/X

Based on the included studies, traffic-related data from Twitter/X was historically collected using two primary methods: REST APIs (programmatic Representational State Transfer) and Streaming APIs. These APIs allowed researchers to define a centroid (latitude, longitude), a radius, and a set of keywords using operators, including AND, OR, and EXCLUDE, in their format (Ali et al. 2017; Ali et al. 2019).

Each API serves distinct purposes and offers unique advantages for researchers and developers. The REST API was typically used for historical data collection, allowing users to query based on specific keywords, locations (centroid + radius), and timeframes. It supported up to 3,200 tweets per request, with a 15-minute rate limit of 350 requests and was suited for historical analysis or offline model training (Gu, Qian, and Chen 2016; Xu, Li, and Wen 2018; Ali et al. 2019). In contrast, the Streaming API offers real-time tweet collection as they are posted, making it ideal for live TED and traffic monitoring (Doguc and Ahmet 2023). However, it only returned a limited sample of tweets and lacked the flexibility of advanced query customisation.

As of 2024, access to Twitter/X's APIs has undergone a significant transformation. Following corporate ownership changes and a strategic shift toward monetisation, Twitter/X now severely restricts free-tier access. Real-time streaming and full historical search functionalities are available only through premium or enterprise-level subscriptions. The cost of full access, which may reach up to $42,000/month, has rendered large-scale data collection impractical for many academic researchers (Murtfeldt et al. 2024).

This change has introduced a new challenge to reproducibility, limiting the ability to replicate earlier studies. Many previously effective crawling methods (e.g. Tweepy, REST API scripts, TWINT) have become unreliable or non-functional under the updated API policies. In response, some researchers have shifted to alternative methods, such as web scraping; however, these approaches raise both ethical and technical concerns (Poudel and Weninger 2024).

In light of restricted Twitter/X API access as of 2024, researchers are encouraged to utilise existing publicly available Twitter/X datasets to support reproducible research. For instance, the CrisisLexT6 dataset includes annotated tweets from natural disasters, some of which are transportation-related (Imran et al. 2015).

The T4SA dataset contains over 4.5 million sentiment-labelled tweets and is accessible via GitHub [https://github.com/codiceSpaghetti/T4SA-2.0]. Table 1 summarises widely available datasets, their domains, access types, and representative studies in which they have been used.

Table 2 provides a comparative overview of past TED studies, highlighting the data crawling methods used and their viability under present-day Twitter/X API policies. It is essential to note that while some researchers have collected upwards of 1 million tweets (e.g. Alomari, Mehmood, and Katib 2019), such volumes are no longer realistically attainable without institutional resources or paid access. Jonnalagadda and Hashemi (2021) collected approximately 10,000 tweets, of which 5,000 were used for developing the deep learning model. Dabiri and Heaslip (2019) retrieved around 50,000 tweets, with 17,437 manually

Table 1. Sample dataset comparison table.

Dataset name	Domain	Language	Access type	Used by
CrisisLexT6	Crisis/Traffic	English	Public via website	(Olteanu, Vieweg, and Castillo 2015; Alam, Ofli, and Imran 2018)
Twevent	Event detection	English	GitHub	(Li et al. 2012)
QCRI datasets	Crisis/Traffic	Multilingual	GitHub	(Alam, Ofli, and Imran 2018)
GeoCOV19	Mobility	English	Public via portal	(Qazi, Imran, and Ofli 2020)
Custom (Dabiri)	Traffic-specific	English	GitHub	(Dabiri and Heaslip 2019)

Table 2. Summary of studies highlighting social media platforms, crawling techniques, tweet collection volume, and data representation methods.

Study	Platform	Data crawling method	No. of Tweets	Feature learning method	Notes on current feasibility
(Ali et al. 2021)	Twitter/X Facebook	standard APIs	60000 tweets 5000 Facebook posts	Word2vec FastText	Standard API access is deprecated or highly limited on Twitter/X as of 2024.
(Dabiri and Heaslip 2019)	Twitter/X	n/a	50,000 tweets	Word2vec FastText	Method unspecified. Assumed pre-2022 feasibility
(Alomari, Mehmood, and Katib 2019)		REST API	1 million	TF-IDF	Large-scale tweet collection using REST API no longer viable (requires enterprise access).
(Ali et al. 2019)	Twitter/X	REST and Streaming APIs	30,000 tweets	String2word Glove2vec Lexicon Features Doc2vec	Streaming API deprecated under v2. Results not reproducible with free-tier access.
(Chen et al. 2018)	Sina Weibo	Sina Weibo crawler	11,000	Continuous Bag of Words (CBOW)	Non-Twitter/X platform. Method not affected.
(Lu et al. 2018)	News articles and Weibo posts,	Network of social sensors	1.15 million texts	Word2Vec CBOW	Non-Twitter/X platform.
(Zhang et al. 2018)	Twitter/X	Streaming API	3 million tweets	A systematic feature selection process	High-volume collection no longer feasible under current Twitter/X API limits.
(Fatichah et al. 2020)	Twitter/X and Images	Twitter/X API	10000 tweets 1000 images	Word Embedding	API usage requires elevated or paid access tiers.
(Jonnalagadda and Hashemi 2021)	Twitter/X	n/a	10,000 tweets	Word2vec	Method unspecified. Presumed historical feasibility.
(Ambastha and Desarkar 2020)	Twitter/X	Streaming API	1887 tweets	TF-IDF Word2Vec	Streaming API deprecated.
(Puangnak and Rachsiriwatcharabul 2022)	Twitter/X	n/a.	3363 tweets	Word Embedding Word Indexing	No crawling method provided.
(Neruda and Winarko 2021)	Twitter/X	Tweepy TWINT	6319 tweets	Bidirectional Encoder Representations from Transformers (BERT)	TWINT is no longer functional due to X's API and frontend obfuscation changes.
(Almassar and Girsang 2022)	Twitter/X	Twitter/X API	4,087 tweets	Word2vec FastText	Tweet API access is limited under new pricing tiers.

identified as traffic-related tweets. Alomari, Mehmood, and Katib (2019) gathered over 1 million tweets, but only an estimated 5,000 were categorised as relevant to traffic events; the remainder reflected broader social or environmental contexts. Neruda and Winarko (2021) attempted to supplement data using TWINT, which is now non-operational due to changes in X's frontend structure and API restrictions.

The variability in Twitter/X data affects model development in multiple ways. While large datasets enhance statistical strength, they often come with noisy labels and higher preprocessing demands (Tsou, Zhang, and Jung 2017; Effrosynidis, Sylaios, and Arampatzis 2024). In contrast, smaller datasets are cleaner but less generalisable and more prone to overfitting, especially when class imbalance exists, as seen in cases where there are fewer tweets related to traffic incidents (Liu et al. 2024). Additionally, geotagged Tweets are rare (found in only about 1–2% of tweets), limiting spatial analysis unless complemented by inferred or external location data (Tsou, Zhang, and Jung 2017).

In summary, while Twitter/X data remains valuable for TED, practical access is now significantly constrained. Researchers must adapt by utilising limited-access APIs, employing ethical web scraping, or leveraging multimodal datasets to compensate for the limitations imposed by the post-API era (Blakey 2024; Poudel and Weninger 2024).

3.2.2 Twitter/X data preprocessing

Preprocessing of raw Twitter/X data is a critical step in TED, as it addresses the platform's non-standard language and reduces ambiguity, abbreviations, and uncertainty (Ramadhani and Goo 2017; Ali et al. 2019; Karthik et al. 2023; Rezaeinia, Ghodsi, and Rouhollah 2017; Safitri et al. 2024). A key challenge is the scarcity of traffic-related tweets relative to general posts, which further necessitates efficient preprocessing to mitigate linguistic ambiguity and noise (Fatichah et al. 2020; Afyouni, Aghbari, and Razack 2022; Li, Dou, and Zhou 2023).

The research addresses three primary goals: noise reduction, improved traffic event extraction, and enhanced dataset generalisability (Garg and Kumar 2016; Ramadhani and Goo 2017; Xu, Li, and Wen 2018; Zhang et al. 2018; Ali et al. 2019; Azhar et al. 2022). However, challenges persist in handling the evolving nature of informal language, slang, emojis, sarcasm, and ambiguity in tweets (Fatichah et al. 2020). Additionally, privacy concerns necessitate adherence to regulations, including data anonymization and obtaining user consent (Ramadhani and Goo 2017; Karthik et al. 2023). Integrating sentiment analysis and contextual information may help address these linguistic challenges, while privacy concerns require appropriate anonymization and consent procedures. Figure 6 illustrates the preprocessing workflow for Twitter/X data for TED.

3.2.3 Semantic feature learning: from traditional vectors to semantic embeddings

A critical challenge in applying DL to traffic event detection TED lies in converting raw text into meaningful, vectorized representations suitable for algorithmic processing. Text representation plays a foundational role, influencing model accuracy and robustness, particularly when dealing with noisy and informal Twitter/X data (Zhang et al. 2018; Dabiri and Heaslip 2019; Neruda and Winarko 2021).

Text representation in TED has evolved across three major phases:

a. frequency-based models including Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF).
b. Word embeddings such as Word2Vec and GloVe.
c. contextual embeddings using transformer-based models such as BERT (discussed further in Section 3.2.4).

a. Traditional Frequency-Based Models: BoW and TF-IDF

Early approaches to representing textual data in TED systems relied on BoW and TF-IDF. The BoW approach converts unstructured tweets into fixed-size numerical vectors based on word occurrence, ignoring context and word order (Dabiri and Heaslip 2019). Each word is assigned a unique index, and tweets are represented as N-dimensional vectors, where N is the vocabulary size. TF-IDF refines this by weighting terms based on their frequency across multiple documents, which emphasises more informative words (Rajaraman and Ullman 2011).

Several studies have employed these methods in analysing traffic-related tweets. D'Andrea et al. (2015) used IDF-based feature selection to classify Italian tweets, while Alomari et al. (2021) developed a Spark-based feature extraction pipeline for Arabic tweets. However, both BoW and TF-IDF suffer from key limitations: they ignore word order, semantic similarity, and contextual meaning. This reduces their effectiveness in handling the informal, slang-heavy, and abbreviation-rich nature of social media data (Deho et al. 2018; Rudkowsky et al. 2018; Dabiri and Heaslip 2019).

As deep learning models require dense, semantically rich input vectors, the limitations of BoW and TF-IDF have led to a shift toward word embedding techniques that can capture syntactic and semantic relationships more effectively. This evolution reflects the broader transition from classical vectorisation to neural embedding in modern TED pipelines.

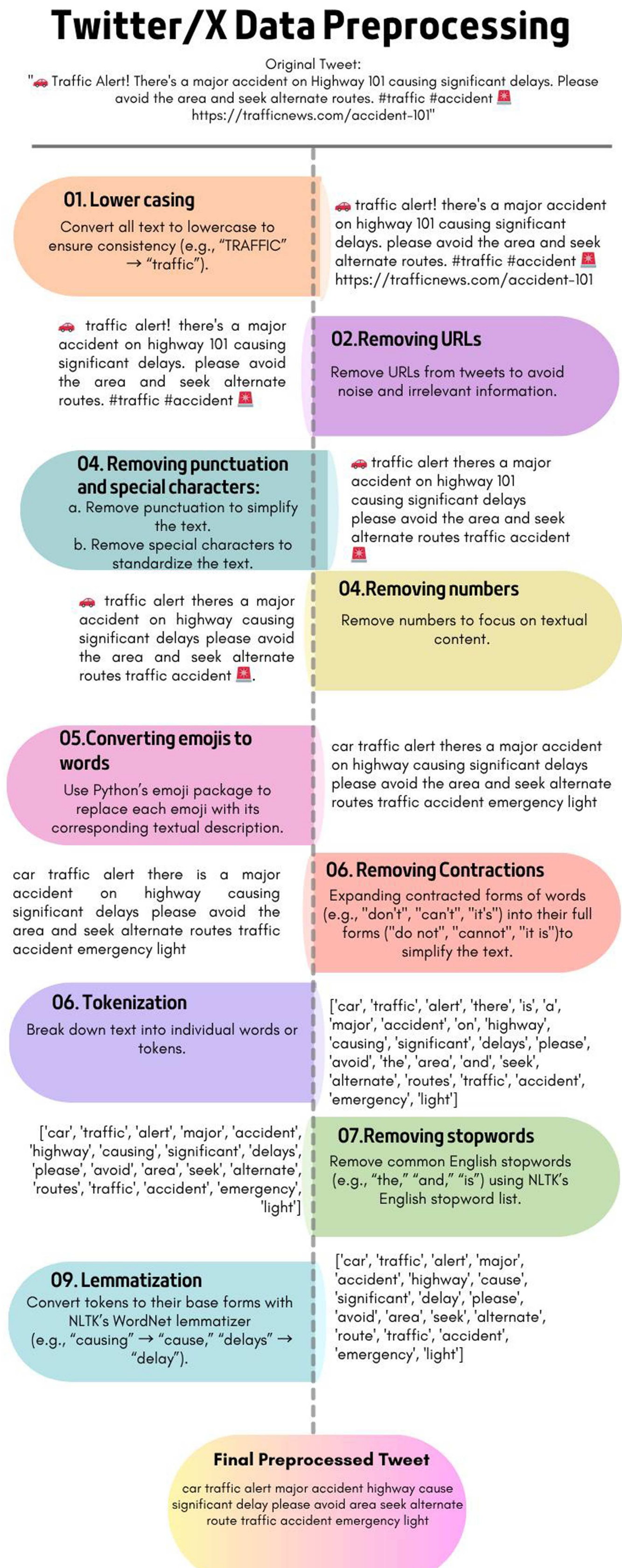


Figure 6. Twitter/X data preprocessing steps applied to a traffic-related tweet, including text cleaning, normalisation, tokenization, and lemmatization for model-ready input.

b. Transition to Word Embeddings

Word embeddings extend beyond traditional vectorisation by capturing contextual relationships between terms (Ali et al. 2019; Dabiri and Heaslip 2019). Widely applied across natural language processing, computer science, artificial intelligence, machine learning, and computational linguistics, these methods map words into low-dimensional vector spaces that encode syntactic, semantic, and distributional meanings (Noori and Mehra 2020; Sampath and Supriya 2023). Common word embedding techniques include String2Vec, Word2Vec, Doc2Vec, GloVe, and FastText, with Word2Vec and FastText being the most frequently used in Twitter/X-based TED. In summary, while Twitter/X data remains valuable for TED, practical access is now significantly constrained. Researchers must adapt by utilising limited-access APIs, employing ethical web scraping, or leveraging multimodal datasets to compensate for the limitations imposed in the post-API era (Blakey 2024; Poudel and Weninger 2024). Table 2 summarises the feature learning methods utilised in this domain.

Word2Vec, introduced by Google in 2013, is an unsupervised DL technique that learns word representation by capturing semantic relations, synonyms, and analogies through analysis of large text corpora like Twitter/X (Bilgin and Şentürk 2017; Mikolov, Quoc V, and Ilya 2013; Mikolov and Sutskever, 2013). Although it generalises well, it struggles with Out-of-Vocabulary words (Bilgin and Şentürk 2017).

FastText, developed by Facebook's AI Research lab, represents words as sub-word *n*-grams. It is effective in handling typos, abbreviations, and informal language commonly found in Twitter/X traffic reports. FastText enhances NER and DL models by capturing morphological similarities (Mannes 2017).

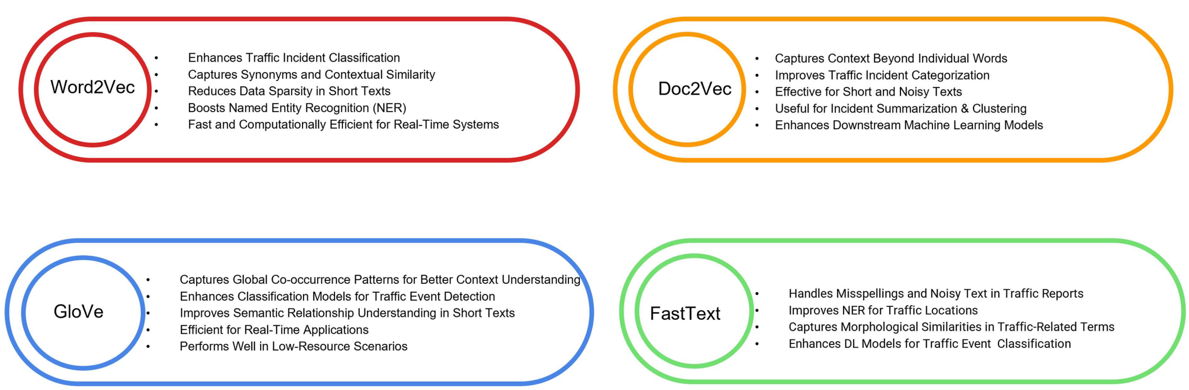
GloVe (Global Vectors for Word Representation) differs from Word2Vec in that it leverages global co-occurrence statistics. While it performs well on similar tasks, it struggles with words that have multiple meanings (Pennington, Socher, and Manning 2014; Abad et al. 2016).

Doc2Vec, an extension of Word2Vec, generates embeddings for full documents, enabling phrase and paragraph-level analysis (Bilgin and Şentürk 2017; Kamkarhaghighi and Makrehchi 2017). Figure 7 presents the efficiency and comparative performance of word embedding methods in TED.

c. Performance and Trade-Offs in TED Applications

Several studies have demonstrated the effectiveness of word embeddings in Twitter/X-based TED, often outperforming traditional BoW and TF-IDF models. Dabiri and Heaslip (2019) found that integrating CNN with Word2Vec achieved better accuracy and F-score. Similarly, Lu et al. (2018) developed a Word2Vec-based approach that outperformed CBOW in terms of traffic incident detection accuracy.

Ali et al. (2019) evaluated String2Word, Word2Vec, Doc2Vec, and GloVe for transportation sentiment analysis, identifying key terms such as 'Crash,' 'Accident,' 'Traffic,' 'Speed,' and 'Event.' While



	Handles Noisy Twitter	*Captures Context*	*Efficiency with limited labeled training*	*Time Efficiency*	*Capabilities for Classification*
String2Vec	✓ Good for typos	× String-level similarities Only	× Limited	✓ Fast	× limited
Word2Vec	× Struggles with typos	✓Good semantic understanding	✓Need Training	✓ Efficient	✓ Improves the models
Doc2Vec	× sensitive to typo	✓Understand full tweet context	✓ Good with limited training data	× Slower than Word2vec	✓ Great for classification
GloVe	× Limited typo handling	✓Captures word relationships well	✓Good with limited training data	✓ Fast for real-time user	✓ Improves DL models
FastText	✓ Best for typos & misspellings	✓ Understand of morphology & and context	✓Good with limited training data	✓Efficient	✓Great for classification

Figure 7. Comparison of word embedding models for traffic text, showing their strengths in context capture, noise handling, and classification. The bottom panel compares typo tolerance, context understanding, data efficiency, speed, and classification performance.

Table 3. The trade-offs between embedding methods in terms of semantic richness and computational efficiency.

Embedding method	Sematic richness	Training time	Inference speed	Computational cost	Dimensionality	Best use case
Word2Vec	Moderate	Fast	Fast	Low	100-300	Real-time systems, short texts
FastText	Moderate–High	Very Fast	Very Fast	Very Low	100-300	Real-time, Out-of-Vocabulary handling
Doc2Vec	High	Slow	Moderate	Medium–High	+ 300	Document-level classification
GloVe	Moderate	Moderate	Fast	Medium	100-300	Sentiment analysis, pre-trained use

String2Word had good accuracy, its high dimensionality increased computational time and reduced precision. Doc2Vec achieved 74% accuracy using logistic regression and 73% with deep learning, outperforming both Word2Vec and GloVe. Finally, GloVe, trained on a transportation corpus, reached 66% accuracy in deep learning tasks, excelling in feature recognition but underperforming in accuracy compared to Word2Vec and Doc2Vec. Almassar and Girsang (2022) tested FastText, achieving 86.33% accuracy and a 96.61 F-score in congestion detection. Fatichah et al. (2020) combined FastText with CNN, LSTM, and C-LSTM for handling out-of-vocabulary terms for incident detection. Ambastha and Desarkar (2020) compared TF-IDF and Word2Vec, using SVM, Naïve Bayes, CNN, and LSTM models. Azhar et al. (2022) proposed an integrated model combining word embeddings with numeric traffic and sentiment layers to enhance TED performance.

Despite their advantages, advanced embeddings can pose computational challenges in real-time systems. Lightweight models, such as Word2Vec and FastText, are widely preferred due to their fast inference and low memory usage, making them ideal for large-scale tweet streams. In contrast, Doc2Vec provides richer semantic representations but requires greater computational resources, including GPU acceleration and longer training times. Embeddings exceeding 300 dimensions can slow processing considerably in high-throughput environments. Thus, choosing an embedding requires balancing semantic accuracy against processing efficiency, particularly in latency-sensitive deployments (Rudkowsky et al. 2018; Ali et al. 2019; Gu et al. 2021).

To improve conceptual clarity while considering practical limitations, we compare embeddings based on both performance and suitability for different use cases. FastText, due to sub-word modelling, is ideal for short and informal text (Bonandrini and Gatti 2024). Conversely, semantically rich models like Doc2Vec or GloVe are better suited to batch processing and archival analysis, where computational latency is less critical (Pita and Pappa 2018).

Word embeddings offer clear benefits but also introduce challenges. One major issue is optimising the embedding dimension, typically set between 100 and 300, which affects both accuracy and computational cost (Asudani et al. 2023). Embeddings also struggle with informal and evolving social media language, such as slang and abbreviations. Additionally, relying on pre-trained embeddings may limit adaptability, thus requiring domain-specific fine-tuning (Raunak 2017; Wilson et al. 2020; Torregrossa et al. 2021; Asudani et al. 2023). Table 3 summarises the trade-offs among various embedding methods in terms of semantic richness and computational demands.

Furthermore, high-dimensional embeddings increase computational complexity, which affects scalability in real-time TED applications (Deho et al. 2018; Rudkowsky et al. 2018; Gu et al. 2021). In summary, the selection of an appropriate embedding technique depends on dataset characteristics, computational constraints, and task-specific requirements (Gu et al. 2021). These considerations are especially important when deploying models in real-time systems, where the trade-off between embedding accuracy and computational efficiency becomes critical.

3.2.4 DL models for TED

To build a solid foundation, it is essential to contextualise the development of TED models from ML to DL and, more recently, to transformer-based architectures. This shift reflects advances in data availability, algorithmic design, and computing power.

Table 4 presents the evolution of TED methodologies from traditional ML approaches to transformer-based models. During the ML era, algorithms such as SVM, Decision Trees, and Random Forests formed the foundation of early TED systems. These models were typically combined with TF-IDF or Bag-of-

Table 4. Evolution of TED methodologies from ML to transformer-based models, highlighting models, data representation, and modelling capabilities.

Methodology era	Models	Data representation	Strengths	Weaknesses
ML	SVM, Decision Trees, Random Forests	TF-IDF, BoW	Simple, interpretable	No semantic/contextual understanding
DL	CNN, LSTM, GRU	Word Embedding	Learns hierarchical features	Requires more data & computing
Transformers	BERT, Multimodal Bi-transformers (MMBT), Vision-and-Language Transformer (ViLT)	Multimodal inputs	Contextual & cross-modal reasoning	Resource intensive

Words (BoW) for feature representation. Prior studies have shown that such approaches lack context sensitivity and require extensive manual feature engineering (Atefeh and Khreich 2013; D'Andrea et al. 2015).

A major shift occurred in the deep learning era, with models such as CNNs, LSTMs, and GRUs emerging as more effective alternatives. These architectures learn hierarchical and temporal patterns directly from raw text, enhancing performance on unstructured tweet data (Dabiri and Heaslip 2019; Almassar and Girsang 2022).

Finally, the transformer era introduced models like BERT, which offer bidirectional contextual understanding, effectively handle multilingual and noisy data, and support multimodal data integration. These models, however, are computationally intensive (Neruda and Winarko 2021; Nirbhaya and Suadaa 2023).

The effectiveness of DL models in TED depends on the quantity and quality of training data, as well as the complexity of the tasks. DL Models like CNN, LSTM, and BERT require large, labelled datasets, whereas simpler ML models can perform reasonably well on smaller datasets. Model selection is also influenced by the application. Complex models tend to improve accuracy, while simpler models offer faster inference. Integrating features such as emotion, weather, and geo-location data can further enhance TED performance. However, no single DL model guarantees optimal results. The choice depends on dataset characteristics, feature engineering, and computational constraints (Azhar et al. 2022; Yang 2022). This section reviews the DL models adopted for TED in the selected studies. Figure 8 illustrates their capabilities for Twitter/X-based TED.

3.2.4.1 Multi-Layer Perceptron (MLP) MLP is a feed-forward neural network comprising an input layer, a hidden layer, and an output layer. Jonnalagadda and Hashemi (2021) compared MLP with SVM, Bi-LSTM, and CNN, finding that MLP underperformed due to its inability to capture sequential dependencies. Similarly, Puangnak and Rachsiriwatcharabul (2022) showed that CNN and CNN + LSTM outperformed MLP, underscoring the importance of sequence modelling. MLP remains useful but is highly dependent on data quality and feature integration.

3.2.4.2 Convolutional Neural Networks (CNNs) Originally developed for image recognition, CNNs have been effectively adapted for natural language processing by applying one-dimensional convolutions to text embeddings. They excel at identifying local patterns, such as *n*-grams, making them suitable for keyword-based incident detection. Integrated with pretrained embeddings like FastText, CNNs perform well even on noisy text. Hybrid models, such as CNN + LSTM, combine spatial pattern recognition with temporal learning, thereby improving overall accuracy (Alifi and Supangkat 2018; Dabiri and Heaslip 2019; Liu et al. 2020; Neruda and Winarko 2021; Jain et al. 2023; Qutaishat and Li 2025b). Almassar and Girsang (2022) found that CNN + FastText outperformed CNN + Word2Vec and SVM. Dabiri and Heaslip (2019) found that CNN outperformed both LSTM and CNN + LSTM. CNN + LSTM hybrids further enhance performance by combining spatial and temporal features. Chen et al. (2018) confirmed this in Sina Weibo traffic analysis, where LSTM–CNN achieved the top F1 score.

3.2.4.3 Recurrent neural networks (RNNs) and long short-term memory (LSTM) RNNs process sequential data but struggle with long-term dependencies due to the vanishing gradient problem. LSTM addresses this issue. Azhar et al. (2022) reported that LSTM achieved 94.2% accuracy in road accident detection, followed by GRU of 91.6% and RNN of 39.7%. Yang (2022) combined stacked autoencoders with LSTM, achieving 98.25% accuracy, which outperforms CNN + LSTM at 96.36%.

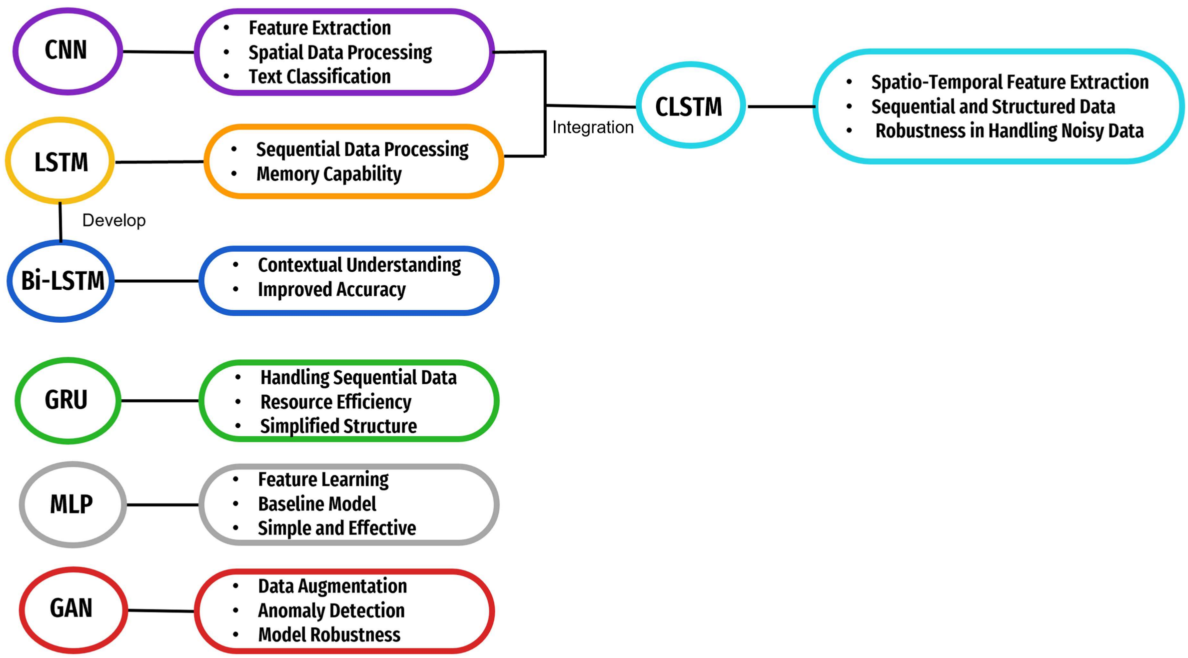


Figure 8. DL models for Twitter/X traffic event detection, highlighting their roles in feature extraction, sequential processing, and noise robustness.

3.2.4.4 Bidirectional LSTM (Bi-LSTM) Bi-LSTM improves upon LSTM by capturing bidirectional dependencies. Puangnak and Rachsiriwatcharabul (2022) reported an accuracy of 93.53% in incident detection using Bi-LSTM, although the accuracy for severity classification was lower, reaching 77.92%. Alifi and Supangkat (2018) developed a hybrid model combining Bi-LSTM and CNN, achieving an F-score of 78.9%.

3.2.4.5 CNN-LSTM(C-LSTM) C-LSTM integrates CNNs for spatial feature extraction with LSTMs for temporal learning. Fatichah et al. (2020) adapt C-LSTM for incident detection, achieving 99.09% accuracy. Combined with VGG16, it yielded the highest multimodal prediction confidence. Zeng et al. (2019) reported C-LSTM outperformed both CNN at 80.27% and LSTM at 80.96% in traffic classification.

3.2.4.6 Generative Adversarial Networks (GANs) GANs consist of a generator and a discriminator working together to improve data quality. They are used for detecting traffic anomalies and generating synthetic, yet realistic, data. Liu et al. (2024) proposed a GAN-Transformer for TED, using GANs to balance datasets while Transformers extracted complex relations. Lin et al. (2020) utilised GANs to mitigate data imbalance, resulting in a 3.2% improvement in accuracy and a 5.65% reduction in false alarms.

3.2.4.7 Gated Recurrent Unit (GRU) GRU, developed by Kyunghyun Cho in 2014, is an RNN variant with update and reset gates to manage information flow. Azhar et al. (2022) demonstrated that GRU achieved an accuracy of 93.7% in detecting accident-related tweets. Suat-Rojas, Gutierrez-Osorio, and Pedraza (2022) proposed a GRU-CNN hybrid that outperforms baseline models in accident prediction.

3.2.4.8 Bidirectional Encoder Representations from Transformers (BERT) BERT was developed in two model sizes: BERTBASE (110 M parameters) and BERTLARGE (340 M parameters). Although studies on BERT for TED are limited, it has proven effective for detecting and analysing traffic incidents in noisy, unstructured social media data (Qutaishat and Li 2025a). Nirbhaya and Suadaa (2023) reported 99.26% accuracy using IndoBERT in Indonesian-language TED, outperforming traditional models. Neruda and Winarko (2021) demonstrated that BERT-CNN outperformed ELMo-CNN and Word2Vec-CNN in terms of performance.

3.2.5 DL model training and validation

To achieve effective training and model validation, Twitter/X-based TED studies commonly divide datasets into three subsets: training, validation, and testing (Raschka 2018). The training set is repeatedly used for model learning, allowing it to discover the underlying patterns, while the validation and test sets assess the model's performance and generalisation.

The proportions of these subsets often vary based on dataset size and characteristics. In social media-based TED, where data sparsity and informal language are common, selecting appropriate splits becomes crucial (Jain et al. 2022; Savvides and Mäkelä 2023; Qiu 2024). As illustrated in Figure 9 although a general protocol for dataset partitioning exists in DL, many studies fail to adhere to it.

Several studies report their data split strategies. Alifi and Supangkat (2018) used a 70:15:15 split for training, testing, and validation. Nirbhaya and Suadaa (2023) employed k-fold cross-validation, dividing the data into 80% training, 10% validation, and 10% testing sets. Almassar and Girsang (2022) used 60% for training, 20% for validation, and 20% for testing. Puangnak and Rachsiriwatcharabul (2022) explored multiple ratios, including 50:50, 60:40, 70:30, 80:20, 90:10, and 95:5, and identified 90:10 and 95:5 as optimal; however, they ultimately adopted 70:30. Nevertheless, their study lacked detail on the validation procedure. Neruda and Winarko (2021) employed a 64:16:20 split and explicitly noted measures to prevent data leakage through preprocessing and feature extraction. Table 5 presents the dataset partitioning strategies across reviewed TED studies.

While many studies report their dataset split ratios, few evaluate how these choices impact model performance. High training proportions, such as 90:10 or 95:5, may boost training accuracy but risk overfitting due to limited evaluation data. In contrast, balanced splits, such as 70:30 or k-fold cross-validation, generally provide more robust estimates of generalisability. For instance, although Puangnak and Rachsiriwatcharabul (2022) reported strong accuracy with 90:10 and 95:5 splits, the absence of a detailed validation procedure limits interpretability. Figure 10 illustrates how reported performance measures vary by split strategy, reinforcing the need for standardised, transparent validation procedures to support fair benchmarking across TED studies.

3.2.6 Model performance measures and evaluations

Based on Table 5, most studies included in this review used accuracy and F-score to evaluate model performance. Accuracy is a common starting point due to its simplicity and clear indication of overall correct

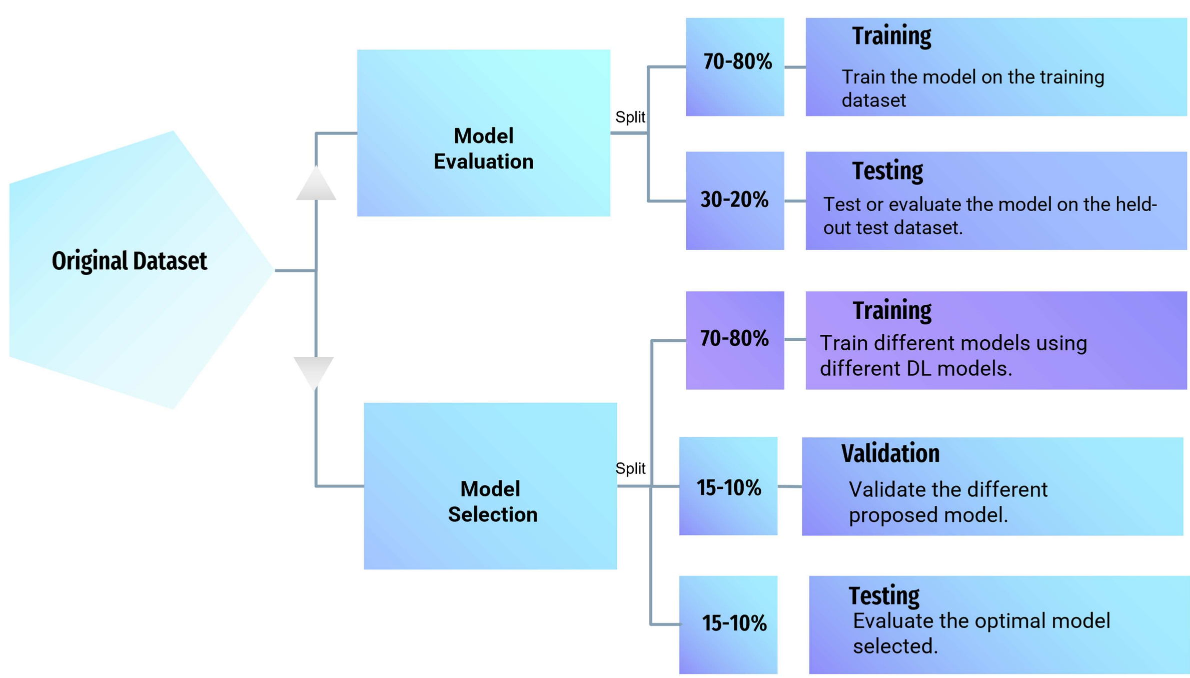


Figure 9. The DL dataset split for model selection and evaluation. Though widely recommended, this protocol is often inconsistently applied in practice.

Table 5. DL-based twitter/X TED studies.

Study author	Study objective	Model	Dat split (train/validation/test) & validation method	Performance measures: Accuracy (%)	F1-score (%)	Precision (%)	Recall (%)	Future work
Almassar and Girsang (2022)	Traffic congestion detection	CNN + Word2Vec, CNN + FastText SVM	60/20/20—Grid Search Cross-Validation	85.79% 86.33% 67.62%	86.11% 96.61% 68.53%	80.59% 81.18% 63.84%	92.45% 92.83% 73.96%	• Improve model accuracy. • Integrate data sources. • Enhance predictive capabilities. • Scale to other regions/languages. • Explore real-time implementation and performance.
Zhang et al. (2018)	Traffic accident detection	Deep Belief Network ANN LSTM SVM	Not reported—5-Fold Cross-Validation	85% 82% 81% 79%	NA NA NA NA	92% 81% 87% 84%	NA NA NA NA	• Investigate underreported accidents on Twitter/X. • Integrate social media data for better traffic jam detection. • Develop a common dataset via community collaboration.
Puangnak and Rachsiriwatchara-bul (2022)	Road traffic incident reports classification	MLP CNN Bi-LSTM LSTM + CNN	70/30—10-Fold Cross-Validation	44.4% 93.24% 77.92% 93.44%	NA NA NA NA	NA NA NA NA	NA NA NA NA	• Automate incident detection. • Reduce reliance on manual surveillance. • Build accurate DL models for traffic classification.
Azhar et al. (2022)	Road accident detection and prediction	GRU RNN LSTM	80/20—k-Fold Cross-Validation (k not specified)	93.7% 91.6% 94.2%	93% 90% 95%	91% 90% 94%	NA NA NA	• Manage large Twitter/X datasets for accident detection. • Improve geo-location accuracy. • Manage incomplete datasets, such as weather info.
(Neruda and Winarko (2021)	TED	BERT + CNN ELMo + CNN Word2Vec + CNN	64/16/20—5-Fold Cross-Validation	NA NA NA	90% 88% 87%	90% 90% 88%	92% 87% 87%	• Enhance data labelling for better TED. • Assess fine-tuning of pre-trained word embeddings on task performance. • Explore advanced oversampling techniques like MLSMOTE. • Test various neural network architectures for detection improvements.
Jonnalagadda and Hashemi (2021)	TED	RF SVM MLP BLSTM CNN	20% Validation—5-Fold Cross-Validation	67% 73.1% 78.7% 88.1% 93.3%	67% 73.4% 78% 88% 93%	NA NA NA NA	NA NA NA NA	• Enhance TED by integrating additional data sources and refining algorithms. • Improve accurate identification of traffic events from social media data. • Refine algorithms for better accuracy and reliability. • Address the vanishing gradient problem RNN to improve long-term dependencies. • Expand TED beyond binary classification to include fine-grained categorisation.
Fatichah et al. (2020)	Incident type prediction	CNN CNN + LSTM	Not reported—Hold-out Validation	98.95% 99.09%	NA NA	NA NA	NA NA	• Enhance real-time processing for live monitoring. • Assess model robustness in varied environments.

Table 5. (*Continued*)

Study author	Study objective	Model	Dat split (train/ validation/test) & validation method	Performance measures: Accuracy (%)	F1-score (%)	Precision (%)	Recall (%)	Future work
								• Analyse data augmentation effects on CNN architectures. • Incorporate audio and video for incident detection. • Explore advanced DL models like transformers for classification.
Dabiri and Heaslip (2019)	Traffic incident detection	CNN + Word2vec LSTM + Word2vec CISTM + Word2vec CNN + FastText LSTM + FastText CISTM + FastText CNN + Random Word vector LSTM + Random Word vector CLSTM + Random word vector	Not reported—5-Fold Cross-Validation	98.6% 98.4% 98.5% 98.6% 98.5% 98.6% 50.1% 50.3% 49.9%	98.6% 98.4% 98.5% 98.6% 98.5% 98.6%3939.8% 50.2% 49.8%	NA NA NA NA NA NA NA NA NA	NA NA NA NA NA NA NA NA NA	• Develop an efficient geocoder to extract traffic event locations from tweets. • Implement a unified Twitter/X-based traffic information system across traffic networks. • Enhance real-time traffic information dissemination to drivers and traffic managers. • Explore integration with transportation agencies and state police for improved traffic flow management
Ambastha and Desarkar (2020)	Traffic congestion detection	CNN LSTM ULMFiT	L-TWITS split (not specified)—Holdout Validation	NA NA NA	78% 82% 90%	70% 76% 84%	64% 74% 89%	• Explore traffic location detection from tweet content. • Expand the L-TWITS dataset by collecting more tweets on diverse traffic incidents. • Address ULMFiT's prediction failures for sarcastic and irrelevant traffic tweets.
Ali et al. (2021)	Traffic Incident Detection and Condition Analysis	OLDA + word2Vec + RNN OLDA + word2Vec + LST-M OLDA + word2Vec + Bi-LSTM OLDA + FastText + RNN OLDA + FastText + LSTM OLDA + FastText + bi-LSTM	70/30—Holdout Validation	80% 85% 91% 85% 92% 97%	77% 84% 89% 85% 92% 97%	83% 83% 85% 82% 94% 97%	71% 86% 94% 88% 92% 97%	• Data Enhancement • Incorporating diverse event types • Improving model performance
Alifi and Supangkat (2018)	Traffic condition recognition	Bi-LSTM + CNN	70/15/ 15—Holdout Validation		78.9%	82.1%	75%	• Combine GANs with other DL models for improved accuracy. • Adapt the framework for real-time traffic data processing. • Test on diverse datasets to ensure robustness. • Address imbalances in incident types and conditions. • Investigate CNNs or RNNs to boost GAN performance.

(*Continued*)

Table 5. (*Continued*)

Study author	Study objective	Model	Dat split (train/validation/test) & validation method	Performance measures				Future work
				Accuracy (%)	F1-score (%)	Precision (%)	Recall (%)	
Lin et al. (2020)	Traffic incident detection	GAN RF SVM	Not reported—10-Fold Cross-Validation	91.87% NA 91.53%	NA NA Na	NA NA NA	NA NA NA	• Combine GANs with other DL models for improved accuracy. • Adapt the framework for real-time traffic data processing. • Test on diverse datasets to ensure robustness. • Address imbalances in incident types and conditions. • Investigate CNNs or RNNs to boost GAN performance.
Nirbhaya and Suadaa (2023)	Traffic incident detection	SVM Navey Bayes Logistic Regression LSTM IndoBERT	80/10/10—5-Fold Cross-Validation	98.81% 93.86% 98.68% 96.14% 99.26%	98.54% 92.49% 98.37% 95.16% 99.10%	98.78% 91.91% 98.77% 95.72% 99.20%	98.31% 93.16% 98% 94.72% 99.02%	• Enhanced data collection • Advanced pre-processing techniques • Incorporating multimodal data

Note: CNN = Convolutional Neural Network; LSTM = Long Short-Term Memory; Bi-LSTM = Bidirectional LSTM; SVM = Support Vector Machine; MLP = Multi-Layer Perceptron; GRU = Gated Recurrent Unit; GAN = Generative Adversarial Network; ULMFiT = Universal Language Model Fine-tuning for Text Classification; RF = Random Forest; L-TWITS = Labelled-TWeets for Indian Traffic Scenario; MLSMOTE = Multi-Label Synthetic Minority Over-sampling Technique. NA = Not Available. Some studies did not report full metric sets or specify validation splits.

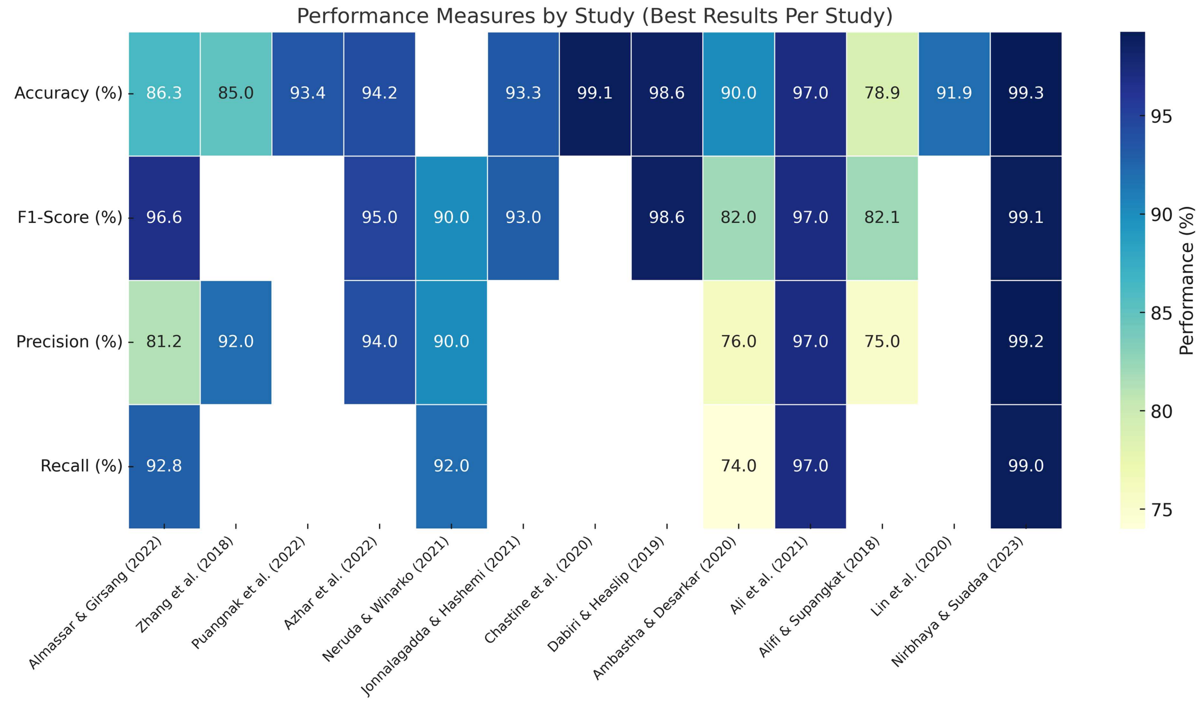


Figure 10. Performance metrics (accuracy, precision, recall, and F1-score) reported across studies using various dataset split strategies for TED.

predictions. However, with imbalanced data, a frequent issue in TED where incident-related tweets are fewer than non-incident ones, accuracy can be misleading. It often prioritises the majority class while overlooking performance on minority classes (Johnson and Khoshgoftaar 2019). To mitigate this, the F-score is widely adopted for its balance of precision, which refers to the proportion of correctly identified positives, and recall, which refers to the proportion of actual positives correctly captured (Dinga et al. 2019).

Nonetheless, comparisons across studies reveal inconsistencies. For example, Almassar and Girsang (2022) reported an accuracy of 86.33% using CNN + FastText, but with a lower precision of 81.18%, indicating a potential misclassification of minority-class events. Similarly, Ali et al. (2021) achieved accuracy scores up to 97%, while some F1-scores ranged from 84% to 89%, reflecting the influence of class imbalance on model reliability.

Most studies mention data augmentation or resampling to address this issue, but rarely specify the methods used. For instance, Neruda and Winarko (2021) referred to synthetic oversampling, and Jonnalagadda and Hashemi (2021) adapted weighted loss functions to emphasise underrepresented labels. Other studies, such as those by Azhar et al. (2022), Puangnak and Rachsiriwatcharabul (2022), and Ambastha and Desarkar (2020), either referred to augmentation broadly or did not report a specific strategy.

Class imbalance can significantly reduce recall and F1-scores for rare events (Johnson and Khoshgoftaar 2019; Walsh and Tardy 2022; Jiang et al. 2025). This lack of transparency in the study's methodology makes it challenging to assess the comparative effectiveness of imbalance-handling techniques. So future work should clearly report the techniques used and their impact on minority-class performance. Table 6 presents common techniques for handling class imbalance in deep learning models (Johnson and Khoshgoftaar 2019; Walsh and Tardy 2022; Jiang et al. 2025).

4 Challenges, open issues, and future directions

Challenges and limitations arise at each stage of the DL-based Twitter/X TED workflow, culminating in the final detection stage. Key issues, open challenges, and future directions are discussed below and summarised in Table 5.

Table 6. Common techniques for handling class imbalance in deep learning models.

Technique	Description	Type	Reported effect
Random oversampling/ undersampling	Duplicate minority class samples or downsample the majority class to achieve a balanced distribution.	Data-level	Improved recall, sometimes at the cost of precision.
Class weighting/cost-sensitive learning	Assign higher loss weights to minority classes during model training.	Algorithm-level	Balanced F1-score, reduced bias towards the majority class.
SMOTE (synthetic minority over-sampling technique)	Generate synthetic minority samples by interpolating the feature space.	Data-level	Increased sensitivity to rare classes; reduced under-detection.
ADASYN/MLSMOTE (advanced SMOTE variants)	Adaptive generation of minority samples, focusing on harder-to-learn examples.	Data-level	Improved minority-class recall, but higher training complexity.
GAN-based data augmentation	Use Generative Adversarial Networks to synthesise realistic minority samples.	Hybrid	Boosted accuracy and F1-scores, especially in rare-event detection.
Ensemble methods	Combine classifiers trained on different balanced subsets.	Algorithm-level	Improved robustness, stable performance across imbalanced datasets.

4.1 Data quality and labelling

Twitter/X's dynamic nature produces noisy data with misspellings, abbreviations, and informal language that hinder accurate TED. Ambiguity and brevity in tweets further complicate the extraction of meaningful information. DL models require substantial labelled data, yet obtaining sufficient traffic-related tweets remains difficult. Manual labelling is time-consuming and resource-intensive. Two solutions have been suggested:

a. Adopting semi-supervised learning to leverage a small volume of labelled data alongside a larger pool of unlabelled data for DL model training.
b. Incorporating external sources such as police reports or traffic authority data to automatically label relevant tweets.

4.2 Imbalanced data distribution

Traffic-related events that occur more frequently, such as 'congestion,' are given more priority than those that rarely occur, such as 'accidents,' which skew the training process. To mitigate this challenge, researchers can use data augmentation, which can be used to generate new data from the existing dataset and train DL models. Additionally, assigning higher weights to misclassify rare events during training can be incorporated to ensure attention is paid when modelling.

4.3 Ethics and data governance

The application of social media data for identifying traffic events presents ethical and governance challenges. Privacy concerns arise when users are unaware that their posts are being analysed, necessitating informed consent and data anonymization (Mredula and Noyon, 2022). Additionally, algorithms trained on such data may reflect existing social biases, which can potentially lead to unfair outcomes. Addressing these issues requires the adoption of fairness-aware models (Chen and Wang 2019; Alomari and Mehmood 2023). Data governance also plays a key role, involving questions about data ownership and compliance with regulations such as the General Data Protection Regulation, which mandates user consent and control over personal data (Chen, Chen, and Qian 2014; Melhem, Abdi, and Meziane 2024). While technical issues are being managed, the focus should be on ethical standards and regulatory frameworks. A responsible approach will require collaboration across technology, ethics, and policy domains to ensure both effectiveness and respect for user rights.

4.4 Multi-modal data integration

Ongoing research explores the integration of traffic sensor data, speed logs, and accident reports to enhance traffic condition analysis. Multimodal fusion techniques combine inputs from multiple sources, boosting detection accuracy. Recent work also emphasises spatial and temporal modelling to better capture the evolving nature of traffic events.

In addition to tweets, recent research emphasises the integration of text, audio, video, and IoT sensor data to enhance TED. These modalities complement Twitter/X's text-based signals by providing real-world, multimodal evidence of incidents or congestion. Each source contributes distinct strengths:

- Text (tweets) offers real-time, crowd-sourced alerts.
- Audio (sirens, honking) reflects ambient traffic conditions.
- Video (CCTV, dashcams) provides visual confirmation of events.
- IoT sensors (GPS, accelerometers, traffic counters) deliver location-specific, quantitative indicators.

To process these diverse inputs, deep learning models such as late fusion CNN-LSTM architectures and attention-based transformers are employed. These models support cross-modal feature learning, enhancing detection robustness, particularly in noisy or ambiguous textual environments. Transformer-based frameworks, such as Multimodal Bi-transformers (MMBT) and Vision-and-Language Transformer (ViLT), are also promising for aligning visual, textual, and sensor-based information within a unified pipeline.

A practical implementation is demonstrated in Fatichah et al. (2020), who developed a TED system that integrates Twitter/X text and images. Their approach used CNNs for image analysis and C-LSTM for text processing. By multimodal data integration, the system improved detection accuracy compared to text-only approaches. This shows the potential of multimodal deep learning for enhancing incident detection by leveraging complementary sources of information available on social media platforms. Figure 11 expands on this framework, illustrating a conceptual pipeline that incorporates additional Twitter/X data such as audio, video, and IoT sensor data alongside Twitter/X text for enhanced TED.

4.5 Deep learning models hardware constraints

The need for extensive hardware resources when deploying deep learning models, especially Transformer-based architectures, is considered a challenge(Burhanuddin 2023; Lyu et al. 2022). For example, achieving inference latency below 500 ms requires GPU acceleration with at least 16 GB of VRAM. For medium- to large-scale

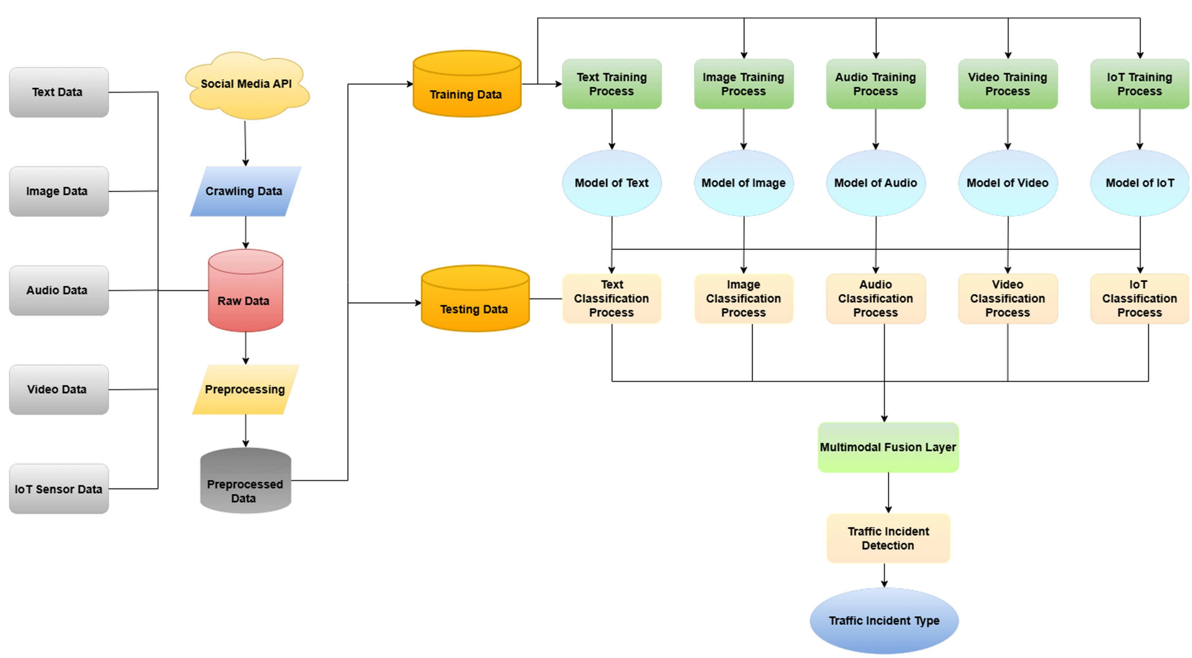


Figure 11. Conceptual framework for multimodal data fusion in traffic event detection (TED), integrating Twitter/X text, audio, video, and IoT sensor data. This figure is an expanded version of the framework proposed in Fatichah et al. (2020), illustrating how deep learning models can jointly process heterogeneous data streams to improve incident detection accuracy and robustness.

processing, such as handling more than 10,000 tweets/second, TPU integration or distributed GPU clusters are recommended to maintain throughput (Cao et al. 2025). A number of memory architecture optimisation techniques were suggested, such as halving the embedding size, parameter pruning, and quantisation (Saai and Vijayakumar 2024). In addition, DeepSpeed-Inference offers a multi-GPU and heterogeneous inference solution that significantly reduces latency and increases throughput (Yazdani et al. 2022). It enables the inference of models up to 25 times larger than those supported by GPU-only solutions.

4.6 Large language models (LLMs) for TED

LLMs, such as ChatGPT and Claude, are not widely adopted in TED research. However, their capabilities in handling informal, multilingual, and ambiguous language show a promising path in future applications. Additionally, LLMs support zero- and few-shot learning, empowering classification, summarisation, and sentiment analysis with minimal labelled data (Wandelt et al. 2024). Still, challenges such as high computational costs, Limited open access to real-time inference, fine-tuning issues on traffic data, and reproducibility issues due to API-based LLMs impede their adaptability in the TED research (Zhang et al. 2024). Hybrid systems combining LLMs for preprocessing with lightweight models for detection may offer an effective path forward. As LLMs evolve, their role in multilingual and multimodal TED systems calls for further exploration (Mahmud et al. 2025).

4.7 Complexity of DL and the need for explainability and interpretability

DL models often function as black boxes, making it essential to improve their interpretability and debugging for reliable predictions. Researchers are developing AI techniques to understand and refine model processes. Common approaches include:

- Attention visualisation, which highlights key parts of a tweet sequence during prediction to assess element significance.
- Counterfactual explanations, or 'what-if' analysis, examining how changes in specific tweet elements affect predictions.
- SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), two model-agnostic interpretability techniques that help explain predictions by attributing them to individual features.

4.8 The constant change of language and behaviour of social media users

Incremental learning techniques can help models continuously learn from new data while supporting previous knowledge and adapting to changing language patterns on social media. Model adaptation involves training on historical tweets with labelled data and applying this knowledge to current tweets where labelled data is limited, enabling knowledge transfer for improved performance.

Future research should focus on transfer learning by using pre-trained models on large text corpora to address the limited labelled Twitter/X data challenge and improve model generalisation and performance. Moreover, Deep fusion techniques can also be explored by integrating multiple DL models with different architectures or data subsets to enhance accuracy and robustness. By addressing these challenges, researchers can enhance DL models for Twitter/X traffic event detection, resulting in improved traffic management, reduced congestion, and real-time commuter awareness.

Author contributions

CRediT: Danya Qutaishat conducted the literature review, analysed the data, draughted the manuscript, and revised it based on feedback. Songnian Li supervised the study, contributed to the conceptual design, critically reviewed the manuscript, and provided intellectual guidance.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was funded by the Natural Sciences and Engineering Research Council of Canada, grant number RGPIN-2017-05950.

Data availability statement

The data supporting the findings of this study come from public datasets, which are described in Section 2.2. Data sharing is not applicable to this article as no new data were created or analysed in this study.

References

Abad, A., O. Alfonso, T. António, M. Carmen García, M.H. Carlos D, P. Fernando, B. Fernando and M. Nuno, eds. 2016. Advances in Speech and Language Technologies for Iberian Languages. In *Lecture Notes in Computer Science*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-49169-1.

Afyouni, I., Z. A. Aghbari, and R. A. Razack. 2022. "Multi-Feature, Multi-Modal, and Multi-Source Social Event Detection: A Comprehensive Survey." *Information Fusion* 79: 279–308. https://doi.org/10.1016/j.inffus.2021.10.013.

Aminabadi Yazdani, R., S. Rajbhandari, A.A. Awan, C. Li, Du Li, E. Zheng, O. Ruwase, R. Y. Aminabadi, S. Smith, M. Zhang, J. Rasley, and Y. He. 2022. DeepSpeed-Inference: Enabling Efficient Inference of Transformer Models at Unprecedented Scale. In *SC22: International Conference for High Performance Computing, Networking, Storage and Analysis,* 1–15. Dallas, TX, USA: IEEE. https://doi.org/10.1109/SC41404.2022.00051

Alam, F., F. Ofli, and M. Imran. 2018. "CrisisMMD: Multimodal Twitter Datasets from Natural Disasters." *Proceedings of the International AAAI Conference on Web and Social Media* 12(1), https://doi.org/10.1609/icwsm.v12i1.14983.

Ali, F., D. Kwak, P. Khan, S. Islam, K. H. Kim, and K. S. Kwak. 2017. "Fuzzy Ontology-Based Sentiment Analysis of Transportation and City Feature Reviews for Safe Traveling." *Transportation Research Part C: Emerging Technologies* 77: 33–48. https://doi.org/10.1016/j.trc.2017.01.014.

Ali, F., A. Ali, M. Imran, R. Ali Naqvi, M. H. Siddiqi, and K.-S. Kwak. 2021. "Traffic Accident Detection and Condition Analysis Based on Social Networking Data." *Accident Analysis & Prevention* 151: 105973. https://doi.org/10.1016/j.aap.2021.105973.

Ali, F., D. Kwak, P. Khan, S. El-Sappagh, A. Ali, S. Ullah, K. H. Kim, and K.-S. Kwak. 2019. "Transportation Sentiment Analysis Using Word Embedding and Ontology-Based Topic Modeling." *Knowledge-Based Systems* 174: 27–42. https://doi.org/10.1016/j.knosys.2019.02.033.

Alifi, M. R. and S. H. Supangkat. 2018. Information Extraction of Traffic Condition from Social Media Using Bidirectional LSTM-CNN. In *2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI),* 637–640. https://doi.org/10.1109/ISRTI.2018.8864265

Almassar, R. R., and A. S. Girsang. 2022. "Detection of Traffic Congestion Based on Twitter Using Convolutional Neural Network Model." *IAES International Journal of Artificial Intelligence (IJ-AI)* 11 (4): 1448–1459. https://doi.org/10.11591/ijai.v11.i4.pp1448-1459.

Alomari, E., I. Katib, A. Albeshri, T. Yigitcanlar, and R. Mehmood. 2021. "Iktishaf+: A Big Data Tool with Automatic Labeling for Road Traffic Social Sensing and Event Detection Using Distributed Machine Learning." *Sensors* 21(9): 2993. https://doi.org/10.3390/s21092993.

Alomari, E. A., and R. Mehmood. 2023. "Smart Cities, Smarter Roads: A Review of Leveraging Cutting-Edge Technologies for Intelligent Event Detection from Social Media." *International Journal of Advanced Computer Science and Applications* 14(11): 364–374. https://doi.org/10.14569/ijacsa.2023.014113.

Alomari, E., R. Mehmood, and I. Katib. 2019. Road Traffic Event Detection Using Twitter Data, Machine Learning, and Apache Spark. In *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI),* 1888–1895. Leicester, UK: IEEE. https://doi.org/10.1109/SmartWorld-UIC-ATC-ScalCom-IOP-SCI.2019.00332

Ambastha, P. and M. S. Desarkar. 2020. Incident Detection From Social Media Targeting Indian Traffic Scenario Using Transfer Learning. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 1–6. Rhodes, Greece: IEEE. https://doi.org/10.1109/ITSC45102.2020.9294295

Anda, C., A. Erath, and P. J. Fourie. 2017. "Transport Modelling in the Age of Big Data." *International Journal of Urban Sciences* 21(sup1): 19–42. https://doi.org/10.1080/12265934.2017.1281150.

Asudani, D., N. K. Suresh, P. Nagwani, and Singh. 2023. "Impact of Word Embedding Models on Text Analytics in Deep Learning Environment: A Review." *Artificial Intelligence Review* 56(9): 10345–10425. https://doi.org/10.1007/s10462-023-10419-1.

Atefeh, F., and W. Khreich. 2013. "A Survey of Techniques for Event Detection in Twitter." *Computational Intelligence* 31(1): 132–164. https://doi.org/10.1111/coin.12017.

Azhar, A., S. Rubab, M. M. Khan, Y. A. Bangash, M. D. Alshehri, F. Illahi, and A. K. Bashir. 2022. "Detection and Prediction of Traffic Accidents Using Deep Learning Techniques." *Cluster Computing* 26(1): 477–493. https://doi.org/10.1007/s10586-021-03502-1.

Bilgin, M. and I. F. Şentürk. 2017. Sentiment Analysis on Twitter Data with Semi-Supervised Doc2Vec. In *2017 International Conference on Computer Science and Engineering (UBMK),* 661–666. Antalya, Turkey: IEEE.https://doi.org/10.1109/UBMK.2017.8093492

Blakey, E. 2024. "The Day Data Transparency Died: How Twitter/X Cut Off Access for Social Research." *Contexts* 23(2): 30–35. https://doi.org/10.1177/15365042241252125.

Bonandrini, R., and D. Gatti. 2024. "FastText (Sub)Word Vectors, *Reference Module in Social Sciences.* https://doi.org/10.1016/b978-0-323-95504-1.00032-6.

Burhanuddin, M. A. 2023. "Efficient Hardware Acceleration Techniques for Deep Learning on Edge Devices: A Comprehensive Performance Analysis." *KHWARIZMIA* 2023: 103–112. https://doi.org/10.70470/khwarizmia/2023/010.

Cao, S., S. Liu, T. Griggs, P. Schafhalter, X. Liu, Y. Sheng, J. E. Gonzalez, M. Zaharia, and I. Stoica. 2025. Moe-Lightning: High-Throughput MoE Inference on Memory-Constrained GPUs. In *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems,* 715–730.https://doi.org/10.1145/3669940.3707267

Chen, Q. and W. Wang. 2019. Multi-Modal Neural Network for Traffic Event Detection. In *2019 IEEE 2nd International Conference on Electronics and Communication Engineering (ICECE)*, 26–30. Xi'an, China: IEEE. https://doi.org/10.1109/ICECE48499.2019.9058508

Chen, P.-T., F. Chen, and Z. Qian. 2014. Road Traffic Congestion Monitoring in Social Media with Hinge-Loss Markov Random Fields. In *2014 IEEE International Conference on Data Mining.*80–89. Shenzhen, China: IEEE. https://doi.org/10.1109/ICDM.2014.139

Chen, Y., Y. Lv, X. Wang, L. Li, and F.-Y. Wang. 2018. Detecting Traffic Information From Social Media Texts With Deep Learning Approaches. In *IEEE Transactions on Intelligent Transportation Systems,*20: 3049–3058. IEEE.https://doi.org/10.1109/TITS.2018.2871269

Cottrill, C., P. Gault, G. Yeboah, J. D. Nelson, J. Anable, and T. Budd. 2017. "Tweeting Transit: An Examination of Social Media Strategies for Transport Information Management During a Large Event." *Transportation Research Part C: Emerging Technologies* 77: 421–432. https://doi.org/10.1016/j.trc.2017.02.008.

D'Andrea, E., P. Ducange, B. Lazzerini, and F. Marcelloni. 2015. Real-Time Detection of Traffic From Twitter Stream Analysis In *IEEE Transactions on Intelligent Transportation Systems*, 16: 2269–2283. IEEE.https://doi.org/10.1109/TITS.2015.2404431

Dabiri, S., and K. Heaslip. 2019. "Developing a Twitter-Based Traffic Event Detection Model Using Deep Learning Architectures." *Expert Systems with Applications* 118: 425–439. https://doi.org/10.1016/j.eswa.2018.10.017.

Deho, B.O., A. William Agangiba, L. Felix Aryeh, and A. Jeffery Ansah. Sentiment Analysis With Word Embedding. In *2018 IEEE 7th International Conference on Adaptive Science & Technology (ICAST),* 1–4.

Dhiman, A., and D. Toshniwal. 2020. "An Approximate Model for Event Detection From Twitter Data." *IEEE Access* 8: 122168–122184. https://doi.org/10.1109/ACCESS.2020.3007004.

Dinesh, L., B. Kuhaneswaran, and N. Ravikumar. 2023. A Multifaceted Machine Learning Approach to Understand Road Accident Dynamics Using Twitter Data. In *Handbook of Research on Advancements of Contactless Technology and Service Innovation in Library and Information Science*, 247–267. https://doi.org/10.4018/978-1-6684-7693-2.ch013

Dinga, R., P. Brenda W. J. H, V. Dick J, S. Lianne, and M. Andre F. 2019. "Beyond Accuracy: Measures for Assessing Machine Learning Models, Pitfalls and Guidelines." *bioRxiv*, 743138. https://doi.org/10.1101/743138.

Doguc, T. B., and A. A. Ahmet. 2023. "Tweet Toplama, Analiz ve Depolama İçin Platform Tasarımı (TweetCASP)." *Computer Science*, https://doi.org/10.53070/bbd.1344271.

Effrosynidis, D., G. Sylaios, and A. Arampatzis. 2024. "The Effect of Training Data Size on Disaster Classification From Twitter." *Information* 15(7): 393. https://doi.org/10.3390/info15070393.

ElSahly, O., and A. Abdelfatah. 2023. "An Incident Detection Model Using Random Forest Classifier." *Smart Cities* 6(4): 1786–1813. https://doi.org/10.3390/smartcities6040083.

Es Swidi, A., S. Ardchir, A. Daif, and M. Azouazi. 2023. "Road Users Detection for Traffic Congestion Classification." *Mathematical Modeling and Computing* 10(2): 518–523. https://doi.org/10.23939/mmc2023.02.518.

Fatichah, C., P. D. Sammy Wiyadi, D. Adni Navastara, N. Suciati, and A. Munif. 2020. Incident Detection Based on Multimodal Data From Social Media Using Deep Learning Methods. In *2020 International Conference on ICT for Smart Society (ICISS),* 1–6. Bandung, Indonesia: IEEE. https://doi.org/10.1109/ICISS50791.2020.9307555

Gannina Kumar, A., A. A. Ram, T. M. Jaffarullah, S. M. Reddy, A. S. Subba Reddy, S. Vikas, V. Mathi, and Ramalingam. 2024. "A New Approach to Road Incident Detection Leveraging Live Traffic Data: An Empirical Investigation." *Procedia Computer Science* 235: 2288–2296. https://doi.org/10.1016/j.procs.2024.04.217.

Garg, M., and M. Kumar. 2016. “Review on Event Detection Techniques in Social Multimedia.” *Online Information Review* 40(3): 347–361. https://doi.org/10.1108/OIR-08-2015-0281.
Gu, Y., Z. (S. Qian, and F. Chen. 2016. “From Twitter to Detector: Real-Time Traffic Incident Detection Using Social Media Data.” *Transportation Research Part C: Emerging Technologies* 67: 321–342. https://doi.org/10.1016/j.trc.2016.02.011.
Gu, W., A. Tandon, Y.-Y. Ahn, and F. Radicchi. 2021. “Principled Approach to the Selection of the Embedding Dimension of Networks.” *Nature Communications* 12(1): 3772. https://doi.org/10.1038/s41467-021-23795-5.
Hall, F. L., Y. Shi, and G. Atala. 1993. “On-Line Testing of the McMaster Incident Detection Algorithm Under Recurrent Congestion.” *Transportation Research Record* 1394: 1–7.
Hasan, S., and S. V. Ukkusuri. 2014. “Urban Activity Pattern Classification Using Topic Models From Online Geo-Location Data.” *Transportation Research Part C: Emerging Technologies* 44: 363–381. https://doi.org/10.1016/j.trc.2014.04.003.
Hasnat, M. M., and S. Hasan. 2018. “Identifying Tourists and Analyzing Spatial Patterns of Their Destinations From Location-Based Social Media Data.” *Transportation Research Part C: Emerging Technologies* 96: 38–54. https://doi.org/10.1016/j.trc.2018.09.006.
Hasnat, M., A. Mehedi, N. Faghih-Imani, S. Eluru, and Hasan. 2019. “Destination Choice Modeling Using Location-Based Social Media Data.” *Journal of Choice Modelling* 31: 22–34. https://doi.org/10.1016/j.jocm.2019.03.002.
Huang, A., L. Gallegos, and K. Lerman. 2017. “Travel Analytics: Understanding How Destination Choice and Business Clusters Are Connected Based on Social Media Data.” *Transportation Research Part C: Emerging Technologies* 77: 245–256. https://doi.org/10.1016/j.trc.2016.12.019.
Hussain, N. Y. 2024. “Deep Learning Architectures Enabling Sophisticated Feature Extraction and Representation for Complex Data Analysis.” *International Journal of Innovative Science and Research Technology (IJISRT)* 9(10): 2290–2300. https://doi.org/10.38124/ijisrt/ijisrt24oct1521.
Imran, M., C. Castillo, F. Diaz, and S. Vieweg. 2015. “Processing Social Media Messages in Mass Emergency.” *ACM Computing Surveys* 47(4): 1–38. https://doi.org/10.1145/2771588.
Jain, E., Neeraja, J., Banerjee, B., and Ghosh, P. 2022. A Diagnostic Approach to Assess the Quality of Data Splitting in Machine LearningarXiv preprint arXiv:2206.11721. https://doi.org/10.48550/arXiv.2206.11721
Jain, S., R. Pankaj, Z. Sharma, and Fatima. 2023. Traffic Rule Violation and Accident Detection Using CNN. In *International Conference on Innovative Computing and Communications*, 867–878.https://doi.org/10.1007/978-981-99-4071-4_66
Jiang, H., and H. Deng. 2020. “Traffic Incident Detection Method Based on Factor Analysis and Weighted Random Forest.” *IEEE Access* 8: 168394–168404. https://doi.org/10.1109/ACCESS.2020.3023961.
Jiang, J., C. Zhang, L. Ke, N. Hayes, Y. Zhu, H. Qiu, B. Zhang, T. Zhou, and G.-W. Wei. 2025. “A Review of Machine Learning Methods for Imbalanced Data Challenges in Chemistry.” *Chemical Science* 16(18): 7637–7658. https://doi.org/10.1039/D5SC00270B.
Johnson, J. M., and T. M. Khoshgoftaar. 2019. “Survey on Deep Learning With Class Imbalance.” *Journal of Big Data* 6(1): 27. https://doi.org/10.1186/s40537-019-0192-5.
Jonnalagadda, J. and M. Hashemi. 2021. A Deep Learning-Based Traffic Event Detection From Social Media. In *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, 1–8. Las Vegas, NV, USA: IEEE. https://doi.org/10.1109/IRI51335.2021.00007
Kamkarhaghighi, M., and M. Makrehchi. 2017. “Content Tree Word Embedding for Document Representation.” *Expert Systems with Applications* 90: 241–249. https://doi.org/10.1016/j.eswa.2017.08.021.
Karthik, B., D. Sai, D. NSSVV, M. A. Bharat, B. T. Alam, M. Sai, S. Rana, S. A. Sharma, and Yadav. 2023. Sentimental Technique Implementation on Textual Data. In *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*. 82–86, Gautam Buddha Nagar, India: IEEE. https://doi.org/10.1109/IC3I59117.2023.10397771
Kim, Y., S. Song, H. Lee, D. Choi, J. Lim, K. Bok, and J. Yoo. 2023. “Regional Traffic Event Detection Using Data Crowdsourcing.” *Applied Sciences* 13(16): 9422. https://doi.org/10.3390/app13169422.
Kisters, P., T. Bauer, W. Posdorfer, and J. Edinger. 2023. Real-Time Traffic Congestion Detection for Driver-Centric Applications. In *2023 IEEE 43rd International Conference on Distributed Computing Systems Workshops (ICDCSW)*, 163–168. Hong Kong, Hong Kong: IEEE. https://doi.org/10.1109/ICDCSW60045.2023.00029
Lee, S., S. Lee, J. Noh, J. Kim, and H. Jeong. 2023. “Special Traffic Event Detection: Framework, Dataset Generation, and Deep Neural Network Perspectives.” *Sensors (Basel, Switzerland)* 23(19): 8129. https://doi.org/10.3390/s23198129.
Li, L., Y. Dou, and J. Zhou. 2023. Traffic Accident Detection Based on Multimodal Knowledge Graphs. In *2023 5th International Conference on Robotics, Intelligent Control and Artificial Intelligence (RICAI)*, 644–647. Hangzhou, China: IEEE.https://doi.org/10.1109/RICAI60863.2023.10489808
Li, R., K. H. Lei, R. Khadiwala, and K. C.-C. Chang. 2012. TEDAS: A Twitter-Based Event Detection and Analysis System In *2012 IEEE 28th International Conference on Data Engineering (ICDE),* 1273–1276. Arlington, VA, USA: IEEE.https://doi.org/10.1109/ICDE.2012.125
Li, Y., L. Li, H. Jing, B. Ran, and D. Sun. 2020. “Examining imbalanced classification algorithms in predicting real-time traffic crash risk.” *Accident Analysis & Prevention* 144: 105610. https://doi.org/10.1016/j.aap.2020.105610.
Lin, L., M. Ni, Q. He, J. Gao, and A. W. Sadek. 2015. “Modeling the Impacts of Inclement Weather on Freeway Traffic Speed: Exploratory Study With Social Media Data.” *Transportation Research Record: Journal of the Transportation Research Board* 2482(1): 82–89. https://doi.org/10.3141/2482-11.

Lin, Yi., L. Li, H. Jing, B. Ran, and D. Sun. 2020. "Automated Traffic Incident Detection with a Smaller Dataset Based on Generative Adversarial Networks." *Accident Analysis & Prevention* 144: 105610. https://doi.org/10.1016/j.aap.2020.105610.

Liu, J., T. Li, P. Xie, S. Du, F. Teng, and X. Yang. 2020. "Urban Big Data Fusion Based on Deep Learning: An Overview." *Information Fusion* 53: 123–133. https://doi.org/10.1016/j.inffus.2019.06.016.

Liu, H., M. Sheng, Z. Sun, Y. Yao, X.-S. Hua, and H.-T. Shen. 2024. "Learning With Imbalanced Noisy Data by Preventing Bias in Sample Selection." *IEEE Transactions on Multimedia* 26: 7426–7437. https://doi.org/10.1109/TMM.2024.3368910.

Liu, Y., X. Shen, Y. Zhang, Z. Wang, Y. Tian, J. Dai, and Y. Cao. 2024. *A Systematic Review of Machine Learning Approaches for Detecting Deceptive Activities on Social Media: Methods, Challenges, and Biases*.arXiv preprint arXiv:2403.20293. https://doi.org/10.48550/arXiv.2403.20293

Lu, H., K. Shi, Y. Zhu, Y. Lv, and Z. Niu. 2018. "Sensing Urban Transportation Events from Multi-Channel Social Signals with the Word2vec Fusion Model." *Sensors (Basel, Switzerland)* 18(12): 4093. https://doi.org/10.3390/s18124093.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. "Distributed Representations of Words and Phrases and Their Compositionality." *Advances in Neural Information Processing Systems* 26: 3111–3119. Red Hook, NY: Curran Associates.

Lyu, B., H. Yuan, L. Lu, and Y. Zhang. 2022. "Resource-Constrained Neural Architecture Search on Edge Devices." *IEEE Transactions on Network Science and Engineering* 9(1): 134–142. https://doi.org/10.1109/TNSE.2021.3054583.

Maghrebi, M., A. Abbasi, and S. Waller. 2016. Transportation Application of Social Media: Travel Mode Extraction. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, 1648–1653. Rio de Janeiro, Brazil: IEEE. https://doi.org/10.1109/ITSC.2016.7795779

Mahmud, D., H. Hajmohamed, S. Almentheri, S. Alqaydi, L. Aldhaheri, R. A. Khalil, and N. Saeed. 2025. "Integrating LLMs with ITS: Recent Advances, Potentials, Challenges, and Future Directions." *IEEE Transactions on Intelligent Transportation Systems* 26(5): 5674–5709. https://doi.org/10.1109/TITS.2025.3528116.

Mannes, J. 2017. Facebook's fastText Library Is Now Optimized for MobileTechCrunch https://TechCrunch.com/2017/05/02/facebooks-fasttext-library-is-now-optimized-for-mobile/

Melhem, W., A. Abdi, and F. Meziane. 2024. "Traffic Detection and Forecasting from Social Media Data Using a Deep Learning-Based Model, Linguistic Knowledge, Large Language Models, and Knowledge Graphs." *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. 235–242. https://doi.org/10.5220/0013066900003838.

Mikolov, T., Quoc V, L., and Ilya, S. 2013. Exploiting Similarities Among Languages for Machine Translation." arXiv preprint arXiv:1309.4168. https://doi.org/10.48550/arXiv.1309.4168

Molloy, J., and R. Moeckel. 2017. "Improving Destination Choice Modeling Using Location-Based Big Data." *ISPRS International Journal of Geo-Information* 6(9): 291. https://doi.org/10.3390/ijgi6090291.

Mredula, M.S., C. Noyon, S.R. Md, M. Imtiaz, and C. You-Ze. 2022. "A Review on the Trends in Event Detection by Analyzing Social Media Platforms' Data." *Sensors (Basel, Switzerland)* 22(12): 4531. https://doi.org/10.3390/s22124531.

Münz, G., L. Sa, and C. Georg. 2007. "Traffic Anomaly Detection Using K-Means Clustering." *GI/ITG Workshop MMBnet* 7 (9). https://www.net.in.tum.de/projects/dfg-lupus/files/muenz07k-means.pdf.

Murtfeldt, R., Naomi, A., Ihsan, K., and Jevin D, W. 2024. RIP Twitter API: A Eulogy to Its Vast Research Contributions." arXiv preprint arXiv:2404.07340 https://doi.org/10.48550/arXiv.2404.07340

Nama, N., M. Sampson, N. Barrowman, R. Sandarage, K. Menon, G. Macartney, K. Murto, et al.2019. "Crowdsourcing the Citation Screening Process for Systematic Reviews: Validation Study." *Journal of Medical Internet Research* 21(4): e12953. https://doi.org/10.2196/12953.

Nejjari, F., L. Benhlima, and S. Bah. 2016. Event Traffic Detection Using Heterogeneous Wireless Sensors Network. In *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA),* 1–6. https://doi.org/10.1109/AICCSA.2016.7945825

Neruda, G. A. and E. Winarko. 2021. Traffic Event Detection From Twitter Using a Combination of CNN and BERT. In *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS),* 1–7. https://doi.org/10.1109/ICACSIS53237.2021.9631334

Ni, M., Q. He, and J. Gao. 2017. "Forecasting the Subway Passenger Flow Under Event Occurrences With Social Media." *IEEE Transactions on Intelligent Transportation Systems* 18(6): 1623–1632. https://doi.org/10.1109/TITS.2016.2611644.

Nirbhaya, M. A. W. and L. H. Suadaa. 2023. Traffic Incident Detection in Jakarta on Twitter Texts Using a Multi-Label Classification Approach. In *2023 International Conference on Computer, Control, Informatics and Its Applications (IC3INA),* 290–295. https://doi.org/10.1109/IC3INA60834.2023.10285731

Noori, M. A. R., and R. Mehra. 2020. "Traffic Congestion Detection from Twitter Using Word2vec." *In ICT Analysis and Applications*, 527–534. https://doi.org/10.1007/978-981-15-8354-4_52.

Olteanu, A., S. Vieweg, and C. Castillo. 2015. What to Expect When the Unexpected Happens: Social Media Communications Across Crises. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing,* 994–1009. https://doi.org/10.1145/2675133.2675242

Page, M. J., J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, et al. 2021. "The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews." *International Journal of Surgery* 88: 105906. https://doi.org/10.1016/j.ijsu.2021.105906.

Page, M. J., J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, and M. David. 2020. "Updating Guidance for Reporting Systematic Reviews: Development of the PRISMA 2020 Statement." *OSF Preprints*, https://doi.org/10.31222/osf.io/jb4dx.

Pennington, J., R. Socher, and C. D. Manning. 2014. GloVe: Global Vectors for Word Representation In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Pita, M. and G. L. Pappa. 2018. Strategies for Short Text Representation in the Word Vector Space. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*. 266–271. https://doi.org/10.1109/BRACIS.2018.00053

platforms' data." 22(12): 4531.

Poudel, A., and T. Weninger. 2024. "Navigating the Post-API Dilemma." *Proceedings of the ACM Web Conference 2024*, 2476–2484. https://doi.org/10.1145/3589334.3645503.

Puangnak, K., and N. Rachsiriwatcharabul. 2022. "Collection of Road Traffic Incidents in Bangkok from Twitter Data Based on Deep Learning Algorithm." *ECTI Transactions on Computer and Information Technology (ECTI-CIT* 16(3): 268–277. https://doi.org/10.37936/ecticit.2022163.248535.

Qazi, U., M. Imran, and F. Ofli. 2020. "GeoCoV19: A Dataset of Hundreds of Millions of Multilingual COVID-19 Tweets With Location Information." *SIGSPATIAL Special* 12(1): 6–15. https://doi.org/10.1145/3404820.3404823.

Qiu, J. 2024. "An Analysis of Model Evaluation With Cross-Validation: Techniques, Applications, and Recent Advances." *Advances in Economics, Management and Political Sciences* 99(1): 69–72. https://doi.org/10.54254/2754-1169/99/2024ox0213.

Qutaishat, D., and S. Li. 2025a. "A Transformer-Based Multi-Feature Fusion Method for Detecting Traffic Events Using Twitter Data." *Big Earth Data*. 1–33. https://doi.org/10.1080/20964471.2025.2564525. October.

Qutaishat, D., and S. Li. 2025b. "Multimodal Spatiotemporal Deep Fusion for Highway Traffic Accident Prediction in Toronto: A Case Study and Roadmap." *ISPRS International Journal of Geo-Information* 14(no. 11): 434. https://doi.org/10.3390/ijgi14110434.

Rajaraman, A. and J. D. Ullman. 2011. Data Mining. In *In Mining of Massive Datasets.* 1–17. Cambridge University Press.https://doi.org/10.1017/CBO9781139058452.002

Ramadhani, A. M. and H. S. Goo. 2017. Twitter Sentiment Analysis Using Deep Learning Methods. In *2017 7th International Annual Engineering Seminar (InAES)*,1–4. Yogyakarta, Indonesia: IEEE. https://doi.org/10.1109/INAES.2017.8068556

Raschka, S. 2018. *Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning*. arXiv preprint arXiv:1811.12808. https://doi.org/10.48550/arXiv.1811.12808

Raunak, V. 2017. *Simple and Effective Dimensionality Reduction for Word Embeddings*. arXiv preprint arXiv:1708.03629. https://doi.org/10.48550/arXiv.1708.03629

Rezaeinia, S. M., Ghodsi, A., and Rouhollah, R. 2017. Improving the Accuracy of Pre-Trained Word Embeddings for Sentiment Analysis." arXiv preprint arXiv:1711.08609 https://doi.org/10.48550/arXiv.1711.08609

Rudkowsky, E., M. Haselmayer, M. Wastian, M. Jenny, &A. Emrich, and M. Sedlmair. 2018. "More Than Bags of Words: Sentiment Analysis With Word Embeddings." *Communication Methods and Measures* 12(2–3): 140–157. https://doi.org/10.1080/19312458.2018.1455817.

Saai, K. P., and Vijayakumar, V. 2024. Resource-Efficient Transformer Architecture: Optimizing Memory and Execution Time for Real-Time Applications." arXiv preprint arXiv:2501.00042 https://doi.org/10.48550/arXiv.2501.00042

Saeed, Z., R. A. Abbasi, O. Maqbool, A. Sadaf, I. Razzak, A. Daud, N. R. Aljohani, and G. Xu. 2019. "What's Happening Around the World? A Survey and Framework on Event Detection Techniques on Twitter." *Journal of Grid Computing* 17(2): 279–312. https://doi.org/10.1007/s10723-019-09482-2.

Saeedi, R., M. Sankaranarayanasamy, R. Vishwakarma, P. Singh, and R. Vennelakanti. 2020. Towards Modular Modeling and Analytic for Multi-Modal Transportation Networks. In *2020 IEEE International Conference on Big Data (Big Data)*, 2426–2432. Atlanta, GA, USA: IEEE. https://doi.org/10.1109/BigData50022.2020.9377806

Safitri, A., I. Sekar, S. Wijayanto, and Hadiyoso. 2024. Improving Classification Accuracy with Preprocessing Techniques for Sentiment Analysis. In *2024 International Conference on Data Science and Its Applications (ICoDSA)*, 487–490. Kuta, Bali, Indonesia: IEEE. https://doi.org/10.1109/ICoDSA62899.2024.10651657

Sahni, A., and N. Raja. 2018. "Analyzation and Detection of Cyberbullying: A Twitter-Based Indian Case Study." *In Data Science and Analytics*, 484–497. https://doi.org/10.1007/978-981-10-8527-7_41.

Samant, A., and H. Adeli. 2000. "Feature Extraction for Traffic Incident Detection Using Wavelet Transform and Linear Discriminant Analysis." *Computer-Aided Civil and Infrastructure Engineering* 15(4): 241–250. https://doi.org/10.1111/0885-9507.00188.

Sampath, K. K., and M. Supriya. 2023. "Traffic Prediction in Indian Cities from Twitter Data Using Deep Learning and Word Embedding Models." *In Multi-Disciplinary Trends in Artificial Intelligence*, 671–682. https://doi.org/10.1007/978-3-031-36402-0_62.

Savvides, R., and J. Mäkelä Kai %J Statistical Analysis Puolamäki, and Data Mining: The ASA Data Science Journal 2023. "Model selection with bootstrap validation." 16 (2): 162–186. https://doi.org/10.1002/sam.11606.

Sethi, V., N. Bhandari, F. S. Koppelman, and J. L. Schofer. 1995. “Arterial Incident Detection Using Fixed Detector and Probe Vehicle Data.” *Transportation Research Part C: Emerging Technologies* 3(no. 2): 99–112. https://doi.org/10.1016/0968-090X(94)00017-Y. April.

Sethurajan, M. R., and N. K. 2023. “An Adept Approach to Ascertain and Elude Probable Social Bots Attacks on Twitter and Twitch Employing Machine Learning Approach.” *MethodsX* 11 (December), 102430. https://doi.org/10.1016/j.mex.2023.102430.

Suat-Rojas, N., C. Gutierrez-Osorio, and C. Pedraza. 2022. “Extraction and Analysis of Social Networks Data to Detect Traffic Accidents.” *Information* 13(no. 1): 26. https://doi.org/10.3390/info13010026.

Torregrossa, F., R. Allesiardo, V. Claveau, N. Kooli, and G. Gravier. 2021. “A Survey on Training and Evaluation of Word Embeddings.” *International Journal of Data Science and Analytics* 11(no. 2): 85–103. https://doi.org/10.1007/s41060-021-00242-8.

Tsou, M.-H., Zhang, H., and Jung, C.-T. 2017. Identifying Data Noises, User BiasesJung, Ming-Hsiang, and System Errors in Geo-Tagged Twitter Messages (Tweets).” arXiv preprint arXiv:1712.02433 https://doi.org/10.48550/arXiv.1712.02433

Walsh, R., and M. Tardy. 2022. “A Comparison of Techniques for Class Imbalance in Deep Learning Classification of Breast Cancer.” *Diagnostics* 13(no. 1): 67. https://doi.org/10.3390/diagnostics13010067.

Wandelt, S., C. Zheng, S. Wang, Y. Liu, and X. Sun. 2024. “Large Language Models for Intelligent Transportation: A Review of the State of the Art and Challenges.” *Applied Sciences* 14(no. 17): 7455. https://doi.org/10.3390/app14177455.

Wilson, S., W. Magdy, B. McGillivray, K. Garimella, and G. Tyson. 2020. Urban Dictionary Embeddings for Slang NLP Applications. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 4764–4773. European Language Resources Association (ELR). https://aclanthology.org/2020.lrec-1.586

Xu, S., S. Li, and R. Wen. 2018. “Sensing and Detecting Traffic Events Using Geosocial Media Data: A Review.” *Computers, Environment and Urban Systems* 72(November): 146–160. https://doi.org/10.1016/j.compenvurbsys.2018.06.006.

Yang, S. U. 2022. Anomaly Traffic Detection Based on LSTM. In *2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, 667–670. Chongqing, China: IEEE, June 17. https://doi.org/10.1109/itaic54216.2022.9836629

Zhang, Z., Q. He, J. Gao, and M. Ni. 2018. “A Deep Learning Approach for Detecting Traffic Accidents from Social Media Data.” *Transportation Research Part C: Emerging Technologies* 86: 580–596. https://doi.org/10.1016/j.trc.2017.11.027.

Zhang, S., D. Fu, W. Liang, Z. Zhang, B. Yu, P. Cai, and B. Yao. 2024. “TrafficGPT: Viewing, Processing and Interacting with Traffic Foundation Models.” *Transport Policy* 150: 95–105. https://doi.org/10.1016/j.tranpol.2024.07.004.