Integrating Social Media and Large Language Models for Real-Time Traffic Incident Detection

CEYLIN ECE, University of Twente, The Netherlands

This paper presents a novel methodology for detecting and classifying traffic incidents from X, formerly known as Twitter, posts and extracting their geographical information in a geocodable format using the Large Language Model (LLM) `Meta-LLama-3-8B-Instruct`. The prompt-based methodology consists of three phases. First, the X posts were classified as either 'Traffic Incident (TI)' or 'Not Traffic Incident (NTI)' and further categorized as 'Ongoing (O)' or 'Past (P)' by the LLM and evaluated against a manually labeled dataset. Second, the LLM was used to geo-parse the TI posts. The geo-parsed posts were geocoded using the `HERE Geocoding & Search API`. Third, the geocoded X posts were validated against traffic incident reports from the `HERE Traffic Incident API`. This methodology does not require preprocessing and training the classifiers. Instead, it presents a cost-effective, efficient, and scalable approach. The methodology achieved 98.2% accuracy in traffic incident classification and 91.6% in categorizing temporal status. Additionally, it geo-parsed 100% of TI X posts, achieving a geocoding accuracy of 98.5%. The methodology identified 61% of the traffic incidents earlier than HERE. The results demonstrate that LLMs are effective tools for detecting traffic incidents from unstructured social media data.

Additional Key Words and Phrases: real-time traffic incident detection, social media, Natural Language Processing, Large Language Models, geocoding

1 INTRODUCTION

1.1 Technical Background on Traffic Incidents

Traffic incidents can be categorized into two types: recurrent and nonrecurrent congestion. In nonrecurring traffic, the cause of congestion is inconsistent, which can be categorized into accidents, roadwork, hazards and weather, events, and obstacles from vehicles [4].

Several approaches have been developed over the years to address issues related to traffic incident detection. Some of these methods include utilizing loop detectors, Support Vector Machine (SVM), Random Forest (RF), and Long Short-Term Memory (LSTM) networks [2]. Additionally, multiple sensors are scattered throughout the transportation networks, and real-time traffic data is gathered by mining, allowing for the identification of traffic incidents [4]. Furthermore, new methods are emerging to gather information about traffic incidents, specifically through social media platforms like X. Considering the vast number of individuals on the website; there is a high likelihood that some users will post about ongoing traffic incidents, which numerous stakeholders, such as emergency responders and traffic management systems, could benefit from.

This paper presents a novel methodology for utilizing LLMs to detect and classify traffic incidents from unstructured X posts, categorize their temporal status, and extract geographical information in a structured format. The results are then validated against the 'ground-truth' provided by the HERE Traffic Incident API.

1.2 Contributions

Main contributions of this paper:

(1) A new and efficient approach for classifying unstructured social media data for traffic incident detection by using LLMs
(2) A new approach for geo-parsing unstructured social media data by using LLMs
(3) An extensive evaluation of the different prompting strategies for the classification of X posts
(4) An illustration of how X posts can be utilized to uncover traffic incidents that can go unnoticed by current traffic incident detection systems, such as HERE
(5) Confirming the early detection ability of X posts

1.3 Motivation

Traffic incidents can lead to increased carbon emissions, additional accidents, and fatalities; therefore, it is essential to detect and respond to these incidents quickly. Through social media data, incidents can be identified and addressed. Using LLMs' NLP capabilities, it is possible to detect traffic incidents. This will enhance the efficiency and effectiveness of responses to traffic incidents, thereby improving overall road safety.

1.3.1 Brief Definition of LLMs. Large Language Models (LLMs) are models built on the Transformer architecture, which undergo a pre-training process involving enormous text datasets, therefore enabling them to perform tasks such as generating human-like text, giving answers to questions presented, help translating and summarizing texts, and complete several NLP tasks in an accurate way [7]. Some examples of LLMs are GPT-4, Llama, BERT, and Mistral.

1.3.2 Description of NLP. Natural Language Processing (NLP) is a process that enables computers to understand human language and form communications with it [11]. For example, some NLP tasks include classification, text processing, text-to-text generation, chatbots, information retrieval, and document ranking.

1.3.3 Limitations of Traditional Traffic Incident Detection Approaches. Algorithmic incident detection is not cost-effective, as it is expensive to deploy sensors to achieve full coverage [4]. One of the most prominent limitations is that it cannot work effectively in local arterials [4].

1.4 Problem Statement

1.4.1 Extended Definition of Integrating Social Media and Large Language Models (LLMs) for Real-Time Traffic Incident Detection. X posts will be used to detect whether a traffic incident is occurring and to identify the location of the incident by using the LLM. LLM's

TScIT 43, July 4, 2025, Enschede, The Netherlands

function in this would be to complete NLP tasks. These tasks will be labeling the posts as a 'Traffic Incident (TI)' or 'Non-Traffic Incident (NTI)' and 'Ongoing (O)' or 'Past (P),' then geo-parsing the post so that it is ready to be geocoded using the HERE Geocoding & Search API. Afterward, it will be validated against the 'ground-truth' obtained from the HERE Traffic Incident API.

1.4.2 Challenges of Using Social Media Data as a Form of Identifying Traffic Incidents. X data is inherently obscure and non-standardized [4]. Unlike influential users (IUs), such as news agencies, who tend to post more standardized and detailed posts, it is unrealistic to expect regular users to provide detailed information and precise location data. Their language often contains slang and grammatical errors. Furthermore, how people describe traffic incidents could vary significantly, making it difficult to extract the information autonomously [4].

1.4.3 Description of Utilizing LLMs for Real-time Traffic Incident Detection Using X Data. This study aims to achieve higher efficiency by leveraging the capabilities of the LLM. The LLM will be asked to classify the X posts either as TI or NTI and O or P. The outcome will be validated against the manually labeled data. Afterward, the LLM will geo-parse the TI posts. Then, the geo-parsed posts will be geocoded using the HERE Geocoding & Search API. Finally, they would be ready to be validated against verified traffic incidents from HERE Traffic Incident API, which is considered the 'ground truth.' Thus, the accuracy could be examined.

1.4.4 The core research problems associated with this basic concept.

(1) How to classify unstructured X data as 'Traffic Incident (TI)' or 'Not-Traffic Incident (NTI)' and further categorize it as 'Ongoing (O)' or 'Past (P)' by using Large Language Models (LLMs) without any preprocessing or training, while validating the classification results against verified traffic incident data from HERE API?
(2) How can structured and segmented strings containing geographical information be effectively extracted from unstructured social media data using Large Language Models (LLMs) to enable geocoding?

1.5 Research objectives

1.5.1 General objective of this work. This study presents a novel methodology that automatically identifies traffic-related X posts, categorizes their temporal status, and extracts relevant geographic information using large language models (LLMs). By leveraging the strong NLP capabilities of such models, this research contributes to the design of a scalable and efficient traffic incident detection system.

1.5.2 Specific objective of this work. To design and evaluate an end-to-end methodology that leverages the LLM to automatically detect and classify traffic incidents from social media posts, extract geographic information, and validate the extracted data against official records.

2 PREVIOUS WORK

Several studies have attempted to leverage Twitter data to detect traffic incidents. Gu et al. [4] first acquired data from Twitter servers. They focused on two areas: Pittsburgh and Philadelphia. Next came adaptive data acquisition. Afterward, tweets were tokenized so that they could be classified as 'Traffic Incident (TI)' or 'Non-traffic Incident (NTI)' by using a Semi-Naive-Bayesian (SNB) Classifier. TI tweets were classified into their respective categories using Supervised Latent Dirichlet Allocation (sLDA). Next, a geo-parser was used to extract geographical information in a structured manner. Then, they were geocoded to generate the latitude and longitude coordinates. They validated the results of their classifier and geocoder against manually labeled tweets and data gathered from the 'Road Condition Report System (RCRS)' and 'Call For Service (CFS).' Finally, they validated the travel time of their data against the HERE API's travel time data. Gu et al.'s main findings were that the tweets they acquired from Twitter servers covered most of the traffic incidents in the existing dataset, provided broader coverage on arterial roads, and that individual users' tweets were less geocodable than those of influential users (IUs).

On the other hand, this study has some limitations; the dictionary of words that grows during the adaptive data acquisition process makes the classifier insensitive to potential typos and slang words. This leads to the inability to detect words and word combinations, resulting in incorrect tweet classification. Moreover, they rely on two different geo-parsers because one performs satisfactorily only in cases where the geographical information, such as road names or highway exit numbers, is clearly stated in the tweet. However, they cannot function properly in cases where only commonly indicated places, such as landmarks, are used; therefore, a second geo-parser is deployed for this specific task. Furthermore, even though the second geo-parser is used for fuzzy words, any typos on the tweets can still lead to potentially unnoticed geographical information that needs to be parsed or wrongly parsed geographical information.

Furthermore, other works, such as those by Salas et al. [9], and Jones et al. [6], acquired tweets from Twitter servers using the Twitter Streaming API from March 1, 2017, to May 31, 2017, while applying a geolocation filter. They looked at the tweets for the West Midlands, UK. They manually labeled their acquired tweets and divided them into training and testing sets. They used NLP to preprocess the tweets, such as tokenization and stop word removal. Afterward, while [9] utilized only the 'Support Vector Machine (SVM)' text classification algorithm to classify tweets, [6] used Ridge Classifier (RC), Naive Bayes (NB), k-Nearest Neighbour (kNN), Multilayer Perceptron (MLP), and Support Vector Machine (SVM). They evaluated their results using accuracy, precision, recall, and F1 score. Salas et al. [9] deduced that an accuracy of 88.27% could be achieved within their test set by using an SVM classifier. Additionally, Jones et al.'s [6] main discovery was that the most accurate classifier for this task was the Ridge Classifier (RC), which achieved an accuracy of 92.86%.

Nevertheless, both studies share similar limitations; the tokenization and stop-word removal can lead to a loss of the intended sentiment and context in the posts. The lost context can lead to misclassification of the posts. Moreover, text classification algorithms must

be explicitly trained, making them less scalable, as they would need to be retrained for newer datasets. Furthermore, both studies utilize small training sets. However, according to Ying [15], a small training set can result in 'noise learning,' meaning there is a high likelihood that the model can learn the noises. During the classification, they may lead to biases [15].

Lastly, Suat-Rojas et al.'s [12] work proposed a traffic incident detection methodology using Spanish tweets for Bogota data collected from October 2018 to July 2019. Their methodology comprised four steps: data collection (handled through Twitter API and manually labeled for training purposes), the classification of accident-related tweets (consisting of preprocessing, feature extraction, and classification by automatic classification model), entity recognition for location and time extraction, (including two steps: preprocessing, and sequence labeling), and finally geolocation of the incidents, (containing geocoding the extracted locations by using an app called 'Batch Geocode'). This study used the Support Vector Machine (SVM) model to classify tweets.

Similar to previous studies, one of the constraints of this study is that SVM is not scalable and requires retraining the model for new domains. Additionally, the preprocessing methodology cannot correct or interpret potential typos, which can result in wrongful classification.

Previous studies share similar limitations: they are labor-intensive, require training, and struggle to understand misspelled words and informal language when performing NLP tasks.

The proposed methodology is novel compared to the current literature and advances the state of the art. Unlike traditional classifiers that require training and data preprocessing, which necessitate significant computational time and resources, this study introduces a cost-effective, training and preprocessing-free prompt-based methodology using the LLM `Meta-Llama-3-8B-Instruct`. This results in a more accurate and faster method for detecting traffic incidents.

The study's novelty lies in its formulation of an LLM-based methodology to detect and classify traffic incidents, categorize temporal status, and geo-parse geographical information from X posts. Additionally, it evaluates the classification results of the LLM against validated traffic incident logs from HERE. This study aims to enhance traffic incident detection on a larger scale by using X posts, which traditional traffic incident detection methods often overlook.

3 BACKGROUND

For optimal outcomes in LLM tasks, the appropriate instructions must be provided [3]. The desired action from the model is to perform the task as intended without any confusion. To achieve this, one has to be as specific and concise as possible. In the case of an instruction that is too broad, the result will likely be unclear and redundant. Since the model lacks sufficient context to generate accurate results, the output will likely be superficial. Being descriptive also helps the model to stay relevant to the desired output. Moreover, [3] mentions 'role-based prompting,' which is LLMs emulating certain roles and providing task-specific responses. This research employs 'static role prompting,' which designates a predetermined role to elicit consistent responses. Additionally, [3] touches on the influence of 'delimiters for separation' to divide sections of the intended prompts so that complex prompts can be understood correctly.

This study leverages multiple prompting strategies. First, zero-shot prompting utilizes the capabilities of pre-trained LLMs while not providing any examples for them to train on [3]. The overarching idea is that one will solely rely on the model's knowledge from pre-training. Moreover, another strategy that [3] mentions is few-shot prompting. In this strategy, in contrast to zero-shot prompting, the model is provided with several examples to achieve better performance due to the now-included background information and guidance. Thirdly, [3] discusses chain-of-thought prompting. This strategy "involves providing intermediate reasoning steps to guide the responses of a model" [3]. This technique aims to enhance the model's capacity to perform complex reasoning tasks by guiding the model to take several intermediate thinking steps [14].

Besides the various prompting strategies, Chen et al. state that 'resampling,' running the exact prompt several times, can improve the possibility of getting higher-quality and more reliable outputs [3].

4 METHODOLOGY

The proposed prompt-based methodology using the LLM consisted of three distinct phases: detecting and classifying traffic incidents, extracting geographical information, and validating the results against real-world incident data.

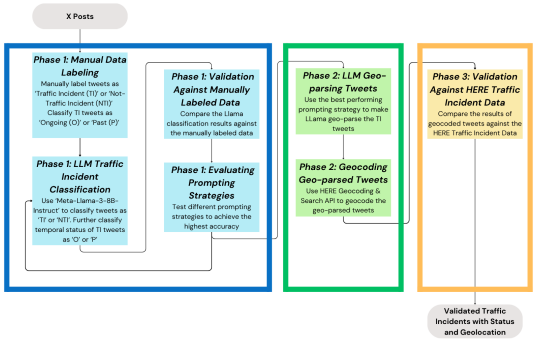

Fig. 1. Overview of the Prompt-Based Traffic Incident Detection Methodology

As illustrated in Figure 1 in blue, Phase 1 was comprised of four sub-phases. Firstly, X posts from the pre-existing dataset were manually labeled. Secondly, the LLM was used to classify traffic incidents and their respective temporal statuses. Thirdly, the results were validated against the manually labeled data. Fourth, various prompting strategies were chosen for testing; therefore, steps two, three, and four were repeated until the most satisfactory outcome was achieved.

Phase 2, indicated in green in Figure 1, included two sub-phases. These phases involved using the LLM to geo-parse the X posts with the best-performing prompting strategy and geocoding the geo-parsed posts with `HERE Geocoding & Search API`.

Finally, Phase 3, as shown in yellow in Figure 1, involved validating the results produced in Phase 2 against official traffic incident records from HERE.

4.1 Phase 1: Manually labeling and classification of X posts by LLM

X posts were acquired from a pre-existing dataset containing posts related to traffic from Los Angeles. The posts were manually labeled as 'Traffic Incident (TI)' or 'Not-Traffic Incident (NTI).' Additionally, they underwent temporal classification and were labeled as either 'Ongoing (O)' or 'Past (P)'. Afterward, several prompting strategies were tested on Llama to classify the posts as either TI or NTI and as O or P. To evaluate the performance of the LLM, 'True Positives (TP),' 'True Negatives (TN),' 'False Positives (FP),' and 'False Negatives (FN)' were measured, and the statistical metrics: accuracy, precision, recall, and f1-score were calculated.

4.1.1 Dataset. Asenov [1] created the pre-existing dataset consisting of X posts by using the 'search_recent_tweets' endpoint of X's Developer Basic API. The X posts were collected from Los Angeles from June 1, 2024, to June 14, 2024, using traffic-related keywords and street names sourced from real-time flow and incident data accessed via the HERE API. The query used was '((traffic OR road) AND (accident OR congestion OR jam)) AND (<street name> (OR <street name>)*).' The original CSV file contained 2,845 X posts and consisted of three columns: `id`, `created_at`, and `text`. These columns included the post ID, the time it was created, and the post itself. Notably, posts in this dataset were not geotagged.

4.1.2 Manually Labeling Data. The X posts gathered from the pre-existing dataset were manually labeled by following the approach taken by [4]. The labeling criterion for posts was as follows: a traffic incident was defined as non-recurrent congestion arising from irregular factors, including traffic accidents, work zones, adverse weather events, and special events [4]. TI posts included information regarding a traffic incident, suggested an anomaly in the transportation system, or signaled a probable significant traffic disruption; otherwise, it was NTI.

A 'Status' column was added to assist model categorization. The temporal status of traffic incidents in posts was defined as 'Past (P)' if the context implied that the incident had occurred or been resolved, 'Ongoing (O)' if the post suggested that the situation was active, used present-tense language, or did not indicate whether the traffic incident was resolved. For the status category, only TI posts were classified; the status of NTI posts was left empty.

At the end of the manual classification, out of 2,845 posts, 2,468 were classified as TI, and 377 were classified as NTI. Of 2,468 TI posts, 2,176 were classified as O, and 292 were classified as P.

4.1.3 Llama Implementation for Phase 1. The model used for this research was 'meta-llama/Meta-Llama-3-8B-Instruct', an open-source LLM. The 'Instruct' version of this model was used for classification since it is instruction-tuned. This model used Hugging Face's 'transformers' pipeline, specifically the 'Transformers AutoModelForCausalLM.'

The X posts were uploaded as a CSV file, and the model processed each row of the CSV file separately, which contained the columns `id`, `created_at`, and `text`. This meant that the model classified posts individually rather than in batches. Once the posts were classified, the model provided predictions in the specified format. Afterward, the model's predictions were parsed to filter out the actual results for its predictions. These outputs were then written into another CSV file containing the new columns: `traffic_incident` and `status`; therefore, they could be evaluated against the manually labeled posts.

4.1.4 Prompt Decision. A two-dimensional approach was employed: first, structural design principles, including task, context, persona, exemplar, format, and tone, and second, prompting strategies, such as zero-shot, few-shot, and chain-of-thought prompting.

The prompts adhered to the basic structural design principles from Chen et al. [3] and were combined with prompting strategies, including 'zero-shot,' 'few-shot,' and 'chain-of-thought' to achieve the most accurate output.

To find the appropriate prompt style for the task, eight different prompts were tested:

(1) Zero-shot without persona B.1.1
   (a) This prompt included context, format, and tone; however, the persona element was not included. Essentially, this prompt did not follow the technique of role-based prompting and tested whether it could still achieve the precision that would be accomplished through a model with a fixed role.
(2) Zero-shot without context B.1.2
   (a) This prompt included persona, format, and tone; however, the context element was not included to test whether the model could adequately complete the task even without providing any context, such as what traffic incident meant and how it was categorized.
(3) Zero-shot - extensive B.1.3
   (a) This prompt included context, persona, format, and tone. The purpose of this prompt was to determine if the model could perform better if it were assigned both a fixed role and provided with the necessary context.
(4) Zero-shot - extensive & shortened B.1.4
   (a) Like strategy 3, this prompt contained context, persona, format, and tone. However, to increase performance, the information provided in each section was made more concise to prevent overwhelming the model.
(5) Zero-shot with chain-of-thought B.1.5
   (a) To build on strategy 4, chain-of-thought prompting was introduced to the prompt, with the line 'Let's think step by step' under a new section called 'Reasoning Steps.'
(6) Few-shot B.1.6
   (a) In this prompt, context, persona, format, and tone were included, as well as exemplars. This prompt assessed whether the examples and guidance could enhance the model's performance. For the first phase, three examples were provided. One for 'Traffic Incident (TI)': 1, and 'Status': Ongoing (O). Another one for 'Traffic Incident (TI)': 1, and 'Status': Past (P). And the last one, for 'Traffic Incident (TI)': 0, and 'Status': NONE. Therefore, the model gained a good grasp of how X posts were classified. For this prompt, the examples

used were removed from the CSV file provided for Llama to classify, thereby preventing evaluation bias.

(7) Few-shot with chain-of-thought B.1.7
   (a) This prompt included a chain-of-thought prompting to build on strategy 6. Similar to strategy 5, the line 'Let's think step by step' under the 'Reasoning Steps' section was included.
(8) Few-shot with chain-of-thought - 2 B.1.8
   (a) The purpose of this trial was to determine whether re-sampling affected accuracy. Therefore, the prompt from strategy 7 was kept and executed a second time.

The metrics used to assess the performance for classifying traffic incidents were accuracy, precision, recall, specificity, and F1-score.

Definition 4.1. Accuracy expresses the division of correctly guessed labels by the total number of labels [8].

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

Definition 4.2. Precision indicates the number of guessed positives that are correct. It looks at the division of rightly classified positives by everything classified as positive [8].

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Definition 4.3. Recall, or true positive rate, demonstrates the division of correctly guessed positives by all actual positives [8].

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

Definition 4.4. Specificity, or true negative rate, demonstrates the division of correctly guessed negatives by all actual negatives [8].

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4}$$

Definition 4.5. F1-score is the harmonic mean between recall and precision. It is a way of demonstrating the precision and robustness of the classifier [8].

$$\text{F1-score} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{5}$$

The metrics used to evaluate the performance for classifying status were accuracy, precision (macro, micro, and weighted), recall (macro, micro, and weighted), and F1-score (macro, micro, and weighted).

Definition 4.6. The micro-averaged score evaluates all classes using all true positives, false negatives, and false positives [5]. (See Appendix A.1.1 for detailed examples of the calculations)

Definition 4.7. The macro-averaged score independently computes the specified metric score for each label and averages them while assigning equal weight to each class [5]. (See Appendix A.1.2 for detailed examples of the calculations)

Definition 4.8. The weighted-averaged score calculates the specified metric score for each class independently, similar to the macro-averaged score, except the average is weighted based on true cases each class contains [5]. (See Appendix A.1.3 for detailed examples of the calculations)

4.2 Phase 2: Geo-parsing and geocoding TI Posts

TI posts from Phase 1 were used for geo-parsing and geocoding. Llama was prompted to extract the geographical information in structured and segmented strings. Since the best performance outcome was observed for 'few-shot w/CoT' in Phase 1, the same approach was taken in this phase to geo-parse the X posts. These strings were passed to the HERE Geocoding API to acquire latitude and longitude information.

4.2.1 Geo-parsing TI posts. To ensure a reliable format for geocoding, the same data structure used in Gu et al.'s [4] geo-parser was employed. The prompt followed the same structure as in Phase 1, a 'few-shot w/CoT' (See Appendix B.2.1 for the detailed prompt).

For the task, the LLM was instructed to geo-parse both the locations of posts that contain road names, intersection names, highway exit numbers, and highway mile markers. In addition, it was instructed to extract the geographical coordinates of posts that only contained points of interest. In the task, the meaning of the keys was explained. Therefore, the model could better understand, for instance, what 'road1:' corresponded to.

Regarding the exemplars, the same X posts were utilized from Phase 1. This phase used the two TI posts to demonstrate what the geo-parsing could look like. Similar to Phase 1, these posts were not included in the CSV file that Llama processed.

4.2.2 Llama Implementation for Phase 2. TI posts were geo-parsed; therefore, a separate CSV file was created beforehand that solely contained TI posts from the output of Phase 1. Afterward, the prompt was run. Pattern matching was applied to extract the values from the Llama output. Consequently, an output CSV file with the columns `road1`, `road2`, `road3`, `hwy1`, `hwy2`, `hwy3`, `hwy1-mm1`, `hwy1-mm2`, `relational-word`, `original-text` was produced in addition to the already existing columns.

4.2.3 Geocoding geo-parsed posts. The successfully geo-parsed posts were filtered out, and the outputs for `road1`, `road2`, `road3`, `hwy1`, `hwy2`, `hwy3`, `hwy1-mm1`, and `hwy1-mm2` columns were merged. Afterward, they were sent to the HERE Geocoding API individually, specifying the city, state, and country from which the posts originate. From the HERE Geocoding API, fields `title`, `id`, `resultType`, `houseNumberType`, `address`, `position`, `access`, `mapView`, and `scoring` were requested.

4.3 Phase 3: Validation against HERE API Data

This Phase evaluated Phase 2, testing the accuracy of the geocoded TI X posts. The posts' latitude, longitude, and time information was compared against the HERE API's verified traffic data, testing spatial and temporal accuracy.

4.3.1 HERE Traffic Incident Dataset. A 'ground truth' obtained through the HERE Traffic API was used to validate the overall performance of the methodology. This dataset contained traffic incident information with the columns: `id`, `hrn`, `originalId`, `originalHrn`, `startTime`, `endTime`, `entryTime`, `roadClosed`, `criticality`, `type`, `codes`, `description`, `summary`, `location_length`, and `location_points`. The specific fields that were focused on are

`startTime`, which indicated the start time of the incident; `entryTime`, which indicated when the incident was recorded in the HERE system; and `location_points`, which contained an array of latitude and longitude information. This dataset had 30,519 traffic incident logs recorded from June 1, 2024, to June 14, 2024.

4.3.2 Validation with Existing HERE Incidents Data. To validate the geocoding output against the existing traffic incident data, Gu et al.'s [4] approach was followed. [4] stated that the decided reporting time discrepancy was 30 minutes, and the spatial radius, meaning the distance, was 1 mile (≈ 1.6 kilometers). Therefore, in this study, the same temporal and spatial radii, 30 minutes and 1.6 kilometers, were used for testing. To validate whether a post accurately reported a traffic incident identified by HERE detection systems, it was ensured that the post's `created_at` fell within the 15 minutes before or after the `startTime` of the traffic incident. Additionally, the traffic incident log was located within 1.6 kilometers of the post. If the post satisfied both conditions, it was declared that the incident matching was successful; otherwise, it was deemed unsuccessful.

Furthermore, this phase also tested the early detection performance of X posts against HERE's incident detection system. If incident matching was a success, the post's `created_at` was compared with HERE's log's `entryTime`. If the `created_at` was earlier than `entryTime`, it was indicated that the X post detected the incident quicker than HERE.

5 RESULTS

5.1 Performance Evaluation for Phase 1

It was observed that the few-shot approaches yielded better outcomes in traffic incident classification, particularly in terms of accuracy, precision, specificity, and F1 score.

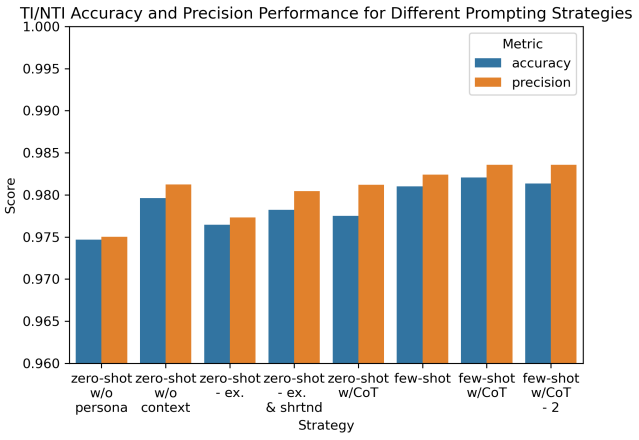

Fig. 2. The accuracy and precision scores of all prompting strategies for traffic incident classification.

Figure 2 highlights the accuracy and precision scores for each prompting strategy, displaying which strategy was more robust in traffic incident classification. Table 1 contains more detailed outputs for the evaluation metrics. A critical detail was the difference in the accuracy of traffic incident classification between 'zero-shot w/o persona,' 'zero-shot w/o context,' and 'zero-shot - extensive.' Moreover, the difference between the outcome for accuracy in 'zero-shot w/o persona' and 'zero-shot w/o context' and the decrease exhibited in the 'zero-shot - extensive' suggests the extensive information supplied overwhelmed the model, causing it to lose its memory of the intended task. Therefore, more detailed descriptions do not necessarily equate to better performance. To support this claim, the outcome of status classification should be examined.

As shown in Table 2, 'zero-shot w/o persona' had lower accuracy compared to 'zero-shot w/o context.' Mainly, concerning macro precision, recall, and F-1 score, it can be observed that there was a substantial decrease. The same decline happened between 'zero-shot w/o context' and 'zero-shot - extensive.' The explanation for this prominent decrease was the existence of 'NoResult' produced by the model. The model produced 'NoResult' for the cases where it could not classify the temporal status as 'O' or 'P.' While the model did not produce any 'NoResult's for 'zero-shot w/o context,' it produced one 'NoResult' for both 'zero-shot w/o persona' and 'zero-shot - extensive,' further solidifying the complication it started to have. To address the misclassification that Llama's confusion may have caused, the persona and context were shortened to make the task clearer, resulting in 'zero-shot - extensive & shortened.' This approach resulted in improved performance regarding traffic incident classification. In terms of temporal classification, it resulted in zero 'NoResult,' eliminating the issue of confusion during classification. Moreover, in the previous three methodologies, the LLM had a substantial overestimation; however, with this newer approach, the overestimation decreased. Finally, for zero-shot approaches, 'zero-shot w/CoT' was tested. Regarding the evaluation metrics, it performed similarly to 'zero-shot - extensive & shortened.' Similarly, Figure 3 represents the confusion matrix for 'few-shot w/CoT-2'.

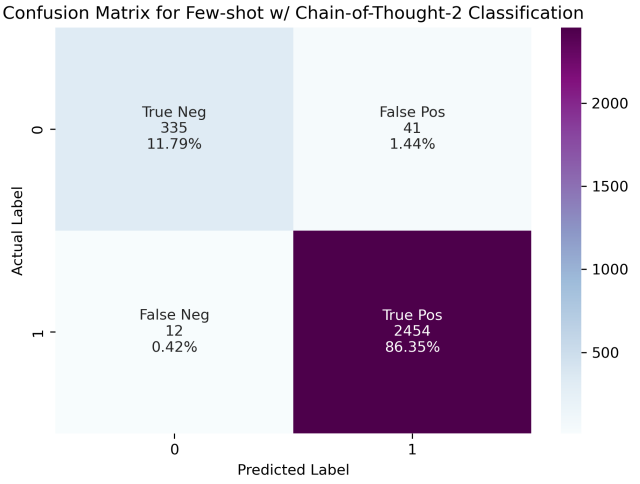

Fig. 3. The confusion matrix of few-shot with chain-of-thought -2 for traffic incident classification.

The few-shot prompts performed better for traffic incident and temporal status classification than for the zero-shot approaches.

Strategy	Accuracy	Precision	Recall	Specificity	F1_score
zero-shot w/o persona	0.975	0.975	0.996	0.833	0.986
zero-shot w/o context	0.980	0.981	0.996	0.875	0.988
zero-shot - extensive	0.976	0.977	0.996	0.849	0.987
zero-shot - extensive & shortened	0.978	0.980	0.995	0.870	0.988
zero-shot w/CoT	0.978	0.981	0.993	0.875	0.987
few-shot	0.981	0.982	0.996	0.883	0.989
few-shot w/CoT	0.982	0.984	0.996	0.891	0.990
few-shot w/CoT - 2	0.981	0.984	0.995	0.891	0.989

Table 1. Performance of Prompting Strategies on Traffic Incident Classification

Strategy	Accuracy	Precision			Recall			F-1 Score		
		Macro	Micro	Weighted	Macro	Micro	Weighted	Macro	Micro	Weighted
zero-shot w/o persona	0.906	0.629	0.906	0.927	0.659	0.906	0.906	0.634	0.906	0.912
zero-shot w/o context	0.917	0.925	0.917	0.917	0.746	0.917	0.917	0.796	0.917	0.903
zero-shot - extensive	0.911	0.632	0.911	0.928	0.657	0.911	0.911	0.638	0.911	0.916
zero-shot - extensive & shortened	0.909	0.837	0.909	0.933	0.896	0.909	0.909	0.853	0.909	0.916
zero-shot w/CoT	0.909	0.833	0.909	0.932	0.894	0.909	0.909	0.851	0.909	0.916
few-shot	0.916	0.847	0.916	0.934	0.896	0.916	0.916	0.862	0.916	0.921
few-shot w/CoT	0.911	0.840	0.911	0.932	0.895	0.911	0.911	0.856	0.911	0.917
few-shot w/CoT - 2	0.913	0.841	0.913	0.932	0.893	0.913	0.913	0.857	0.913	0.919

Table 2. Performance of Prompting Strategies on Status Classification

A nuance emerged with the comparison of 'few-shot w/CoT' and 'few-shot w/CoT-2'. In terms of the evaluation metrics, 'few-shot w/CoT' had higher accuracy for traffic incident classification with 98.2% compared to 'few-shot w/CoT-2"s 98.1% as seen from Figure 3; however, as for overall performance, 'few-shot w/CoT-2' performed better, with slightly higher results in temporal status classification, supporting the claim that resampling results in quality and dependable outcomes.

5.2 Performance Evaluation for Phase 2

As a result of the geo-parsing prompt, the model delivered results for all 2,495 of the TI posts, indicating that it encountered no complications, even with the fuzzy words, and could produce at least some geographical information.

Regarding geocoding, a post was considered successfully geocoded if the API returned a response to the requested fields. To track the success of the geocoding, a new field, `geocoding_success`, was created, and it was indicated as 'yes' if the API returned a response and 'no' if the API failed to do so. As a result, 2,457 of 2,495 (98.5%) geo-parsed posts were successfully geocoded.

5.3 Performance Evaluation for Phase 3

In the final phase, the comparison with the HERE-verified traffic incidents resulted in an 'incident matching' success rate of 236 (9.6%) out of 2,457. Therefore, 9.6% of the successfully geocoded posts matched HERE-verified incident data records and were confirmed by HERE. This demonstrated that by utilizing X posts, more traffic incidents were identified; specifically, 90.4% of traffic incidents in the X post dataset were not identified by HERE with the predetermined parameters (see section 4.3.2).

For the early detection performance, out of 236 matching incidents, X posts detected 145 (61%) of the incidents earlier than HERE.

6 DISCUSSION

6.1 Interpretation of Results

In terms of traffic incident classification, this study's approach yields higher performance compared to previous literature, with the highest accuracy of 98.2% with the 'few-shot w/CoT' approach and the lowest accuracy of 97.5% with the 'zero-shot w/o persona' strategy. For instance, with their Semi-Naive-Bayes classifier, Gu et al. [4] achieved an accuracy of 90.5% for traffic incident classification. Furthermore, Salas et al. [9] employed a Support Vector Machine (SVM) classifier, achieving an overall accuracy of 88.27% on their test dataset. Moreover, Jones et al. [6] reported an overall accuracy of 92.86% when using a Ridge Classifier on their test data. Additionally, Suat-Rojas et al. [12] reported achieving 96.8% accuracy using a Support Vector Machine (SVM).

A key contribution of this study is saving the valuable time and resources expended on preprocessing and training, thereby simplifying the traffic incident detection task and potentially making the response to traffic incidents faster and more scalable.

Overall, few-shot approaches yield better outcomes than zero-shot approaches in traffic incident and temporal status classification. However, upon reviewing previous studies, zero-shot approaches remain an effective means of traffic incident detection, offering improvements over previously suggested methods.

Moreover, zero-shot approaches demonstrate that even though adequate context is necessary, the surplus of information results in complications and a decrease in performance, as the ample instructions lead to the model's confusion about its primary task. Specifically, the low performance in 'zero-shot w/o persona' and 'zero-shot - extensive' compared to 'zero-shot w/o context' and 'zero-shot - extensive & shortened' illustrates that the prompts should include all the information needed but should be concise.

The study contributes to the prompt engineering literature by assessing the performance of various prompting strategies and evaluating the results against the manually labeled dataset. The results demonstrate how different approaches to prompting yielded varying outcomes.

For geo-parsing and geocoding of the TI posts, the 'few-shot w/CoT' strategy for geo-parsing yields promising results, with Llama managing to geo-parse 100% of the TI posts. This performance demonstrates that Llama is a suitable replacement for the two geo-parsers proposed by Gu et al. [4]. In terms of geocoding, this study demonstrates that with Gu et al.'s [4] geo-parsing fields and the inclusion of city, state, and country, X posts can be successfully geocoded using the HERE Geocoding API. With a 98.5% success in geocoding, Llama's reliability as a geo-parser is solidified.

This study enables the LLM to handle the geo-parsing task; therefore, the trade-off between extracting geographical information, such as road names and numbers, and extracting geographical information from points of interest becomes nonexistent. Both are extracted from unstructured data without requiring any sets or multiple geo-parsers.

Ultimately, it is demonstrated that X posts can identify the traffic incidents uncovered by HERE. However, they can also find more unidentified traffic incidents. With 90.4% of additional cases where X posts successfully detect traffic incidents using the predetermined parameters (see Section 4.3.2), it can be stated that X data can be used as an additional form of traffic incident detector. In addition to uncovering newer instances, X posts serve as an earlier means of detecting incidents, identifying 61% of incidents earlier than HERE.

6.2 Limitations

Although this study yielded promising results, it had several limitations. One of the shortcomings of the proposed system was that the 'ground truth' in incident classification assessment was a manually labeled post dataset. The manual labeling process can introduce bias and subjectivity, therefore affecting the accuracy assessment of the LLM. Additionally, this step hinders the processing of larger datasets, as the manual labeling process is time-consuming. Furthermore, the X post classification was performed using a dataset limited to Los Angeles, covering the period from June 1, 2024, to June 14, 2024. Therefore, the conclusions cannot be generalized for data in which posts are in different languages, include different slang words, or exhibit dissimilar X post patterns and incident types. Moreover, even though the LLM-based methodology can process the input X post dataset in real-time, this study did not test the real-time incoming data, meaning that observing the methodology's behavior and assessing its performance for live post streams was not feasible.

6.3 Future Work

The proposed methodology can be further improved by assessing the scalability limitations of manual labeling through the use of active learning with human assistance. For instance, the model can be trained using a small dataset. Afterward, by manually labeling the posts that the model is unsure of, including them in the labeled dataset, and then retraining the model, the need for manually labeling the entire dataset can be overcome. Furthermore, in future research, the generalizability of the current approach can be tested by using posts from various locations. Additionally, by deploying the X Streaming API, the behavior of the methodology can be examined. Furthermore, LLM can be tasked with classifying incident types for further clarity on the incident. Moreover, for future research in newer and broader directions, LLMs can be utilized to make traffic flow predictions beyond detecting traffic incidents. The LLM can be tasked with predicting the jamFactor, which ranges from 0 to 10, based on the contents of the posts. Afterward, this prediction can be evaluated against the verified HERE reports, which contain jamFactor and are attained from the HERE Traffic API. Depending on the results, the potential fields in which LLMs can be deployed can grow.

7 CONCLUSIONS

This research proposed an efficient and novel approach for detecting traffic incidents from unstructured X posts. It aimed to enhance incident detection by leveraging social media data with Large Language Models (LLMs). The performance of the LLM was evaluated against manually labeled posts and verified traffic incident data from the HERE API, which served as the 'ground truth.'

Regarding the first research question, the unstructured X data was classified as 'Traffic Incident (TI)' or 'Not Traffic Incident (NTI),' as well as further categorized as 'Ongoing (O)' or 'Past (P)' by utilizing a prompt-based LLM methodology. Regarding the various strategies tested in the study, the approach with the highest performance was 'few-shot w/CoT-2', yielding 98.1% accuracy in traffic incident classification and 91.3% accuracy in status categorization against the manually labeled data. The results were validated against verified traffic incident logs from HERE API and produced a 9.6% match.

As for the second research question, a few-shot approach combined with a chain-of-thought, 'few-shot w/CoT,' effectively extracted structured and segmented strings containing geographical information. The LLM replaced the need for a geo-parser. It performed this process meticulously, extracting information such as road names and highway exit numbers in a format suitable for geocoding for 100% of social media data. 98.5% of social media data could be successfully geocoded using the HERE Geocoding API.

In summary, the research outcomes demonstrated that LLMs were effective tools for detecting traffic incidents. In addition to identifying traffic incidents, they were powerful tools for categorizing temporal status. They also excelled at geo-parsing tasks that required multiple geo-parsers to be used individually, indicating that they could replace traditional methods and produce reliable results. The proposed methodology also illustrated that X data could be utilized as a valuable early detection tool for traffic incident detection.

REFERENCES

[1] D. Asenov. 2024. A Data-driven Approach to Study the Influence of Social Media on Human Behaviours in Transportation. http://essay.utwente.nl/100967/

[2] Meryem Ayou, Youssef Trardi, Loqman Chakir, and Jaouad Boumhidi. 2023. Data-driven traffic incident detection in urban roads based on machine learning algorithms. In *E3S Web of Conferences*, Vol. 469. EDP Sciences, 00102.

[3] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2025. Unleashing the potential of prompt engineering for large language models. *Patterns* (2025).

[4] Yiming Gu, Zhen Sean Qian, and Feng Chen. 2016. From Twitter to detector: Real-time traffic incident detection using social media data. *Transportation research part C: emerging technologies* 67 (2016), 321–342.

[5] Maria Cristina Hinojosa Lee, Johan Braet, and Johan Springael. 2024. Performance metrics for multilabel emotion classification: comparing micro, macro, and weighted f1-scores. *Applied Sciences* 14, 21 (2024), 9863.

[6] Angelica Salas Jones, Panagiotis Georgakis, Yannis Petalas, and Renukappa Suresh. 2018. Real-time traffic event detection using Twitter data. *Infrastructure Asset Management* 5, 3 (2018), 77–84.

[7] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences* 103 (2023), 102274.

[8] Jude Chukwura Obi. 2023. A comparative study of several classification metrics and their performances on data. *World Journal of Advanced Engineering Technology and Sciences* 8, 1 (2023), 308–314.

[9] Angelica Salas, Panagiotis Georgakis, and Yannis Petalas. 2017. Incident detection using data from social media. In *2017 IEEE 20th International conference on intelligent transportation systems (ITSC)*. IEEE, 751–755.

[10] Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information processing & management* 45, 4 (2009), 427–437.

[11] Cole Stryker and Jim Holdsworth. 2024. What Is NLP (Natural Language Processing)? | IBM.

[12] Nestor Suat-Rojas, Camilo Gutierrez-Osorio, and Cesar Pedraza. 2022. Extraction and analysis of social networks data to detect traffic accidents. *Information* 13, 1 (2022), 26.

[13] Kanae Takahashi, Kouji Yamamoto, Aya Kuchiba, and Tatsuki Koyama. 2022. Confidence interval for micro-averaged F 1 and macro-averaged F 1 scores. *Applied Intelligence* 52, 5 (2022), 4961–4972.

[14] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[15] Xue Ying. 2019. An overview of overfitting and its solutions. In *Journal of physics: Conference series*, Vol. 1168. IOP Publishing, 022022.

A APPENDIX A

A.1 Micro, Macro, and Weighted Averaged Scores

A.1.1 Micro-Averaged Scores. The micro-averaged precision is computed as shown [13]:

$$Precision_{\text{micro}} = \frac{\sum_{i=1}^{r} TP_i}{\sum_{i=1}^{r} (TP_i + FP_i)} \tag{6}$$

The micro-averaged recall is computed as shown [13]:

$$Recall_{\text{micro}} = \frac{\sum_{i=1}^{r} TP_i}{\sum_{i=1}^{r} (TP_i + FN_i)} \tag{7}$$

The micro-averaged F1-score is computed as shown [5]:

$$F1_{\text{micro}} = \frac{2 \cdot \text{Precision}_{\text{micro}} \cdot \text{Recall}_{\text{micro}}}{\text{Precision}_{\text{micro}} + \text{Recall}_{\text{micro}}} \tag{8}$$

A.1.2 Macro-Averaged Scores. The macro-averaged precision is computed as shown [13]:

$$Precision_{\text{macro}} = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FP_i} \tag{9}$$

The macro-averaged recall is computed as shown [13]:

$$Recall_{\text{macro}} = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FN_i} \tag{10}$$

The macro-averaged F1-score is computed as shown [5]:

$$F1_{\text{macro}} = \frac{1}{N} \sum_{i=1}^{N} \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{11}$$

A.1.3 Weighted-Averaged Scores. The weighted-averaged precision is computed as shown [10]:

$$Precision_{\text{weighted}} = \sum_{i=1}^{N} w_i \cdot \text{Precision}_i \tag{12}$$

The weighted-averaged recall is computed as shown [10]:

$$Recall_{\text{weighted}} = \sum_{i=1}^{N} w_i \cdot \text{Recall}_i \tag{13}$$

The weighted-averaged F1-score is computed as shown [5]:

$$F1_{\text{weighted}} = \sum_{i=1}^{N} w_i \cdot \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{14}$$

where w_i is defined as [5]:

$$w_i = \frac{TP_i + FN_i}{\sum_{j=1}^{N} (TP_j + FN_j)} \tag{15}$$

B APPENDIX B

B.1 Prompts used for Phase 1

B.1.1 zero-shot w/o persona. """«CONTEXT» According to Gu et al. (2016), traffic incidents are 'non-recurrent congestion is induced by non-recurring causes, such as traffic accidents, work zones, adverse weather events, and special events, which takes about half of the total congestion.' 'A TI tweet is one contains information about a traffic incident, implies abnormality on the transportation infrastructure, or indicates a (potentially) significant disruption to the traffic'(Gu et al., 2016). There are 5 categories of traffic incidents: 1. Accidents: traffic accidents such as collision. 2. Road work: the scheduled or unplanned road work. 3. Hazards & Weather. 4. Events: special events such as Marathon. 5. Obstacle Vehicles. The temporal status of traffic incidents in tweets is defined as: - 'Past' if the context implies that the incident has occurred or been resolved. - 'Ongoing' if the tweet implies that the situation is active, uses present-tense language, or does not indicate if the traffic incident is resolved.

«TASK» Your task is: 1. Classify the tweets into 'Traffic-Incident' or 'Not Traffic-Incident' by labeling them as 1 if 'Traffic-Incident' and 0 if 'Not Traffic_incident'. 2. For the 'Traffic-Incident' tweets, meaning 1, classify them as 'Ongoing' or 'Past' by labeling them as O if 'Ongoing' and P if 'Past'; for 'Not Traffic_Incident' tweets, meaning 0, label them as NONE.

«TONE» Use a confident and formal tone. Avoid uncertain answers.

«FORMAT» Follow the format: TI: [0 or 1], Status: [O, P, or NONE] """

B.1.2 zero-shot w/o context. """«PERSONA» You are an operator in a Traffic Management Center (TMC), and your job entails detecting and analyzing traffic incidents from posted tweets in real-time. Your job requires you to do more than just traditional monitoring; it requires real-time social media monitoring, incident detection, and verification.

«TASK» Your task is: 1. Classify the tweets into 'Traffic-Incident' or 'Not Traffic-Incident' by labeling them as 1 if 'Traffic-Incident' and 0 if 'Not Traffic_incident'. 2. For the 'Traffic-Incident' tweets, meaning 1, classify them as 'Ongoing' or 'Past' by labeling them as O if 'Ongoing' and P if 'Past'; for 'Not Traffic_Incident' tweets, meaning 0, label them as NONE.

«TONE» Use a confident and formal tone. Avoid uncertain answers.

«FORMAT» Follow the format: TI: [0 or 1], Status: [O, P, or NONE] """

B.1.3 zero-shot - extensive. """«PERSONA» You are an operator in a Traffic Management Center (TMC), and your job entails detecting and analyzing traffic incidents from posted tweets in real-time. Your job requires you to do more than just traditional monitoring; it requires real-time social media monitoring, incident detection, and verification.

«CONTEXT» According to Gu et al. (2016), traffic incidents are 'non-recurrent congestion is induced by non-recurring causes, such as traffic accidents, work zones, adverse weather events, and special events, which takes about half of the total congestion.' 'A TI tweet is one contains information about a traffic incident, implies abnormality on the transportation infrastructure, or indicates a (potentially) significant disruption to the traffic'(Gu et al., 2016). There are 5 categories of traffic incidents: 1. Accidents: traffic accidents such as collision. 2. Road work: the scheduled or unplanned road work. 3. Hazards & Weather. 4. Events: special events such as Marathon. 5. Obstacle Vehicles. The temporal status of traffic incidents in tweets is defined as: - 'Past' if the context implies that the incident has occurred or been resolved. - 'Ongoing' if the tweet implies that the situation is active, uses present-tense language, or does not indicate if the traffic incident is resolved.

«TASK» Your task is: 1. Classify the tweets into 'Traffic-Incident' or 'Not Traffic-Incident' by labeling them as 1 if 'Traffic-Incident' and 0 if 'Not Traffic_incident'. 2. For the 'Traffic-Incident' tweets, meaning 1, classify them as 'Ongoing' or 'Past' by labeling them as O if 'Ongoing' and P if 'Past'; for 'Not Traffic_Incident' tweets, meaning 0, label them as NONE.

«TONE» Use a confident and formal tone. Avoid uncertain answers.

«FORMAT» Follow the format: TI: [0 or 1], Status: [O, P, or NONE] """

B.1.4 zero-shot - extensive & shortened. """«PERSONA» You are an operator in a Traffic Management Center (TMC), and your job entails detecting and analyzing traffic incidents from posted tweets in real-time. Your job requires more than just traditional monitoring; it requires real-time social media monitoring, incident detection, and verification.

«CONTEXT» According to Gu et al. (2016), traffic incidents are 'non-recurrent congestion is induced by non-recurring causes, such as traffic accidents, work zones, adverse weather events, and special events, which takes about half of the total congestion.' 'A TI tweet is one contains information about a traffic incident, implies abnormality on the transportation infrastructure, or indicates a (potentially) significant disruption to the traffic'(Gu et al., 2016). The temporal status of traffic incidents in tweets is defined as: - 'Past' if the context implies that the incident has occurred or been resolved. - 'Ongoing' if the tweet implies that the situation is active, uses present-tense language, or does not indicate if the traffic incident is resolved.

«TASK» Your task is: 1. Label the tweets: - 1 if it is 'Traffic-Incident' - 0 if it is 'Not Traffic_incident' 2. For the 'Traffic-Incident' tweets, meaning 1, label their temporal status as: - O if it is 'Ongoing' - P if it is 'Past' 3. For 'Not Traffic_Incident' tweets, meaning 0, label their temporal status as: - NONE

«TONE» Use a confident and formal tone. Avoid uncertain answers.

«FORMAT» Follow the format: TI: [0 or 1], Status: [O, P, or NONE] """

B.1.5 zero-shot w/CoT. """«PERSONA» You are an operator in a Traffic Management Center (TMC), and your job entails detecting and analyzing traffic incidents from posted tweets in real-time. Your job requires more than just traditional monitoring; it requires real-time social media monitoring, incident detection, and verification.

«CONTEXT» According to Gu et al. (2016), traffic incidents are 'non-recurrent congestion is induced by non-recurring causes, such as traffic accidents, work zones, adverse weather events, and special events, which takes about half of the total congestion.' 'A TI tweet is one contains information about a traffic incident, implies abnormality on the transportation infrastructure, or indicates a (potentially) significant disruption to the traffic'(Gu et al., 2016). The temporal status of traffic incidents in tweets is defined as: - 'Past' if the context implies that the incident has occurred or been resolved. - 'Ongoing' if the tweet implies that the situation is active, uses present-tense language, or does not indicate if the traffic incident is resolved.

«TASK» Your task is: 1. Label the tweets: - 1 if it is 'Traffic-Incident' - 0 if it is 'Not Traffic_incident' 2. For the 'Traffic-Incident' tweets, meaning 1, label their temporal status as: - O if it is 'Ongoing' - P if it is 'Past' 3. For 'Not Traffic_Incident' tweets, meaning 0, label their temporal status as: - NONE

«REASONING STEPS» Let's think step by step.

«TONE» Use a confident and formal tone. Avoid uncertain answers.

«FORMAT» Follow the format: TI: [0 or 1], Status: [O, P, or NONE] """

B.1.6 few-shot. """«PERSONA» You are an operator in a Traffic Management Center (TMC), and your job entails detecting and analyzing traffic incidents from posted tweets in real-time. Your job requires more than just traditional monitoring; it requires real-time social media monitoring, incident detection, and verification.

«CONTEXT» According to Gu et al. (2016), traffic incidents are 'non-recurrent congestion is induced by non-recurring causes, such as traffic accidents, work zones, adverse weather events, and special events, which takes about half of the total congestion.' 'A TI tweet is one contains information about a traffic incident, implies abnormality on the transportation infrastructure, or indicates a (potentially)

significant disruption to the traffic'(Gu et al., 2016). The temporal status of traffic incidents in tweets is defined as: - 'Past' if the context implies that the incident has occurred or been resolved. - 'Ongoing' if the tweet implies that the situation is active, uses present-tense language, or does not indicate if the traffic incident is resolved.

«TASK» Your task is: 1. Label the tweets: - 1 if it is 'Traffic-Incident' - 0 if it is 'Not Traffic_incident' 2. For the 'Traffic-Incident' tweets, meaning 1, label their temporal status as: - O if it is 'Ongoing' - P if it is 'Past' 3. For 'Not Traffic_Incident' tweets, meaning 0, label their temporal status as: - NONE

«EXEMPLARS» Tweet: Accident in #Broadway-SlavicVlg on Broadway Ave at Mcbride Avenue. Reported by C-COMS #traffic https://t.co/IXfI2jI9uh TI: 1, Status: O

Tweet: Accident cleared in #DowntownLA on US-101 (Santa Ana Freeway) NB at Alameda St, stopped traffic back to the 5 #LAtraffic https://t.co/e2e6MWlmlg TI: 1, Status: P

Tweet: @BPL_Transport Having experienced significant congestion delays due to traffic signal timings at Talbot Road/Dickson Road junction eastbound can I suggest that services 3/3A divert across Talbot Rd to Abingdon Street to turn right into Queen St and left onto Dickson Rd. TI: 0, Status: NONE

«TONE» Use a confident and formal tone. Avoid uncertain answers.

«FORMAT» Follow the format: TI: [0 or 1], Status: [O, P, or NONE] """

B.1.7 few-shot w/CoT. """«PERSONA» You are an operator in a Traffic Management Center (TMC), and your job entails detecting and analyzing traffic incidents from posted tweets in real-time. Your job requires more than just traditional monitoring; it requires real-time social media monitoring, incident detection, and verification.

«CONTEXT» According to Gu et al. (2016), traffic incidents are 'non-recurrent congestion is induced by non-recurring causes, such as traffic accidents, work zones, adverse weather events, and special events, which takes about half of the total congestion.' 'A TI tweet is one contains information about a traffic incident, implies abnormality on the transportation infrastructure, or indicates a (potentially) significant disruption to the traffic'(Gu et al., 2016). The temporal status of traffic incidents in tweets is defined as: - 'Past' if the context implies that the incident has occurred or been resolved. - 'Ongoing' if the tweet implies that the situation is active, uses present-tense language, or does not indicate if the traffic incident is resolved.

«TASK» Your task is: 1. Label the tweets: - 1 if it is 'Traffic-Incident' - 0 if it is 'Not Traffic_incident' 2. For the 'Traffic-Incident' tweets, meaning 1, label their temporal status as: - O if it is 'Ongoing' - P if it is 'Past' 3. For 'Not Traffic_Incident' tweets, meaning 0, label their temporal status as: - NONE

«EXEMPLARS» Tweet: Accident in #Broadway-SlavicVlg on Broadway Ave at Mcbride Avenue. Reported by C-COMS #traffic https://t.co/IXfI2jI9uh TI: 1, Status: O

Tweet: Accident cleared in #DowntownLA on US-101 (Santa Ana Freeway) NB at Alameda St, stopped traffic back to the 5 #LAtraffic https://t.co/e2e6MWlmlg TI: 1, Status: P

Tweet: @BPL_Transport Having experienced significant congestion delays due to traffic signal timings at Talbot Road/Dickson Road junction eastbound can I suggest that services 3/3A divert across Talbot Rd to Abingdon Street to turn right into Queen St and left onto Dickson Rd. TI: 0, Status: NONE

«REASONING STEPS» Let's think step by step.

«TONE» Use a confident and formal tone. Avoid uncertain answers.

«FORMAT» Follow the format: TI: [0 or 1], Status: [O, P, or NONE] """

B.1.8 few-shot w/CoT - 2. """«PERSONA» You are an operator in a Traffic Management Center (TMC), and your job entails detecting and analyzing traffic incidents from posted tweets in real-time. Your job requires more than just traditional monitoring; it requires real-time social media monitoring, incident detection, and verification.

«CONTEXT» According to Gu et al. (2016), traffic incidents are 'non-recurrent congestion is induced by non-recurring causes, such as traffic accidents, work zones, adverse weather events, and special events, which takes about half of the total congestion.' 'A TI tweet is one contains information about a traffic incident, implies abnormality on the transportation infrastructure, or indicates a (potentially) significant disruption to the traffic'(Gu et al., 2016). The temporal status of traffic incidents in tweets is defined as: - 'Past' if the context implies that the incident has occurred or been resolved. - 'Ongoing' if the tweet implies that the situation is active, uses present-tense language, or does not indicate if the traffic incident is resolved.

«TASK» Your task is: 1. Label the tweets: - 1 if it is 'Traffic-Incident' - 0 if it is 'Not Traffic_incident' 2. For the 'Traffic-Incident' tweets, meaning 1, label their temporal status as: - O if it is 'Ongoing' - P if it is 'Past' 3. For 'Not Traffic_Incident' tweets, meaning 0, label their temporal status as: - NONE

«EXEMPLARS» Tweet: Accident in #Broadway-SlavicVlg on Broadway Ave at Mcbride Avenue. Reported by C-COMS #traffic https://t.co/IXfI2jI9uh TI: 1, Status: O

Tweet: Accident cleared in #DowntownLA on US-101 (Santa Ana Freeway) NB at Alameda St, stopped traffic back to the 5 #LAtraffic https://t.co/e2e6MWlmlg TI: 1, Status: P

Tweet: @BPL_Transport Having experienced significant congestion delays due to traffic signal timings at Talbot Road/Dickson Road junction eastbound can I suggest that services 3/3A divert across Talbot Rd to Abingdon Street to turn right into Queen St and left onto Dickson Rd. TI: 0, Status: NONE

«REASONING STEPS» Let's think step by step.

«TONE» Use a confident and formal tone. Avoid uncertain answers.

«FORMAT» Follow the format: TI: [0 or 1], Status: [O, P, or NONE] """

B.2 Prompt used for Phase 2

B.2.1 geo-parsing prompt. """«PERSONA» You are an operator in a Traffic Management Center (TMC), and your job entails geo-parsing the traffic-incident tweets in real time.

«CONTEXT» "A geo-parser is a machine that receives input of a string and produces a structured and segmented strings that contain only geographical information" (Gu et al., 2016).

«TASK» Your task is: - Geo-parse the tweets, so that they are ready for geocoding. - Process the names of points of interest referred to in tweets and parse those words relevant to locations. Data structure

of the geoparser (Keys: Meaning): road1: The road name mentioned road2: The second road name mentioned road3: The third road name mentioned hwy1: The highway name mentioned hwy2: The second highway name mentioned hwy3: The third highway name mentioned hwy1-mm1: The starting mile-marker/exit number of the highway hwy1-mm2: The ending mile-marker/exit number of the highway relational-word: The relational word used like “near”, “cross”, “intersection”, etc. original-text: The original tweet text

«EXEMPLARS» Tweet: Accident in #Broadway-SlavicVlg on Broadway Ave at Mcbride Avenue. Reported by C-COMS #traffic https://t.co/IXfI2jI9uh road1: Broadway Ave road2: Mcbride Avenue road3: hwy1: hwy2: hwy3: hwy1-mm1: hwy1-mm2: relational-word: at original-text: Accident in #Broadway-SlavicVlg on Broadway Ave at Mcbride Avenue. Reported by C-COMS #traffic https://t.co/IXfI2jI9uh

Tweet: Accident cleared in #DowntownLA on US-101 (Santa Ana Freeway) NB at Alameda St, stopped traffic back to the 5 #LAtraffic https://t.co/e2e6MWlmlg road1: Alameda St road2: road3: hwy1: US-101 (Santa Ana Freeway) NB hwy2: hwy3: hwy1-mm1: hwy1-mm2: relational-word: at original-text: Accident cleared in #DowntownLA on US-101 (Santa Ana Freeway) NB at Alameda St, stopped traffic back to the 5 #LAtraffic https://t.co/e2e6MWlmlg

«REASONING STEPS» Let’s think step by step.

«TONE» Use a confident and formal tone. Avoid uncertain answers.

«FORMAT» Follow the format: road1: [...] road2: [...] road3: [...] hwy1: [...] hwy2: [...] hwy3: [...] hwy1-mm1: [...] hwy1-mm2: [...] relational-word: [...] original-text: [...] """

C APPENDIX C

C.1 Usage of AI

ChatGPT and Grammarly were used during the thesis process to identify grammatical errors and inconsistencies in writing. Furthermore, ChatGPT is utilized for assisting in coding the graphs, which are Figure 2 and Figure 3. All ideas, data analysis, and interpretations presented in the thesis belong to the author.