SAT-Geo: A Social Sensing based Content-only Approach to Geolocating Abnormal Traffic Events using Syntax-based Probabilistic Learning

Lanyu Shang[a,*], Yang Zhang[b,*], Christina Youn[b], Dong Wang[a,**]

[a]*School of Information Sciences*
University of Illinois Urbana-Champaign, Champaign, IL, USA
[b]*Department of Computer Science and Engineering*
University of Notre Dame, Notre Dame, IN, USA

Abstract

Social sensing has become an emerging and pervasive sensing paradigm to collect timely observations of the physical world from human sensors. In this paper, we study the problem of geolocating abnormal traffic events using social sensing. Our goal is to infer the location (i.e., geographical coordinates) of the abnormal traffic events by exploring the location entities from the content of social media posts. Two critical challenges exist in solving our problem: i) how to accurately identify the location entities related to the abnormal traffic event from the content of social media posts? ii) How to accurately estimate the geographic coordinates of the abnormal traffic event from the set of identified location entities? To address the above challenges, we develop a Social sensing based Abnormal Traffic Geolocalization (SAT-Geo) framework to accurately estimate the geographic coordinates of abnormal traffic events by exploring the syntax-based patterns in the content of social media posts and the geographic information associated with the location entities from the social media posts. We evaluate the SAT-Geo framework on three real-world Twitter datasets collected from New York City, Los Angeles, and London. Evaluation results demonstrate

*The first two authors contributed equally to this work.
**Corresponding author
Email addresses: lshang3@illinois.edu (Lanyu Shang), yzhang42@nd.edu (Yang Zhang), cyoun@nd.edu (Christina Youn), dwang24@illinois.edu (Dong Wang)

that SAT-Geo significantly outperforms state-of-the-art baselines by effectively identifying location entities related to the abnormal traffic events and accurately estimating the geographic coordinates of the events.

Keywords: Syntax-based Learning, Abnormal Detection, Geolocalization, Social Sensing

---

1. Introduction

With the proliferation of mobile devices and the ubiquitous network connections, social sensing has become an emerging and pervasive sensing paradigm to collect timely observations of the physical world from human sensors [1]. Examples of social sensing applications include post-disaster damage assessment with social media user posts [2], urban environment monitoring using input from citizen scientists [3], and smart health condition tracing using wearable devices [4]. Real-time traffic monitoring is an important application of social sensing in intelligent transportation systems (ITS), where timely social media posts are collected to acquire real-time traffic situation awareness (e.g., road congestion, traffic accident) of an urban area. Comparing to traditional infrastructure-based solutions (e.g., surveillance cameras, radar sensors), social sensing provides an infrastructure-free solution that is more pervasive and scalable [5]. In this paper, we focus on the problem of identifying the geographic coordinates (i.e., latitude and longitude coordinates) of abnormal traffic events reported on social media. We refer to this problem as *social sensing based abnormal traffic event geolocalization*. The identified geographic coordinates information of abnormal traffic events can be utilized to provide effective precautions (e.g., traffic accident alerts) and timely responses (e.g., emergency medical rescue for severe traffic accidents) for improving traffic safety and efficiency [6].

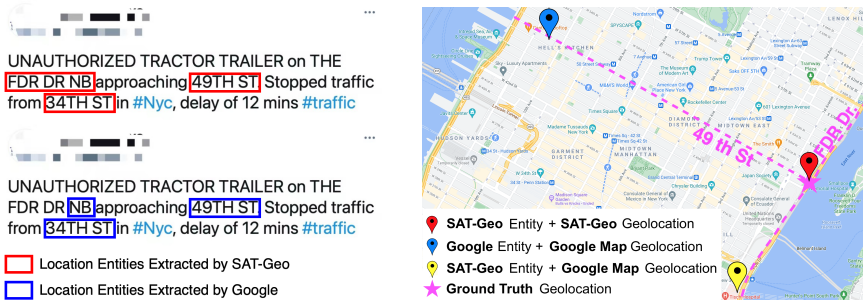
Many efforts have been made to study the problem of event localization using social media data [7, 8, 9, 10, 11, 12, 13, 14]. These solutions can be mainly categorized into two categories: *geotagging-based solutions* [7, 8, 15] and *content-based solutions* [9, 10, 11, 12, 13, 14]. However, these solutions are insufficient to

fully address the problem of fine-grained abnormal traffic event geolocalization. First, the geotagging-based solutions that leverage the geotagging information associated with social media posts (e.g., "coordinates" field of a tweet[1]) often suffer from two critical limitations. On one hand, the geotagging information of social media posts is sparse due to the privacy concerns of users (e.g., fewer than 0.5% tweets have geotags [16]). On the other hand, the geotagging information of a social media post may not always represent the real geolocation of the reported event (e.g., a user may travel a few blocks away from the accident site after he/she finishes editing the post) [17]. Second, existing content-based solutions are also impractical to solve our problem. This is because these solutions often require auxiliary information that is not always available (e.g., private user activities, user's previous posts), and the inferred event locations are often inaccurate (e.g., the estimated event location in current solutions can only reduce the average error distance to about 10 km [15]). Therefore, the problem of geolocalizing abnormal traffic events using social sensing data feeds remains to be a challenging problem to be addressed.

In this paper, we develop a social sensing based solution to directly infer the geographic coordinates of abnormal traffic events from the content of social media data (e.g., tweets). Our design is to first identify the location entities in the social media post (i.e., the named entities in a social media post that indicate the location of the abnormal traffic event) by exploring the syntax of the post content. We then leverage the identified location entities to accurately infer the geographic coordinates of the abnormal traffic event by investigating the geographic information of these location entities and their relations. An example of our abnormal traffic event geolocalization problem is shown in Figure 1. Our goal is to infer the event geolocation (e.g., the geographic coordinates marked with the red pin in Figure 1(b)) using the location entities identified in the text of the tweet (e.g., the location entities highlighted in red boxes in Figure 1(a)). However, it is not a trivial task to accurately geolocate the abnormal traffic

[1]https://developer.twitter.com/en/docs/tutorials/filtering-tweets-by-location.html

event from the content of social media posts due to two important challenges that are elaborated on below.



(a) Location Entities Extraction (b) Accident Geolocation Estimation

Figure 1: Example of Abnormal Traffic Event Geolocalization

Content-only Location Entity Inference. The first challenge lies in how we can accurately identify the location entities related to the abnormal traffic event from the content of social media posts. A possible approach to address the aforementioned issue of sparse geotagging information is to infer the event location by analyzing the content of social media posts [18]. However, the limited and unstructured content in a social media post (e.g., 280 characters in a tweet) makes the location entity inference problem challenging [19]. For example, the essential location entity "FDR DR NB" in Figure 1(a) is misidentified as "NB" (in blue box) by the state-of-the-art entity extraction method *Google Named Entity Detection service*[2] and leads to the inaccurate geolocation (i.e., the blue pin in Figure 1(b)). In addition, existing solutions for event localization often require external information (e.g., using the content of abnormal traffic event posts to retrieve the geotagging information in geotagged tweets with similar content [8]). However, such external information may not always be available. For example, our case study in New York City shows that most traffic incidents are only reported by a single tweet. Therefore, these solutions are insufficient to fundamentally address the content-only location entity inference challenge in

[2]https://cloud.google.com/natural-language/

abnormal traffic event geolocalization.

Fine-grained Geolocation Estimation. The second challenge lies in how to accurately estimate the geographic coordinates of the abnormal traffic event by leveraging the location entities identified in a social media post. Existing solutions for geolocation estimation often utilize a grid-based method that divides the map area of interest into a set of grids of equal size and estimates the event geolocation in terms of the grid (e.g., the center of the estimated grid) [20, 15]. However, the estimated geolocation of interest is often coarse-grained and is not precise enough for estimating the geographic coordinates of abnormal traffic events. For example, a single grid in an urban area (e.g., New York City in Figure 1(b)) could contain more than a hundred roads and intersections. It is challenging to accurately identify the geographic coordinates of the abnormal traffic event given such a non-trivial amount of roads and intersections in the grid. Therefore, it remains a challenging task to accurately estimate the fine-grained geolocation of the abnormal traffic event.

To address the above challenges, we develop a Social sensing based Abnormal Traffic Geolocalization framework (SAT-Geo) that can accurately estimate the geographic coordinates of abnormal traffic events using syntax-based probabilistic learning. In particular, to address the first challenge, we design a syntax-based pattern learning module to extract the syntax-based patterns and encode the extracted patterns using a novel probabilistic representation method. To address the second challenge, we develop a distance-aware geolocation estimation module to effectively estimate the geographic coordinates of the abnormal traffic event location using a point-based map representation. We evaluate SAT-Geo on three real-world Twitter datasets collected from the three large cities in the world: New York City, Los Angeles, and London. Evaluation results show that SAT-Geo significantly outperforms state-of-the-art baselines by effectively extracting location entities related to the abnormal traffic events and accurately estimating the corresponding geolocation.

A preliminary version of this work has been published in ASONAM 2019 [21] to study the problem of identifying location entities for abnormal traffic event

localization which is an initial step in geolocating abnormal traffic events. This paper is a significant extension of our conference paper (i.e., SyntaxLoc) in the following aspects. First, we focus on an abnormal traffic event geolocalization problem where the goal is to infer the geographic coordinates from the fuzzy description in social media posts instead of only extracting the location entities as we studied in the conference paper (Section 1 and 3). Second, we develop a new SAT-Geo framework to address the fine-grained geolocation estimation challenge by developing a distance-aware geolocation estimation model to accurately estimate the geographic coordinates of the abnormal traffic events (Section 4). Third, we conduct a set of new experiments to comprehensively evaluate the geolocation estimation performance of the proposed SAT-Geo framework comparing to the state-of-the-art baseline methods (Section 5). Fourth, we extend the related work by reviewing recent works on intelligent transportation systems (Section 2).

2. Related Work

2.1. Social Sensing

Social sensing has become as an emerging sensing paradigm to observe the physical world by exploring the "wisdom of the crowd" on social media [5, 22]. Social sensing has been adopted in a wide range of application domains [23, 24, 25, 26, 27, 28], including damage assessment in the aftermath of a disaster using social media data [25], cross-modal data fusion using crowdsourcing intelligence [26], and environment and urban infrastructure monitoring with inputs from citizen scientists [28]. The problem of abnormal traffic event geolocalization remains to be an important challenge that has not been well-addressed in social sensing applications. Specifically, the goal of abnormal traffic event geolocalization is to accurately identify the geographic coordinates of abnormal traffic events reported on social media. The identified geographic coordinates can be leveraged to provide effective precautions and timely responses for enhancing the safety and performance of the traffic systems. In this paper, we

develop SAT-Geo, a social sensing approach to effectively estimate the location entities associated with the abnormal traffic events and accurately estimate the corresponding geographic coordinates.

2.2. Location Inference in Social Sensing

A good amount of efforts have been made towards addressing the location inference problems in social sensing [16, 9, 10, 11, 29, 12, 13, 14]. For example, Li *et al.* designed a unified discriminative influence scheme that utilizes users' social network activities to estimate their home location [10]. Kinsella *et al.* developed a probabilistic-based language model to infer the city-level locations of social media posts using a training dataset of geotagged tweets [11]. Shahrakia *et al.* proposed an event localization solution that leverages Dempster–Shafer theory to estimate event locations on social media using a combination of user profiles, post content, and geotagging information [15]. However, the above solutions cannot be adapted to address our problem of geolocating abnormal traffic events since they either require prior knowledge or external information (e.g., users' private online activities, complete gazetteer database), or are insufficient to perform fine-grained geolocation estimation (e.g., average error distance is about 10-100 miles). In contrast to existing solutions, we design a novel SAT-Geo scheme that focuses on exploring the syntax patterns in the textual content of social media posts and estimates the geographic coordinates of the reported abnormal traffic event via a probabilistic-based learning approach.

2.3. Probabilistic Learning Technique

Our SAT-Geo framework is related to the probabilistic learning method in machine learning. Probabilistic learning has been applied to a wide range of studies, including natural language processing, computer vision, and information retrieval [30, 31, 32]. For example, Li *et al.* developed a probabilistic image annotation framework to estimate the image-to-word correlation using multi-correlation probabilistic matrix factorization [30]. Zettlemoyer *et al.* proposed

a structured classification model that leverages probabilistic categorical grammars to learn the mapping from sentences to logical forms [31]. Danelljan *et al.* designed a probabilistic regression approach to track the state of the target object in visual frames of video [33]. However, none of these approaches is designed to study the syntax patterns in the short and informal text of social media posts for geolocating abnormal traffic events. In contrast, the proposed SAT-Geo framework develops a syntax-based probabilistic learning approach to explicitly explore syntax patterns of social media content and effectively identify the relevant location entities for the accurate estimation of the abnormal traffic event's geographic coordinates.

2.4. Intelligent Transportation Systems

Our proposed work for abnormal traffic event geolocalization is closely related to intelligent transportation systems (ITS) [34] and can benefit many applications in ITS (e.g., improving traffic management efficiency [35], enhancing public transportation safety [36]). Examples of intelligent transportation systems include traffic monitoring, traffic congestion/accident detection, and public transportation management in urban planning [37]. For example, Barmpounakis *et al.* designed an urban traffic monitoring system using the sensing data collected from drones to monitor traffic congestion in the urban area [38]. Celesti *et al.* developed an Internet-of-Things (IoT) cloud system that utilizes mobile traffic sensors installed in public transportation and volunteer vehicles to monitor traffic conditions and detect vehicular accidents [39]. Tian *et al.* developed an infrastructural traffic monitoring solution to estimate traffic conditions by combining the Light Detection and Ranging (LiDAR) data with visual information captured by surveillance cameras [40]. Kalamaras *et al.* designed a unified interactive visual analytic platform for ITS management in urban planning [41]. To the best of our knowledge, the SAT-Geo framework is the first infrastructure-free solution to address the problem of geolocating abnormal traffic events at a fine-grained level using social media data.

3. Problem

We present the problem of geolocating abnormal traffic event in social sensing. First of all, we define a few key terms in our problem formulation.

Definition 1. ***Social Media Posts*** *(P): We define the social media posts P as a set of S social media posts (e.g., tweets) that are posted by social media users to report abnormal traffic events. Specifically, we define P as* $P = \{P^1, P^2, ..., P^S\}$ *where* P^s, $\forall\ 1 \leq s \leq S$, *denotes a social media post reporting abnormal traffic event.*

Definition 2. ***Location Entities*** *(L): The location entities (L) is defined as a set of named entities that are associated with the geolocation of the abnormal traffic event reported in a social media post. For example, "FDR Dr NB", "49th St", and "34th St" are the location entities in the social media post shown in Figure 1(a). Specifically, we define* $L^s = \{L_1^s, L_2^s, ..., L_C^s\}$ *to be the set of C location entities from the post* P^s.

Definition 3. ***Event Geographic Coordinates*** *(G): We define the Event Geographic Coordinates (G) to be the longitude and latitude coordinates of the abnormal traffic event depicted in a social media post. In particular, we define* $G^s = (g_{lat}^s, g_{long}^s)$ *to be the geographic coordinates of the abnormal traffic event in post* P^s.

The goal of our problem is to precisely estimate the abnormal traffic event geolocation by accurately identifying all location entities from a social media post. We formally formulate our problem as below:

$$\arg\min_{\widehat{G^s}}\ \mathcal{D}(\widehat{G^s},\ G^s \mid P^s),\ \forall\ 1 \leq s \leq S \tag{1}$$

where $\widehat{G^s}$ and G^s are the *estimated* and *ground-truth* geographic coordinates of the traffic event reported in social media post P^s, respectively. $\mathcal{D}(\cdot)$ is the application-specific distance measurement.

4. Solution

In this section, we present the SAT-Geo framework to address the problem of geolocating abnormal traffic events using social sensing. An overview of the SAT-Geo framework is shown in Figure 2. In particular, SAT-Geo consists of three modules: i) a *Syntax-based Pattern Learning (SPL)* module that effectively learns the syntax-based patterns in social media posts and embeds the learned patterns with two novel probability representations; ii) a *Probabilistic-based Entity Extraction (PEE)* module that identifies the location entities of an input social media post utilizing the syntax-based patterns from the SPL module using a principled probabilistic model; iii) a *Distance-aware Geolocation Estimation (DGE)* module that estimates the geographic coordinates of the abnormal traffic event location. We elaborate on the above modules in detail below.

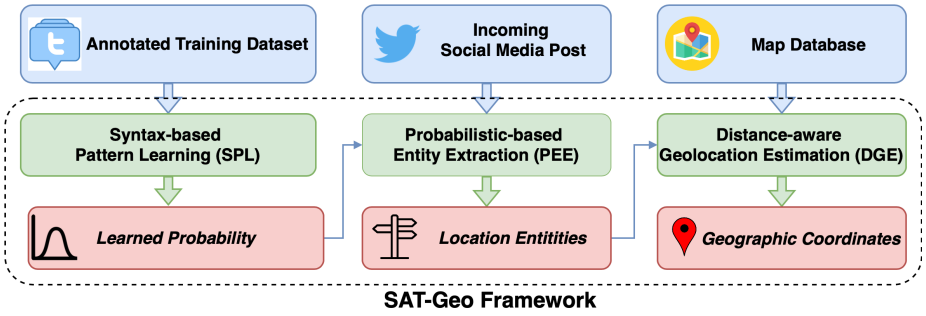


Figure 2: An Overview of SAT-Geo Framework

4.1. Syntax-based Pattern Learning (SPL)

The SPL module is designed to effectively learn the syntax-based patterns in social media posts for identifying the location entities. First, we define several key terms that will be used in this module.

Definition 4. Entity (e): We define an *entity* e to be a sequence of words that belongs to the same *part-of-speech* (e.g., "Conduit Ave"). In particular,

Table 1: Example of Syntax Model of a Tweet

Social Media Post S	"Accident on Conduit Ave approaching Sutter Ave"
Entity e (Syntax)	"Accident ($NOUN$)", "on (ADP)", "Conduit Ave ($NOUN$)", "approaching ($VERB$)", "Sutter Ave ($NOUN$)"
2-Syntax Model $M^{(2)}$	$T_1^{(2)}$: $NOUN$+ADP ["Accident ($NOUN$)","on (ADP)"] $T_2^{(2)}$: ADP+$NOUN$ ["on (ADP)", "Conduit Ave ($NOUN$)"] $T_3^{(2)}$: $NOUN$+$VERB$ ["Conduit Ave ($NOUN$)", "at ($VERB$)"] $T_4^{(2)}$: $VERB$+$NOUN$ ["approaching ($VERB$)", "Sutter Ave ($NOUN$)"]
3-Syntax Model $M^{(3)}$	$T_1^{(3)}$: $NOUN$+ADP+$NOUN$ ["Accident ($NOUN$)", "on (ADP)", "Conduit Ave ($NOUN$)"] $T_2^{(3)}$: ADP+$NOUN$+$VERB$ ["on (ADP)", "Conduit Ave ($NOUN$)", "approaching ($VERB$)"] $T_3^{(3)}$: $NOUN$+$VERB$+$NOUN$ ["Conduit Ave ($NOUN$)", "approaching ($VERB$)", "Sutter Ave ($NOUN$)"]
4-Syntax Model $M^{(4)}$	$T_1^{(4)}$: $NOUN$+ADP+$NOUN$+$VERB$ ["Accident ($NOUN$)", "on (ADP)", "Conduit Ave ($NOUN$)", "at (ADP)"] $T_2^{(4)}$: ADP+$NOUN$+$VERB$+$NOUN$ ["on (ADP)", "Conduit Ave($NOUN$)", "approaching (ADP)", "Sutter Ave ($NOUN$)"]

we use the state-of-the-art natural language tool[3] to annotate the *part-of-speech* for each social media post.

Definition 5. n-Syntax Pattern ($T^{(n)}$): We define an *n-Syntax Pattern* $T^{(n)}$ to be a contiguous syntax sequence of n entities in a given social media post. For example, the 3-entity sequence "*Accident+on+Conduit Ave*" has a *3-syntax pattern* of "*NOUN+ADP+NOUN*".

Definition 6. n-Syntax Model ($M^{(n)}$): We define an *n-Syntax Model* $M^{(n)}$ as the set of all possible n-Syntax patterns $T^{(n)}$.

Table 1 shows a simplified example that includes a social media post and the related entities, n-Syntax patterns, and n-Syntax models as defined above. In addition, we also define two types of probabilities that will be used to extract location entities from the social media post.

Definition 7. Pattern Probability: Pattern probability represents the probability of an n-Syntax pattern $T^{(n)}$ in an n-Syntax model $M^{(n)}$ that is defined

[3]https://cloud.google.com/natural-language/docs/analyzing-syntax

as:

$$\Pr(T^{(n)}|M^{(n)}) = \frac{|T^{(n)}|}{|M^{(n)}|} \tag{2}$$

where $|T^{(n)}|$ is the number of occurrences of the n-Syntax pattern $T^{(n)}$ in a given set of social media posts. $|M^{(n)}|$ is the total number of all n-Syntax patterns.

Definition 8. Index Probability: Index probability represents the probability of a location entity index $i^{(n)}$ in an n-Syntax pattern $T^{(n)}$ which is defined as:

$$\Pr(i^{(n)}|T^{(n)}) = \frac{|i^{(n)}|}{|T^{(n)}|} \tag{3}$$

where $|i^{(n)}|$ is the number of the location entities in the i^{th} entity given the n-Syntax pattern $T^{(n)}$.

With the key concepts defined above, our next goal is to leverage the learned *pattern probability* $\Pr(T^{(n)}|M^{(n)})$ and *index probability* $\Pr(i^{(n)}|T^{(n)})$ to effectively extract location entities in the unlabeled social media posts in next subsection.

4.2. Probabilistic-based Entity Extraction (PEE)

The PEE module aims to effectively extract the location entities from the content of the social media posts using the *pattern probability* and *index probability* learned in the SPL module. In particular, we first measure the likelihood of entity e to be a location entity from the *pattern probability* and *index probability* defined in Equation 2 and Equation 3, respectively. Formally, the likelihood of entity e being a location entity is as follows.

$$\Pr(e \in L|i^{(n)}, T^{(n)}, M^{(n)}) = \Pr(i^{(n)}|T^{(n)}) \times \Pr(T^{(n)}|M^{(n)}) \times \Pr(M^{(n)}) \tag{4}$$

where $\Pr(i^{(n)}|T^{(n)})$ and $\Pr(T^{(n)}|M^{(n)})$ denote the index probability and pattern probability, respectively. $\Pr(M^{(n)})$ is the weight of n-Syntax model that represents the importance of each n-Syntax model in extracting the location entities. $\Pr(M^{(n)})$ is often set to be a small value if we do not have prior knowledge.

In addition, we note that an entity e often appears in multiple n-Syntax patterns $T^{(n)}$ with different index $i^{(n)}$ in different n-Syntax model $M^{(n)}$. For example, "Conduit Ave ($NOUN$)" occurs in different n-Syntax patterns (i.e., 2, 3 and 4 syntax patterns), as the example shown in Table 1. Thus, we aggregate the likelihood of each entity over different n-Syntax patterns as below:

$$\Pr(e \in L) = \sum_{M^{(n)}} \sum_{(i^{(n)}, T^{(n)})} \Pr(e \in L | i^{(n)}, T^{(n)}, M^{(n)}) \tag{5}$$

Finally, an entity e is classified to be a location entity if the likelihood $\Pr(e \in L)$ is greater than a predefined threshold Δ[4]. Specifically,

$$\begin{cases} \mathbf{1} : \{\Pr(e \in L) > \Delta\} \\ \mathbf{0} : \{\Pr(e \in L) \leq \Delta\} \end{cases} \tag{6}$$

where "1" (i.e., true) indicates entity e is classified as a location entity and "0" (i.e., false) otherwise. The classified location entities are to be used as the input to effectively estimate the corresponding geographic coordinates in the DGE module that will be elaborated in next subsection.

4.3. Distance-aware Geolocation Estimation (DGE)

The DGE module is developed to accurately estimate the geographic coordinates of the abnormal traffic event using the location entities identified in the PEE module. First, we design a point-based map representation method to accurately extract the geographic coordinates of the location entities associated with the abnormal traffic event. Current solutions for geolocation estimation often adopt a grid-based approach that divides the geological areas of interest into grids and identifies the grid covering the abnormal traffic event [15, 8]. However, the precision of such an approach is limited by the size/area of the grid, which often ranges from 1 square mile (e.g., event location identification using social media data [15]) to 1000 square miles (e.g., global localization for

[4] Δ is an application specific parameter. We present a robustness study of the variation of Δ in the evaluation section.

the origin of social media posts [20]). In addition, the grid-based approach often uses the center of the grid to represent the estimated geographic coordinates for the identified location and is sub-optimal to effectively geolocate abnormal traffic events in urban areas with dense traffic where each grid contains multiple roads and intersections. For example, there are more than 100 intersections per square mile in Manhattan, New York [42].

In light of such a limitation, we design a point-based map representation approach to effectively model the geographic coordinates associated with the location entities identified in PEE. We first define the point-based map database that will be used in the DGE module to estimate the geographic coordinates of the abnormal traffic event.

Definition 9. Map Database ($\mathcal{Q}$): We define the map database as a set of H road entities, $\mathcal{Q} = \{R_1, R_2, \cdots, R_H\}$, where each road entity is associated with a location entity identified in PEE.

In particular, we formally define the road entity in the map database $\mathcal{Q}$ as follows.

Definition 10. Road Entity (R_h): We define a road entity $R_h \in \mathcal{Q}$ to be a sequence of K_h geographic points sampled from the road that is associated with each location entity in L. In particular, for a social media post P^s, we define $R^s = \{R_1^s, R_2^s, \cdots, R_C^s\}$ to be the set of C road entities associated with the location entities in L^s. Formally, each road entity is defined as $R_h = [v^1, v^2, \cdots, v^{K_h}]\ \forall\ 1 \leq h \leq H$, where each geographic point $v^k \in R_h$ is denoted as its geographic coordinates (i.e., $v^k = (g_{lat}^k, g_{long}^k)$).

With the map database $\mathcal{Q}$ defined above, our goal is to find the geographic point that is closest to the abnormal traffic event location from a set of candidate geographic points sampled from the road entities corresponding to the location entities of the reported abnormal traffic event. However, it is not a trivial task to accurately infer a geographic point that is closest to the abnormal traffic event location from multiple road entities (i.e., multiple sequences of geographic

points) identified in the social media post. We observe that the abnormal traffic events reported on social media often contain more than two location entities (e.g., the location entities "FDR DR NB", "49TH ST", and "34TH ST" shown in Figure 3). Therefore, we assume there are at least two road entities in R^s. Otherwise, we output the geographic point at the midpoint of the road entity as the estimated event geographic coordinates G^s. However, the coordinates of the estimated geographic point cannot be automatically identified by simply finding the geographic coordinates of the intersections of road entities. This is especially true when two or more intersections exist. For example, as Figure 3 shows, there are two intersections that exist among the road entities associated with the abnormal traffic event reported on social media (i.e., the intersection of "FDR Dr NB" and "49th St", and the intersection of "FDR Dr NB" and "34th St"). However, only the intersection of "FDR Dr NB" and "49th St" is the accurate estimation of the geographic coordinates of the reported abnormal traffic event.

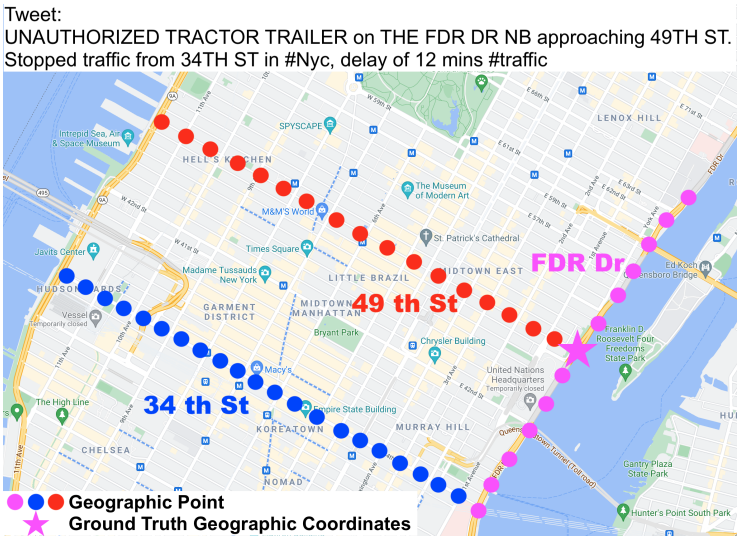


Figure 3: Example of an Abnormal Traffic Event

To address such a challenge, we design a distance-aware geolocation estimation method to accurately infer the geographic coordinates of the abnormal

traffic event. We jointly consider the distance between each geographic point and its neighborhood road entities (i.e., the road entities that co-appear in the same social media post), and the syntax-based relations of road entities related to the abnormal traffic event. In particular, we define the distance-driven weight of each geographic point in the identified road entities of the abnormal traffic event.

Definition 11. Distance-driven Weight: For each road entity $R_i^s \in R^s$, the distance-driven weight w^{k_i} of each $v_i^{k_i} \in r_i$ is defined as

$$w^{k_i} = \sum_{\substack{R_j^s \in R^s \\ R_j^s \neq R_i^s}} \frac{1}{\delta(v_i^{k_i}, R_j^s) + \epsilon} \tag{7}$$

where $\delta(\cdot)$ is distance function that measures the shortest Euclidean distance between a geolocation node $v_i^{k_i}$ and road entity R_j^s, and ϵ is a small constant to avoid the zero value in the denominator.

We observe that the location of the abnormal traffic event location often appears to be also close to the road entities describing the location of the abnormal traffic event in the social media post. For example, the traffic event location in Figure 3 (marked with a pink star) has the shortest distance to the road entities "FDR Dr" and "49th St" and is reasonably close to the road entity "34th St". Therefore, we compute the distance-driven weight of each geographic point to measure the distance between the geographic point and the geographic coordinates of the abnormal traffic event.

In addition, we observe that relations between the location entities are also critical in inferring the traffic event geolocation. For example, in the social media post (i.e., "Accident on Conduit Ave approaching Sutter Ave") shown in Table 1, we can effectively infer that the traffic event geographic coordinates belong to the road entity "Conduit Ave" according to the adposition "on" in the 2-syntax pattern "on+Conduit Ave". Therefore, we further identify the important location entities from the location entities in a tweet based on the adpositions in the syntax patterns. In particular, we add a relation indicator τ_i

to the distance-driven weight w^{k_i} and update w^{k_i} as follow:

$$w_{\tau}^{k_i} = \tau_i \cdot w^{k_i} \tag{8}$$

where $\tau_i = 1$ if the location entity associated with r_i co-appears with an adposition (ADP) in the 2-syntax and 3-syntax pattern, otherwise $\tau_i = 0$. In particular, we focus on the adpositions *on, at, approaching, after* based on the empirical observation. Finally, the geographic point with the highest distance-driven weight is output as the estimated traffic event geographic coordinates $\hat{G}^s$.

A summary of the distance-aware geolocation estimation (DGE) module is summarized in Algorithm 1. The input to the DGE module is the set of location entities $\widehat{L^s}$ of a social media post P^s from the PEE module. The output of the DGE module is the estimated abnormal traffic event geographic coordinates $\widehat{G^s}$.

Algorithm 1 Distance-aware Geolocation Estimation (DGE)

for each e in $\widehat{P^s}$ **do**
    retrieve the road entity r from $\mathcal{Q}$
    add r to the road entity set R^s
end for
for each r_i in R^s **do**
    for each $v_i^{k_i}$ in r_i **do**
        compute w^{k_i} and $w_{\tau}^{k_i}$
    end for
end for
if $\max(w_{\tau}^{k_i}) \neq 0$ **then**
    assign the geolocation node $v_i^{k_i}$ corresponds to $\max(w_{\tau}^{k_i})$ to $\widehat{G^s}$
else
    assign the geolocation node $v_i^{k_i}$ corresponds to $\max(w^{k_i})$ to $\widehat{G^s}$
end if
output $\widehat{G^s}$ for P^s

4.4. Summary of SAT-Geo Framework

The pseudocode of the SAT-Geo framework is summarized in Algorithm 2. The input of our SAT-Geo framework is a set of social media posts P that depict abnormal traffic events on social media. The output of our SAT-Geo framework is the estimated geographic coordinates G^s of the abnormal traffic event reported in each social media post P^s.

Algorithm 2 Summary of the SAT-Geo Framework

input: a set of N social media posts P, a map database $\mathcal{Q}$
output: the estimated geolocation $\widehat{G^s}$ for each $P^s \in P$
compute *pattern probability* $\Pr(T^{(n)}|M^{(n)})$ and *index probability* $\Pr(i^{(n)}|T^{(n)})$ using SPL
for each P^s in P **do**
    for each e in P^s **do**
        classify e using PEE
        if e is a location entity **then**
            $L^s \leftarrow e$
        end if
    end for
    estimate the geolocation $\widehat{G^s}$ from L^s using DGE module
    output $\widehat{G^s}$ for P^s
end for

v

5. Evaluation

In this section, we evaluate the performance of the proposed SAT-Geo framework on three real-world Twitter datasets collected from three cities. In particular, we first compare the location entity identification accuracy of SAT-Geo in comparison to state-of-the-art baseline methods. In addition, we also evaluate the geolocation estimation performance of SAT-Geo. Evaluation results show that SAT-Geo achieves significant performance gains compared to state-of-the-art baselines in terms of accurately identifying location entities associated

with abnormal traffic events and estimating the geographic coordinates of the abnormal traffic event.

5.1. Dataset

First, we describe the real-world Twitter datasets we collected from three major cities in the world, namely New York City (NYC), Los Angeles (LA), and London. In particular, the social media posts (i.e., tweets) on abnormal traffic events are collected from Twitter using the crawler *Get Old Tweets*[5] with a set of keywords and hashtags (e.g., "slow traffic", "accident", city names). We manually select 200 tweets from each dataset for our study, and verify that each tweet contains a unique abnormal traffic event (i.e., 1 tweet per event)[6]. The reported abnormal traffic events in our datasets can be mainly categorized into the following types: traffic accidents (e.g., collision, broken down vehicles), infrastructure incidents (e.g., out-of-order traffic signal, falling trees), and unusual road conditions (e.g., road closure, road construction). Each dataset is randomly split into 80% training set and 20% testing set. We manually annotate the location entities and the traffic event's geographic coordinates in each post to obtain the ground-truth annotations.

A summary of these three datasets are reported in Table 2. In particular, there are 2,412 entities in the NYC datasets and 20.4% of them are location entities. The LA dataset contains 2,851 location entities and 16.7% of them are location entities. The London dataset contains 2,483 location entities and 19.6% of these entities are location entities. We observe similar syntax patterns in social media posts among different English-speaking countries (e.g., United States and United Kingdom). For example, "Accident on Grand Ave SB at CR-12" and "Collision on Greenford Road Northbound at Daryngton Drive" are reported on social media in New York and London, respectively. We also show the distribution of the abnormal traffic events across each studied city

[5]https://github.com/Mottl/GetOldTweets3

[6]The number of tweets is mainly limited by the human labor of annotation.

by presenting the heatmap of the abnormal traffic event geolocations in Figure 4. Additionally, since the geotagging information associated with each tweet in our dataset is not necessarily available (due to privacy and legal concerns), we also invited independent human annotators to annotate the geographic coordinates associated with each tweet for evaluating the performance of geographic coordinates estimation. In particular, we annotate the ground-truth geographic coordinates by manually assessing the abnormal traffic event described in each tweet and finding the geographic coordinates of the traffic event location using an online map service[7].

Table 2: Data Trace Statistics

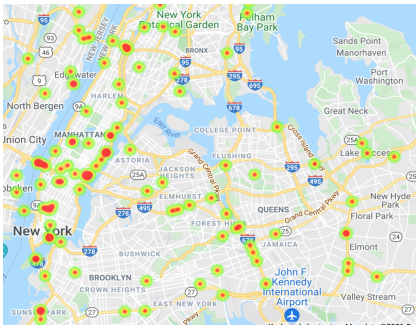
City	New York City	Los Angeles	London
Number of Abnormal Traffic Events	200	200	200
Number of Abnormal Traffic Events Related to Traffic Accident	144	151	142
Number of Abnormal Traffic Events Related to Infrastructure Incident	25	11	28
Number of Abnormal Traffic Events Related to Unusual Road Conditions	31	38	30
Number of Entities	2,412	2,851	2,483
Number of Location Entities	492	476	487

5.2. Baselines

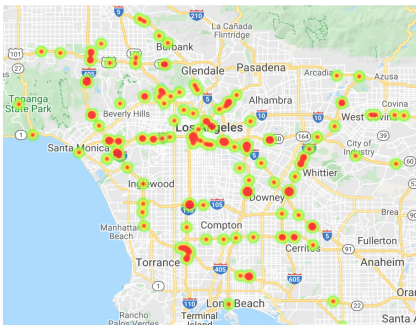
We compare SAT-Geo with a set of state-of-the-art baseline methods in *location entity identification* and *geolocation estimation*.

- **Google Named Entity Detection[8] (GoogleNE):** Google Named Entity Detection is the advanced commercial entity recognition service that extracts entities with the corresponding entity types (e.g., location entity) using a set of pre-trained natural language models.
- **Stanford CoreNLP (StanfordNLP) [12]:** Stanford CoreNLP is an integrated natural language processing toolkits that can be applied to

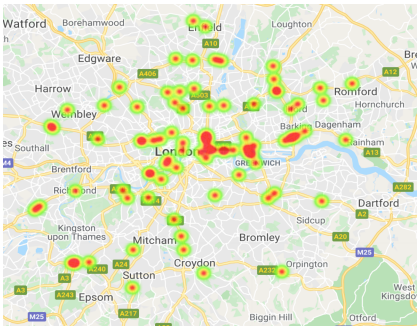
[7]https://www.google.com/maps
[8]https://cloud.google.com/natural-language/



(a) New York City



(b) Los Angeles



(c) London

Figure 4: Heatmap of Abnormal Traffic Event Location

identify location entities from text documents.

- **Spacy [43]:** Spacy is an industrial natural language processing framework that detects named entities in text document with a set of well-trained entity recognition models.

For all the baseline methods, we use the Google Maps Geocoding API[9] (Geocoding API) to convert the extracted location entities to the geographic coordinates of the corresponding traffic event location. In particular, we first concatenate the extracted location entities and pass them to the Geocoding API. The geographic coordinates returned by the Geocoding API are used as

[9]https://developers.google.com/maps/documentation/geocoding/overview

the estimated geographic coordinates for each corresponding baseline.

5.3. Evaluation Metrics

In evaluating the performance of location entity identification (i.e., location entity v.s. non-location entity), we adopt the following metrics that are commonly used for binary classification: *Accuracy*, *Precision*, *Recall*, and *F1-score*.

In evaluating the performance of geolocation estimation, we adopt the *Mean Error Distance* and *Median Error Distance* that are commonly used to evaluate the error distance in geolocation estimation [20]. In particular, the error distance d (in miles) between the estimated and ground-truth geographic coordinates is computed using the Haversine formula [44] as:

$$d = 2r \sin^{-1}\left(\sqrt{\sin^2\left(\frac{\hat{g}^b_{long} - g^b_{long}}{2}\right) + \cos(\hat{g}^b_{long})\cos(g^b_{long})\sin^2\left(\frac{\hat{g}^b_{lat} - g^b_{lat}}{2}\right)}\right) \tag{9}$$

where r is the radius of the earth. $(\hat{g}^b_{lat}, \hat{g}^b_{long})$ and (g^b_{lat}, g^b_{long}) are the estimated and ground-truth geographic coordinates of the abnormal traffic event reported in social media post S^b, respectively. If the ground-truth geographic coordinates of the traffic event is a single point, we measure the error distance in terms of the distance between the estimated and ground-truth event geographic coordinates. If the ground-truth geolocation is a road segment, we measure the error distance in terms of the perpendicular distance (in miles) between the estimated geographic coordinates and the road segment.

5.4. Evaluation Results

5.4.1. Location Entity Identification Performance

In the first set of experiments, we evaluate the performance of location entity identification. In particular, we vary the threshold Δ (defined in Equation 6) from 0.4 to 0.6 for the SAT-Geo scheme (e.g., SAT-Geo$_{0.6}$ represents the SAT-Geo scheme with $\Delta = 0.6$). The evaluation results on the NYC, LA, and London datasets are reported in Table 3, Table 4, and Table 5, respectively.

Table 3: Evaluation Results (NYC)

	Accuracy	Precision	Recall	F1-Score
SAT-Geo$_{0.4}$	**0.8568**	**0.6147**	**0.8152**	**0.7009**
SAT-Geo$_{0.5}$	**0.8702**	**0.6545**	**0.7826**	**0.7128**
SAT-Geo$_{0.6}$	**0.8724**	**0.6881**	**0.6956**	**0.6918**
GoogleNE	0.7807	0.4659	0.4456	0.4555
StanfordNLP	0.7203	0.1333	0.0652	0.0875
Spacy	0.7897	0.4838	0.3260	0.3896

We observe that the SAT-Geo scheme consistently outperforms all baselines under all evaluation metrics on all datasets. In particular, SAT-Geo achieves performance gains of 9.2%, 22.2%, 36.9%, and 25.7% comparing to the best-performing baseline in NYC (i.e., GoogleNE) in terms of accuracy, precision, recall, and F1-score, respectively. We observe similar performance gains on the LA and London datasets. The significant performance gains achieved by SAT-Geo demonstrate the effectiveness of judicious syntax patterns learning and the accurate location entity extraction in the principled probabilistic learning framework. We also note that SAT-Geo also outperforms all baseline methods as the Δ value varies in all datasets. Such consistent performance improvements again show the robustness of SAT-Geo with respect to the Δ parameter in the PEE module.

We also evaluate the performance of the SAT-Geo framework by varying the training set ratio from 60% to 80% for the NYC, LA, and London datasets. The results of SAT-Geo are shown in Figure 5. We observe a stable performance of SAT-Geo over different sizes of the training set across all cities in our study.

5.4.2. Geolocation Estimation Performance

We also study the geolocation estimation accuracy of SAT-Geo and the compared baselines. The results of the geolocation estimation performance on the NYC, LA, and London datasets are shown in Table 6, Table 7, and Table 8,

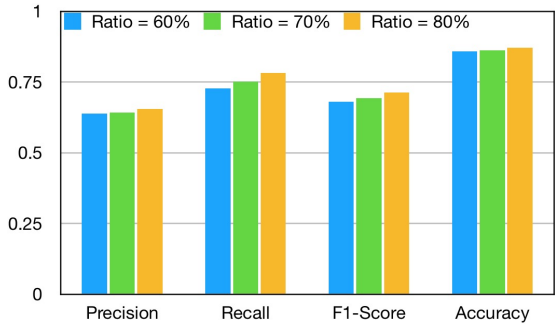
Table 4: Evaluation Results (LA)

	Accuracy	Precision	Recall	F1-Score
SAT-Geo$_{0.4}$	**0.8648**	**0.5447**	**0.7790**	**0.6411**
SAT-Geo$_{0.5}$	**0.8846**	**0.6122**	**0.6976**	**0.6521**
SAT-Geo$_{0.6}$	**0.8918**	**0.6444**	**0.6744**	**0.6590**
GoogleNE	0.8342	0.4680	0.5116	0.4889
StanfordNLP	0.7873	0.1190	0.0581	0.0781
Spacy	0.8162	0.3571	0.2325	0.2816

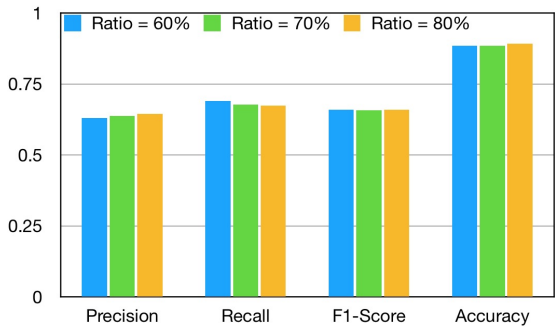
Table 5: Evaluation Results (London)

	Accuracy	Precision	Recall	F1-Score
SAT-Geo$_{0.4}$	**0.8379**	**0.5618**	**0.7529**	**0.6439**
SAT-Geo$_{0.5}$	**0.8526**	**0.6014**	**0.7221**	**0.6562**
SAT-Geo$_{0.6}$	**0.8687**	**0.6473**	**0.7138**	**0.6788**
GoogleNE	0.7633	0.4182	0.3379	0.3738
StanfordNLP	0.7062	0.2175	0.1627	0.1863
Spacy	0.7735	0.4317	0.3308	0.3746

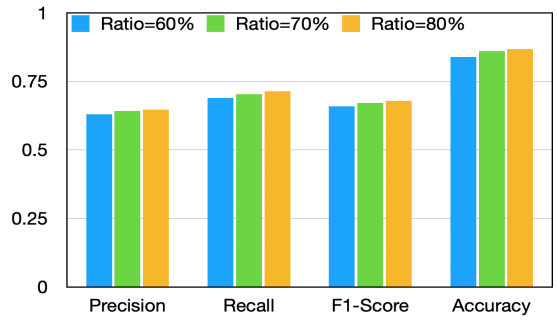
respectively. We observe that SAT-Geo consistently outperforms all the baseline methods on all datasets. In particular, the SAT-Geo framework achieves a mean error distance of 2.26 miles (i.e., SAT-Geo$_{0.6}$) on the NYC dataset which is 56.8% less than the mean error distance of the best performing baseline method (i.e., GoogleNE). Similarly, the mean error distance of the SAT-Geo framework is 37.2% and 58.1% less than the best-performing baseline method (i.e., GoogleNE) on the LA and London datasets, respectively. In addition to the effective entity extraction in SAT-Geo, we also attribute the performance gains to the accurate distance-aware geolocation estimation that jointly models the distance between each geographic point and the road entities reported in the same post, and the different syntax-based relation between the road entities related to the



(a) New York City



(b) Los Angeles



(c) London

Figure 5: Performance of SAT-Geo with Different Training Ratio

abnormal traffic event location.

Table 6: Geolocation Estimation Results (NYC)

	Mean Error Distance (mile)	Median Error Distance (mile)
SAT-Geo$_{0.4}$	**2.7231**	**0.6337**
SAT-Geo$_{0.5}$	**2.3172**	**0.6021**
SAT-Geo$_{0.6}$	**2.2613**	**0.5964**
GoogleNE	5.2346	1.0107
StanfordNLP	10.3018	4.7173
Spacy	6.3055	3.6543

Table 7: Geolocation Estimation Results (LA)

	Mean Error Distance (mile)	Median Error Distance (mile)
SAT-Geo$_{0.4}$	**4.2634**	**1.5733**
SAT-Geo$_{0.5}$	**4.1509**	**1.3061**
SAT-Geo$_{0.6}$	**3.9328**	**1.1275**
GoogleNE	6.2598	2.6154
StanfordNLP	10.7582	10.3306
Spacy	8.4526	6.5121

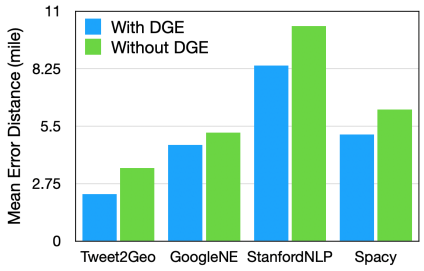
5.4.3. Ablation Study for Geolocation Estimation

Finally, we carry out an ablation study to investigate the geolocation estimation effectiveness of the DGE module in the SAT-Geo framework. In particular, we consider the following variations of SAT-Geo and the baseline methods: i) *with DGE:* using DGE as the geolocation estimation module to estimate the traffic event geographic coordinates using location entities identified by SAT-Geo and the baseline methods; ii)*without DGE:* using Google Map Geocoding as the geolocation estimation module to estimate the traffic event geographic coordinates using location entities identified by SAT-Geo and the baseline methods. The evaluation results are summarized in Figure 6 for the NYC, LA, and London datasets. We observe that the SAT-Geo *with DGE* achieves the best performance on all three datasets in terms of the mean error distance. In addi-

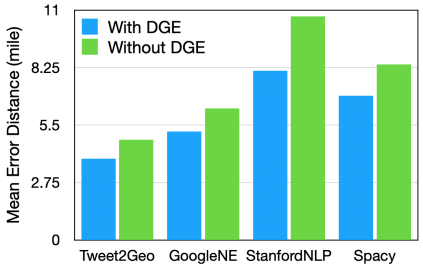
Table 8: Geolocation Estimation Results (London)

	Mean Error Distance (mile)	Median Error Distance (mile)
SAT-Geo$_{0.4}$	**3.2091**	**0.7276**
SAT-Geo$_{0.5}$	**2.9762**	**0.6949**
SAT-Geo$_{0.6}$	**2.6138**	**0.6852**
GoogleNE	6.2351	1.5913
StanfordNLP	9.6765	4.8502
Spacy	6.9336	4.1537

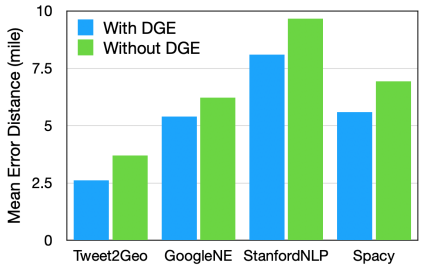
tion, we also observe that the incorporation of the DGE module also enhances the performance of all baselines by effectively measuring the distance-aware weight of each geographic point in the road entities related to the abnormal traffic event.



(a) New York City



(b) Los Angeles



(c) London

Figure 6: Performance for Ablations of DGE Module

6. Discussion

In this study, we focus on estimating the abnormal traffic event geolocation associated with social media posts. In our experiments, we only use a single tweet to geolocate each abnormal traffic event due to the high manual label cost [45]. However, the performance of the proposed SAT-Geo framework can be further enhanced by leveraging multiple data sources (e.g., multiple social media users reporting the same abnormal traffic event) to improve SAT-Geo's robustness against misinformation on social media [46, 47]. One critical challenge to leverage multiple data sources to geolocate the abnormal traffic event is that the reliability of different data sources are often unknown *a priori*, where the tweets posted by the unreliable social media users could lead to inconsistent and inaccurate geolocating results [1]. To address this challenge, we plan to leverage the estimation theoretical truth discovery solutions [23, 48] that are designed to jointly estimate the reliability of each studied social media user as well as the credibility of their posts to help us cross-validate the geolocating results and improve the abnormal traffic location estimation accuracy.

In addition, we collect tweets from both traffic authority accounts (i.e., the Twitter accounts managed by traffic authorities to publish traffic-related information) and general Twitter user accounts. Our current framework does not explicitly explore the authoritativeness of the Twitter accounts as we manually verify the abnormal traffic events reported in the collected tweets and choose the credible ones as the input to SAT-Geo. This is mainly due to the high labor cost of manually verifying the authoritativeness of all Twitter accounts involved in the study [45]. It will also be interesting to further investigate the authoritativeness of Twitter accounts by modeling the *reliability* of these accounts as well as the *credibility* of their posts about abnormal traffic events. However, it is not a trivial task to rigorously model the reliability of different Twitter accounts in the SAT-Geo framework. The reason is that the reliability is often not known for all Twitter accounts *a priori*, and the Twitter accounts with unknown/uncertain reliability may report inconsistent or conflicting infor-

mation about the same abnormal traffic event [1]. To address such a challenge, we plan to utilize the estimation theoretical methods in truth discovery [23] [48] to jointly estimate the reliability of the studied Twitter accounts and the credibility of posts associated with these accounts to improve the geolocation estimation performance of the SAT-Geo framework. As this line of effort is beyond the scope of this paper, we plan to implement it in our future work.

We also acknowledge that there is a limitation of using an identified set of tweets relevant to abnormal traffic events in our experiments that is labor-intensive and not scalable. In our future work, we plan to integrate the SAT-Geo framework with abnormal traffic event detection methods [49, 50] to automate the process of retrieving traffic-related tweets from real-time data streams. In particular, the social media posts retrieved by keywords/hashtags can be fed into a pre-trained abnormal traffic event detection model to classify whether a tweet contains the description related to an abnormal traffic event. The identified tweets will then be used as the input to our SAT-Geo framework for estimating the geolocation of abnormal traffic events. However, such a pre-trained abnormal traffic event detection model often requires a non-trivial amount of annotated ground-truth labels of the social media posts that report a diverse set of abnormal traffic events across different cities [50]. We plan to implement the abnormal traffic event detection model in our future work by leveraging the crowdsourcing platforms (e.g., Amazon MTurk) to collect sufficient ground-truth labels to train the detection model.

We note that the scalability of the SAT-Geo framework in the inference phase is expected to be linear to the size of the dataset. In particular, the time complexity of location entity extraction in the PEE module is $O(N)$ where N is the total number of social media posts [51]. The time complexity of geolocation estimation in the DGE module is also $O(N)$ given the time complexity of estimating the geolocation of each tweet is $O(1)$ (i.e., the time complexity of retrieving road entity from the map database is $O(1)$ and the time complexity of computing the max distance-aware weight is also $O(1)$) [52] . Therefore, the SAT-Geo framework can be deployed to efficiently estimate geolocations of

abnormal traffic events for a large-scale dataset. In addition, the efficiency of estimating abnormal traffic event geolocation for a large amount of tweets can be further boosted by deploying the trained SAT-Geo model at different computing nodes in distributed systems or cloud computing platforms to perform the geolocation estimation in parallel. In particular, our SAT-Geo framework does not involve any model training during the geolocation estimation phase. Instead, it utilizes the learned optimized model instance to infer the geolocation from each tweet. Hence, we can distribute the learned model instances to different computing nodes to process multiple subsets of the entire dataset in parallel to significantly improve the computational efficiency.

In this work, we focus on the social sensing based abnormal traffic event geolocation problem in large cities with high traffic volume (e.g., New York City, Los Angeles, London). In general, our model is more feasible for cities of a large size and population. This is because a large city is more likely to have a higher occurrence of different types of abnormal traffic events and there are more active social media users in a large city to post different abnormal traffic events in time [53]. As a result, our SAT-Geo model can be trained to detect different abnormal traffic events by leveraging the rich set of reported abnormal traffic events in the studied cities. For small or middle size cities, we expect the detection accuracy of our scheme would decrease because both the occurrence of abnormal traffic events and the chances of them being reported on social media decrease as the size of the city shrinks, leading to insufficient training data for our SAT-Geo model. One possible solution to address the above problem is to apply the transfer learning techniques [54] [55] to train our SAT-Geo model in a large city (e.g., NYC) and transfer the trained model to locate abnormal traffic events in a smaller city (e.g., El Paso, TX). However, it is challenging to effectively adapt the trained model across cities of different sizes. This is especially the case when the training data at the smaller city is sparse or unavailable [54]. To address this challenging problem, we plan to leverage the deep transfer learning techniques (e.g., adversarial transfer learning) to capture the latent feature of the syntax patterns from the social media posts reported

in the large city (e.g., NYC). The extracted latent features can then be applied to identify location entities in the posts reporting abnormal traffic events at the smaller city (e.g., El Paso, TX).

The training phase of our SAT-Geo framework in a new city (i.e., different from the city that SAT-Geo is trained with) will depend on the availability of the training data in the new city. In the case that a sufficient amount of training data is available in the new city, SAT-Geo can be re-trained to achieve the desired performance. However, if the amount of training data of the new city is sparse or insufficient, we can use the limited training data to fine-tune the SAT-Geo framework that has been pre-trained with the training data from the original city. Lastly, if the training dataset of the new city is not available at all, we can integrate SAT-Geo with the aforementioned transfer learning techniques [54, 55] to transfer the syntax pattern features learned from the original city with sufficient training data to geolocate the abnormal traffic events in the new city.

Another limitation of our work lies in the adaptability of our scheme to geolocate abnormal traffic events reported in regions where the primary language is not English (e.g., Arabic countries, Germany, Portugal, China). Our model does not directly apply to languages other than English. This is mainly due to the fundamental difference of grammar and syntax patterns between English and other languages [56]. For example, descriptive adjectives are often placed after nouns in Spanish which is opposite to the syntax pattern in English. In our future work, we consider two possible solutions to address the above problem. The first solution is to leverage the state-of-the-art machine translations models to translate the non-English posts to English and apply the SAT-Geo framework to geolocate the abnormal traffic events reported in the translated posts. Alternatively, our second solution aims to modify the n-Syntax Patterns and n-Syntax Models in the SPL module of SAT-Geo to accommodate the language-specific patterns in non-English languages [57]. We will investigate both solutions and compare their performance on non-English case studies. In particular, we plan to further evaluate the performance of SAT-Geo in geolo-

cating abnormal traffic events in Arabic-speaking countries (e.g., Saudi Arabia) and Portuguese-speaking countries (e.g., Portugal) in our future work.

7. Conclusion

In this paper, we develop SAT-Geo, a syntax-based probabilistic learning approach to geolocate abnormal traffic events using social sensing. The SAT-Geo framework is designed to estimate the geographic coordinates of the abnormal traffic events from the content of social media posts. In particular, we first identify the location entities associated with the abnormal traffic event location in social media posts by developing a syntax-based probabilistic learning approach. In addition, we design a distance-aware geolocation estimation method to accurately estimate the geographic coordinates associated with the reported abnormal traffic event. We evaluate the SAT-Geo framework on two real-world Twitter datasets. Results show that our SAT-Geo framework achieves significant performance gains comparing to state-of-the-art baseline methods in terms of accurately estimating the geographic coordinates of abnormal traffic events using social media data. The SAT-Geo framework can be further generalized and applied to a broader range of applications in fine-grained geolocalization using social media input (e.g., geolocating natural disasters or public safety events).

CRediT authorship contribution statement

Lanyu Shang: Conceptualization, Methodology, Formal analysis, Writing - original draft. **Yang Zhang:** Conceptualization, Methodology, Formal analysis, Writing - original draft. **Christina Youn:** Data curation. **Dong Wang:** Supervision, Writing - review & editing.

Acknowledgment

This research is supported in part by the National Science Foundation under Grant No. CHE- 2105032, IIS-2008228, CNS-1845639, CNS-1831669, Army Re-

search Office under Grant W911NF-17-1-0409. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] D. Wang, T. Abdelzaher, L. Kaplan, Social sensing: building reliable systems on unreliable data, Morgan Kaufmann, 2015.

[2] B. Predic, D. Stojanovic, Enhancing driver situational awareness through crowd intelligence, Expert Systems with Applications 42 (11) (2015) 4892–4909.

[3] L. Wang, D. Zhang, A. Pathak, C. Chen, H. Xiong, D. Yang, Y. Wang, Ccs-ta: Quality-guaranteed online task allocation in compressive crowdsensing, in: Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing, 2015, pp. 683–694.

[4] Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, A. Alamri, Health-cps: Healthcare cyber-physical system assisted by cloud and big data, IEEE Systems Journal 11 (1) (2015) 88–95.

[5] D. Wang, B. K. Szymanski, T. Abdelzaher, H. Ji, L. Kaplan, The age of social sensing, Computer 52 (1) (2019) 36–45.

[6] J. Lingad, S. Karimi, J. Yin, Location extraction from disaster-related microblogs, in: Proceedings of the 22nd international conference on world wide web, ACM, 2013, pp. 1017–1020.

[7] S. Wongcharoen, T. Senivongse, Twitter analysis of road traffic congestion severity estimation, in: 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), IEEE, 2016, pp. 1–6.

[8] J. D. G. Paule, Y. Sun, Y. Moshfeghi, On fine-grained geolocalisation of tweets and real-time traffic incident detection, Information Processing & Management 56 (3) (2019) 1119–1132.

[9] A. Schulz, A. Hadjakos, H. Paulheim, J. Nachtwey, M. Mühlhäuser, A multi-indicator approach for geolocalization of tweets, in: Seventh international AAAI conference on weblogs and social media, 2013.

[10] R. Li, S. Wang, H. Deng, R. Wang, K. C.-C. Chang, Towards social user profiling: unified and discriminative influence model for inferring home locations, in: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2012.

[11] S. Kinsella, V. Murdock, N. O'Hare, I'm eating a sandwich in glasgow: modeling locations with tweets, in: Proceedings of the 3rd international workshop on Search and mining user-generated contents, ACM, 2011, pp. 61–68.

[12] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, D. McClosky, The stanford corenlp natural language processing toolkit, in: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, 2014, pp. 55–60.

[13] L. Backstrom, E. Sun, C. Marlow, Find me if you can: improving geographical prediction with social and spatial proximity, in: Proceedings of the 19th international conference on World wide web, ACM, 2010, pp. 61–70.

[14] D. Y. Zhang, D. Wang, H. Zheng, X. Mu, Q. Li, Y. Zhang, Large-scale point-of-interest category prediction using natural language processing models, in: 2017 IEEE International Conference on Big Data (Big Data), IEEE, 2017, pp. 1027–1032.

[15] Z. K. Shahraki, A. Fatemi, H. T. Malazi, Evidential fine-grained event localization using twitter, Information Processing & Management 56 (6) (2019) 102045.

[16] Z. Cheng, J. Caverlee, K. Lee, You are where you tweet: a content-based approach to geo-locating twitter users, in: Proceedings of the 19th ACM international conference on Information and knowledge management, ACM, 2010, pp. 759–768.

[17] L. Sloan, J. Morgan, Who tweets with their location? understanding the relationship between demographic characteristics and the use of geoservices and geotagging on twitter, PloS one 10 (11) (2015) e0142209.

[18] O. Ajao, J. Hong, W. Liu, A survey of location inference techniques on twitter, Journal of Information Science 41 (6).

[19] T. B. N. Hoang, J. Mothe, Location extraction from tweets, Information Processing & Management 54 (2).

[20] M. Hulden, M. Silfverberg, J. Francom, Kernel density estimation for text-based geolocation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 29, 2015.

[21] Y. Zhang, X. Dong, D. Zhang, D. Wang, A syntax-based learning approach to geo-locating abnormal traffic events using social sensing, in: 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2019, pp. 663–670.

[22] D. Y. Zhang, N. Vance, D. Wang, When social sensing meets edge computing: Vision and challenges, in: 2019 28th International Conference on Computer Communication and Networks (ICCCN), IEEE, 2019, accepted.

[23] D. Wang, L. Kaplan, H. Le, T. Abdelzaher, On truth discovery in social sensing: A maximum likelihood estimation approach, in: Proceedings of the 11th international conference on Information Processing in Sensor Networks, 2012, pp. 233–244.

[24] D. Wang, T. Abdelzaher, L. Kaplan, C. C. Aggarwal, Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing appli-

cations, in: 2013 IEEE 33rd international conference on distributed computing systems, IEEE, 2013, pp. 530–539.

[25] D. Y. Zhang, Y. Zhang, Q. Li, T. Plummer, D. Wang, Crowdlearn: A crowd-ai hybrid system for deep learning-based damage assessment applications, in: Distributed Computing Systems (ICDCS), 2019 IEEE 39th International Conference on, IEEE, 2019.

[26] Y. Zhang, Y. Lu, D. Zhang, L. Shang, D. Wang, Risksens: A multi-view learning approach to identifying risky traffic locations in intelligent transportation systems using social and remote sensing, in: 2018 IEEE International Conference on Big Data (Big Data), IEEE, 2018, pp. 1544–1553.

[27] N. Vance, T. Rashid, , D. Y. Zhang, D. Wang, Towards reliability in online high-churn edge computing: A deviceless pipelining approach, in: The 5th IEEE International Conference on Smart Computing (SMARTCOMP 2019), IEEE, 2019.

[28] Y. Zhang, D. Zhang, N. Vance, D. Wang, Optimizing online task allocation for multi-attribute social sensing, in: 2018 27th International Conference on Computer Communication and Networks (ICCCN), IEEE, 2018, pp. 1–9.

[29] M. T. Rashid, D. Y. Zhang, Z. Liu, H. Lin, D. Wang, Collab-drone: A collaborative spatiotemporal-aware drone sensing system driven by socialsensing signals, in: 2019 28th International Conference on Computer Communication and Networks (ICCCN), IEEE, 2019, accepted.

[30] Z. Li, J. Liu, X. Zhu, T. Liu, H. Lu, Image annotation using multi-correlation probabilistic matrix factorization, in: Proceedings of the 18th ACM international conference on Multimedia, ACM, 2010, pp. 1187–1190.

[31] L. S. Zettlemoyer, M. Collins, Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars, arXiv preprint arXiv:1207.1420.

[32] Y. Huang, Q. Liu, S. Zhang, D. N. Metaxas, Image retrieval via probabilistic hypergraph ranking, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 3376–3383.

[33] M. Danelljan, L. V. Gool, R. Timofte, Probabilistic regression for visual tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7183–7192.

[34] L. Zhu, F. R. Yu, Y. Wang, B. Ning, T. Tang, Big data analytics in intelligent transportation systems: A survey, IEEE Transactions on Intelligent Transportation Systems 20 (1) (2018) 383–398.

[35] V. Kapitanov, V. Silyanov, O. Monina, A. Chubukov, Methods for traffic management efficiency improvement in cities, Transportation Research Procedia 36 (2018) 252–259.

[36] G. N. Kouziokas, The application of artificial intelligence in public administration for forecasting high crime risk transportation areas in urban environment, Transportation research procedia 24 (2017) 467–473.

[37] M. Bernas, B. Płaczek, W. Korski, P. Loska, J. Smyła, P. Szymała, A survey and comparison of low-cost sensing technologies for road traffic monitoring, Sensors 18 (10) (2018) 3243.

[38] E. Barmpounakis, N. Geroliminis, On the new era of urban traffic monitoring with massive drone data: The pneuma large-scale field experiment, Transportation research part C: emerging technologies 111 (2020) 50–71.

[39] A. Celesti, A. Galletta, L. Carnevale, M. Fazio, A. Ĺay-Ekuakille, M. Villari, An iot cloud system for traffic monitoring and vehicular accidents prevention based on mobile sensor data processing, IEEE Sensors Journal 18 (12) (2017) 4795–4802.

[40] Y. Tian, H. Liu, T. Furukawa, Reliable infrastructural urban traffic monitoring via lidar and camera fusion, SAE International Journal of Passenger Cars-Electronic and Electrical Systems 10 (2017-01-0083) (2017) 173–180.

[41] I. Kalamaras, A. Zamichos, A. Salamanis, A. Drosou, D. D. Kehagias, G. Margaritis, S. Papadopoulos, D. Tzovaras, An interactive visual analytics platform for smart intelligent transportation systems management, IEEE Transactions on Intelligent Transportation Systems 19 (2) (2017) 487–496.

[42] F. Tupin, B. Houshmand, M. Datcu, Road detection in dense urban areas using sar imagery and the usefulness of multiple views, IEEE Transactions on Geoscience and Remote Sensing 40 (11) (2002) 2405–2414.

[43] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spaCy: Industrial-strength Natural Language Processing in Pythondoi:10.5281/zenodo.1212303.
URL https://doi.org/10.5281/zenodo.1212303

[44] C. C. Robusto, The cosine-haversine formula, The American Mathematical Monthly 64 (1) (1957) 38–40.

[45] M. Karimzadeh, A. M. MacEachren, Geoannotator: A collaborative semi-automatic platform for constructing geo-annotated text corpora, ISPRS International Journal of Geo-Information 8 (4) (2019) 161.

[46] Z. Kou, L. Shang, Y. Zhang, C. Youn, D. Wang, Fakesens: A social sensing approach to covid-19 misinformation detection on social media, in: 2021 17th International Conference on Distributed Computing in Sensor Systems (DCOSS), IEEE, 2021.

[47] D. Y. Zhang, L. Shang, B. Geng, S. Lai, K. Li, H. Zhu, M. T. Amin, D. Wang, Fauxbuster: A content-free fauxtography detector using social media comments, in: 2018 IEEE International Conference on Big Data (Big Data), IEEE, 2018, pp. 891–900.

[48] D. Wang, M. T. Amin, S. Li, T. Abdelzaher, L. Kaplan, S. Gu, C. Pan, H. Liu, C. C. Aggarwal, R. Ganti, et al., Using humans as sensors: an

estimation-theoretic perspective, in: IPSN-14 proceedings of the 13th international symposium on information processing in sensor networks, IEEE, 2014, pp. 35–46.

[49] J. He, W. Shen, P. Divakaruni, L. Wynter, R. Lawrence, Improving traffic prediction with tweet semantics, in: Twenty-Third International Joint Conference on Artificial Intelligence, 2013.

[50] D. A. Kurniawan, S. Wibirama, N. A. Setiawan, Real-time traffic classification with twitter data mining, in: 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE), IEEE, 2016, pp. 1–5.

[51] G. E. Pibiri, R. Venturini, Handling massive n-gram datasets efficiently, ACM Transactions on Information Systems (TOIS) 37 (2) (2019) 1–41.

[52] V. Singh, Replace or retrieve keywords in documents at scale, arXiv preprint arXiv:1711.00046.

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

[54] Y. Zhang, H. Wang, D. Zhang, D. Wang, Deeprisk: A deep transfer learning approach to migratable traffic risk estimation in intelligent transportation using social sensing, in: 2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS), IEEE, 2019, pp. 123–130.

[55] Y. Zhang, D. Zhang, D. Wang, On migratable traffic risk estimation in urban sensing: A social sensing based deep transfer network approach, Ad Hoc Networks 111 (2021) 102320.

[56] L. R. Naigles, P. Terrazas, Motion-verb generalizations in english and spanish: Influences of language and syntax, Psychological Science 9 (5) (1998) 363–369.

[57] S. Ravi, K. Knight, Deciphering foreign language, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 12–21.