Road Data Enrichment Framework Based on Heterogeneous Data Fusion for ITS

Paulo H. L. Rettore, Bruno P. Santos, Roberto Rigolin F. Lopes, Guilherme Maia, Leandro A. Villas, and Antonio A. F. Loureiro

***Abstract*—In this work, we propose the Road Data Enrichment (RoDE), a framework that fuses data from heterogeneous data sources to enhance Intelligent Transportation System (ITS) services, such as vehicle routing and traffic event detection. We describe RoDE through two services: (i) Route service, and (ii) Event service. For the first service, we present the Twitter MAPS (T-MAPS), a low-cost spatiotemporal model to improve the description of traffic conditions through Location-Based Social Media (LBSM) data. As a case study, we explain how T-MAPS is able to enhance routing and trajectory descriptions by using tweets. Our experiments compare T-MAPS' routes against Google Maps' routes, showing up to 62% of route similarity, even though T-MAPS uses fewer and coarse-grained data. We then propose three applications, Route Sentiment (RS), Route Information (RI), and Area Tags (AT), to enrich T-MAPS' suggested routes. For the second service, we present the Twitter Incident (T-Incident), a low-cost learning-based road incident detection and enrichment approach built using heterogeneous data fusion. Our approach uses a learning-based model to identify patterns on social media data which is then used to describe a class of events, aiming to detect different types of events. Our model to detect events achieved scores above 90%, thus allowing incident detection and description as a RoDE application. As a result, the enriched event description allows ITS to better understand the LBSM user's viewpoint about traffic events (e.g., jams) and points of interest (e.g., restaurants, theaters, stadiums).**

***Index Terms*—ITS, heterogeneous data fusion, data enrichment, LBSM, incident detection, VANETs.**

I. Introduction

NOWADAYS, the intelligent planning and management of transportation systems are fundamental tasks to promote the sustainable growth of modern cities. Governments, researchers, and industries have been developing and deploying systems to record and understand mobility patterns within a city to support solutions to reduce traffic issues and incident events [1]. In this context, the ITS emerges as a feasible way to improve real-time decision-making by leveraging the availability of information and communication technologies, thus providing applications and services to improve transportation systems. ITS depends on the availability of users' data (from different sources and covering the whole city) and communication technologies to support both data sharing and access to services. However, the lack of a timely access to relevant data may present a limitation to the real-time traffic analysis performed by those systems, since only a set of companies have access to such data (e.g., data from inductive loops, traffic cameras, semaphores, GPS track, and origin-destination matrix) or it is often outdated. This happens due to the commercial value that transportation-related data have for companies, and also to the deprecated infrastructure employed to deliver such data to end users. These facts become a barrier to better understand urban mobility and the transportation scenario, thus requiring alternative solutions like the one being proposed in this investigation.

Currently, information about road traffic and road events available to the users are outdated or have a low quality descriptions, which limit the efficiency of services that can be provided, such as route management and flow control. As a result, the spread of detailed and useful descriptions of critical events is also compromised. Overcoming these issues demands multidisciplinary expertise to leverage transportation system data to improve traffic efficiency and safety. Recent investigations use location-based data from different sources (e.g., LBSM, maps service and *ad hoc* transportation systems). Therefore, integrating multiple data sources is a fundamental process to provide consistent, accurate and useful information. Such a process is called Data Fusion, which is particularly challenging when dealing with heterogeneous and asynchronous data that may include noise and errors. Furthermore, spatiotemporal aspects increase the complexity of fusing these heterogeneous data sources [2].

Nevertheless, traffic data from most of our cities is still gathered and shared by few private/public institutions in a centralized fashion, i.e. with restricted data access for third parties and also with limited description of pertinent events. Consequently, LBSM (e.g., Twitter, Instagram and Foursquare) combined with navigation systems (e.g., Google Maps, Here WeGo and Bing Maps) has become an alternative data source to study urban mobility. Social media platforms allow users to share their thoughts, viewpoints and activities related to their feelings about almost everything including the perceived

Manuscript received December 28, 2018; revised July 24, 2019 and October 14, 2019; accepted January 23, 2020. Date of current version March 27, 2020. The Associate Editor for this article was B. F. Ciuffo. *(Corresponding author: Paulo H. L. Rettore.)*

Paulo H. L. Rettore is with the Department of Computer Science, Federal University of Minas Gerais, Belo Horizonte 31270-901, Brazil, and also with the Communication Systems, Fraunhofer FKIE, 53177 Bonn, Germany (e-mail: rettore@dcc.ufmg.br; paulo.lopes.rettore@fkie.fraunhofer.de).

Bruno P. Santos, Guilherme Maia, and Antonio A. F. Loureiro are with the Department of Computer Science, Federal University of Minas Gerais, Belo Horizonte 31270-901, Brazil (e-mail: bruno.ps@dcc.ufmg.br; jgmm@dcc.ufmg.br; loureiro@dcc.ufmg.br).

Roberto Rigolin F. Lopes is with the Communication Systems, Fraunhofer FKIE, 53177 Bonn, Germany (e-mail: roberto.lopes@fkie.fraunhofer.de).

Leandro A. Villas is with the Institute of Computing, University of Campinas, Campinas 13083-970, Brazil (e-mail: leandro@ic.unicamp.br).

Digital Object Identifier 10.1109/TITS.2020.2971111

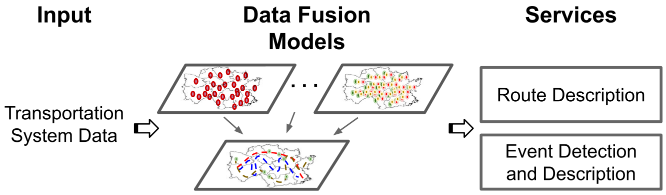


Fig. 1. The design of RoDE.

traffic conditions (different modern life issues subject of recent research can take advantage of LBSM as a low-cost data source [1], [3]–[5]).

With this in mind, this investigation compiles the current traffic status in a given city using LBSM and navigation platforms. In order to compile the status, we introduce a robust framework named RoDE, which is based on heterogeneous data fusion. Our framework, depicted in Fig. 1, aims at delivering enriched event descriptions as input to navigation systems, road planners and the general public. Our description of routes and events replace the usual binary feedback from transportation systems, such as whether the current traffic is good or not, or whether there is or isn't an incident. For example, after a set of data source passes through our data fusion models, the outcome is used as an input to specialized services dealing with routing enrichment, and event detection and description.

In particular, RoDE takes as input two different classes of transportation system data sources, which are categorized as Infrastructure as Vehicular Sensor (InfraVS), and Media as Vehicular Sensor (MVS), as described in our previous work [6]. The former class, InfraVS, acquires data from navigation systems, such as routes from Google Maps, traffic jams and incidents from Here WeGo, and from Bing Maps. The latter class, MVS, acquires data from pertinent LBSM, such as the user's viewpoint from Twitter and Points of Interest from TripAdvisor. It is important to notice that the use of LBSM can also introduce user' bias to the system. However, RoDE implements a methodology to pre-process LBSM by taking into consideration, in most of cases, reliable sources, such as Twitter accounts managed by the Department of Transportation, Traffic Department, Police Department, road managers, news and so on. After providing RoDE with these two types of data, the system uses appropriate data fusion models to treat and fuse them, based on two specific goals, which are: (i) enrich the description of routes generated by Google Maps; and (ii) detect road events (incident and non-incident) and enrich the description of those events.

The contributions of the RoDE framework are summarized into two fundamental services using enriched data from heterogeneous data sources as follows:

- *Route Services* (Twitter MAPS (T-MAPS)): is a low-cost spatiotemporal model to improve the description of traffic conditions based on tweets. Our quantitative experiments compare T-MAPS routes with Google maps routes and show a high route similarity, even though T-MAPS uses few and coarse-grained data. As a result, the similarity is used to enrich the route description within three new services over T-MAPS: Route Sentiment (RS), Route Information (RI), and Area Tags (AT) aiming at enhancing the route information;
- *Event Services* (Twitter Incident (T-Incident)): is a low-cost learning-based road incident detection model which also enriches the incident description using heterogeneous data fusion techniques implemented as RoDE services.

This paper is organized as follows. Section II presents the related work. Section III describes the data acquisition process to propose RoDE services. Section IV motivates the use of LBSM data to enhance and complement the conventional ways to see traffic and transit in urban areas. Section V describes LBSM data aspects such as data imprecision, user bias and so on. Section VI provides details about the Route Service. Thereafter, Section VII describes the Event Service architecture and its quantitative evaluation through experiments. Finally, Section VIII concludes the paper and also lists future work.

II. Related Work

The growth of the Internet and the proliferation of LBSM have enabled scientific investigations supported by an enormous amount of data generated every single day in urban areas. When considering the traffic and transit perspective, several studies have analyzed traffic conditions using LBSMs [7]. Many other studies focus on event detection and diagnostics using Natural Language Processing (NLP) techniques [4], [8], [9]. Rettore et al. [6] survey the recent literature using such data on ITS context. Moreover, they outline a taxonomy, according to the different data sources, highlighting the MVS.

Other studies perform sentiment analysis using LBSM data [10], [11]. Kim et al. [5] proposed SocRoutes, a safe route recommending system, based on Twitter data. Unusual traffic events, based on social media, was investigated in [12]. Septiana et al. [13] categorized road conditions with an accuracy of up to 92%. Gu et al. [14] explored tweets text to extract traffic incident information and to provide a low-cost solution to existing data sources. They validated the Twitter-based incidents using data from RCRS (Road Condition Report System) incident, 911 Call For Service (CFS) incident, and Here WeGo travel time.

Yazici et al. [15] showed that tweets collected from regular accounts are more likely to be irrelevant, though they can capture events that have just happened. On the other hand, tweets from specialist accounts are more valuable and structured, therefore, better identifying incident events. Also, they showed that the combination of both sources, leads to better results when dealing with event detection. In the same way, Zhang et al. [16] complemented the incident detection scenario by using social media data. They showed that social media data can be useful as an alternative way to improve traditional methods to detect traffic events in real-time.

Nguyen et al. [17] developed the TrafficWatch, a real-time Twitter-based system designed to leverage traffic-related information to allow incident analysis and incident visualization in Australia. They also developed a case study to detect road incidents before the Transport Management Centre (TMC) Log Time and also detecting incidents not reported by TMC.

Pereira et al. [18] used a reliable media provided by traffic management centers, NLP techniques, featuring topic modeling and text analysis to improve the accuracy measuring the duration time of an incident. They showed that the use of this data source improves the prediction of an incident by 28% when compared to a base line experiment not using the reliable media.

The present investigation extends our previous study in [19] suggesting that LBSM posts can improve the comprehension of road traffic events. Differently from most of the related work discussed above, we take a step forward by providing a model to clarify the ongoing traffic condition, based on heterogeneous data fusion, by adding extra information to current navigation systems. Complementing most of the recent literature discussed above, we start with the hypothesis that a model to fuse data from heterogeneous data sources can add extra information to current navigation system improving the feedback to the users. In particular, RoDE provides route description services, such as Route Sentiment (RS), Route Information (RI), and Area Tags (AT). Moreover, RoDE also provides event detection and description services. We also detail the spatiotemporal grouping algorithm, the feature extraction process, as well as the ground truth for the route service (Google maps routes), and the event service (Here and Bing incidents) to conduct our learning-based model with LBSM data.

III. Data Acquisition

Nowadays, most of our cities are still missing a digital platform to collect and share mobility data with their inhabitants and third parties. Thus, the lack of available data in urban transport environments is one of the greatest challenges for those developing (ITS). Researchers are often restricted to theoretical studies or to limited public data in both coverage and quality. Luckily, the current increase of online platforms, such as LBSM, makes it possible for people to share their daily urban mobility and opinions regarding a variety of aspects.

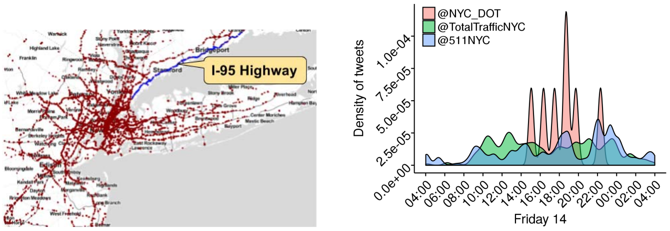
In this sense, we conducted the data acquisition process to support our ideas in developing and evaluating the RoDE services. We executed a daily data acquisition process of LBSM to support the development and evaluation of two RoDE services: T-MAPS and T-Incident. Based on previous experiences, we have chosen New York City (NYC) as our use case because of the large amount of data generated in that area, which increases the chances to have more spatiotemporal data coverage. As a result, Tab. I summarizes the data collected during this investigation for both RoDE services, the first dataset addressed to the Route, and the second dataset addressed to the Event service.

The motivation behind these RoDE services come from the desire to expand the knowledge about the traffic conditions by providing a more detailed description about a given scenario, instead of the usual binary feedback from transportation system which categorizes if the current traffic is good or bad. Using social media data, it is possible to describe the traffic scenarios such as the indication of the route's condition, the intensity of accidents and more detailed information about road events. This type of information may enrich the user's transportation experience, providing better assistance for decision-makers when dealing with urban mobility.

TABLE I

Data Acquired With RoDE Framework

	Source	Goal	Total Sample	Filtered Sample	Temporal Interval	Spatial Location
Route	Twitter	Route Description	353,807	38,112	2016-10-07 to 2016-12-17	Manhattan, NYC
Event	Twitter	Event Detection	158,413	158,413	2018-08-28 to 2018-09-17	
	Here WeGo	Incident	9,784	9,784		
	Bing Maps	Incident	1,924	1,924		
	Trip Advisor	Non-Incident	50	50		



(a) Tweets on NYC (b) Hourly tweeting density

Fig. 2. Dataset coverage in new york city and neighborhood.

A. Route Service

Our first dataset to evaluate the T-MAPS consists of 353,807 tweets from 21 manually selected users' accounts. Those accounts are maintained by transport departments, traffic specialists and transit reports such as news channels or dedicated companies. The main reason to carefully select these accounts is to introduce reliable data to our experiments. While there is no restriction on subject in regular users' accounts, the specialist accounts have a specific subject and usually a high-quality description. Indeed, the specialists' accounts gather data from different sources (for example, radio, television, transport and police departments, podcasts, automotive data[1]). The number of tweets with geotagging is 307,020, most of them in NYC. Here, we explored the Manhattan region, which has 38,112 tweets. The dataset was collected during the last three months of 2016. Tab. I summarizes the data collected to achieve the Route Service goals. The dataset does not contain regular users due to the high user bias in their tweets regarding traffic feelings. Moreover, some aspects involving the use of LBSM data are highlighted later in Sec. V.

To visualize our first dataset, Fig. 2(a) shows the spatial coverage of tweets in our dataset. Most tweets are about the road network, i.e., if we zoom in, it is possible to see the I-95 highway with tweets along its extension. From the temporal viewpoint, Fig. 2(b) shows the tweets' density along the hours for @NYC_DOT, @TotalTrafficNYC, and @511NYC users' accounts. Note that some peaks of tweets appear during rush times confirming the reliability of these accounts reporting traffic jams or other traffic-related events likely to occur during rush hours. Confirmation of the reliability of these accounts is required because large cities like NYC are known to suffer from traffic jams and accidents during rush hours (further details about the data acquisition process is provided in [19]).

[1] https://www.ttwnetwork.com

An important question emerges from the inherent subjectivity of enriching the description of traffic events. To the best of our expertise, there is no ground truth for the best route within urban areas. For that reason, many tools offer their own traffic viewpoint, such as Google Maps, Here Wego, Bing Maps, and TomTom maps. Our motivation to develop T-MAPS was to demonstrate the potential usage of LBSM data as a complement to traffic data. Also, our goal is to encourage the design of new applications, models, and analysis of urban mobility using LBSM.

B. Event Service

T-Incident is a service to accurately identify traffic events (incident and non-incident) and enrich their descriptions. The data acquisition process aims at combining different data sources, such as Here WeGo, Bing Maps[2], Tripadvisor[3] and Twitter[4] in both temporal and spatial dimensions to achieve these goals.

The second dataset used in this investigation consists of 158,413 tweets acquired from 2018-09-14 to 2018-11-06. During this period, we extracted data from Twitter filtering tweets by a set of words related to incident events, such as *congestion, accident, construction, planned event, road hazard, disabled vehicle, traffic, jam, car and weather* (most of them are used by navigation tools to describes types of incident). All collected tweets are geolocated and most of them are in Manhattan, NYC. Moreover, we were interested in tweets from both regular (common accounts) and specialist (professional accounts controlled by corporations) users. We also discarded tweets posted as *retweet*, aiming to collect the user's impressions and ignoring the spread of information.

LBSM data poses several challenges, as discussed later in Sec. V. To collect as much incident events as possible, we acquired data from two different data sources: Here WeGo and Bing Maps. The incidents gathered from both platforms have temporal granularity of one hour from 2018-09-14 to 2018-11-06; therefore overlapping temporally and spatially. We have collected 9,784 distinct incidents acquired from Here WeGo and 1,924 distinct incidents acquired from Bing Maps. To use those incident data, we fused both data sources, filling the gaps in a data source with entries from each other, and vice-versa. We also combined common incidents to these two data sources because the descriptions are, usually, different and can complement each other (see Sec. VII-A for details of this process).

The first step to detect incidents is to define what is not an incident. In order to do it, we picked a set of places with no traffic incidents from a data source dealing with touristic places, called Tripadvisor. Tripadvisor is a travel website that compiles reviews of hotels and restaurants, together with other multimedia travel-related content. Next, we extracted data from Tripadvisor to compile a set of the most popular places ranked by the tourists, such as museums, observatories, parks, pubs, and theaters. Tab. I summarizes the data collected to achieve the Event Service goals and Fig. 3 shows the spatial data coverage of each data source used to develop the T-Incident service.

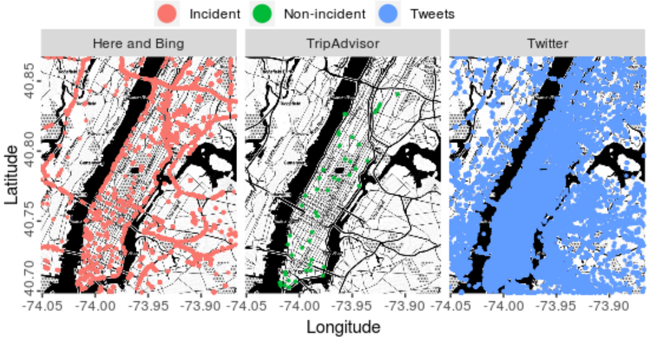


Fig. 3. The data sources spatial coverage.

IV. Twitter as a Traffic Sensor

To reveal the potential of LBSM data to enhance and complement the conventional ways of seeing traffic and transit, it is fundamental to understand how tweets are related to the traditional traffic sensor. The relation between tweets and conventional traffic sensors highlights the potential of the former to enhance the latter. For example, if a conventional traffic sensor detects an anomalous event, can tweets explain such a typical event? This section presents direction to answer questions like this.

First, it is required to get access to classic traffic measurement data, such as inductive loop detector counts, traffic cameras, vehicle GPS traces on road network, or origin-destination matrices, among others. With these data sources, traffic specialists can study demand and supply aspects of the transportation systems. Demand can be seen as vehicles and pedestrians trying to reach a particular place while supply is related to streets, highways, sensors and control devices [1]. Thus, it is possible to study the interactions between demand and supply, and eventually develop efficient transportation systems, which optimize urban mobility and decrease transit congestion.

Unfortunately, the access to raw traditional sensor data is a challenge for the regular community. Raw traffic data are kept locked by government entities or large companies. Usually, the traditional sensors sense three variables of interest: velocity, density, and flow. These quantities are relate to each other allowing traffic behavior analyses and visualizations [1], [20]. On the other hand, LBSMs is accessible on the Internet and is an alternative to support recent urban mobility studies like the ones reported in [1], [3]. Also, it is common that users share their thoughts, viewpoints, and activities on LBSM platforms. Such personal posts expand the sensing capacity by capturing the users' perspective about the situation.

Naturally, raw data holders perform some data fusion process and present the result in their online services or in periodical statistics. For example, Google gathers heterogeneous data such as GPS traces, cameras, and inductive loops. Thus, Google Maps implements a process to fuse data sources adding a colored layer to highlight traffic conditions on a map. In that way, companies like Google, HERE WeGo, Bing, and TomTom allow access to the resulting data fusion process,

[2] https://bing.com/maps

[3] https://tripadvisor.com/

[4] https://developer.twitter.com/en/docs

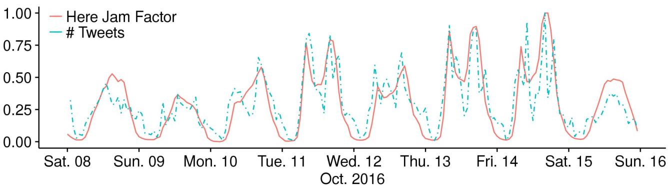


Fig. 4. Tweets frequency and here Jam factor time series.

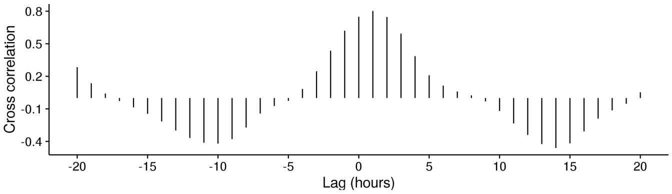


Fig. 5. Cross-correlation between Jam Factor and #tweets time series.

or part of it. Here, we reuse the Jam Factor (JF) from HERE WeGo API as a traditional traffic sensor aggregated to our framework. According to the HERE documentation, the JF is a fused representation of traditional heterogeneous data varying from 0 (free) to 10 (street blocked). We choose the jam factor from HERE WeGo because it is, currently, the only API providing such metric.

Fig. 4 shows the correlation between HERE JF and correlated traffic tweets in the dataset along a week in Oct. 2016. The time series in blue is the aggregated HERE JF, and the orange curve corresponds to the number of tweets. We re-scale the tweet and JF time series to lie between 0 and 1, and aggregated each series hourly. This explains why the two curves look similar in the plot. We also compute the Spearman's rank (ρ), a nonparametric correlation coefficient, to identify relationship between two variables. The ρ has a value between -1 and $+1$, where -1 means that the observations are entirely dissimilar and $+1$ the opposite. We apply Spearman's rank in the time series resulting in $\rho = +0.81$. It is possible to interpret that the #tweets tend to increase when the JF increases. Notice that, our methodology (data acquisition and data preparation) guarantees that we are using traffic data, avoiding correlation between JF and any other non-related subject.

Applying the cross-correlation technique, it is possible to figure out where time series match [21]. Fig. 5 shows on the y-axis the cross-correlation between JF and #tweets, and on the x-axis the lag between the time series in hours, we use JF as the test waveform. The highest correlation (0.8) appears when the lag is $+1$ meaning that #tweets curve is 1 hour ahead of JF. One can interpret these results as an indication that tweets appear on the platform before JF increases, but note that the time series were hourly aggregated. Then, we are not able to claim that tweets can predict traffic jam, but these both correlations give us enough information to consider our dataset as reliable source to support the RoDE proposal. In the analyses presented in Fig. 4 and Fig. 5 we used the total sample of the acquired tweets to evaluate the Route Service (see Tab. I). Fig. 4 and Fig. 6 complement each other, where the former shows a macro viewpoint in days of the correlation between tweet and JF and the latter does the same hourly providing a micro viewpoint of events. It is possible to see the same shape in both graphics, highlighting the hourly frequency of tweets.

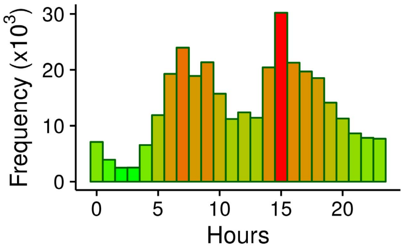


Fig. 6. The frequency of tweets per hour.

After we verified the potential of using Twitter as a traffic sensor, we characterized the problems in the next section. Because we have to deal with these problems before proposing solutions using Twitter as a data source for ITS like RoDE.

V. LBSM Data Aspects

Social media is a fundamental data source that can be used as an input for RoDE because it complements the set of heterogeneous sources (e.g., traffic and navigation data) potentially enriching the transportation scenario with descriptive data compiled by the end user. In this section, we discuss what we have learned from the LBSM data aspect while designing/implementing RoDE. For example, data from Twitter poses many challenges when related to traffic events. Here, we classify the data aspects into four classes, namely: Data imprecision, User bias, Spatiotemporal assignment, and Inconsistencies. More extensive taxonomies can be found in [2], [22].

LBSM data comes with a certain degree of imprecision mainly because people can express their opinions freely. Often, the data imprecision presents at least one of the following characteristics: incomplete data, vagueness and granularity effects. For instance, let us consider the following tweet:

> *"Now 8:00AM an accident at 100 W 33rd St #NYC #BadTraffic #creepedOut"*.

One can obtain relevant knowledge about the event, e.g., the user's sentiment, traffic condition, date and time. However, the tweet lacks fundamental information, such as geotag and event severity, therefore it can be classified as *incomplete*. There are some techniques to mitigate data incompleteness in LBSM. For instance, Pinto et al. [23] proposed a record linkage approach to enrich incomplete data. Dubois and Prade [24] and Yagger [25] used possibility theory and the probability of fuzzy events to handle imperfect data.

The *Vagueness* corresponds to an unclear description without the context information related to an event. The above tweet shows vagueness due to the inability to precisely define the extension, position, cause or even those involved in the accident. Usually, a way to deal with vagueness is matching and fusing data from different sources to make the context surrounding an understandable event by systems and users.

The *Granularity* ranges from fine-grained to coarse-grained data related to an event. Fine-grained data contains enough information to accurately describe the event location, the affected direction(s), the severity of the accident, and so on. On the other hand, coarse-grained data provides a macro view of events with a broad description.

A. User Bias

The user bias comes from the diversity of user's experience and background while interpreting and reporting the current traffic conditions. For instance, suppose that Alice and Bob are both tweeting about the same traffic jam. However, Alice has been facing such traffic conditions for many years, but Bob just moved to the city few months ago. There is a high probability of Alice tweeting the traffic jam as a normal daily event whereas Bob may interpret the regular traffic situation as a chaotic one. Consequently, the user's perception may lead to bias introduction on data traffic from LBSMs

One way of reducing user bias is to use specialist's accounts (professional accounts) which are also reporting traffic information in a given area. The specialist may also introduce his own bias sharing information for a specific audience or place. However, data from regular users demands more pre-processing treatment before fed as input to RoDE. In the present investigation, we selected the specialists' accounts manually to reduce users' bias within the RoDE Route Services. We also checked that the data from these accounts are reliable, as detailed later in Sec. III-A. Then, the RoDE Event Services balance the reliability to obtain incident description from specialists also gathering the description of non-incident events (say cultural events and so on) from regular users.

B. Spatiotemporal Assignment

Tagging data, from heterogeneous sources, to space and time is a fundamental task for RoDE. Because geographical location and temporal tagging allow intelligent transportation systems to study and characterize a region at any instant or time interval. In this section, we discuss the challenges to extract spatiotemporal information from LBSM.

The *Spatial* aspect assigns geographical location (latitude, longitude and altitude) to the data, therefore, allowing RoDE to compile the context surrounding the data. Twitter users are able to geo-reference their tweets in three primary methods: (i) geographical references in a tweet message; (ii) tweets geo-tagged by the user or client application; and (iii) account profile ('home' location) set by a user. However, deriving this information, even when present, is not always a trivial task. For example, a tweet may contain the spatial location in written form instead of a geo-tag, requiring an algorithm to identify and extract a textual address, and also convert to latitude and longitude. Although such algorithms already exist, the inherent unstructured form and freedom of writing (e.g., abbreviations within a limit of only 280 characters) on LBSMs turn the spatial textual extraction a challenge. Moreover, such challenges often result in ambiguous information subject to misinterpretation. There are research efforts investigating these challenges as follows. Liu et al. [26] and Finkel et al. [27] used Natural Language Processing (NLP) techniques to obtain parts of speech and entity recognition to label sequences of words that are the proper nouns. Li et al. [28] optimized NLP techniques to tweets text.

Information availability is another challenge for spatial data assignment. Some regions will have more coverage than others because of factors such as the density of inhabitants per square meter and number of tourists. For instance, large cities tend to have higher spatial coverage of LBSMs than smaller towns because of differences in the number of users, penetration of smart phones, young population, tourism, companies and traffic related information, or other complex social aspects out of scope of this investigation.

The *Temporal* aspect is the key to understand the past, present, and, possibly, the future scenario of the transportation system. LBSM platforms usually assign a timestamp when users input data to the system. However, this markup may not represent the same moment as when the event occurred. Thus, some open questions about temporal assignment are the following: *What is the validity of data published by a user of LBSM? How can we characterize the delay between the event and the data input on LBSM platforms?*

C. Inconsistencies

This subsection discusses data inconsistency challenges we faced while using data from LBSM platforms. We focus on the inconsistencies related to conflicts and out of order aspects because previous literature already addressed other inconsistencies e.g., [2], [22].

The *Conflict* on LBSMs appears when two or more data sources diverge about a specific event. For instance, suppose that Alice and Bob share their feelings about the same traffic event. Alice reports that nothing serious happened and the traffic flows well, while Bob reports that a severe accident happened which promotes a negative impact on the traffic. Based only on these two points of view, it is difficult to determine what happened. In the literature, the Dempster-Shafer evidence theory has gained notoriety in reducing data source divergences [29], [30]. Also, it is possible to give a reputation weighting to users' accounts, and then apply rules to decide on the most credible information. In this work, we acquired data from regular and specialist accounts in order to reduce the chances of conflict aspects. Moreover, the analysis of frequent terms tend to reduce conflict when most of the users agree with the description of traffic state, varying just in its intensity but not in reporting a completely different situation. For instance, users can report traffic at different levels, but they usually do not report that there is no traffic.

The *Out of order* occurs because of the freedom offered by LBSM platforms which allow users to enter traffic and transit information out of sequence into the system. These data appear as inconsistent to systems like RoDE because they are posted hours/days after the event (temporal dimension). Therefore, we have to consider how to use such data properly. Usually, the trivial solution is to discard the out of sequence data. However, if the data was identified correctly and then sorted, it may be used as a feedback data at the cost of more processing and storage resources.

VI. RoDE: Route Service

In order to provide a useful route service, we started with the hypothesis that it is possible to provide a route service

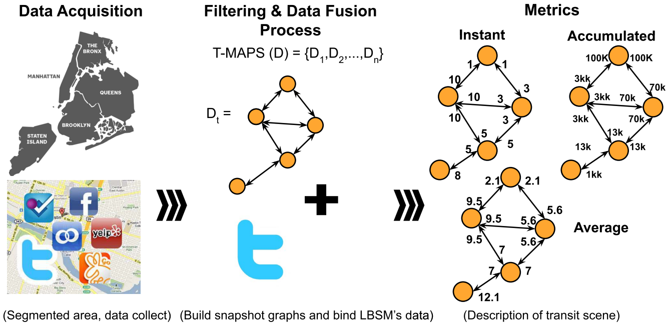


Fig. 7. T-MAPS modeling process.

using data from LBSM. In order to verify this hypothesis, we studied the relation between real traffic jams and the tweets about these jams. In the next sections we describe how we collected and characterized LBSM data to develop Twitter MAPS (T-MAPS). Twitter MAPS (T-MAPS) enhances the navigation context by adding the users' viewpoint in a new layer compiled from LBSM data in different ways. For example, by evaluating tweet's frequency or users' perspective on a region of interest.

A. Modeling Process of Twitter MAPS (T-MAPS)

The T-MAPS is a low-cost spatiotemporal model to enrich traffic events (jams and accidents) using tweets from the same area posted online at around the same time the traffic events occur. This model allows the representation of the traffic scenario highlighting different aspects by considering instantaneous or historical data, and also using text mining techniques over relevant tweets. Below, we present the three steps of our modeling process as discussed in [19], see Fig. 7.

1) Data Acquisition: this step consists of segmenting the area of interest and retrieving geotagged data from LBSM platforms. We use the geographic shape of a given area and its neighborhood segmentation to develop the T-MAPS model.

2) Filtering and Data Fusion Process: this step aims to filter and bind LBSM data to the segmented region. We use a weighted time-varying digraph as a model to map these areas and data. The time-varying digraph is represented as a series of static networks, one for each time step. Formally, let R be the set of segments of the region, then a snapshot digraph is defined as $D_t = (V, E, m)$, where $V = \{r | r \in R\}$ denotes the segmented region, and $E = \{(u, v) \in V | u$ is adjacent to v in R segmentation$\}$ denotes the directed edges between physically connected regions, and m is the weights (discussed below). The T-MAPS time-varying digraph is a sequence of snapshot digraphs, thus T-MAPS$(D) = \{D_{t=t_{\min}}, D_{t+\Delta}, \dots, D_{t_{\max}}\}$, where $t_{\min}$ and $t_{\max}$ are the start and end time of the available dataset, and Δ can be adjusted conveniently. Notice that our digraph connect each region with a bidirectional edge.

3) Metrics: it consists of assigning cost weights to the directed edges. Formally, $m(u, w) : E \rightarrow value$, where $m(u, w)$ is a function mapping the directed edges to a cost metric. The metric function represents a particular traffic scenario analyzed using the LBSM data posted within the same geographical region. Fig. 7 illustrates an example of the T-MAPS modeling process. First, we segmented the NYC map into five regions of interest, then we collected LBSM available data posted from these regions. Next, we obtained the digraph $G = (V, E, m)$, where V is the set of regions, and E the directed edges between adjacent regions. Then, we bound Twitter's traffic data to the resulting regions graph. Finally, the weights were assigned to the edges using different metric functions. The resulting time-varying digraph allows us to analyze the traffic scenario taking into account both conditions and descriptions of events. The main metrics are the following:

- *Instant:* this metric considers all tweets at a given instant of time t. This strategy corresponds to a snapshot of the traffic scenario. Usually, instantaneous data are sparse with limited coverage of the region of interest. However, these data may highlight an event at a given time.
- *Accumulated:* this metric considers all data acquired within a given time frame. It requires two parameters, t_{start} and $t_{\text{reference}}$, where $t_{\text{start}} < t_{\text{reference}}$ and must respect the temporal dataset availability, which stores all data between t_{start} and $t_{\text{reference}}$. One can interpret this metric as a historical metric looking into the past until the reference time point. In our experiments $t_{\text{start}} = t_{\min}$.
- *Average:* it uses a similar same approach if compared to the *Accumulated* metric described above. However, the values assigned to the edges are the average of tweets' occurrences over time, compiled daily, monthly and yearly. This information is the mandatory input to the metric function. One can interpret it as a typical traffic condition metric computed within a given time frame.

In short, the T-MAPS modeling process acquires social media posts that are represented as a graph matching both space and time of a given area. Our model also includes three metrics used to evaluate different routes using feedback from the users (tweets).

B. A Case Study

We conducted a case study to demonstrate the potential of T-MAPS. We compared T-MAPS recommendations against Google Direction[5] routes (GD) computing the similarity between the output of both systems. Later, we present three route description services demonstrating the potential of T-MAPS to enhance and describe the routes suggested as well as to give an overview of the traffic scenario. The Manhattan region was segmented into 29 official neighborhoods[6]. Consequently, the T-MAPS digraph snapshot contains 29 vertices. Besides, the minimum time interval between two consecutive T-MAPS graphs corresponds to a $\Delta = 1$ hour. Although T-MAPS was designed to accommodate both data resolutions (micro and macro), the case study used a macro viewpoint due to data coverage limitation.

1) T-MAPS Applicability: We evaluated the T-MAPS applicability by comparing its similarity, in recommended routes, with GD. Note that the T-MAPS route suggestion considers a macro resolution of the regions on the map, but our model is flexible enough to encompass fine-grained resolution

[5] We consider Google Directions as the most accurate representation of the traffic scenario, although this assumption is not a verified fact.

[6] www1.nyc.gov/site/planning/data-maps/open-data/districts-download-metadata.page

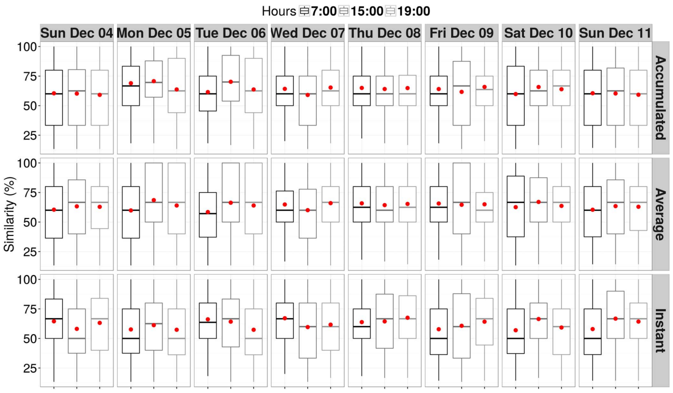


Fig. 8. Route recommendation similarity between T-MAPS and google directions (dots represent the mean).

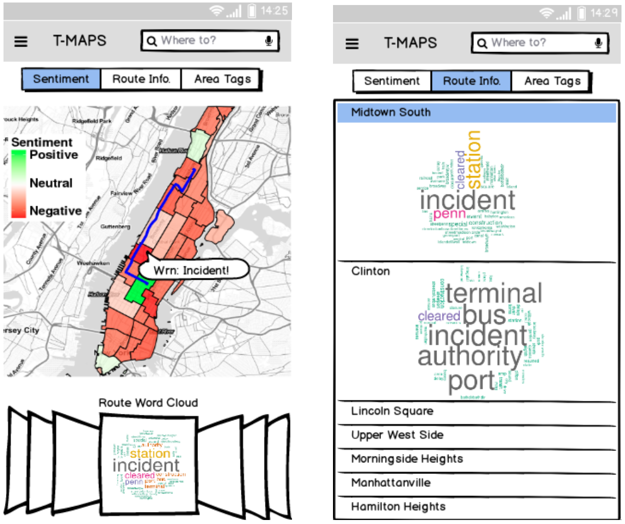
if there is enough data for this. From a macro resolution, T-MAPS aims to recommend regions which have the best conditions regarding the applied metrics. The aim here is to find the similarity between the routes suggested by GD and T-MAPS, in order to enrich these routes with a description based on the set of tweets in that path.

We query the T-MAPS and GD 812 times to recommend routes within Manhattan neighborhoods. The routes were derived from the combination $2 \times C_k^n$, where $n = 29$ (Manhattan neighborhoods) and $k = 2$ (origins and destinations). Note that we considered routes like $A \rightarrow B$ and $B \rightarrow A$. The routes start and end at the center of the region. Also, we rule out routes that start and end at the same local. We query the routes in three different moments (7:00 am, 3:00 pm and 7:00 pm) of a day throughout a week. Those moments were chosen purposely, based on its rush hour representation, and its higher volume of tweets in the dataset. Besides this, the Jam Factor from Here WeGo increases on those moments as well as the frequency of tweets per hour in the dataset, see Fig. 6.

The similarity was computed matching areas where the recommended routes by both T-MAPS and GD passed through (using Dijkstra's algorithm). Fig. 8 displays the similarity between routes along eight days in the dataset, considering three metric functions. The box-plots summarize 58,464 routes analyzed. T-MAPS with *Instant* metric showed a high variation of similarity rate, its median ranges from 50% up to 66.7%, while *Accumulated* metric shows 60% to 70% and *Average* metric 60% to 66.7%. It means that more than half of the evaluated routes overlapped the GD. We expected that *Instant* metric would pose the lowest similarity due to its intrinsic disparity with other metrics since it does not consider the historical data. As a global evaluation, the median of route similarity reached 62% with Google Directions. Note that T-MAPS uses a macro view, while GD does not, which implies in fewer regions per route by T-MAPS than GD. The upper quartile (1/4 of the routes) until the maximum value exhibited a similarity of 75% to 100% between the routes suggested by T-MAPS and GD.

C. Route Description Services

The applicability results demonstrated that it is possible to aggregate information to route recommendation, so we moved further to explore tweets. Considering that Google Maps does not deliver a detailed description of the routes, then any extra information added in that path will enrich the current scenario. Initially, we performed the cleaning phase in the tweet (lowercase transformation, accent removal, token extraction, and filtering stops words, links, and special characters). Then, we applied three types of text mining techniques to build the description services over the T-MAPS model: Route Sentiment (RS), Route Information (RI), and Area Tags (AT). Fig. 9 depicts the graphical user interface of a prototype to access the T-MAPS services.



(a) Route Sentiment (RS). (b) Route Information (RI).

Fig. 9. Route sentiment based on the tweets text analysis.

In Fig. 9(a), the RS service allows the user to observe the users' feelings (positive to negative) at a given area (for more detail see [19]). The RI service explores each area providing a word cloud, Figure 9(b), where the word size indicates its high-frequency over the route. The spread information enables the users to see the big picture of highlighted events in each area. Finally, we developed the AT service, Fig. 10. For that service, we used the Term Frequency (TF) and Inverse Document Frequency (IDF) – (TF-IDF) – method to measure how important is a word to a set of tweets in given area of Manhattan. This technique allowed us to find words which are unique for those explored area.

T-MAPS used the *Accumulated* metric to characterize Manhattan within our observation window. Any other metric can be applied to provide a different description, achieving a different goal. With these services (sentiment, route information and area tags), the T-MAPS can enrich the current route recommendation systems, indicating to the users an extra path description or even providing routing based on these descriptions. For instance, the user may choose a route which expresses good feelings and beautiful environment or alternatively, containing cultural activities.

D. Discussion

In summary, the results of *Route Services* showed the value of using social media data to enrich the data from transportation systems. We showed that the median of route

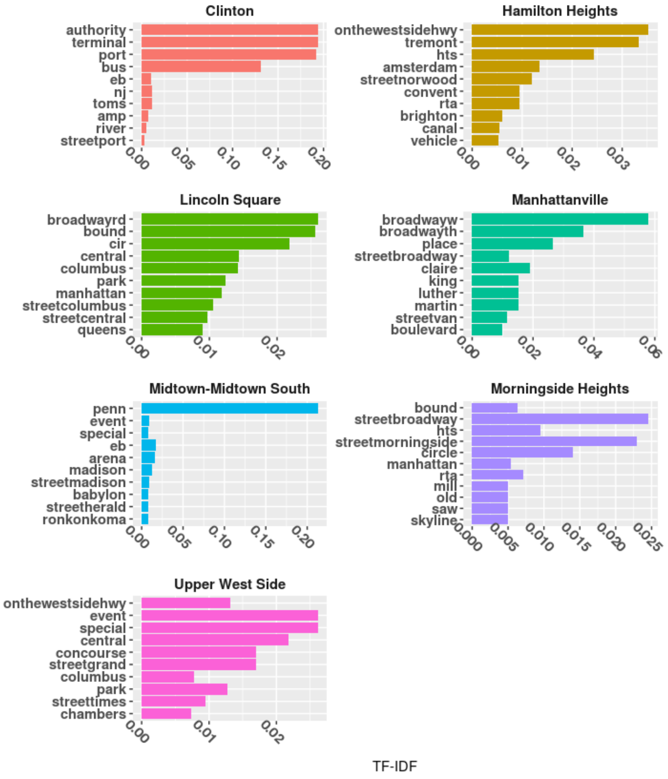


Fig. 10. The area' Tags (AT) of each region of the path.

similarity between our approach and Google Maps reached 62%, where T-MAPS uses region granularity while GD uses street granularity. For a quarter of the evaluated trajectories, the similarities achieved up to 100%. Based on that, we presented three route description services, aiming to enrich the current route services and navigation tools with a user's viewpoint. Using natural language processing techniques, we were able to create Route Sentiment (RS), Route Information (RI), and Area Tags (AT).

VII. RoDE: Event Service

After the experiments with our route services, the next challenge was to improve the detection and description of incidents in those routes. So we develop a new service called T-Incident which is a low-cost learning-based event detection and enrichment mechanism built using heterogeneous data fusion techniques. For this purpose, we designed a spatiotemporal grouping algorithm that fuses the incident data from two different data sources (i.e., HERE WeGo and Bing Maps), resulting in a new incident layer with more data coverage. Then, by using the same approach, we fuse (i) non-incident data (acquired from TripAdvisor), (ii) LBSM data (acquired from Twitter), and (iii) the new incident data layer obtained in the previous step. Moreover, we apply refined methods of NLP to extract patterns from social media data that may describe the incident event and its surrounding. Finally, we use a learning-based model to identify these patterns and detect the event types automatically. The results show the best setup of our T-Incident approach, achieving scores above 90%. Allowing the conception of incident detection and event description services using LBSM.

Algorithm 1 Spatiotemporal Grouping

```
Input: tweets, road events, radius
Result: tweets grouped by event, incident Id, and
        incident Type
/* The previous step splits each
   dataset into x slices, reducing the
   computation                                    */
initialization;
for each tweets do
    currentIncidentId ← 0;
    currentIncidentTmp ← None;
    currentDistance ← $\infty$; /* larger than radius
    */
    for each incidents do
        if equal(tweets.sec, incidents.sec) or diff(tweets.sec,
        incidents.sec) is (+ 1 or - 1) then
            /* Tweets between the incid. time
               */
            if TemporalFilter(incidents.starttime,
            incidents.endtime, tweets.timestamp) then
                /* Distance from the radius    */
                distance ← SpatialFilter(tweets.coord,
                incidents.coord, currentDistance, radius);
                /* Record the less distance    */
                if distance < currentDistance then
                    currentIncidentId ← incidents.Id;
                    currentIncidentTmp ← incidents.Type;
                    currentDistance ← distance;
                end
            end
        end
    end
    /* Assigning the event type
       (Incident, Non-Incident, Unknown)
       for each tweet                              */
end
```

A. Incident Data Fusion

In this section, we present a method to increase the coverage of incident data and enrich its description by fusing data from different sources. We argue that the greater the number of incidents used, the more tweets can be grouped, benefiting our learning-based approach. After acquiring data from HERE WeGo and Bing Maps platforms, we pre-processed them to standardize their features.

Thereafter, we conducted a spatiotemporal grouping (see Sec. VII-B.1 and Algorithm 1 for more details). However, the goal here was to identify an incident event reported by both data sources, thus representing the same event. In this case, the temporal interval and the spatial location of them must be very close. We assume that two events are close, and, therefore, the same, if they start on the same day and hour but are also located at most 10 meters apart from each other. We named these same events as *Intersection*. Fig. 11 shows the frequency of each incident type by a given data source.

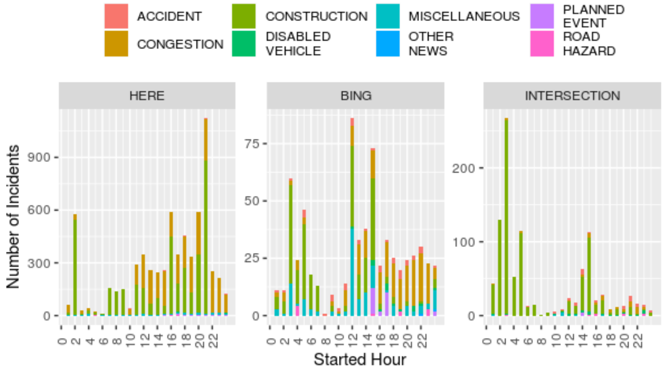


Fig. 11. Hour of an incident by data source and the intersection of them.

Moreover, we can see the same events reported by both sources in the *Intersection* graphic.

We also evaluated the similarity of incident types from the *Intersection*. We found that the incident type similarity between HERE and Bing reached 99.83%. In other words, both data sources labeled the incidents almost similarly. As a final step, we created a *New Incident Layer*, which combines the data coverage from both data sources and increases the information description about incidents, using the intersection of them. Since each data source has its own way of reporting incident events (e.g., detailing the road name or providing a short textual description) the fusion enriches the whole context.

Fig. 12 shows the spatial data coverage of each data source and the intersection between them, during the process of data acquisition (2018-09-14 to 208-11-06). It also shows the data representativeness for each source. Let us consider HERE WeGo as H and Bing Maps as B, we have $(H \cup B) = 100\%$ of the data, $H = 80.53\%$ of the whole data, while $B = 8.31\%$ and the *Intersection* $(H \cap B) = 11.16\%$. The *New Incident Layer* covers 100% of the entire data collected, where more than 11% of matching incidents could be enriched based on a short description from HERE WeGo and the information of street intersection from Bing Maps.

B. T-Incident Design Architecture

This section presents a learning-based incident detection approach based on heterogeneous data fusion. We started with the hypothesis that LBSM can provide valuable information about the traffic conditions and eventual incidents, as previously discussed in [19].

Given the different data sources used as input to our design, we created a spatiotemporal grouping algorithm to combine together these different data sources (see Sec. III-B) in both temporal and spatial dimensions. In sequence, we extracted the pertinent features to compile the user's viewpoint around each event previously grouped by our algorithm. Then, we developed a learning-based model to identify potential incidents considering the user's viewpoint. Finally, we evaluated our approach using different spatial grouping modes. This section describes in detail each stage of T-Incident as depicted in Fig. 13.

1) Spatiotemporal Grouping: The grouping approach considers the heterogeneity of the data sources used and its spatiotemporal coverage variation. Therefore, we proposed an approach which merges the incident/non-incident data layers with the tweets layer based on both dimensions. To do that, we considered the incident as an event regardless of its type (i.e. accident, construction, road hazard, disabled vehicle and traffic), therefore also combining the different types in only one event – *Incident*. Each incident has a start location, an end location, and a time duration. Our model considers only the incident start location as same to the events named – *Non-Incident* (point of interest). Another characteristic of our data preparation consists of setting the non-incident time interval with the same time interval of the data acquisition. In other words, there is no time duration for non-incident events. Its duration starts and ends with the data acquisition process.

Based on the incident and non-incident dataset, we are able to conduct a temporal filter which looks for the intersection

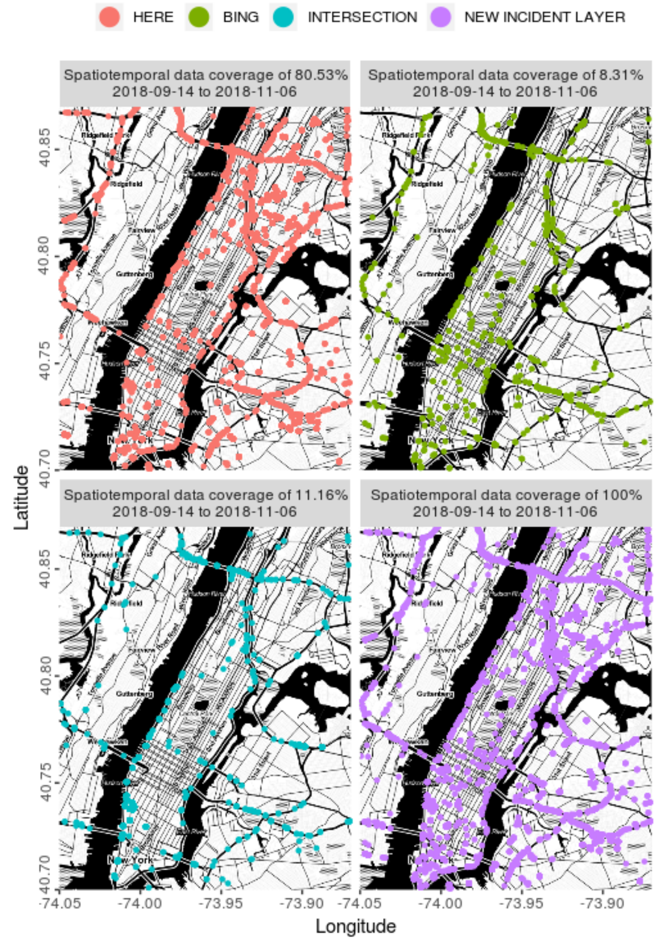


Fig. 12. Spatial incident coverage per data layer.

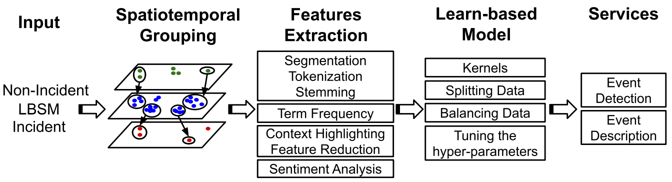


Fig. 13. Design of T-Incident.

TABLE II

NUMBER OF TWEETS FOR EACH SPATIOTEMPORAL GROUPING MODEL

Event	**Radius (km)**						
	0.01	0.05	0.1	0.2	0.3	0.4	0.5
Incident	121	959	3098	9467	30085	63853	68877
Not Incident	260	3161	6522	13060	20699	30492	35786

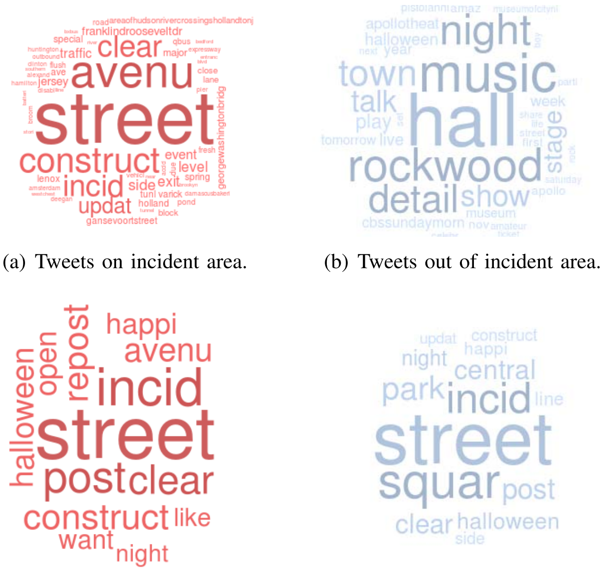
between events and tweets. After merging those data, we perform a spatial filter based on the radius of each event location. We created a set of radii, aiming to identify the better grouping mode since we are dealing with user bias and the vast amounts of unrelated data. This methodology enabled T-Incident to group a different number of tweets around the event (see Tab. II), and, thus, the information surrounding the event can be more valuable to the context or more generalized to it. Notice that we used tweets with coordinates provided by a GPS. GPS needs line-of-sight with the sky to compute the coordinates therefore, urban barriers may interrupt the sensor. Thus, we discarded tweets without geotags.

Even though the spatiotemporal grouping could be conducted in different ways (e.g., based on streets segment, neighborhoods and a grid dividing the geographical area), we chose the use of different radii around the incident, as our initial approach. Tweets, which were not grouped by that algorithm, were labeled as *Unknown* and removed. We noticed a trade-off to choose the radius size and the relevance of information floating around the event. In other words, a small radius implies fewer data grouping, but relevant information about the event. A larger radius results in more data grouped, but less descriptive information about the event. This situation becomes a challenging task when there are reduced amounts of data acquired.

We describe the spatiotemporal grouping in Algorithm 1. The inputs to the grouping algorithm are the tweets, road events (incidents and non-incidents) and the radius. The expected result is an updated tweet dataset containing the *event*, *incident id*, and *incident type*. We also developed an optimization process dividing the geographic area, latitudinally, in x sections, aiming to reduce the number of operations conducted in large areas with large amounts of data. After that, for each tweet and incident, we tested if they were in the same section or near with one hop up or down (Line 7). Once satisfied, the tweet must be between the incident start and end time (Line 8). Then, we measure the distance between the tweet and the incident, aiming to find the minimum distance to assign its new attributes (Lines 9-14).

2) Feature Extraction: We assume that the information of interest floats around the observation location. Stressing the model based on a radius around the event, makes it an intuitive and powerful approach, as shown in Sec. VII-C. However, data from LBSM brings issues that can lead to other challenges such as data imprecision and user bias. In that way, the feature extraction role aims to clean the tweet and provide a set of words which describes the event's surrounding better.

We first applied for each grouping model and event class a set of NLP methods such as lowercase transformation, accent removal, token extraction, and filtering stop words, links, and special characters. After that, we reduced inflectional and derivational forms of a word to a common base form. Then, we analyzed the Term Frequency (TF) from the event, extracting a matrix of the most frequent words mentioned in that area. Moreover, we filtered that matrix based on the sparsity, i.e., we removed terms that were sparse lesser than 0.98%.



(a) Tweets on incident area.

(b) Tweets out of incident area.

(c) Tweets on incident areas.

(d) Tweets out of incident areas.

Fig. 14. Spatiotemporal grouping based on a radius of 0.01 km ((a) and (b)) and 0.5 km ((c) and (d)).

We also introduced a context highlighting the step for a specialist to reduce non-related words of a given event. This is because, even though we conducted the previous steps, our LBSM dataset still contained non-related words (noise) which had to be removed. Most of the noise comes from regular user's account, which we used to increase the description of incidents, but specially to describe non-incident events. We noticed, by experiments, that the Term Frequency-Inverse Document Frequency (TF-IDF) approach does not stress the words which describe each event' class accurately. Then, this analysis was not enough to derive valuable information about the event.

At the end of this process, we gathered the set of most important words posted by regular and specialist Twitter users. Fig. 14 shows an example of a set of words grouped by radius between 0.01 km and 0.5 km. This indicated how specific or general was the information shared within the event radius. Figs. 14(a) and 14(b) show more words clouds, weighting them differently and reducing the intersection between incident and non-incident events. However, upon increasing the radius we can see fewer words with higher weights stressing common words between both classes [see Figs. 14(c) and 14(d)]. Our goal is to understand this behavior and train an algorithm to automatically identify these classes.

a) Feature reduction: The number of features obtained from the last stage may be large enough to introduce computational barriers such as the processing time, memory and storage capacities. We conducted a method to reduce the number of features based on their importance and frequency. In other words, we initially developed two approaches to

TABLE III

RELEVANT FEATURES BASED ON RADIUS OF 0.01 km

Event	Most Frequent Features					
Incident	traffic	side	exit	contruct	incid	accid
	avenu	street	updat	georgewashingtonbridg	clear	
	jersey	event	major	franklindrooseveltdr	level	
Non-Incident	town	night	year	apollotheat	show	
	detail	hall	music	halloween	stage	
	week	play	live	rockwood	talk	

achieve that goal. The first one was the Principal Component Analysis (PCA) to extract a set of relevant features. This process identifies the most variable information from a multivariate dataset and expresses it as a set of new features – Principal Components (PCs). These PCs represent the directions along which the variation in the data is maximal. The second one was based on the ranking of the most frequent words.

Both methods output the results to the specialist who makes the decision. We noticed that the PCA did not catch a good set of words such as the use of most frequent words. When the tweet dataset was acquired without a *track of words* (any tweet, without specific words), the PCA performs better than the use of the most frequent words. On the other hand, PCA was not suitable for tweets with a set of specific track words as mentioned in Sec. III-B. As a result, we performed the feature reduction for each grouping model and event class, extracting only the most representative set of words from the previous step. Tab. III shows an example of features obtained after ranking the most frequent words from the spatiotemporal grouping using a radius of 0.01 km.

b) Sentiment analysis: The sentiment analysis was conducted for each tweet, grouping and event class, allowing us to extract the feelings that Twitter users shared about a particular event. To derive the sentiment from the tweet's text, we used a dictionary of words and its associated feelings [31]. The sentiment depends on the occurrences of the number of words/feelings to calculate the score, and we can then associate a sentiment (positive or negative) to the tweet. As a result, for each tweet we extracted the set of words corresponding to feelings and its frequencies, binding them with the set of words processed in the previous stage, for that same tweet, increasing the features which describe somehow these events.

3) Learning-Based Model: The last stage was responsible for extracting useful information which better describes a given class of event and feeds our learning-based model with a set of features labeled by the event. In this way, we started to deal with a classification problem. First, we chose the most common classification algorithms (*kernels*), used in the same context of this investigation, based on the literature review [7]. To conduct this step, we used the following *kernels*: Support Vector Machine (SVM), k-Nearest Neighbors (KNN) and Random Forest Classifier (RF).

Next, we split the data into two sets, following the convention of the most machine learning approaches: Training Set, corresponding to 70% of the entire dataset; and Test Set, corresponding to 30% of the entire dataset. To validate the training process, we applied the cross-validation considering 10 folds split in 70% and 30% of the training and test datasets, respectively. Our goal was to evaluate the training curve and the testing curve, avoiding possible overfitting and under-fitting. That partition was conducted for each grouping model.

As expected, the dataset was unbalanced because the number of tweets around the non-incident areas is bigger than around the incident ones. In this case, we explored the re-sampling techniques which aim to balance classes either increasing the frequency of the minority class or decreasing the frequency of the majority class. Our goal was to obtain approximately the same number of observations for both classes.

We used a random under-sampling, aiming to balance the class distribution by randomly picking and eliminating the majority of class examples. This strategy helps to improve run-time and storage by reducing the number of training data samples once the training is large enough, considering LBSM data. However, the classifier may suffer hard consequences since the potential useful information can be discarded. For this reason, this step is not limited to that approach, as it always depends on the quality and quantity of LBSM data acquired.

After balancing the dataset, tuning the hyper-parameter became a challenging task and an exploratory approach was adopted to deal with. We used a *GridSearchCV* class from Scikit-Learn API [32], which takes a set of parameters and values to exhaustively combine them, aiming to find the best configuration. Knowing that the complexity of such search grows exponentially with the number of parameters, we defined a set of parameters for each *kernel* following some guidelines. For the SVM, we based on [33], and for the other ones, we followed the user's guide for Auto-WEKA [34].

4) Services: The results of the learning-based model allowed us to understand the best spatiotemporal grouping and the set of NLP methods to filter the LBSM texts, and, then, to accurately outline the events. Based on that, we were able to output the incident and non-incident event detection service and the event description service.

Once we identified an event, we started to analyze its context. To do that, we conducted a text summarization process, aiming to create a short and coherent version of a longer document. We considered a document a set of tweets grouped by incident type, i.e., we applied the text summarization to a group of tweets labeled by incident type and hour, and by incident id. This process provides a short description for each group, providing the users and traffic planners the viewpoint of the LBSM users regarding the transit events and points of interest.

In that area, there are two methods of text summarization: *Extractive* and *Abstractive*. The first one selects the tweets, ranks their relevant phrases and chooses only those which are meaningful to the event. The abstractive method aims to generate entirely new sentences to capture the meaning of the event. For this version of T-Incident, we developed the event description service, using the extractive method.

C. Evaluation

In this section, we describe T-Incident performance evaluation against the set of classification algorithms and

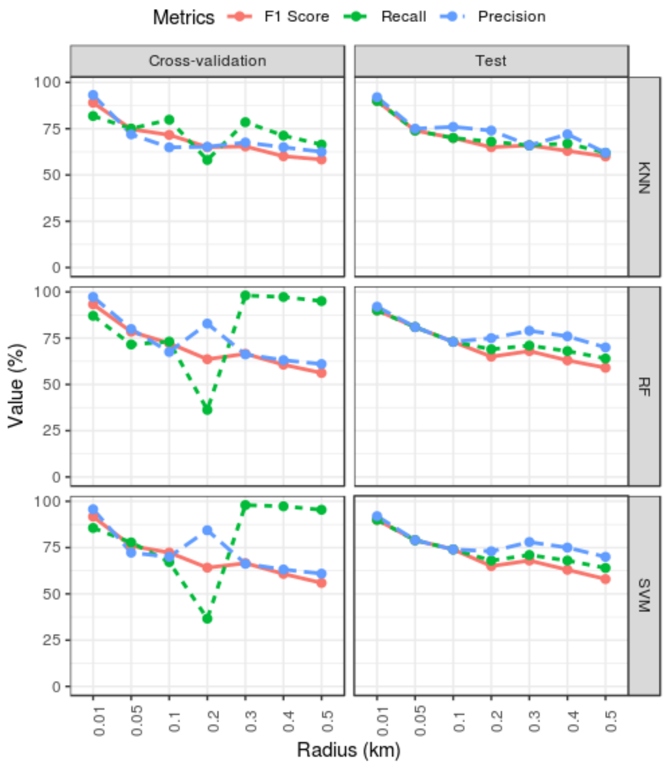


Fig. 15. Classification results based on different kernels and metrics.

spatiotemporal grouping modes as outlined in Sec. VII-B. Then, we present services to detect and enrich the event description.

1) Event Detection: Our incident detection approach was based on an exploratory analysis of classification algorithms, hyper-parameters and radius. Fig. 15 shows the results regarding a Training and Test process. We validate our training process performing a *Cross-validation approach* which aims to split the training set in training and validation sets among 10 folds. Fig. 16 shows the learning curve of each kernel performing on a spatiotemporal grouping with radius of 0.01 km and 0.5 km, as an example. The main goal here is to study the generalization of a given model, avoiding over-fitting and under-fitting, and find out the best spatiotemporal grouping. We noticed that the radius of 0.01 km [Figs. 16(a), 16(b), and16(c)] delivers the best score, around 90%, in most kernels after 140 training samples where we see the curves converging and the model stabilization. However, the reduced amount of data limits the exploration of event description service.

Once we increase the radius, we were able to see the curves decreasing as depicted in Figs. 16(d), 16(e), and 16(f). Using a 0.5 km radius, we observed a score between 58% and 65%. Decreasing the radius to 0.4 km, we noticed averaged scores above 61% and below 65%. A radius between 0.3 km and 0.2 km showed very close results as scores above 65% and below 70%, in average. Using 0.1 km, we obtained scores around 70%, and between 75% and 80% considering the radius of 0.05 km.

We deal with a trade-off between higher radius (more grouped data and smaller scores) and lower radius (fewer data and and higher scores). The important lesson learned here is the application of a consistent methodology that was able to provide a generalization model to detect incidents. Next, we evaluated three metrics from the Cross-validation and Test:

i) *F1 Score:* is the weighted average of Precision and Recall. This score takes both false positives and false negatives into account $(2 \times Recall \times Precision)/(Recall + Precision)$;

ii) *Recall:* measures how good a test is at detecting the positives $(TP/TP + FN)$;

iii) *Precision:* is the ratio of correct predicted positive observations to the total predicted positive observations $(TP/TP + FP)$.

Fig. 15 shows the best set of parameters (kernel and radius) that can feed the T-Incident service. As suggested in the learning curves, the better spatiotemporal grouping could be the radius of 0.01 km which shows a Test score above 90% in all metrics evaluated. However, we achieved very good scores, above 70%, due to the quality of LBSM data. Taking this fact in consideration, we can even use a 0.1 km radius keeping the *F1 sore*, *Recall* and *Precision* around 75% on average. After the spatiotemporal grouping, we observed

a) the correlation among those scores and the radius sizes, and

b) the decrease of scores, which can be explained by the increase of intersection between the incident and the non-incident set of features.

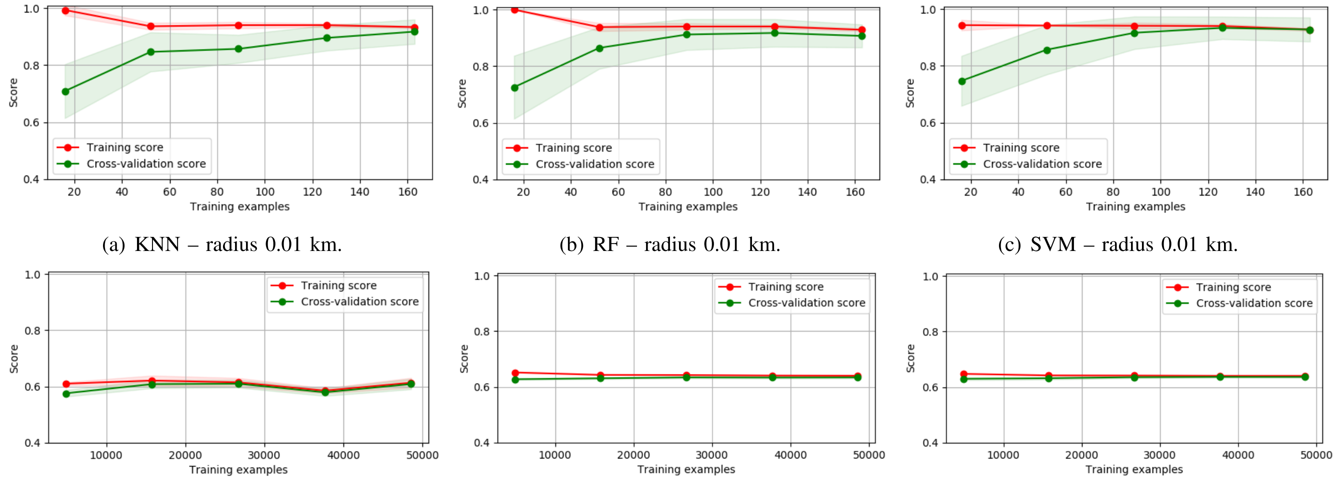
2) Event Description: The results observed in the detection stage, allowed us to identify the best spatiotemporal grouping which accurately outlines the event. In this sense, we conducted a text summarization process, based on the *Extractive* method, creating a short and coherent version of the event. Notice that we used for this analysis the spatiotemporal grouping with both radii 0.01 km and 0.1 km, based on the trade-off between accuracy and size of the data sample.

As an example of the T-Incident description service with a radius of 0.01 km, the Text 1 summarizes a specific incident event on *Franklin D Roosevelt Drive*. We highlighted the words to make this text clear for the reader to understand what happened there. With that analysis at hand, we aim to enable users and road managers to understand and decide what can be done about it.

Cleared: Construction on #FranklinDRooseveltDrive SB from Exit 9 - East 42nd Street to 34 street; **Updated**: Incident on #FranklinDRooseveltDrive SB at Exit 9 - East 42nd Street; **Cleared: Incident** on #FranklinDRooseveltDrive SB at Exit 9 - East 42nd Street; **Incident** on #FranklinDRooseveltDrive SB at Exit 9 - East 42nd Street; **Closure** on #FranklinDRooseveltDrive NB at Exit 9 - East 42nd Street; **Cleared**: Closure on #FranklinDRooseveltDrive NB at Exit 9 - East 42nd Street; **Construction** on #FranklinDRooseveltDrive Both directions at Exit 9 - East 42nd Street

Text 1: Incident description with a radius of 0.01 km.

At the same time, using the spatiotemporal grouping with a radius of 0.1 km, for instance, we analyzed a specific non-incident event, the *Town Hall* and its surroundings. The Text 2 summarizes that area, highlighting the top trends of



(a) KNN – radius 0.01 km. (b) RF – radius 0.01 km. (c) SVM – radius 0.01 km.

(d) KNN – radius 0.5 km. (e) RF – radius 0.5 km. (f) SVM – radius 0.5 km.

Fig. 16. The learning curve of a given kernel and spatiotemporal grouping.

Open House New York Sunday Stop 1! **Town Hall**. It was never taken over by the Broadway **theatre** giants because ther..; #30DaysForMyArt DAY 16: ”Go see a **broadway show**.” It’s simple: There is NOTHING like a broadway show. I’ve lived in; **Beastie Boys Book**: Live; Direct with Adam Horovitz; Michael Diamond: The **Town Hall**; Good morning **Times Square**. Bad I have to leave today! (**@Millennium Broadway Hotel** - @millenniumpr in New York, NY); Good night! (**@Millennium Broadway Hotel** - @millenniumpr in New York, NY); YEP! I Like Wrestling Podcast #45: WWE Super Show-Down Predictions, Raw; Smack; Show Time! (@Beautiful: The Carole King Musical in New York, NY); **Mooch’s book party**. Really. (@Hunt; **Fish Club** in New York, NY); Head over heels wPeppermint!!! (@Hudson Theatre - @hudsonbway for Head Over Heels in New York, NY); I had the heirloom tomato lobster salad. **Kristine had the burger** (@Burger; Lobster in New York, NY);

Text 2: Non-Incident description with a radius of 0.1 km.

Cleared: Construction on **#GeorgeWashingtonBridge** WB from New York SideLower Level to New Jersey SideLower Level; **Cleared**: Construction on **#WLine** Both directions from Whitehall Street-South Ferry Station to Ditmars Boulevard-Astoria Station; **Updated**: Construction on **#WLine** Both directions from Whitehall Street-South Ferry Station to Ditmars; **Cleared**: Construction on **#NY9A** SB from West 42nd Street to West 38th Street; **Cleared**: Closure on **#RiversideDrive** Both directions from West 145th Street to West 155th Street; **Cleared**: Construction on **#FranklinDRooseveltDrive** SB from Exit 9 - East 42nd Street to 34 street; **Cleared**: Construction on **#M42Bus** Both directions at 42 St at 12 Av and the 42 St Pier; **Closed** in #NewYork on **42nd St WB between Lexington Ave and Madison Ave**, **stop and go traffic** back to 3rd Ave #traffic; **Accident**, center lane blocked in #HudsonRiverCrossingsGwb on The G.W.B. Upper Level Outbound after The Harlem Riv; **Accident**, left **lane blocked in #HudsonRiverCrossingsGwb** on The G.W.B. Upper Level Outbound after The Harlem River

Text 3: Incident description in Manhattan at 5 am with a radius of 0.05 km.

places which were extracted by users’ impressions. In that way, it is possible to find out cultural places; where to book a hotel’ room and where to dine in that area.

Moreover, the T-Incident description service provides an overview of incident events in each area and at a given day and hour. The Text 3 was summarized considering the spatiotemporal grouping with a radius of 0.05 km in Manhattan at 5 am, for instance. It delivers to the users and road managers a feasible and low-cost way to understand areas which may be avoided or even to take careful attention during that hour. Notice that, our analysis aims to focus on the top trends of incident events at a given day and hour, enriching the current context and delivering to the public a very short and summarized information.

D. Discussion

The *Event Services* delivers real-time incident alerts using a learning based approach, which considers the tweets historical “trends”. After the training process, the model can act in real-time, delivering to the user and managers the incident event (incident or non-incident), based on one or more tweets. Besides, the description service can group that event with its surrounding delivering a summarized description. We also noticed that, navigation tools provide incident events in different time interval compared to user reports. This event services also enables to evaluate the quality of data on transportation system scenario.

This study does not show the case where there is incident data reported by navigation tools but not LBSM data surrounding it. In these cases, our methodology is not able to detect or even enrich them. However, RoDE aims to complement the current transportation system data, and not replace them. In this sense, the navigation tools still use incidents with poor description, when there is a lack of LBSM data availability. On the other hand, RoDE can detect incidents not reported by navigation tools and also enrich incidents with a summarized description.

In summary, the results showed the best set of parameters that can feed our T-Incident approach, leading to the event detection and event description services. The better spatiotemporal grouping mode considered the radius of 0.01 km, showing that incident detection scores above 90% in all evaluated metrics. However, we considered that a very good result presents scores above 70% due to the quality of LBSM data. As a result, the event description service allowed us to provide a summarized description for each group, providing users and traffic planners the viewpoint of the

LBSM users regarding the transit events and points of interest.

VIII. Conclusion

This investigation introduced the Road Data Enrichment (RoDE) framework, a low-cost solution to ITSs based on Heterogeneous Data Fusion. RoDE delivers high-level information through two sets of services, namely *Route Services* and *Event Services*. As a result, navigation systems, road planners and the general public can access a more descriptive and enriched transportation system data. We have been trying to validate the hypothesis that the viewpoint from LBSM users can complement current transportation system tools. Therefore, the effort we spent to develop these methodologies allowed us to quantitatively evaluate how much we can enrich (qualitatively) the suggested information from traditional traffic systems. We did a qualitative analysis by converting textual data into traffic/incident information (sentiments, route and incident descriptions) and quantitative analysis by converting words frequency, number of incidents, and jam factors into traffic/incident information.

The results discussed here may serve as a basis for further exploration of new research ideas. As future work, we plan to increase the quality of the route description by implementing different learning strategies within T-MAPS. T-MAPS could also use data from regular LBSM users' accounts and use reputation models to handle conflicting information. We also plan to extend the event services by developing strategies to eliminate user intervention in the feature extraction stage. In addition, we have noticed that LBSM provides data with some time differences when compared to navigation tools. In other words, Twitter users report incidents sometimes early or later compared to the incident start and end time of navigation tools, respectively. In this sense, T-Incident can be extended to evaluate the incident duration time considering the LBSM, or even evaluate when a given user message is no longer valid based on the incident event duration.

References

[1] A. L. Bazzan and F. Klügl, *Introduction to Intelligent Systems in Traffic and Transportation*. San Rafael, CA, USA: Morgan & Claypool, 2013.

[2] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Inf. Fusion*, vol. 14, no. 1, pp. 28–44, Jan. 2013.

[3] J. Yin and Z. Du, "Exploring multi-scale spatiotemporal Twitter user mobility patterns with a visual-analytics approach," *ISPRS Int. J. Geo-Inf.*, vol. 5, no. 10, p. 187, Oct. 2016.

[4] S. S. Ribeiro, Jr., C. A. Davis, Jr., D. R. R. Oliveira, and W. Meira, Jr., "Traffic observatory: A system to detect and locate traffic events and conditions using Twitter," in *Proc. 5th ACM SIGSPATIAL*, 2012, pp. 5–11.

[5] J. Kim, M. Cha, and T. Sandholm, "SocRoutes: Safe routes based on tweet sentiments," in *Proc. 23rd Int. Conf. World Wide Web*, 2014, pp. 179–182.

[6] P. H. Rettore, G. Maia, L. A. Villas, and A. A. F. Loureiro, "Vehicular data space: The data point of view," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2392–2418, Apr. 2019.

[7] S. Xu, S. Li, and R. Wen, "Sensing and detecting traffic events using geosocial media data: A review," *Comput., Environ. Urban Syst.*, vol. 72, pp. 146–160, Nov. 2018.

[8] A. Crooks, A. Croitoru, A. Stefanidis, and J. Radzikowski, "Earthquake: Twitter as a distributed sensor system," *Trans. GIS*, vol. 17, no. 1, pp. 124–147, 2013.

[9] M. Hasan, M. A. Orgun, and R. Schwitter, "A survey on real-time event detection from the Twitter data stream," *J. Inf. Sci.*, vol. 44, no. 4, pp. 443–463, Aug. 2018.

[10] K. Z. Bertrand, M. Bialik, K. Virdee, A. Gros, and Y. Bar-Yam, "Sentiment in New York City: A high resolution spatial and temporal view," 2013, *arXiv:1308.5010*. [Online]. Available: https://arxiv.org/abs/1308.5010

[11] A. Giachanou and F. Crestani, "Like it or not: A survey of Twitter sentiment analysis methods," *ACM Comput. Surv. CSUR*, vol. 49, no. 2, pp. 1–41, 2016.

[12] P. Giridhar *et al.*, "ClariSense+: An enhanced traffic anomaly explanation service using social network feeds," *Pervas. Mobile Comput.*, vol. 41, pp. 381–396, Oct. 2017.

[13] I. Septiana, Y. Setiowati, and A. Fariza, "Road condition monitoring application based on social media with text mining system: Case study: East Java," in *Proc. Int. Electron. Symp. (IES)*, Sep. 2016, pp. 148–153.

[14] Y. Gu, Z. S. Qian, and F. Chen, "From Twitter to detector: Real-time traffic incident detection using social media data," *Transp. Res. C, Emerg. Technol.*, vol. 67, pp. 321–342, Jun. 2016.

[15] M. A. Yazici, S. Mudigonda, and C. Kamga, "Incident detection through Twitter: Organization versus personal accounts," *Transp. Res. Rec.*, vol. 2643, no. 1, pp. 121–128, Jan. 2017.

[16] Z. Zhang, Q. He, J. Gao, and M. Ni, "A deep learning approach for detecting traffic accidents from social media data," *Transp. Res. C, Emerg. Technol.*, vol. 86, pp. 580–596, Jan. 2018.

[17] H. Nguyen, W. Liu, P. Rivera, and F. Chen, "TrafficWatch: Real-time traffic incident detection and monitoring using social media," in *Proc. PAKDD*. Cham, Switzerland: Springer, 2016, pp. 540–551.

[18] F. C. Pereira, F. Rodrigues, and M. Ben-Akiva, "Text analysis in incident duration prediction," *Transp. Res. C, Emerg. Technol.*, vol. 37, pp. 177–192, Dec. 2013.

[19] B. P. Santos, P. H. Rettore, H. S. Ramos, L. F. M. Vieira, and A. A. F. Loureiro, "Enriching traffic information with a spatiotemporal model based on social media," in *Proc. IEEE ISCC*, Natal, Brazil, Jun. 2018, pp. 464–469.

[20] M. Keyvan-Ekbatani, A. Kouvelas, I. Papamichail, and M. Papageorgiou, "Exploiting the fundamental diagram of urban networks for feedback-based gating," *Transp. Res. B, Methodol.*, vol. 46, no. 10, pp. 1393–1403, Dec. 2012.

[21] R. N. Bracewell and R. N. Bracewell, *The Fourier Transform and Its Applications*, vol. 31999. New York, NY, USA: McGraw-Hill, 1986.

[22] P. H. Rettore, B. P. Santos, A. B. Campolina, L. A. Villas, and A. A. Loureiro, "Towards intra-vehicular sensor data fusion," in *Proc. 19th Int. Conf. ITS*, 2016, pp. 126–131.

[23] C. Pinto *et al.*, "Probabilistic integration of large Brazilian socioeconomic and clinical databases," in *Proc. IEEE 30th Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2017, pp. 515–520.

[24] D. Dubois and H. Prade, "Possibility theory and data fusion in poorly informed environments," *Control Eng. Pract.*, vol. 2, no. 5, pp. 811–823, Oct. 1994.

[25] R. R. Yager, "Generalized probabilities of fuzzy events from fuzzy belief structures," *Inf. Sci.*, vol. 28, no. 1, pp. 45–62, Oct. 1982.

[26] X. Liu, S. Zhang, F. Wei, and M. Zhou, "Recognizing named entities in tweets," in *Proc. 49th AMACL*, 2011, pp. 1–9.

[27] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by Gibbs sampling," in *Proc. 43rd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2005, pp. 363–370.

[28] C. Li and A. Sun, "Fine-grained location extraction from tweets with temporal awareness," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2014, pp. 43–52.

[29] L. A. Zadeh, "Review of a mathematical theory of evidence," *AI Mag.*, vol. 5, no. 3, p. 81, 1984.

[30] M. C. Florea, A.-L. Jousselme, É. Bossé, and D. Grenier, "Robust combination rules for evidence theory," *Inf. Fusion*, vol. 10, no. 2, pp. 183–197, Apr. 2009.

[31] M. Jockers. (2017). *Syuzhet: Extracts Sentiment and Sentiment-Derived Plot Arcs From Text*. [Online]. Available: https://cran.r-project.org/web/packages/syuzhet/

[32] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[33] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, *A Practical Guide to Support Vector Classification*. Citeseer, 2003. [Online]. Available: https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

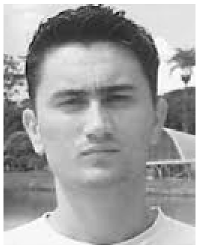
[34] L. Kotthoff, C. Thornton, and F. Hutter, "User guide for auto-WEKA version 2.6," Dept. Comput. Sci., BETA Lab., Univ. Brit. Columbia, Vancouver, BC, Canada, Tech. Rep. 2, 2017.

Paulo H. L. Rettore received the B.Sc. degree in computer science, in 2009, and the Ph.D. degree in computer science from the Federal University of Minas Gerais (UFMG), in 2019. He is currently a Scientist with Fraunhofer FKIE, Bonn, Germany. His research interests include computer networks, distributed systems, tactical networks, ubiquitous computing, the Internet of Things, intelligent transportation systems, and smart mobility.

Bruno P. Santos received the bachelor's degree from the Universidade Estadual de Santa Cruz (UESC), and the M.S. and Ph.D. degrees in computer science from the Universidade Federal de Minas Gerais (UFMG). He is currently a Professor of computer science with the Universidade Federal de Ouro Preto (UFOP). His research interests include computer networks, distributed systems, ubiquitous computing, the Internet of Things, intelligent transportation systems, and smart mobility.

Roberto Rigolin F. Lopes is currently a Scientist with Fraunhofer FKIE, Bonn, Germany. Sitting at the Communication Systems Department, he has been attacking problems in computer networks and distributed systems, with particular interest in the performance bounds of tactical systems. His education/experience as a scientist includes three universities in Brazil (UFMT, UFSCar, and USP), one in The Netherlands (UTwente), one in Norway (NTNU), and many books.

Guilherme Maia received the Ph.D. degree in computer science from the Federal University of Minas Gerais, Brazil, in 2013. He is currently an Assistant Professor in computer science with the Federal University of Minas Gerais. His research interests include distributed algorithms, mobile computing, wireless sensor networks, and vehicular ad hoc networks.



Leandro A. Villas received the master's degree in computer science from the Federal University of Sao Carlos, Brazil, in 2007, and the Ph.D. degree in computer science from the Federal University of Minas Gerais, Brazil, in 2012. He has been a Visiting Ph.D. Student with the PARADISE Research Laboratory, University of Ottawa, Canada, in 2011. He is currently a Professor of computer science with the State University of Campinas (UNICAMP), Brazil. In the last five years, he has published over 100 papers in international conferences and journals, and presented several papers in international conferences. His research interests include distributed algorithms, routing algorithms, wireless sensor networks, and vehicular ad hoc networks.

Antonio A. F. Loureiro received the B.Sc. and M.Sc. degrees in computer science from the Universidade Federal de Minas Gerais (UFMG), and the Ph.D. degree in computer science from The University of British Columbia, Canada. He is currently a Full Professor with UFMG, where he leads the research group on mobile ad hoc networks. He was a recipient of the 2015 IEEE Ad Hoc and Sensor (AHSN) Technical Achievement Award and the Computer Networks and Distributed Systems Interest Group Technical Achievement Award of the Brazilian Computer Society. His main research areas include ad hoc networks, mobile computing, and distributed algorithms. In the last 15 years, he has published regularly in international conferences and journals related to those areas, and have also presented tutorials at international conferences.