Exploring the Potential of Twitter to Understand Traffic Events and Their Locations in Greater Mumbai, India

Rahul Deb Das and Ross S. Purves

***Abstract*— Detecting traffic events and their locations is important for an effective transportation management system and better urban policy making. Traffic events are related to traffic accidents, congestion, parking issues, to name a few. Currently, traffic events are detected through static sensors e.g., CCTV camera, loop detectors. However they have limited spatial coverage and high maintenance cost, especially in developing regions. On the other hand, with Web 2.0 and ubiquitous mobile platforms, people can act as social sensors sharing different traffic events along with their locations. We investigated whether Twitter – a social media platform can be useful to understand urban traffic events from tweets in India. However, such tweets are informal and noisy and containing vernacular geographical information making the location retrieval task challenging. So far most authors have used geotagged tweets to identify traffic events which accounted for only 0.1%-3% or sometimes less than that. Recently Twitter has removed precise geotagging, further decreasing the utility of such approaches. To address these issues, this research explored how ungeotagged tweets could be used to understand traffic events in India. We developed a novel framework that does not only categorize traffic related tweets but also extracts the locations of the traffic events from the tweet content in Greater Mumbai. The results show that an SVM based model performs best detecting traffic related tweets. While extracting location information, a hybrid georeferencing model consists of a supervised learning algorithm and a number of spatial rules outperforms other models. The results suggest people in India, especially in Greater Mumbai often share traffic information along with location mentions, which can be used to complement existing physical transport infrastructure in a cost-effective manner to manage transport services in the urban environment.**

***Index Terms*— Georeference, jaccard distance, placename, toponym, traffic, tweet, vernacular geography, machine learning (ml), natural language processing (NLP), geographical information science (GIS).**

I. Introduction

UNDERSTANDING traffic conditions, both in real time and historically can help transport authorities manage transport infrastructure and vehicular movement effectively. Key are information about both the type of traffic events and their locations. Traditional solutions include the use of physical sensors, such as CCTV cameras and inductive loop detectors, but since these are static, they provide limited spatial coverage and incur high installation and maintenance costs. One emerging approach to addressing this limitation is the use of user-generated content, treating individuals as sensors who can both passively and actively report locations and events. Passive data, for example in the form of GPS locations and derived speeds and events, can be analysed at scale to provide information about perturbations in a system (e.g. vehicles slowing on a highway) [85]. However, they lack any semantics describing the nature of events, which are thus algorithmically inferred. A further potential source are actively generated data, the subject of this paper, where individuals actively report on events through social media posts or microblogs [63], thus acting as social sensors, providing information about ongoing events, ranging from cultural festivals [7] through natural disasters [6] to the subject of this paper, traffic events [8], in a dynamic way [49]. The value of these data lies in their semantic richness: a single tweet reporting an accident at a specific location provides rich information about current and historical traffic conditions.

Among social media platforms, Twitter[1] is a popular source of user-generated contents (UGC). Twitter posts, or tweets, are limited to 280 characters. Twitter had 316 million active users globally with 500 million tweets per day in 2016 [64]–[66], and had further increased to 336 million active users in 2018 [67]. There are three primary sources of location information in tweets, the user's home location mentioned in their profile, location provided in the tweet metadata, and location mentioned in the tweet content [26], [51]. Much previous work has used the location information from tweet metadata, in terms of coordinates, e.g., latitude, longitude [8], [15], [17], [50], although only a very small proportion of tweets are actually furnished with this information (Figures ranging from 0.1% to 3% of the total volume of tweets have been reported [24], [72], [73]) and Twitter announced in June 2019 that it would no longer share precise locations in metadata. Since profile-based information is essentially static, extracting loca-

Manuscript received April 4, 2019; revised August 31, 2019 and October 6, 2019; accepted October 15, 2019. Date of publication November 26, 2019; date of current version November 30, 2020. This work was supported by the Swiss National Science Foundation under Grant 166788. The Associate Editor for this article was D. Sun. *(Corresponding author: Rahul Deb Das.)*

R. D. Das was with the Department of Geography, University of Zurich, 8057 Zürich, Switzerland. He is now with IBM Germany R&D, 71032 Böblingen, Germany (e-mail: das.rahuld@gmail.com).

R. S. Purves is with the Department of Geography, University of Zurich, 8057 Zürich, Switzerland (e-mail: ross.purves@geo.uzh.ch).

Digital Object Identifier 10.1109/TITS.2019.2950782

[1] https://twitter.com/

tion information from tweet content has been the focus of a wide range of research using a variety of methods.

Twitter has already been recognised as a potential source of information in identifying and locating traffic events [11]. However most studies have focussed on traffic event detection, and thus extracting locations have typically either used explicitly geotagged tweets, or methods based on gazetteer lookup [7], [18]. In this paper, we explore the Twitter as a source to detect both traffic events and their locations in Greater Mumbai in India. We developed a complete pipeline, which firstly identifies potentially relevant tweets, before locating these tweets using their content. The selection of our study area is motivated by a number of reasons. Firstly, Mumbai is a growing megacity in one of the fastest developing nations with numerous transportation issues. A recent survey shows Mumbai is one of the major cities in terms of traffic index [68]. Moreover, a lack of traditional traffic sensors, and the growing use of Twitter in India make it a particularly interesting location for our study [9], [67]. Finally, English is a widely spoken language in India, with official status. However, it is often used in combination with other official languages such as Hindi or Marathi. This is particularly true in informal communications such as tweets, as illustrated by the following examples:

- Example tweet 1: *Your favorite #SEO instructor is back Learn how to climb the #search rankings and get more #traffic with ClickMinded..*
- Example tweet 2: *Traffic complete standstill at CST*
- Example tweet 3: *@MumbaiPolice the Chatrapati Shivaji Terminus bus stop near Azad Maidan where bus no. 28 stops is getting creapier [sic] by the day.*
- Example tweet 4: *Just because of 2 checking stands kept on road opposite to chitra signal after dadar tt flyover, there is a whole lot of traffic emerging from matunga end!! @MumbaiPolice This is there from the Time i went to work in the morning and till now …*
- Example tweet 5: *Cars parked illegally outside Nisarg hotel .. LBS Marg Mulund west ..The whole LBS road is no parking zone.. but this hotel has valet parking with special privileges ... Informed Mumbai traffic control an hour back but no action taken @MumbaiPolice @mtptraffic @mulund_info*

The first example contains a potentially relevant keyword *traffic* used in a semantically irrelevant sense. Our first task is therefore to discard such examples, and only retain relevant tweets such as examples 2-5. Having done so, we wish to relate traffic-related tweets to specific locations. Tweets 2 and 3 both refer to the same location, which was formerly known as Victoria Terminus, renamed as Chatrapati Shivaji Terminus in 1996 (often abbreviated to CST). In 2017, the name was again changed to Chatrapati Shivaji Maharaj Terminus (abbreviated to CSMT). All of these names should be assigned to the same physical location. In the fourth example, rich information about the location of a traffic jam is given, using spatial language to identify a location and describing the cause of the disruption. We further observe the use of various spatial terms specific to a given regional language when describing events on Twitter [81]. Such spatial terms used in regional language are known as vernacular names, which are often used as a part of a placename. For example, in the final example, Hindi is used to impart more information, with the term *marg* here referring to a road.

Our contribution is thus threefold - we develop a classifier which can deal with informal natural language to extract relevant traffic events, and having done so locate these, taking into account vernacular uses of spatial language. We do so for a location where a clear need exists for such information, and where traditional sources are sparse. Furthermore, by evaluating our results in detail we demonstrate the potential of these methods and discuss the challenges of adapting them to other local contexts.

The remainder of the paper is organized as follows. In Section II the state-of-the-art is presented. In Section III the methodology is explained in two stages. The first stage explains tweet classification and the second stage explains tweet georeferencing. We discussed how the data was collected and pre-processed in Section III-A. The model is evaluated in Section IV. In Section V strengths and limitations of our approach are discussed, before we briefly conclude the paper and suggest potential further work in Section VI.

II. State-of-the-Art

Twitter has been used previously to understand human mobility patterns and traffic conditions at different granularities. For example, Hawelka and colleagues explored the potential of Twitter data to understand human mobility patterns at a global scale [65]. Liu and colleagues used Twitter data to understand how people navigate at country scales [76], while Gu and others investigated the potential mobility patterns and transport services at a city level [24]. Our work builds on this contribution by investigating the usefulness of Twitter data in understanding traffic events at a city level. Exploring traffic events can also help in understanding human mobility patterns.

Most classifiers for detecting traffic events share some common (and basic) tasks, e.g., data collection, data pre-processing, feature generation, model development. Tweets related to traffic can be collected randomly from public users within a given spatial extent [7], [12], [15] or by using a relevant keyword search [11], [24], or by simply following specific official accounts [18], [33]. Tweets collected using a spatial extent are geotagged whereas tweets collected by keywords or by following some specific user accounts are typically not geotagged. Some researchers [8], [22] used multiple strategies to collect traffic related tweets. For example, Wang and colleagues [8] collected tweets from official accounts, using pre-defined road names and using circular search areas along the road network to collect geotagged tweets near roads.

Once tweets are collected, pre-processing is generally performed to remove noise (typos, unwanted punctuation, white space, non-ASCII characters and emoticons) and stop-words in the text. Feature generation then generally involves tokenization, lemmatization, and converting string to word vectors (to assign weights to each word token). D'Andrea and others used inverse document frequency (IDF) as features [11]. Similarly, other researchers used a single word tokens (unigram) and multiple word tokens (n-gram) and their associated term frequencies (TF) as feature vectors [7], [8], [12], [14], [15], [22], [24]. Gu and colleagues used only a selected number of unigrams pertaining to traffic incidents in the US [24]. Similarly, Klaithin and colleagues used words in Thai from three official sources indicating road names, direction of traffic, location, and traffic state [18].

Andrea *et al.* [11] developed a hierarchical classifier, which can identify a traffic related tweet followed by a more specific category of traffic events. For the first classification task they used 1,330 tweets and for the second classification task they used 999 tweets. They showed that an SVM-based classifier could achieve 95.75% and 88.89% accuracy for two class and three class classification respectively through an n-fold cross validation. Similarly Salas and colleagues developed a single layered tweet classification model using a SVM to categorize tweets into traffic or non-traffic [12]. To train the model, [12] used a balanced data set consists of 871 traffic related tweets and 871 non-traffic related tweets. While testing, they used 290 traffic and 290 non-traffic related tweets. Accuracy varied from 60.12% to 90.71% while using different n-gram features. Klaithin and colleagues [18] developed a Naive Bayes classifier to categorize tweets into six different categories, e.g., accident, announcement, question, orientation, request, sentiment. In training, 4,637 tweets were used and 1,494 tweets were used for evaluation. The classifier achieved 76.40% average accuracy. Gu and others presented a real-time traffic incident detection model, evaluated in Philadelphia and Pittsburgh in the US [24]. They developed the model based on a semi-Naive Bayes classifier and achieved 90.50% accuracy. Zhang and colleagues investigated how well a deep learning technique, e.g., Deep Belief Network (DBN) can perform while detecting tweets related to traffic accidents in Northern Virginia and New York in the US [15]. They found accuracy increases when using bigrams in tweet content instead of unigrams. Zhang et al showed DBN can achieve 85% accuracy outperforming other machine learning approaches e.g., LSTM, ANN, SVM and sLDA [15].

In contrast to these supervised classification models, a number of works have used unsupervised models to explore the semantics behind traffic related tweets. For example, [7], [8], [14], [23] proposed a Latent Dirichlet Allocation (LDA) to extract different traffic related topics from tweets.

Given issues related to user bias and representativity in a single source [65], efforts to triangulate using different data sources to develop multi-modal solutions have been increased over time. For example, In [32], Tostes and colleagues investigated using different types of social media data, e.g., Foursquare check-ins and geotagged Instagram photos to understand traffic congestion in a city. To compare the distribution, they also collected traffic flow data over different road segments from Bing Map [77]. They found a correlation between the number of check-ins from social media platforms and the traffic congestion from Bing Map [32]. Similarly, Bichu and Panangadan investigated how tweets correlate with vehicular traffic in Los Angeles County and Orange County [31]. Bichu and Panangadan collected two different types of data sets: tweets based traffic related keywords and traffic count data from inductive loops on four major freeways in Los Angeles County and Orange County. Similar to [31], [32] showed both that data sets are periodic in nature and correlate, demonstrating that social media could replicate more traditional traffic measurements.

Since only a limited number of tweets are geotagged (0.1%–0.77%) [24], [26], there is an increasing effort to understand location of events from Twitter data either using tweet metadata [21], network information [20] or tweet content [41]. Although the issue of location extraction from text has been addressed by a wide range of researchers, many models are developed to extract location information from formal text [26], [29]. However, tweets are very short text (limited to 280 characters), often written in an informal way, contain typos, abbreviations and vernacular uses of language [66]. Wing and Baldridge showed geolocating tweets are more challenging than for Wikipeida articles, with median prediction errors of 479 km - indicating a great challenge in geographic information extraction from informal text [19].

Although location extraction from tweets has been addressed in other domains, disaster management, for example [41], [47], there are very few works have been done that can retrieve location from tweets in the context of traffic detection. Existing work [17], [18], [24], [33] used predefined knowledge bases and rule-based techniques to extract the location information from traffic related tweets. Some researchers considered a number of official Twitter accounts, which are either maintained by radio channels or police department or traffic authority in the given regions. Such tweets are more formal and systematic in their syntactic structure, and can be easily parsed to extract location information [17]. Such models are not generalisable to more general (and informal) tweet content.

In summary, most earlier works used geotagged tweets to detect traffic events primarily by developing a model that can classify tweets into traffic and non-traffic category. In this work, we aim to use ungeotagged tweets to understand traffic events in one of the major metro cities in India. Motivated by some of the previous works, e.g., [11], [24], we investigate whether a Twitter based approach can be useful in an Indian city to understand road traffic conditions while handling informal and local placenames. Instead of using regular expressions, we use a number of generalisable spatial rules based on spatial prepositions and parts-of-speech tagging and a lexicon based approach to find a number of pre-defined vernacular names. In contrast to [31], we aim to extract all the traffic locations from the relevant tweets while dealing with the informal nature of location mentions in the tweet content. Our work goes beyond [11], [12], [15] and not only detects traffic relevant tweets but also geolocates traffic events using relevant tweet content. To geolocate traffic events, we developed a hybrid georeferencing model, partly motivated by [41], but trained and tested in the context of traffic events in Greater Mumbai in India.

III. Methodology

To detect traffic events and their locations we developed an integrated model, which consists of three phases. In the first phase, tweets are collected and pre-processed. In the second phase, tweets are classified as either a traffic relevant tweet or non-traffic relevant tweet using a supervised model. Finally, once a tweet is classified as relevant to traffic event, a georeferencing module is used to identify the locations mentioned in the tweet content to understand where the traffic event is happening.

A. Data Collection and Preparation

Since we aim to investigate if people mention placenames while tweeting about a traffic event in Greater Mumbai, we collected ungeotagged tweets from GNIP enabled PowerTrack 2.0 from Twitter repository using the premium service of DiscoverText[2] [79]. PowerTrack provides a more exhaustive Twitter Search option to retrieve historical data. This is not possible using the standard Twitter Search API, which can retrieve tweets that are only one week old or even using a Streaming API, which can retrieve tweets in real time. To retrieve traffic related tweets in English language from Greater Mumbai, we used a query containing a wide variety of search keywords potentially related to traffic, from users mentioning *Mumbai* in their profile location, which are identified as being in English, and which are not retweets (thus removing duplicates).[3]

Data were collected in two phases and split for training (Phase 1) and testing purposes (Phase 2). Importantly, since data were split temporally, the same events were not present in both samples (though of course regularly recurring events may be).

- Phase 1: 18th June, 2018 – 1st July, 2018
- Phase 2: 2nd July, 2018 – 10th July, 2018

A total of 29,000 tweets were collected, giving some indication of the potentially relevant volume of information. Since annotation is time consuming, we selected subsamples of similar sizes to work reported in the state-of-the-art, and retained 3,548 tweets (2,035 from Phase 1; 1,513 from Phase 2), containing keywords which are highly relevant to traffic, e.g., *traffic, roadblock, accident, barricade, collision*. To annotate tweets (tweet labels) we used the crowdsourcing service of Figure-Eight platform.[4] For every tweet, three annotators were recruited. Each tweet is labelled based on the majority voting of the three annotators. In the training data (Phase 1) 57% of tweets were labelled as traffic. In the testing phase 54% tweets were labelled as traffic, suggesting that our initial search retrieved a high number of potentially relevant tweets.

Having labelled tweets, we then annotated locations in the traffic related tweets from both training and test data, to train and validate our georeferencing module. We did not use crowdsourcing for this task, as previous research has demonstrated the importance of local knowledge in toponym annotation [86], and thus the first author carried out this work.

B. Tweet Classification Module

A tweet classifier typically removes a variety of content (e.g. emoticons, non ASCII characters, white space) from the text. Thus, pre-processing is performed through the following steps.

- *Cleaning*: To clean the tweets, we removed non-ASCII characters including special symbols, e.g., @, #, and emoticons. This may reduce overfitting and improve the generalization capabilities of the tweet classification module (e.g. reducing the influence of hashtags about the same events).
- *Tokenization* splits a tweet into discrete unigram tokens where each token is a word in the tweet.
- *Stemming* reduces each word to its base form. The aim is to bring words with similar semantics to a common form to train the model more effectively. In this paper, we used Lovins Stemming algorithm [13], which consists of 294 endings, 29 conditions, and 35 transformation rules.
- *Feature generation*: A bag-of-word (BOW) model is used to generate input vectors. To generate input vectors, processed tweet text is converted to a numerical form where each word is assigned a weight based on its term frequency-inverse document frequency (TF-IDF), a standard NLP weight which considers the frequency of a term both in an individual tweet and the corpus as a whole.

$$TF = T_t \tag{1}$$

$$IDF = log[\frac{N}{(1 + D_t)}] \tag{2}$$

$$TF - IDF = T_t \times log[\frac{N}{(1 + D_t)}] \tag{3}$$

  where T_t is the total count of term ‘t’ in document ‘D’. N is the total number of documents in the corpus and D_t is the total number of documents containing the term ‘t’. In this case, each tweet is a document and the collection of all the tweets constitutes a corpus.
- *Classification*: Previous work has shown that a number of supervised classification methods can be effective in identifying traffic related tweets based on their content. As proposed by [27] we trained a number of standard classifiers capable of handling text features including a Decision Tree (DT), a Support Vector Machine (SVM), a Naive Bayes classifier (NB), and a K-Nearest Neighbor (KNN) classifier to identify the classifier offering the best performance on our data. For KNN, we considered K=3. We also tested an ensemble classifier (EC), e.g., an Adaptive Boosting (AdaBoost) with a DT as base classifier.

[2] https://discovertext.com/

[3] The following query was used to retrieve raw tweets with the given keywords from the users mentioning “Mumbai, India” in their profile information: [*(traffic OR trafic OR toll plaza OR express way OR expressway OR accident OR dead OR death OR pothole OR barricade OR casualty OR road OR collision OR collided OR street OR parking OR parked OR injured OR delay OR jam OR southbound OR eastbound OR westbound OR car OR taxi OR truck OR transportation OR transport OR travel OR train OR metro OR rail OR bus OR platform) bio_location:“Mumbai, India” -is:retweet (lang:en)*].

[4] https://www.figure-eight.com/

C. Georeferencing Module

In the context of geographic information retrieval a georeferencing task first retrieves the potential placenames from a given text, followed by mapping that placename to a unique spatial footprint on the earth. Retrieving the placename from the text is known as toponym recognition, whereas mapping the placename to a unique set of coordinates is known as toponym resolution. To develop the georeferencing module a hybrid approach is followed. The georeferencing module has two layers. The first layer is based on a supervised model, whereas the second layer is based on a number of spatial rules that can retrieve location entities. To develop the first layer, two supervised machine learning models, e.g., Maximum Entropy (MaxEnt) [35] and a linear chain Conditional Random Field (CRF) [35] are used. Since, tweets contain many peculiarities, for example, variable number of tokens to refer to the same placename (cf. Section I), local placenames, abbreviations, typos, the MaxEnt model is retrained with the placenames annotated in the tweets. And the CRF model is pre-trained on formal text data. Thus, the two models trained on both formal and informal text data can handle the varied extent of informal aspect in the tweet content.

To further strengthen the performance of the model, rule based layers are developed. Although location entities are proper nouns, we observed that the words in a location entity can be identified as proper nouns or common nouns (or sometimes adjectives) depending on the capitalization of the words and the way the parts-of-speech (POS) tagger is trained. Location entities often appear after spatial prepositions, e.g., *at*, *near*, *towards*, *from*. For example, consider the following example sentences. Each word in the following example sentences is followed by its POS tag. We used Penn Treebank style to label the POS tags (cf. [4]).[5]

- Sentence 1: We[PRP] are[VBP] travelling[VBG] towards[IN] Woodhouse[NNP] road[NN]
- Sentence 2: We[PRP] are[VBP] travelling[VBG] towards[IN] woodhouse[NN] road[NN]

In the first and second example sentences the location entity is *Woodhouse road* with 'W' being capitalized and non-capitalized respectively. When a pre-trained POS tagger (OpenNLP using MaxEnt algorithm) is used, in the first sentence *Woodhouse* is identified as proper noun (NNP) and in the second sentence it is identified as common noun (NN). In both cases *road* has been identified as common noun. Thus, in our algorithm we considered that words identified as either proper or common nouns may be potential placenames (or a part of a location entity).

We also observed that, in our test data, placenames often end with the spatial object types in terms of their affordances or functionalities [82], [83]. For example, a road offers an affordance of navigation and transportation. A building offers an affordance of performing certain activities in indoor built environment, e.g., sleeping, working. Such object types are often mentioned either in English or in a local language (as in Hindi or Marathi in Greater Mumbai). Our rule base therefore includes 85 object types (in English and Hindi) that may occur after a placename (e.g. hospital, road and clinic).

TABLE I
TWEET CLASSIFICATION ACCURACY ON TEST DATA

Model	**Traffic**			**Non-traffic**		
	Precision	Recall	F1	Precision	Recall	F1
DT	76	70	0.73	62	68	0.65
SVM	77	**81**	**0.79**	**78**	74	**0.76**
NB	**95**	66	0.78	42	**88**	0.57
AdaBoost (DT)	69	78	0.73	76	67	0.71
KNN	83	69	0.76	56	73	0.63

To identify a placename using the rule based layer, we first pre-processed and tokenize tweets, then used a POS tagger to detect prepositions. A predefined knowledge base of 29 spatial prepositions classifies these as spatial or non-spatial types. If the current word (W_i) is a spatial preposition and if the next word (W_{i+1}) is a proper noun or common noun then there is a high chance that the W_{i+1} is, or is a part of, a placename, and W_{i+1} is added to a potential placename candidate set. In the subsequent phase, it is assessed if the next to next word (W_{i+2}) is a proper noun or common noun. If the subsequent word (W_{i+2}) is not a proper noun or common noun then it is checked if it is a spatial object type either in English or in Hindi language. If it is a spatial object type then that word (indicating a spatial object type) is also added to the candidate set. Thus the entire ordered set of words (e.g., W_{i+1}, W_{i+2}) added in the candidate set is the retrieved placename. If no such word that relates to an object type is found, the scanning process stops after three iterations.

To maximise recall, all tweets are passed to both layers (supervised machine learning and a rule-base). Since both layers may retrieve the same name, a duplication check is performed to return only unique placenames. Finally, placenames were assigned spatial coordinates by passing them to the OpenStreetMap (OSM) Nominatim API[6] for geocoding.

IV. Evaluation

To evaluate the model we used three metrics – precision, recall and F1. Table I shows the results for the five different classifiers on an independent test data. The NB-model has very high precision 95% for traffic-related tweets. If our aim was simply to be very confident about the relevance of tweets this classifier would be well-suited. However, precision is very low (42%) for non-traffic tweets meaning that many irrelevant tweets are wrongly detected as relevant. Of the other classifiers, SVM has the best overall performance for traffic and non-traffic related tweets (F1 is 0.79 and 0.76 respectively).

In the next stage, to evaluate the performance of the georeferencing module, the hybrid geoparser is used on a labelled test data (Phase 2 data). The hybrid geoparser is composed of two layers.

[5] DT = Determiner; EX = Existential *there*; IN = Preposition; JJ = Adjective; NN = Noun-singular or mass; NNP = Proper noun-singular; PRP = Personal pronoun; VBG = Verb-gerund; VBP = Verb-non 3rd person singular present; VBZ = Verb-3rd person singular present.

[6] https://nominatim.org/release-docs/develop/api/Search/

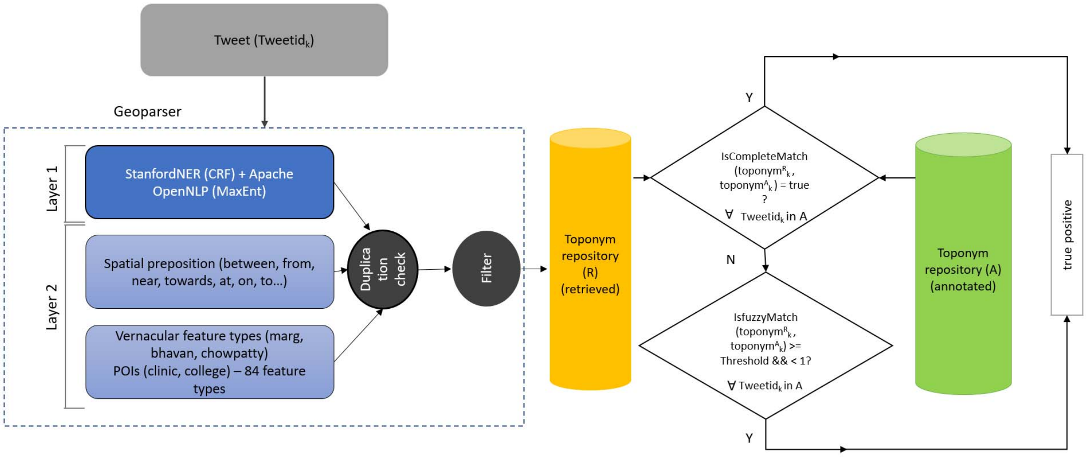
Tweet (Tweetid$_k$)
Geoparser
Layer 1
StanfordNER (CRF) + Apache OpenNLP (MaxEnt)
Layer 2
Spatial preposition (between, from, near, towards, at, on, to…)
Vernacular feature types (marg, bhavan, chowpatty) POIs (clinic, college) – 84 feature types
Duplication check
Filter
Toponym repository (R) (retrieved)
IsCompleteMatch (toponym^R_k, toponym^A_k) = true ? ∀ Tweetid$_k$ in A
Y
N
IsfuzzyMatch (toponym^R_k, toponym^A_k) >= Threshold && < 1? ∀ Tweetid$_k$ in A
Y
Toponym repository (A) (annotated)
true positive

Fig. 1. A hybrid multi-layered Geoparser. Layer 1 consists of supervised location retriever. Layer 2 consists of spatial rules based on spatial prepositions, and vernacular placenames and spatial objects.

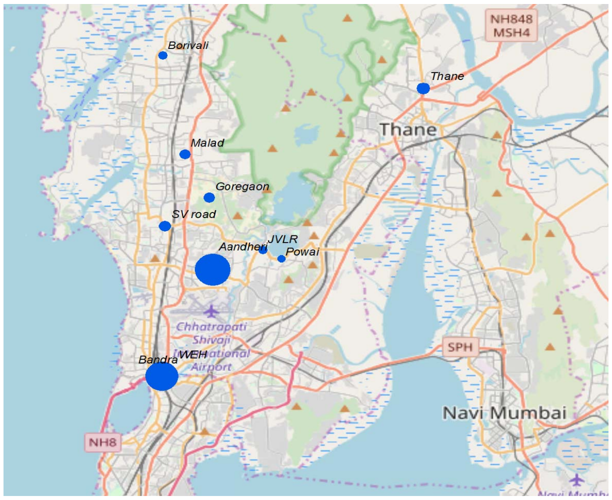

Fig. 2. Map shows top ten most congested location mentions in Mumbai in Twitter. (Basemap © OpenStreetMap contributors CC-BY-SA (www.openstreetmap.org/copyright)).

TABLE II

ACCURACY ACHIEVED BY DIFFERENT SETUPS

	Model Accuracy			
Model	**Setup 1**	**Setup 2**	**Setup 3**	**Setup 4**
Precision (%)	58.97	53.40	52.39	**79.1**
Recall (%)	60.95	**79.14**	64.00	7.75
F1 (0-1)	0.59	**0.63**	0.57	0.14

Layer 1 consists of an MaxEnt (Apache OpenNLP) and a CRF model (Stanford NER). However, we retrained only the MaxEnt model using Phase 1 data (cf. Section III-A) to deal with the more informal tweets, and we used a default CRF model, which is pre-trained on formal text from CoNLL-2003 data set [74] to handle more formal tweets. CoNLL-2003 data set consists of eight text files in English and German language. The CRF model used in this paper is trained on English corpus populated with Reuters articles collected from August 1996 to August 1997, which contains 946 articles and 7,140 placenames for training, 216 articles and 1,837 placenames for validation [74].

Layer 2 is composed of the spatial rules. Since, as previously discussed, tweets are informal, the placenames found therein contain various peculiarities, and retrieval accuracy partly depends on the ability of the POS tagger to correctly identify the tokens that are common noun or proper noun, we assumed that the retrieved placenames may not completely match with the ground truth data. For example, *Chatrapati Shivaji Terminus* may be retrieved as *Shivaji Terminus*. To handle such partially matched situations, we used two evaluation metrics based on complete (perfect) and fuzzy matching. To quantify fuzzy matches, Jaccard distances are calculated, where the shorter the distance, the more similar two terms are.

Jaccard distance (*jd*) is a function of Jaccard coefficient (*jc*), which was calculated as follows, with distances of less than 0.7 being treated as fuzzy matches:

$$jc = \frac{(term_A \cap term_R)}{(term_A \cup term_R)} \quad (4)$$

$$jd = 1 - jc : \forall jd, \quad 0 <= jd <= 1 \quad (5)$$

where $term_R$ is a retrieved term and $term_A$ a ground truth term.

To retrieve location entities, four experimental setups were tested using different combinations, with different levels of customisation of layer 1, where we retrained the OpenNLP model (MaxEnt) using data from Phase 1. We also experimented with using only the machine learning layer 1 (Setups 3 & 4) and the inclusion of our rule base (Setups 1 & 2) (Table II). The best performing model included both local training data and our rule base (Setup 2)[7]. Here, of 2,733 annotated placenames 2,163 placenames are retrieved giving a recall of 79%. Of these 2,163 retrieved placenames, 1,533 placenames match completely with the annotated placenames, whereas 630 placenames partially match with the annotated ones. However, precision of the model was around 53% meaning that Type I errors were fairly common

[7] Source code is available at https://github.com/rddspatial/georeferencing

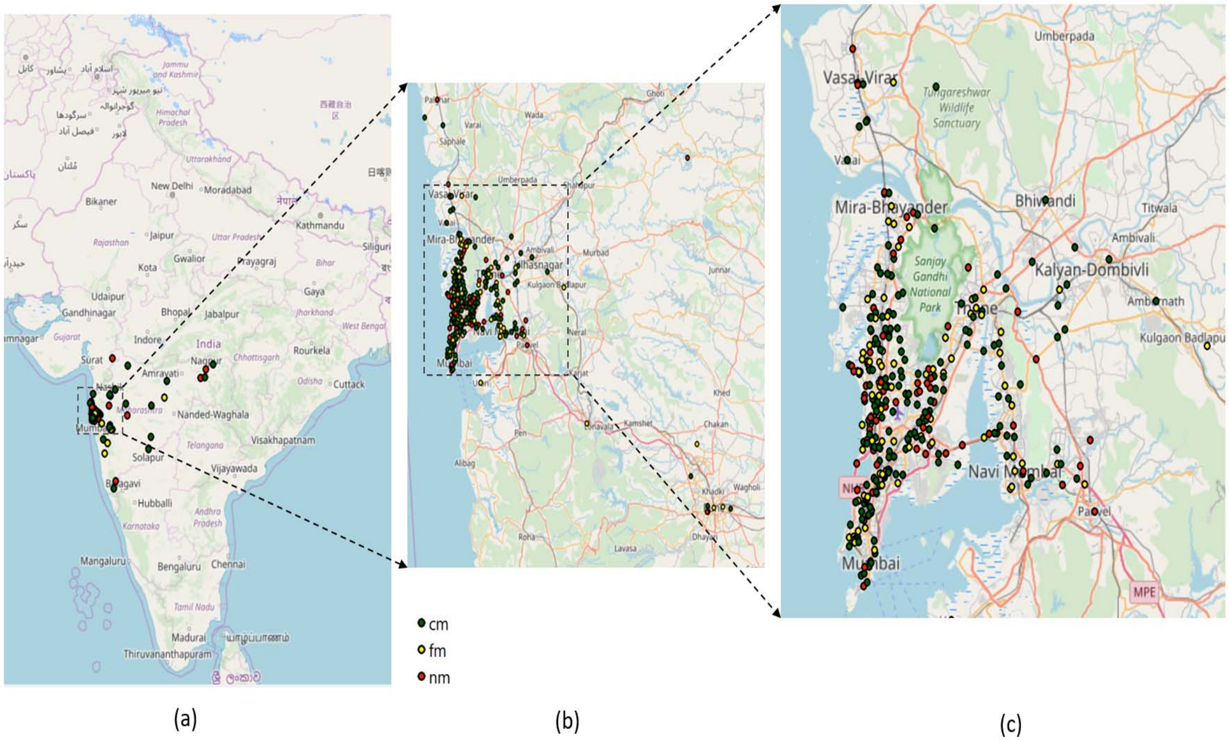

Fig. 3. In this figure locations of traffic events are retrieved using the hybrid multi-layered georeferencing module. Placenames that completely match with the ground truth ones labelled as *cm*. Placenames which partially match with the ground truth placenames are labelled as *fm* and placenames which are not retrieved are labelled as *nm*. Figure a, b, c show spatial distribution of traffic events at different granularity. (Basemap © OpenStreetMap contributors CC-BY-SA (www.openstreetmap.org/copyright)).

(i.e. generic terms such as *home* were wrongly classified as toponyms). In this context a default supervised model is already pre-trained on formal text data, whereas a retrained model is trained on informal tweet contents. We tested with different combinations of the models trained on both formal and informal text to evaluate the best combination to deal with the traffic related tweets.

- Setup 1: default (pre-trained) Stanford NER (CRF) + default (pre-trained) OpenNLP (MaxEnt) + rule base
- Setup 2: default (pre-trained) Stanford NER (CRF) + retrained OpenNLP (MaxEnt) + rule base
- Setup 3: retrained OpenNLP (MaxEnt)
- Setup 4: default (pre-trained) OpenNLP (MaxEnt)

In a final step, we mapped the 2,163 correctly identified toponyms using locations extracted from OpenStreetMap (OSM). From these, 1,465 locations could be geocoded (in comparison with the 2,733 annotated locations). We could not geocode all locations, since they were not found in OSM due to lack of coverage in India.

Figure 3 shows the location of all traffic events retrieved from the tweet contents after passing all the traffic tweets to the georeferencing module along with the nature of the complete match (cm) and fuzzy match (fm) found. Cases with no match (nm) are locations annotated, but not retrieved by the georeferencing module.

We observe that selecting only users with *Mumbai* as their profile location results in many traffic events related to Greater Mumbai, though we also find other traffic events across the state of Maharashtra (Fig 3a). Our data in general relate well to locations known for congestion in Mumbai, including SV Road, Western Express Highway (WEH), Eastern Express Highway (EEH).

To explore the data in more detail, we mapped the ten most reported locations for traffic events (Figure 2). One single location, Andheri accounts for 27% of top ten locations in the tweets, and was the subject of a major incident (the collapse of a bridge) during the study period. Although no baseline data are available, we note that our results are plausible based on local knowledge of traffic in Mumbai, and captured this significant event well.

V. Discussion

In this paper we set out to develop an end-to-end framework, capable of identifying and mapping traffic relevant tweets in Mumbai area. In doing so, we also aimed to generate wherever possible a generic framework, transferrable to other regions, whilst including local idiosyncrasies in the use of placenames and language. Our approach is particularly important since Twitter is in a process of disabling the ability to georeference tweets with precise geocoordinates used in previous work (e.g. [8]).

Our first important decision was to use only tweets whose profile was related to Mumbai. Although this reduced greatly the number of tweets we processed, it also provided a simple spatial filter, especially important for our approach which only used textual content. Future work might also take advantage of commonly found locations to iteratively build a set of search terms to query the initial Twitter stream.

Our classifier used simple textual features, and performed well in further filtering tweets to only those relevant to traffic, using standard machine learning approaches. Although NB gave the best precision, recall was the lowest, and an SVM-based model was found to give the best balance of precision and recall, both on traffic and non-traffic related tweets.

The main area of innovation in our paper was the development of a toponym recognition tool tuned to identify traffic relevant locations using a mixture of machine learning (including locally annotated data for training) and rules incorporating vernacular language. By building a relatively small corpus of 85 vernacular uses common in Mumbai, we improved model performance over purely machine-learning based approaches, and were able to increase recall by 15% over the next best model setup (Table II). However, our approach resulted in somewhat reduced precision, since false positives containing generic terms were more likely to be identified by the POS tagger. The performance can be improved by using more training data in the supervised learning phase and emerging state-of-the-art approaches such as BERT [87]. Applying this, or other deep learning approaches would however require larger training datasets.

Perhaps the main limitation of our approach is in the toponym resolution phase, where coordinates are assigned to the identified toponyms. We noted that the OSM Nominatim API we used was unable to resolve local Indian placenames from our annotated data in around 30% of cases. This points to the importance of creating detailed local gazetteers for tasks such as that reported in this paper. However, having identified commonly used toponyms, this is a relatively straightforward task given local knowledge.

Our primary objective was to retrieve traffic locations from informal tweets in India, using a binary classification of tweets into traffic and non-traffic. However, future work could classify traffic related tweets into more fine grained categories, for example, to observe in our data set tweets related to traffic congestion, accidents, grievance or monetary compensation due to unlawful driving, or even related to parking issues in India. There is also a growing trend when people attach various media information (e.g., photo, video) to their tweets while mentioning about traffic events to provide more dynamic information or to strengthen the credibility of their tweets.

Our georeferencing model can retrieve any location entity, however, it does not consider qualitative spatial relationships [55], geometric properties of spatial objects and the topology of locations. For example, if a tweet mentions a traffic congestion from $location_A$ to $location_B$ along a road network, the model can retrieve $location_A$ and $location_B$, but cannot retrieve the edge or a subgraph between $location_A$ and $location_B$ where traffic congestion is occurring or how $location_A$ and $location_B$ are spatially related to each other or inferring their geometrical properties at different granularities (e.g., point, line or polygon).

Our classification model suggests whether a tweet is relevant depending on the textual patterns, but it cannot assess how true or legitimate the information contained in the tweet is. Research shows fake or misleading information has adverse effect in decision making process [52]. To detect fake or misleading information, existing research leverages user's profile information, user's social interaction, activity patterns and textual patterns [16]. In the context of transportation management, a future work should investigate the characteristics of fake or ambiguous tweets and how to deal with them.

The current traffic surveillance system in Mumbai heavily relies on static sensors and the traffic police, who are constrained by limited resources. Many locations are yet not well monitored in Mumbai, e.g., Girgaum Chowpatty, King's circle, Juhu Tara road, Bandra Worli sea link, airport region, Hindmata, to name a few [71], which require closer surveillance system. The model developed in this paper shows a cost-effective alternative and can be used to understand traffic conditions at those locations which lack proper infrastructure and surveillance system in Mumbai. The model can also be extended to extract various reasons behind traffic issues at various locations using topic modelling [80].

VI. Conclusion

As behaviour and the abilities of infrastructure to meet transport comes under more strain [1], [10], so does the need to develop methods which can allow both real time and historical understanding of not only where transport events occur, but also their nature. In this work we report on the use of Twitter to extract and locate traffic related events in the area of Greater Mumbai. We build a complete pipeline capable of identifying and locating such events on a map, and allowing analysts to explore the nature and emergence of events. This in turn can help urban planners and policy makers in their decision making processes.

By using a combination of machine learning, simple rules and lists of local terms, we were able to build a hybrid georeferencing model which would be easily customisable for other locations, and which offers good performance. Our approach extends previous work by considering all aspects of the pipeline, from initial stream of tweets to locations on a map, and incorporates local language(s) to improve performance. Performance in Mumbai could be improved by building a more comprehensive gazetteer of local placenames, while in general as more training data are generated, the use of state-of-the-art machine learning approaches should be considered. Since tweets often contain rich spatial language, future work should also seek to analyse this to better locate events on a transport network in other resource constraint regions.

Acknowledgment

The authors would like to thank the reviewers for their insightful comments, which improved the clarity of the article.

References

[1] J. Pucher, N. Korattyswaropam, N. Mittal, and N. Ittyerah, "Urban transport crisis in India," *Transp. Policy*, vol. 12, no. 3, pp. 185–198, 2005.

[2] T. H. Rashidi, A. Abbasi, M. Maghrebi, S. Hasan, and T. S. Waller, "Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges," *Transp. Res. C, Emerg. Technol.*, vol. 75, pp. 197–211, Feb. 2017.

[3] D. Pojani and D. Stead, “Sustainable urban transport in the developing world: Beyond megacities,” *Sustainability*, vol. 7, pp. 7784–7805, Jun. 2015.

[4] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, “Building a large annotated corpus of English: The penn treebank,” *Comput. Linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[5] L. Steg, “Can public transport compete with the private car?” *IATSS Res.*, vol. 27, no. 2, pp. 27–35, Jan. 2003.

[6] T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake shakes Twitter users: Real-time event detection by social sensors,” in *Proc. 19th Int. Conf. World Wide Web*, Raleigh, NC, USA, Apr. 2010, pp. 851–860.

[7] Y. Zhou, S. De, and K. Moessner, “Real World City event extraction from Twitter data streams,” *Procedia Comput. Sci.*, vol. 98, pp. 443–448, Jan. 2016.

[8] D. Wang, A. Al-Rubaie, S. S. Clarke, and J. Davies, “Real-time traffic event detection from social media,” *ACM Trans. Internet Technol.*, vol. 18, no. 1, 2017, Art. no. 9.

[9] J. Poushter, C. Bishop, and H. Chwe. Social Network Adoption Varies Widely by Country. Pew Research Center, Global Attitudes and Trends, Accessed: Nov. 7, 2018. [Online]. Available: http://www.pewglobal.org/2018/06/19/3-social-network-adoption-varies-widely-by-country/

[10] R. Cervero and A. Golub, “Informal transport: A global perspective,” *Transp. Policy*, vol. 14, no. 6, pp. 445–457, 2007.

[11] E. D’Andrea, P. Ducange, B. Lazzerini, and F. Marcelloni, “Real-time detection of traffic from Twitter stream analysis,” *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 2269–2283, Aug. 2015.

[12] A. Salas, P. Georgakis, and Y. Petalas, “Incident detection using data from social media,” presented at the IEEE 20th Int. Conf. Intell. Transp. Syst. (ITSC), Yokohama, Japan, Oct. 2017.

[13] J. B. Lovins, “Development of a stemming algorithm,” *Mech. Transl. Comput. Linguistics*, vol. 11, pp. 22–31, Mar. 1968.

[14] A. F. Hidayatullah and M. R. Ma’arif, “Road traffic topic modeling on Twitter using latent Dirichlet allocation,” presented at the Int. Conf. Sustain. Inf. Eng. Technol. (SIET), Batu, Indonesia, Nov. 2017. [Online]. Available: https://www.overleaf.com/project/5d95a6e4db77d00001ce0702

[15] 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.

[16] I. Inuwa-Dutse, M. Liptrott, and I. Korkontzelos, “Detection of spam-posting accounts on Twitter,” *Neurocomputing*, vol. 315, pp. 496–511, Nov. 2018.

[17] S. Wongcharoen and T. Senivongse, “Twitter analysis of road traffic congestion severity estimation,” presented at the 13th Int. Joint Conf. Comput. Sci. Softw. Eng. (JCSSE), Khon Kaen, Thailand: Khon Kaen Univ., Jul. 2016.

[18] S. Klaithin and C. Haruechaiyasak, “Traffic information extraction and classification from Thai Twitter,” presented at the 13th Int. Joint Conf. Comput. Sci. Softw. Eng. (JCSSE), Khon Kaen, Thailand: Khon Kaen Univ., Jul. 2016.

[19] B. Wing and J. Baldridge, “Simple supervised document geolocation with geodesic grids,” presented at the 11th Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol., Portland, OR, USA, Jun. 2011.

[20] W. Hua, K. Zhneg, and X. Zhou, “Microblog entity linking with social temporal context,” presented at the Proc. ACM SIGMOD Int. Conf. Manage. Data, Melbourne, VIC, Australia, May 2015.

[21] J. Mahmud, J. Nichols, and C. Drews, “Where is this tweet from? Inferring home locations of Twitter users,” presented at the Proc. 6th Int. AAAI Conf. Weblogs Social Media, Dublin, Ireland, May 2012.

[22] D. A. Kurniawan, S. Wibirama, and N. A. Setiawan, “Real-time traffic classification with Twitter data mining,” presented at the 8th Int. Conf. Inf. Technol. Electr. Eng. (ICITEE), Yogyakarta, Indonesia, Oct. 2016.

[23] R. Rahman, K. C. Roy, M. Abdel-Aty, and S. Hasan, “Sharing real-time traffic information with travelers using Twitter: An analysis of effectiveness and information content,” *Frontiers Built Environ.*, vol. 5, no. 83, Jun. 2019.

[24] 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.

[25] X. Zheng, J. Han, and A. Sun, “A survey of location prediction on Twitter,” *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 9, pp. 1652–1671, Sep. 2018.

[26] F. Melo and B. Martins, “Automated geocoding of textual documents: A survey of current approaches,” *Trans. GIS*, vol. 21, no. 1, pp. 3–38, 2017.

[27] G. Aurelien, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Newton, MA, USA: O’Reilly Media, 2017.

[28] C. Li and A. Sun, “Extracting fine-grained location with temporal awareness in tweets: A two-stage approach,” *J. Assoc. Inf. Sci. Technol.*, vol. 68, no. 7, pp. 1652–1670, Jul. 2017.

[29] G. DeLozier, B. Wing, J. Baldridge, and S. Nesbit, “Creating a novel geolocation corpus from historical texts,” presented at the Proc. 10th Linguistic Annotation Workshop, Berlin, Germany, Aug. 2016.

[30] S. Malmasi and M. Dras, “Location mention detection in tweets and microblogs,” presented at the Proc. Comput. Linguistics, Beijing, China, 2015.

[31] N. Bichu and A. Panangadan, “Analyzing social media communications for correlation with freeway vehicular traffic,” presented at the IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov., Aug. 2017.

[32] A. I. J. Tostes, T. H. Silva, F. Duarte-Figueiredo, and A. A. F. Loureiro, “Studying traffic conditions by analyzing foursquare and instagram data,” presented at the Proc. 11th ACM Symp. Perform. Eval. Wireless Ad Hoc, Sensor, Ubiquitous Netw., Montreal, QC, Canada, Sep. 2014.

[33] S. K. Endarnoto, S. Pradipta, A. S. Nugroho, and J. Purnama, “Traffic condition information extraction & visualization from social media Twitter for Android mobile application,” presented at the Int. Conf. Electr. Eng. Informat., Bandung, Indonesia, Jul. 2011.

[34] 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*, Ann Arbor, MI, USA, Jun. 2005, pp. 363–370.

[35] A. Ratnaparkhi, “A simple introduction to maximum entropy models for natural language processing,” Inst. Res. Cogn. Sci., Univ. Pennsylvania, Philadelphia, PA, USA, Tech. Rep. IRCS-97-08, 1997.

[36] A. Abbasi, T. H. Rashidi, M. Maghrebi, and S. T. Waller, “Utilising location based social media in travel survey methods: Bringing Twitter data into the play,” presented at the Proc. 8th ACM SIGSPATIAL Int. Workshop Location-Based Social Netw., Bellevue, WA, USA, 2015.

[37] F. Atefeh and W. Khreich, “A survey of techniques for event detection in Twitter,” *Comput. Intell.*, vol. 31, no. 1, pp. 132–164, 2015.

[38] M. Kumar, S. Singh, A. T. Ghate, S. Pal, and S. A. Wilson, “Informal public transport modes in India: A case study of five city regions,” *IATSS Res.*, vol. 39, no. 2, pp. 102–109, 2016.

[39] F. Ali, S. El-Sappagh, and D. Kwak, “Fuzzy ontology and LSTM-based text mining: A transportation network monitoring system for assisting travel,” *Sensors*, vol. 19, no. 2, p. 234, 2019.

[40] K. R. Pandhare and M. A. Shah, “Real time road traffic event detection using Twitter and spark,” presented at the IEEE Int. Conf. Inventive Commun. Comput. Technol. (ICICCT), Coimbatore, India, 2017.

[41] J. Gelernter and S. Balaji, “An algorithm for local geoparsing of microtext,” *GeoInformatica*, vol. 17, no. 4, pp. 635–667, Oct. 2013.

[42] M. F. Goodchild, “Citizens as sensors: The world of volunteered geography,” in *The Map Reader: Theories of Mapping Practice and Cartographic Representation*, M. Dodge, R. Kitchin and C. Perkins, Eds. Hoboken, NJ, USA: Wiley, 2007.

[43] S. E. Middleton, G. Kordopatis-Zilos, S. Papadopoulos, and Y. Kompatsiaris, “Location extraction from social media: Geoparsing, location disambiguation and geotagging,” *ACM Trans. Inf. Syst.*, vol. 36, no. 4, p. 40, 2018.

[44] M. W. Berry, *Survey of Text Mining*. Berlin, Germany: Springer-Verlag, 2003.

[45] 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.

[46] 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.

[47] M. Imran, C. Castillo, F. Diaz, and S. Vieweg, “Processing social media messages in mass emergency: A survey,” *ACM Comput. Surv.*, vol. 47, no. 4, 2015, Art. no. 67.

[48] C. Collins, S. Hasan, and S. V. Ukkusuri, “A novel transit rider satisfaction metric: Rider sentiments measured from online social media data,” *J. Public Transp.*, vol. 16, no. 2, pp. 21–45, 2013.

[49] G. Anastasi *et al.*, “Urban and social sensing for sustainable mobility in smart cities,” presented at the Sustain. Internet ICT Sustainab. (SustainIT), Palermo, Italy, 2013.

[50] S. Hajrahnur, M. Nasrun, C. Setianingsih, and M. A. Murti, “Classification of posts Twitter traffic jam the city of Jakarta using algorithm C4.5,” presented at the Int. Conf. Signals Syst. (ICSigSys), Bali, Indonesia, 2018.

[51] O. Ajao, J. Hong, and W. Liu, “A survey of location inference techniques on Twitter,” *J. Inf. Sci.*, vol. 41, no. 6, pp. 855–864, 2015.

[52] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media: A data mining perspective,” *ACM SIGKDD Explorations Newslett.*, vol. 19, no. 1, pp. 22–36, 2017.

[53] R. S. Purves, P. Clough, C. B. Jones, M. H. Hall, and V. Murdock, *Geographic Information Retrieval: Progress and Challenges in Spatial Search of Text*. Boston, MA, USA: Now, 2018.

[54] W. Zhang and J. Gelernter, “Geocoding location expressions in Twitter messages: A preference learning method,” *J. Spatial Inf. Sci.*, vol. 9, pp. 37–70, Dec. 2014.

[55] M. Vasardani, L. F. Stirling, and S. Winter, “The preposition at from a spatial language, cognition, and information systems perspective,” *Semantics Pragmatics*, vol. 10, no. 3, Jun. 2017.

[56] Y. Hu, “EUPEG: Towards an extensible and unified platform for evaluating geoparsers,” in *Proc. 12th Workshop Geographic Inf. Retr.*, Seattle, WA, USA, Nov. 2018, Art. no. 3.

[57] B. Alex, K. Byrne, C. Grover, and R. Tobin, “Adapting the Edinburgh geoparser for historical georeferencing,” *Int. J. Humanities Arts Comput.*, vol. 9, no. 1, pp. 15–35, 2015.

[58] M. Karimzadeh *et al.*, “GeoTxt: A Web API to leverage place references in text,” in *Proc. 7th Workshop Geographic Inf. Retr.*, Orlando, FL, USA, 2013, pp. 72–73.

[59] L. Hollenstein and R. Purves, “Exploring place through user-generated content: Using Flickr tags to describe city cores,” *J. Spatial Inf. Sci.*, vol. 1, pp. 21–48, Jul. 2010.

[60] F. A. Twaroch, C. B. Jones, and A. I. Abdelmoty, “Acquisition of a vernacular gazetteer from Web sources,” in *Proc. 1st Int. Workshop Location web*, Beijing, China, 2008, pp. 61–64.

[61] T. R. Babu, A. Chatterjee, S. Khandeparker, A. V. Subhash, and S. Gupta, “Geographical address classification without using geolocation coordinates,” in *Proc. 9th Workshop Geographic Inf. Retr.*, Paris, France, Nov. 2015, Art. no. 8.

[62] K. Bontcheva, L. Derczynski, A. Funk, M. A. Greenwood, D. Maynard, and N. Aswani, “TwitIE: An open-source information extraction pipeline for microblog text,” in *Proc. Int. Conf. Recent Adv. Natural Lang. Process.*, Hisarya, Bulgaria, Sep. 2013, pp. 83–90.

[63] N. F. N. Rajani, K. McArdle, and J. Baldridge, “Extracting topics based on authors, recipients and content in microblogs,” in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Gold Coast, QLD, Australia, Jul. 2014, pp. 1171–1174.

[64] A. Cebeillac and Y.-M. Rault, *Contribution of Geotagged Twitter Data in the Study of a Social Group'S Activity Space*. Beijing, China: Netcom, 2016.

[65] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, and C. Ratti, “Geo-located Twitter as proxy for global mobility patterns,” *Cartogr. Geograph. Inf. Sci.*, vol. 41, no. 3, pp. 260–271, 2014.

[66] Y. Li, Q. Li, and J. Shan, “Discover patterns and mobility of Twitter users—A study of four us college cities,” *Int. J. Geo-Inf.*, vol. 6, no. 2, p. 42, 2017.

[67] J. Poushter, C. Bishop, and H. Chwe. Leading Countries Based on Number of Twitter Users as of October 2018 (in Millions). Statista: The Statistics Portal, Accessed: Nov. 26, 2018. [Online]. Available: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

[68] Numbeo. *Traffic Index 2018*. Accessed: Dec. 27, 2018. [Online]. Available: https://https://www.numbeo.com/traffic/rankings.jsp/

[69] NDTV. *Bridge Collapses At Mumbai's Andheri Station, 5 Injured*. Accessed: Jul. 13, 2019. [Online]. Available: https://www.ndtv.com/mumbai-news/part-of-bridge-collapses-in-andheri-station-in-mumbai-trains-on-western-line-affected-1876853

[70] A. Khalid. *Twitter Removes Precise Geo-Tagging Option From Tweets*. Accessed: Jul. 13, 2019. [Online]. Available: https://www.engadget.com/2019/06/19/twitter-removes-precise-geo-tagging/

[71] N. Natu. *Police to Monitor 110 Locations and Ensure Traffic Keeps Moving*. Accessed: Jul. 13, 2019. [Online]. Available: https://timesofindia.indiatimes.com/city/mumbai/police-to-monitor-110-locations-and-ensure-traffic-keeps-moving/articleshow/65016719.cms

[72] F. Laylavi, A. Rajabifard, and M. Kalantari, “A multi-element approach to location inference of Twitter: A case for emergency response,” *ISPRS Int. J. Geo-Inf.*, vol. 5, no. 5, p. 56, 2016.

[73] Z. Cheng, J. Caverlee, and K. Lee, “You are where you tweet: A content-based approach to geo-locating twitter users,” in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manage.*, Toronto, ON, Canada, Oct. 2010, pp. 759–768.

[74] E. F. T. K. Sang and F. De Meulder, “Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition,” in *Proc. 7th Conf. Natural Lang. Learn. HLT-NAACL*, Edmonton, NSR, Canada, 2003, pp. 142–147.

[75] B. Mohit, “Named entity recognition,” in *Natural Language Processing of Semitic Languages*, I. Zitouni, Ed. Berlin, Germany: Springer, 2014, pp. 221–245.

[76] J. Liu, K. Zhao, S. Khan, M. Cameron, and R. Jurdak, “Multi-scale population and mobility estimation with geo-tagged tweets,” in *Proc. 31st IEEE Int. Conf. Data Eng. Workshops (ICDEW)*, Seoul, South Korea, Apr. 2015, pp. 83–86.

[77] M. Levin, “Bing maps and finite-dimensional maps,” *Fundamenta Mathematicae*, vol. 151, no. 1, pp. 47–52, 1996.

[78] M. F. Goodchild, “Maximum entropy model and conditional random field,” in *Machine Learning for Multimedia Content Analysis*, M. Dodge, R. Kitchin, and C. Perkins, Eds. New York, NY, USA: Springer, 2007, pp. 201–233.

[79] S. W. Shulman, “DiscoverText: Software training to unlock the power of text,” in *Proc. 12th Annu. Int. Digit. Government Res. Conf., Digit. Government Innov. Challenging Times*, College Park, MD, USA, 2011, p. 373.

[80] S. Hasan, K. C. Roy, M. M. Hasnat, and M. Abdel-Aty, “Sharing real-time traffic information with travelers using social media: An analysis of Twitter activities, influence, and effectiveness,” in *Proc. Transp. Res. Board 97th Annu. Meeting*, Washington, DC, USA, 2018, p. 7.

[81] D. Mocanu, A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignani, “The Twitter of Babel: Mapping world languages through microblogging platforms,” *PLoS ONE*, vol. 8, no. 4, 2013, Art. no. e61981.

[82] J. J. Gibson, “Affordances and behavior,” in *Reasons for Realism: Selected Essays of James J. Gibson*, E. Reed and R. Jones, Eds. Hillsdale, MI, USA: Lawrence Erlbaum, 1975.

[83] C. M. Raymond, M. Kyttä, and R. Stedman, “Sense of place, fast and slow: The potential contributions of affordance theory to sense of place,” *Frontiers Psychol.*, vol. 8, p. 1674, Sep. 2017.

[84] T. O'Reilly, *What is Web 2.0*. Newton, MA,USA: O'Reilly Media, 2009.

[85] E. D'Andrea and F. Marcelloni, “Detection of traffic congestion and incidents from GPS trace analysis,” *Expert Syst. Appl.*, vol. 73, pp. 43–56, May 2017.

[86] F. O. Ostermann, M. Tomko, and R. S. Purves, “User evaluation of automatically generated keywords and toponyms for geo-referenced images,” *J. Amer. Soc. Inf. Sci. Technol.*, vol. 64, pp. 480–499, Mar. 2013.

[87] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2019, pp. 4171–4186.

Rahul Deb Das received the B.S. degree in mining engineering from the Indian Institute of Engineering Science and Technology in 2010, the M.S. degree in geoinformatics from IIT Bombay in 2012, and the Ph.D. degree in geomatics engineering from the University of Melbourne, Australia, in 2017.

He is currently working with the Data Science Elite Team, IBM Research and Development, Germany, as a Data Scientist. Prior to this, he was a Post-Doctoral Researcher with the University of Zurich, Switzerland, where he led a number of transportation related research as a part of a Swiss National Science Foundation (SNSF) Project. His research interests include artificial intelligence, intelligent transportation systems, and geographical information science and its varied applications in urban analytics.

Ross S. Purves received the B.Sc. degree in technological physics from Glasgow University, Scotland, in 1991, and the Ph.D. degree in physics from Heriot-Watt University, Scotland, in 1995.

He is currently a Professor of geographic information science with the University of Zurich, Switzerland. His research interests focus on the use of unstructured data and social media to extract and understand geographic information.