Towards the usefulness of user-generated content to understand traffic events

Das, Rahul Deb ; Purves, Ross S

Abstract: This paper explores the usefulness of Twitter data to detect traffic events and their geographical locations in India through machine learning and NLP. We develop a classification module that can identify tweets relevant for traffic authorities with 0.80 recall accuracy using a Naive Bayes classifier. The proposed model also handles vernacular geographical aspects while retrieving place information from unstructured texts using a multi-layered georeferencing module. This work shows Mumbai has a wide spread use of Twitter for traffic information dissemination with substantial geographical information contributed by the users.

DOI: https://doi.org/10.4230/LIPIcs.GIScience.2018.25

Towards the Usefulness of User-Generated Content to Understand Traffic Events

Rahul Deb Das[1]
Department of Geography, University of Zurich
Winterthurerstrasse 190, 8057 Zurich, Switzerland
rahul.das@geo.uzh.ch
https://orcid.org/0000-0002-3379-3516

Ross S. Purves
Department of Geography, University of Zurich
Winterthurerstrasse 190, 8057 Zurich, Switzerland
ross.purves@geo.uzh.ch

Abstract

This paper explores the usefulness of Twitter data to detect traffic events and their geographical locations in India through machine learning and NLP. We develop a classification module that can identify tweets relevant for traffic authorities with 0.80 recall accuracy using a Naive Bayes classifier. The proposed model also handles vernacular geographical aspects while retrieving place information from unstructured texts using a multi-layered georeferencing module. This work shows Mumbai has a wide spread use of Twitter for traffic information dissemination with substantial geographical information contributed by the users.

2012 ACM Subject Classification Information systems → Geographic information systems, Information systems → Information retrieval, Computing methodologies → Natural language processing, Computing methodologies → Artificial intelligence, Human-centered computing → Ubiquitous and mobile computing

Keywords and phrases Urban mobility, traffic, UGC, tweet, event, GIR, geoparsing

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.25

Category Short Paper

Funding We are grateful to the Swiss National Science Foundation (SNSF) grant number 166788.

Acknowledgements We would like to thank Alan MacEachren, Natalia Andrienko, Gennady Andrienko, Liao Din, Martin Schorcht, Florian Lautenschlager, and Stefan Kasberger for their valuable comments during the initial stage of this project in VGIScience Summer School'17 in Dresden.

1 Introduction

Retrieving geographical information pertaining to events is important for planning and decision making processes, for instance in identifying locations that demand special attention. With the emergence of user-generated content (UGC), it is now possible to detect various urban events and their geographical locations more ubiquitously. Events may be related to, for example, urban mobility [6], natural disasters [13, 3] or environmental conditions [17].

[1] Corresponding author

UGC derived from social-media platforms are often unstructured and pose challenges if we are to relate vague and ambiguous references in natural language to specific locations [7]. This paper introduces a framework to deal with such challenges while detecting traffic events for managing urban resources and transportation infrastructure.

Currently traffic information is collected through static, and physical sensors e.g., loop detectors or CCTV cameras installed at different locations in a city [8]. Since these sensors are static, they provide limited spatial coverage and come with high installation and maintenance costs. In order to address these issues, this paper leverages the concept of *citizens as sensors* [5] where the citizens contribute information (in)voluntarily, which can be used to characterize traffic events.

We use Twitter to both analyze traffic in real time and gain insights into patterns over time. Our contributions are as follows.

- Unlike previous works [6, 9] we leverage ungeotagged tweets to extract the locations of traffic events through text analysis.
- We assess the usefulness of UGC (e.g., Twitter) to detect traffic events in India where many of the metro cities are highly congested [15] with limited physical traffic infrastructure.
- We develop a hybrid multi-layered geoparser that can retrieve traffic event locations from unstructured texts tweeted in India where place names are often mentioned in local languages.

In Section 2 we briefly review the state of the art. Section 3 and 4 explain the framework and its evaluation, before Sections 5 and 6 discuss some limitation of our approach and propose directions for future work.

2 Related work

Twitter is a ubiquitous UGC source where people post information, reactions and opinions about a vast array of topics [2]. In the past Twitter has been used to detect traffic events, however, mostly identifying traffic related information from geotagged tweets. For example, D'Andrea and colleagues developed a model that could detect traffic related tweets in real time in Italy using Support Vector Machine (SVM) with an accuracy of 95.75% [1]. They used a balanced data set with 665 instances each in training and testing. Kurniawan and colleagues developed a real-time tweet classification model using geotagged tweets in order to provide traffic related information in Indonesia [9]. Similarly Salas and others developed a SVM based supervised model to detect incidents in London using Twitter data [14]. They used a balanced data set for training and testing. In these papers traffic events and their locations are assumed to be the location in the tweet metadata.

Wanichayapong and colleagues developed a model to classify tweets as traffic or non-traffic related through syntactic analysis in Thailand. They also classified traffic related tweets in point and link category depending on the location of the traffic events [16]. They achieved 76.85% accuracy for point category and 93.23% accuracy for link category. Gu and others presented a real-time traffic incident detection model which was evaluated in Philadelphia and Pittsburgh in the USA. They developed the model based on a semi-Naive Bayes classifier and achieved 90.5% accuracy [6]. Since most of tweets are not explicitly georeferenced, various models have been proposed to extract locations from tweet content and metadata [4]. For example, Gelernter and Balaji proposed a hybrid model to georeference tweets in New Zealand [4]. In a slightly different work Pereira and others used text analysis to predict incident durations from the authoritative structured text [12].

In this work we propose a model that goes beyond existing traffic detection models that leverage geotagged tweets. Instead, we use ungeotagged tweets to understand traffic conditions through a hybrid multi-layered geoparser (c.f [4]) by applying a mixture of spatial rules and localized spatial references evaluated in Indian context.

3 Methodology

We propose a hierarchical model that can detect tweets relevant to traffic and then extract spatial information from the tweet to provide more information about the traffic event. The methodology is divided into three stages.

3.1 Data collection

To evaluate the model a data set has been collected in Mumbai [2] using a keyword *traffic* from 1st January to 28th February, 2017. Manual annotation was performed to label whether a tweet was related to traffic. Since all the tweets contain the keyword *traffic*, a number of criteria were set during the annotation process. A tweet is labelled as a relevant tweet if it contains information about a traffic event along with either a spatial reference (*where*) or a temporal reference (*when*) or a cause (*why*). Through this process, 2614 tweets were annotated over two months in Mumbai where the count of traffic related tweets was 755. Another manual annotation was performed to extract the place names mentioned in the traffic related tweets to use them as ground truth to evaluate the performance of the georeferencing module (c.f Section 3.3). A tweet may have more than one place name. In that case all the unique place names are annotated.

3.2 Tweet classification

After preprocessing (to eliminate emoticons and non-ASCII characters and trim white space from tweet content) three different classifiers were tested. Those were rule-based (PART), tree-based (Decision Tree (DT)), and a probabilistic classifier (Naive Bayes (NB)).

To create the features to train the classifiers the 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) as follows.

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

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

$$tf - idf = tf * idf \tag{3}$$

Where T_t is the total count of term 't' in tweet 'D'. 'N' is the total number of tweets in the corpus and D_t is the total number of tweets containing the term 't'.

[2] `https://www.numbeo.com/traffic/rankings.jsp`, last accessed April, 2018

3.3 Georeferencing module

In the third stage a 3-tier tweet georeferencing module (GM) was developed that can retrieve geographical information from the traffic relevant tweets. Initially a pre-trained supervised geoparser e.g., StanfordNLP [11] was used (1st tier). However, due to lack of training on the local data set (as in Mumbai) two more rule-based layers have been implemented. In the first rule-based layer (2nd tier) if a token is a proper noun (NNP) or a common noun (NN) and if it is preceded by a spatial preposition then the token is deemed to be a place name. We used 17 spatial prepositions e.g., *towards*, *from*, *to*, *at*, *on*, *near*. The second rule-based layer (3rd tier) considers vernacular place names in Mumbai (e.g., *naka: toll plaza*, *marg: road*, *bhavan: building*, *chowpatty: fishermen's colony*) and various spatial object types in English (e.g., *building*, *park*, *flyover*). Any NNP or NN token that is followed by one of these vernacular names or an object type is deemed to be a place name. We extracted 84 vernacular names and object types.

Once the spatial references are retrieved place names are resolved by assigning coordinates using OpenStreetmap (OSM). To disambiguate place names *Maharashtra* (the local state name) was used as a spatial context (c.f [10]).

4 Evaluation and results

4.1 Detecting traffic related tweets

The models are evaluated using 3-fold cross validation. While detecting traffic related tweets, a NB classifier performs best with 0.80 recall and 0.52 precision, while a rule-based model (PART) yields 0.67 precision and 0.57 recall and the DT model gives precision 0.65 and recall 0.57. For non-traffic tweets a NB classifier yields 0.89 precision whereas a PART and DT yield 0.88 and 0.87 recall respectively.

4.2 Performance of tweet georeferencing module

As the texts in tweets are often unstructured – involve abbreviation and typos, while measuring the accuracy of georeferencing module, first a complete match was performed. If the retrieved place name in tweet_k does not completely match with any of the annotated place names in the same tweet, then a fuzzy matching was performed. If the cosine similarity (CoSim) between the retrieved place name and the annotated place name is greater than a threshold (0.4) then the retrieved place name is considered as a true positive.

When using the StanfordNLP alone without the second and third tiers (rule base) over two months of traffic related tweets, the georeferencing module yields precision of 0.60 and recall 0.34. However, using all the tiers the georeferencing module yields 0.71 precision and 0.61 recall. Using all the three tiers total 451 places are resolved from the retrieved place names out of 767 resolved places from the annotated ones (Fig 1). As can be seen the proposed model effectively retrieves locations that are subject to traffic events with 58.88% places being successfully resolved (Fig 1).

5 Discussion

Currently the model uses tweets that contain only the keywords *traffic*, but in future more keywords will be incorporated. The georeferencing module presented in this work sometimes fails to detect place names consisting of two tokens followed by a vernacular name or object type. For example, *teen Haath Naka* has been recognized as *Haath Naka*, which is detected

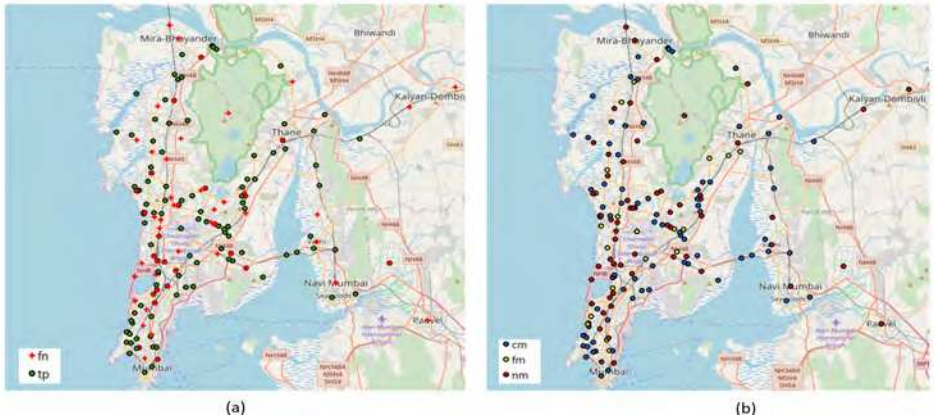

Figure 1 (a) Locations of the traffic events resolved from correctly retrieved tweets i,e true positives (tp) and location of annotated events that could not be retrieved i,e false negatives (fn); (b) Retrieved locations of the traffic events that completely match with the annotated place names (cm), fuzzy match (fm) and annotated places that could not be retrieved (nm) by the model.

Table 1 Spatial granularity in the text.

Tweet	Place Type	Geometry Type
@MumbaiPolice heavy traffic at bkc near income tax office...	Building (income tax office)	Point
Traffic movement on S V Road at Andheri and Jogeshwari is lot better today.	Road name (S V Road), Region (Andheri, Jogeshwari)	Polyline, Polygon

through fuzzy matching. We also observed that people use geographical information at different granularities while tweeting about traffic events (Table 1).

Similar to the earlier works while classifying the tweets, a k-fold cross validation has been used to evaluate the tweet classification model over two months data. It has been observed while tweeting people in Mumbai react in two ways, either they report or share traffic events or they request respective authority (e.g., *@MumbaiPolice*) to resolve a traffic issue. An extension of this work will investigate if the model performs equally well on a data set collected separately in a different time period.

Although in this research a small number of tweets have been analyzed based on only a single keyword, the approach is scalable and adaptive to more traffic related keywords and more tweets. In terms of the size of the data set past studies have also showed promising results with small data sets [1, 14]. The main focus of this paper was on detecting traffic relevant tweets and their respective locations. However, it is also important to identify the reasons behind the traffic events, which requires more complex syntactic and semantic analysis of the text.

6 Conclusions

In this paper a traffic event detection model has been introduced. The model can be useful both in real-time as well as in historical manner and can detect tweets relevant to traffic authorities, urban planners and daily commuters to understand the traffic events and their geographical locations both for short-term and long-term planning. In this research we showed Twitter has potential for detecting traffic events in Indian cities if we build a georeferencing model capable of dealing with unstructured, vague and vernacular text in natural language.

An important limitation of our work is that India is a multi-lingual country and our analysis focused on English. Nonetheless, vernacular terms are often used while communicating about an event with a spatial reference in English. Here the implemented multi-layered geoparser shows its effectiveness in resolving 58.88% of local places that are subject to have traffic events, which was not possible using a pre-trained NER due to lack of local traffic related corpora.

Future work will consider tweets with more traffic related keywords and explore temporal patterns of tweeting behavior reacting to traffic events. Although the study is performed in India, but the same approach can be useful to other places.

References

1. E. D. Andrea, P. Ducange, B. Lazzerini, and F. Marcelloni. Real-time detection of traffic from twitter stream analysis. *IEEE Transactions on Intelligent Transportation Systems*, 16(4):2269–2283, 2015.
2. Farzindar Atefeh and Wael Khreich. A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1):132–164, 2015.
3. Andrew Crooks, Arie Croitoru, Anthony Stefanidis, and Jacek Radzikowski. #Earthquake: Twitter as a distributed sensor system. *Transactions in GIS*, 17(1):124–147, 2013.
4. Judith Gelernter and Shilpa Balaji. An algorithm for local geoparsing of microtext. *GeoInformatica*, 17(4):635–667, 2013.
5. Michael F. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, 2007.
6. Yiming Gu, Zhen Qian, and Feng Chen. From twitter to detector: Real-time traffic incident detection using social media data. *Transportation Research Part C: Emerging Technologies*, 67:321–342, 2016.
7. Livia Hollenstein and Ross S Purves. Exploring place through user-generated content: Using flickr to describe city cores. *Journal of Spatial Information Science*, 1(1):21–48, 2010.
8. Akira Kinoshita, Atsuhiro Takasu, and Jun Adachi. Real-time traffic incident detection using a probabilistic topic model. *Information Systems*, 54:169–188, 2015.
9. D. A. Kurniawan, S. Wibirama, and N. A. Setiawan. Real-time traffic classification with twitter data mining. In *8th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pages 1–5, Yogyakarta, Indonesia, 2016.
10. Jochen L. Leidner, Gail Sinclair, and Bonnie Webber. Grounding spatial named entities for information extraction and question answering. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references - Volume 1*, Stroudsburg, USA, 2003.
11. Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, Maryland, USA, 2014.
12. Francisco C. Pereira, Filipe Rodrigues, and Moshe Ben-Akiva. Text analysis in incident duration prediction. *Transportation Research Part C: Emerging Technologies*, 37:177–192, 2013.
13. Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World Wide Web*, pages 851–860, Raleigh, North Carolina, USA, 2010. ACM.
14. A. Salas, P. Georgakis, and Y. Petalas. Incident detection using data from social media. In *IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 751–755, Yokohama, Japan, 2017.

15 Azeem Uddin. Traffic congestion in indian cities: Challenges of a rising power, draft. Report, General Motors India, 2009.

16 Napong Wanichayapong, Wasawat Pruthipunyaskul, Wasan Pattara-Atikom, and Pimwadee Chaovalit. Social-based traffic information extraction and classification. In *IEEE 11th International Conference on ITS Telecommunications*, St. Petersburg, Russia, 2011.

17 Yuchao Zhou, Suparna De, and Klaus Moessner. Real world city event extraction from twitter data streams. *Procedia Computer Science*, 98:443–448, 2016.