



Detection of Damage and Failure Events of Road Infrastructure Using Social Media

Aibek Musaev^(✉), Zhe Jiang, Steven Jones, Pezhman Sheinidashtegol, and Mirbek Dzhumaliev

The University of Alabama, Tuscaloosa, AL 35487, USA
aibek@cs.ua.edu

Abstract. We study the problem of estimating the state of road infrastructure, which is the backbone of transportation system. Road infrastructure can suffer from various issues, including structural failures, such as potholes, and non-structural issues, such as broken traffic lights. However, it is infeasible to cover all roads with physical sensors for monitoring purposes. Instead, we propose to use social sensor big data to detect and estimate these issues based on the public's activity. As a demonstration, we generate a map of detected road problems based on tweets. The map displays the currently detected hotspots, where for each hotspot we compute the overall sentiment provided by the public. In addition, we identify the peak of public activity during the evaluation period and investigate the key drivers of that peak. Finally, we analyze the most influential users using an extension of PageRank. The proposed approach adds a novel perspective on the state of road infrastructure and may be used to help guide decisions related to road infrastructure funding.

Keywords: Road infrastructure · Social Media · Sentiment analysis
PageRank

1 Introduction

Social Media platforms have become a part of many people's lives. According to Pew Research Center's report, the use of social networking sites by American adults increased from 7% to 65% within a 10 year period [29]. The explosion of user generated content in Social Media published from mobile devices has led to the notion of mobile social reporting [5]. This in turn has led to the concept known as "citizen sensing", whereby a mass of these citizen reports can be gathered and mined for meaningful information about the event being observed [34]. In citizen sensing, users are acting as social sensors and the information must be extracted from the texts of their status updates, photo and video captions. For example, [32] use Twitter posts to detect earthquake events, while [24] introduce a multi-service composition system for landslide detection based on social and physical sensors.

We study the problem of estimating the state of road infrastructure using citizen sensing. Roads are the backbone of transportation system; however road conditions are deteriorating as the need for improvements far outpaces the amount of funding available. The Congressional Budget Office projects an annual gap of around \$15 billion between the anticipated flow of revenue into the highway trust fund and the cost of maintaining current public transportation programs [19]. Hence, the decision makers at the federal and state levels are faced with difficult choices regarding the allocation of limited funds for transportation improvements.

We propose a novel perspective on the state of road infrastructure that may be used to help guide those decisions. Specifically, we introduce a comprehensive framework that detects damage and failure events based on Social Media data. See Fig. 1 for actual examples of public Twitter posts reporting various road conditions.

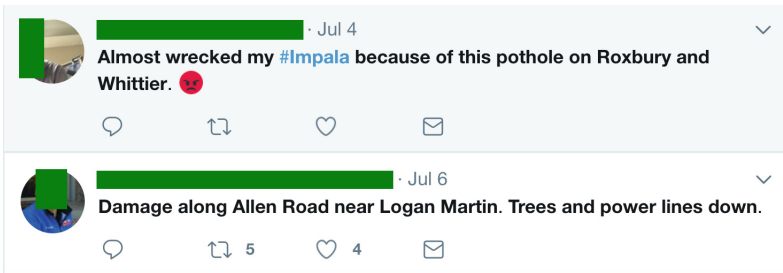


Fig. 1. Examples of tweets reporting failure events of road infrastructure

Note, that both tweets are posted by regular users, which is why their screen names and pictures are hidden. Also observe, that both tweets have a negative sentiment and contain a reference to geographical locations where the events have occurred.

The proposed framework collects Twitter posts using a set of road infrastructure related keywords, namely *pothole* and *road damage*, and processes them in a series of steps. The raw Twitter data are analyzed at the individual and group levels to detect spatial hotspots. For each detected hotspot we compute its aggregate sentiment, such that the hotspots with the highest amount of negative sentiment represent the most critical areas of concern as reported by the public.

The remaining text is organized as follows. Section 2 summarizes related work and provides an overview of the proposed framework and its components. In Sect. 3, we present an evaluation of failure events detection using real data. Section 4 concludes this paper and describes our future work.

2 Overview of Related Work and Proposed Framework

The proposed framework is a comprehensive solution that estimates the state of road infrastructure using Social Media messages. Specifically, using the data

collected from Twitter based on search keywords, we want to determine the hotspots of road infrastructure problems. The framework can be broken into 4 parts as follows:

- (A) Data collection
- (B) Processing of individual messages
- (C) Processing of groups of messages
- (D) Spatial hotspot detection

We start by collecting the data from Twitter using a set of search keywords, namely *pothole* and *road damage*. Then we process each individual message in a series of steps. If a tweet does not contain geographic coordinates then we attempt to retrieve the mentions of geographic places and assign the coordinates to the tweet based on those places. Next, we assign a *relevant* or *irrelevant* label to each geotagged tweet based on its relevance to our topic of interest, which is the state of road infrastructure. Finally, we compute the tweet's *happy* or *sad* label using sentiment analysis.

Given a set of individual tweets with geographic coordinates, relevance and sentiment labels, we estimate the potential failure events next. We begin by grouping messages based on their geographic coordinates. Each such group contains multiple tweets in it. For each group, we determine an aggregate relevance and an aggregate sentiment based on the relevance and sentiment of the individual tweets.

Lastly, we detect spatial hotspots among the estimated events. Note, that the output of our system is not only a list of detected hotspots of road infrastructure issues, but also a set of all tweets associated with each hotspot. We describe each of these steps in detail next.

2.1 Data Collection

Numerous researchers have collected Social Media as a source for real time data on traffic incidents and crashes, e.g., [6, 14]. [20] examined the usefulness of Twitter as a means of disseminating transportation-related information.

Information derived from social media has been shown to be useful in understanding travel patterns [30]. [18] illustrated the usefulness of social media, specifically Twitter, in understanding relationships between travel behavior and the spatial distribution of commercial land uses. While others, e.g., [4] have illustrated the use of Twitter information in understanding how public transportation is used.

We use Twitter's Streaming API to collect public messages related to road infrastructure failures in near real-time¹. The keywords that we apply are *pothole* and *road damage*. Any time a user posts a tweet containing any of the specified keywords, such tweet gets pushed to our streaming API client that stores the message for subsequent analysis.

¹ <https://dev.twitter.com/streaming/overview>.

Note, that the search keywords that we use, are in English, so the tweets containing them are also in English. In order to detect road infrastructure issues in other parts of the world, the framework must be extended to support additional languages. Furthermore, we apply a fixed set of keywords, which is not sufficient to cover all kinds of road infrastructure issues. We plan to add support for other languages and to expand the search keywords as part of the future work.

2.2 Processing of Individual Messages

The proposed system processes each tweet in a series of steps, including geotagging, relevance analysis and sentiment analysis. See Fig. 2 for an overview of the part of the pipeline that is responsible for processing individual messages.

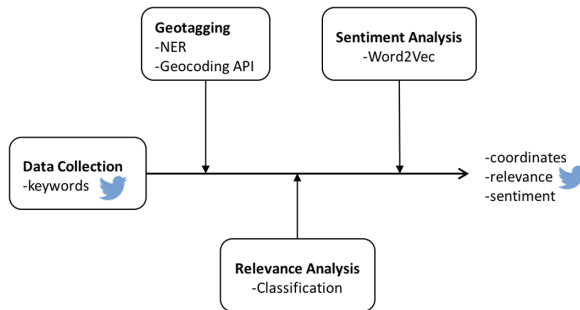


Fig. 2. Processing of individual messages

Geotagging. Like many other modern Social Media platforms, Twitter allows users to add location context to the messages they post, e.g., in their user profile pages. [16] show, however, that 34% of users provide fake locations, and those that input real locations, mostly specify them at a coarse level of granularity, such as a city level. Twitter also allows users to disclose their location when they post a tweet. However, according to [3], less than 0.42% of all Twitter users enable this functionality. In the evaluation dataset that we collected, only 0.01% of all tweets have location context: see Table 1 for more information. Furthermore, even when users disclose their location, they frequently discuss events that occur elsewhere, especially in case of major events, such as presidential elections. [37] identify information in Social Media that may contribute to situational awareness. Their study includes the analysis of the use of geo-location information in Twitter data during natural hazards events. [24] propose an automated approach that looks for mentions of geographic places in Social Media texts using a natural language processing technique called Named Entity Recognition (NER). We apply the same approach to retrieve location entities mentioned in tweets using Stanford

NER library [9] and subsequently convert the detected locations to geographic coordinates by invoking the Google Geocoding API [12].

It should be noted that the location entities used to determine geographic coordinates may be incorrectly retrieved from Social Media messages as illustrated in the following example:

- *Thanks to City of Boroondara for their response to Pothole - Side Road in Highfield Road Camberwell VIC Aust. . .*

In this example, the NER library incorrectly extracts “side road” instead of “Highfield Road Camberwell VIC”. However, the NER library correctly retrieves location entities in the majority of tweets. [25] use this observation to implement a semantic clustering approach that removes outliers from clusters of tweets grouped together based on semantic distance.

Finally, even correctly retrieved entities may have multiple geographic locations and it may not be clear which specific location the user is referring to. As an example, there are multiple cities named *Moscow* in USA alone in addition to the similarly titled capital of Russia. We plan to address this problem in future work, but possible approaches include the analysis of the user’s location context and location’s popularity.

Relevance Analysis. One of the challenges in the analysis of Social Media data is the amount of noise that is present there. This is mainly due to the ambiguity of the search keywords used to collect them. Here is an example of noise in Twitter containing search keywords *road* and *damage*:

- *Social Media Crisis - the Road to Recovery Social Media Crisis and Damage Control What do you do whe. . .*

Such posts can be easily identified based on the presence of the phrase “road to recovery.” Similarly, the following tweets illustrate some irrelevant meanings of the word “pothole”:

- *I call opening the first weed coffee shop in Denver and calling it The Pothole.*
- *Just played: Too Late To Die Young - Luke Spurr Allen - Pothole Heart/2017 (Chicken Little Publishing).*

These examples cannot be easily identified and require a more sophisticated approach for labeling them. [32] suggest a machine learning technique called text classification to determine the relevance of Twitter posts to earthquake

Table 1. Overview of evaluation dataset

Social media	Raw data	Data geo-tagged by user	Data geo-tagged by system
Twitter	14,947	176 (0.01%)	1,740 (11.64%)

events. They devise a classifier based on common statistical features, such as the position of the keywords in a tweet, the number of words, and their context. However, these features do not produce robust classification results; hence, other approaches were suggested. [25] propose a reduced explicit semantic analysis (ESA) approach to be used for classification features. ESA was introduced by [11] as a method for computing semantic relatedness using Wikipedia articles as knowledge repository.

[21] present an automated framework for harvesting a range of transportation-related information from Twitter. They conclude that it is a potentially useful source of information but note the challenges in classifying and interpreting the tweets – suggesting that an appropriate domain ontology be developed.

The majority of tweets containing *pothole* and *road damage* keywords in our dataset are related to road infrastructure issues. Using a trivial rule, such as the presence of a stop word or stop phrase, we are able to identify most of the irrelevant posts in our dataset. We plan to add support for text classification of Social Media in future work.

Sentiment Analysis. In addition to learning the relevance of each tweet to our topic of interest, we also want to determine whether the tweet expresses a positive or negative sentiment. Our objective is not only to find the locations related to road infrastructure issues, but also to find out which ones affect the public the most based on the opinion they share in their posts.

Opinion mining and sentiment analysis is an extensive area of research: see [28] for a generic overview of the work in this field. With proliferation of Social Media, however, this area started to attract an even higher amount of interest among researchers. For example, [31] proposes to use emoticons, such as “:-)” and “:-(” to automatically build a training set for sentiment analysis by collecting texts from Usenet newsgroups. [13] use emoticons to generate a training set based on tweets and find that classification algorithms, such as Naive Bayes, show high accuracy when trained with emoticon data. [27] expand on this idea and study the use of n-grams as features for an automated generation of the sentiment classifier.

We reuse the idea of generating a training set based on the presence of positive and negative emoticons. However, instead of using n-grams as features, we propose to apply the Continuous Bag-of-Words and Skip-gram model introduced by [23]. This approach can effectively compute semantic word similarity and it is able to achieve such results with a relatively small number of vector dimensions. This is important as vectors with smaller dimensions are faster to compute. The published model contains pre-trained 300-dimensional vectors for 3 million words and phrases².

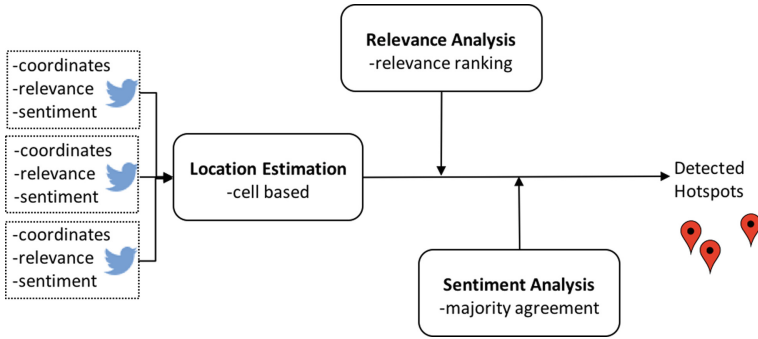
Given a training set of emoticon data, we convert each tweet in this set to a centroid vector [15]. As each vector has a *happy* or *sad* label, we can now build a classification model to predict the label for unseen posts. See Sect. 3 for details

² <https://code.google.com/p/word2vec/>.

on the generation of the training set for sentiment analysis and the evaluation of the proposed sentiment analysis model.

2.3 Processing of Groups of Messages

After the processing of the individual tweets in previous steps is finished, each remaining tweet in our set has a geographic coordinate together with the relevance and sentiment labels. We estimate potential hotspot locations next. Then for each found location, we determine an aggregate relevance and sentiment labels. See Fig. 3 for an overview of the part of the pipeline that is responsible for processing groups of messages.



$$column = (180^\circ + E)/(2.5'/60') = (180^\circ + E) * 24 \quad (2)$$

This approach allows for a resolution of roughly 2.7 miles in latitude and longitude, while being fast to compute. Although the total number of cells in this approach is large, we only need to consider non-empty cells, i.e. cells containing tweets in them.

Aggregate Relevance and Sentiment Analysis. Given a set of tweets with relevance and sentiment labels in each cluster, we want to compute the clusters' aggregate relevance and sentiment scores. These scores would represent our confidence in the aggregate decision labels that we assign to them. They could also be used to rank the clusters, such that the top N clusters above a threshold are preserved for further analysis.

Multiple approaches exist to implement a ranking strategy. In decision theory, weighted sum is a common method for evaluating multiple alternatives based on a number of criteria, where each criteria is given a certain weight [8]. On Twitter, users are not created equal and different weights can be assigned to their posts. For example, Twitter users have varying amount of influence [1]. Also, they may have varying degree of expertise in the subject: a post by an official State Department of Transportation account on road conditions will be highly relevant and should have a higher weight than a post sent by a random user with no history of interest in this subject and no connections with such users.

In this project, we use a simple decision rule of majority agreement, which is used in many domains, e.g. [35]. This approach effectively assigns all users the same weight and, thus, has limitations [33]. In the current implementation, we compute a majority relevance and sentiment label for each cluster and plan to introduce the weighted sum approach in the next iteration of the project.

2.4 Spatial Hotspot Detection

Given a set of spatial objects (e.g. points) in a study area, the problem of *spatial hotspot detection* aims to find regions where the number of objects is unexpectedly or anomalously high. Spatial hotspot detection is different from spatial partitioning or clustering, since spatial hotspots are a special kind of cluster whose intensity is “significantly” higher than the outside. Application domains for spatial hotspot detection range from public health to criminology. For example, in epidemiology finding disease hotspots allows officials to detect an epidemic and allocate resources to limit its spread [22]. In criminology finding a ring-shaped crime hotspot may help officials locate a serial criminal [2, 7].

As we describe in Sect. 2.2, the detected location entities found in tweets are converted to geographic coordinates by a process called geocoding. Hence, the tweets containing the same locations are converted to the same coordinates. As there are many tweets mentioning the same locations, especially in case of critical failure events, they all share the same coordinates. Based on this observation, we currently use the Heatmap Layer of the Google Maps Javascript API to visualize the potential hotspot locations³. The advantage of this mapping tool is that multiple points with the same coordinates produce a higher intensity. We also plan to experiment with the kernel density estimation approach for spatial hotspot detection.

³ <https://developers.google.com/maps/documentation/javascript/heatmaplayer>.

3 Evaluation Using Real Data

Evaluation of sentiment analysis.

Training set. To generate a training set for sentiment analysis, we use “:-)” and “:-(” emoticons to collect Twitter posts and automatically label them as *happy* and *sad* accordingly. As these emoticons are used internationally, we only preserve the tweets that are written in English. Since the number of *happy* tweets is significantly larger than the number of *sad* tweets, we randomly select *happy* tweets, so that the number of both *happy* and *sad* tweets in the training set is equal to 5,000.

Before we build a classifier model, such as Support Vector Machines, we need to convert the tweets in the training set to their vector representations. As we describe in Sect. 2.2, we utilize the Word2Vec repository that contains 300-dimensional vectors for 3 million words and phrases. Specifically, given a tweet in the training set, we split it into words and for each word we retrieve a corresponding vector. As we only consider tweets that contain multiple words, we compute the centroid of the vectors representing the words.

Evaluation set. For evaluation of the proposed sentiment analysis approach, we use Twitter data collected using *pothole* and *road damage* as search keywords. We observe that the majority of posts containing these words have a negative sentiment as shown below:

- *Wheelchair user falls into middle of road outside Letchworth railway station after hitting pothole* <https://t.co/CUaZRaDX3J>
- *Significant tree damage north of Tintern on Tintern road in Niagara.* [#onstorm #DTCWX #NRP pic.twitter.com/6HX8eJi9YX](https://twitter.com/6HX8eJi9YX)

Hence, we define the positive class for our classification task as tweets annotated with the *sad* label and the negative class as tweets annotated with the *happy* label. Based on these definitions, we compute precision and recall, which are standard metrics in information retrieval.

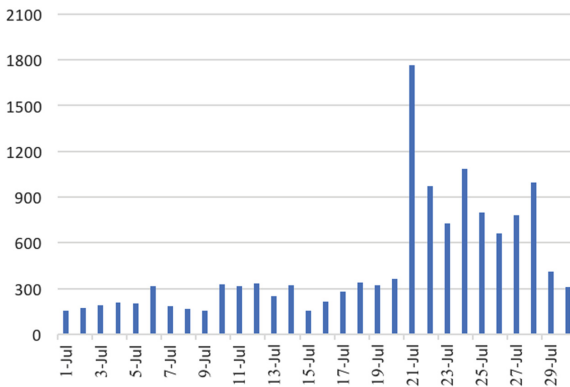


Fig. 4. Daily Twitter activity on potholes in July, 2017

Again, as with the training set, we convert each tweet in the evaluation set to its vector representation using Word2Vec repository for feature generation. Based on the classification model generated using the training set, the precision is equal to 0.92, while recall is equal to 0.43. Although the recall is low, the high value of precision prompts further investigation of this approach.

Hotspot Detection Analysis. We use *pothole* and *road damage* keywords to collect Twitter posts: see an overview of the evaluation set in Table 1. Using the Stanford NER library we retrieve the mentions of places for each tweet and then convert them to geographic coordinates by invoking the Google Geocoding API. Given a set of geographic coordinates, we compute their cell values according to Eqs. 1 and 2 and group tweets into clusters by their cell values. We keep the clusters whose aggregate relevance and sentiment labels are *relevant* and *sad* based on majority agreement. Finally, we map the points using the Heatmap layer of the Google Maps API to determine hotspot locations.

See Fig. 5 for a map of currently detected road infrastructure failures in USA, where major cities, such as New York, Boston, Chicago, and Houston, are highlighted. Note, that the presented framework is not tied to a specific geographic location. Due to ubiquity of mobile devices in all parts of the world and the explosion of user generated content in Social Media, citizen sensing can be also used to estimate the state of road infrastructure in other countries. Here is an example of how it can be applied to India, which recently had 2 events that attracted a high amount of public activity. In the first event a pothole problem led to a fatal accident:

- *Mumbai woman biker gets crushed under truck while trying to avoid pothole* <https://t.co/J8Mnrfdcd> #NewInVids <https://t.co/KOilyb6vIn>.

The second event became a focus of public attention due to a local celebrity involved in the event:

- *BMC officer en-route to serve notice to RJ Malishka gets injured after his bike falls into pothole* <https://t.co/a7Ud3yJo0D>.

Analysis of Twitter Activity on Pothole Issues. We generate a daily time series chart using tweets posted in July, 2017 (Fig. 4).

Note, that there is a spike of activity on July 21st. To understand the reasons for this peak, we show the text of the most retweeted tweets posted on July 21, 2017 in Table 2 together with the number of retweets and relevance to pothole as a road infrastructure issue. These results illustrate the terms used by the most popular social media messages.

Next we analyze the most important terms among all tweets in our dataset. Specifically, we generate a word cloud using the tweets with at least 2 retweets, such that there are 890 of them during July, 2017. Word clouds generated for a body of text can serve as a starting point for a deeper analysis [17]. Thus, we use

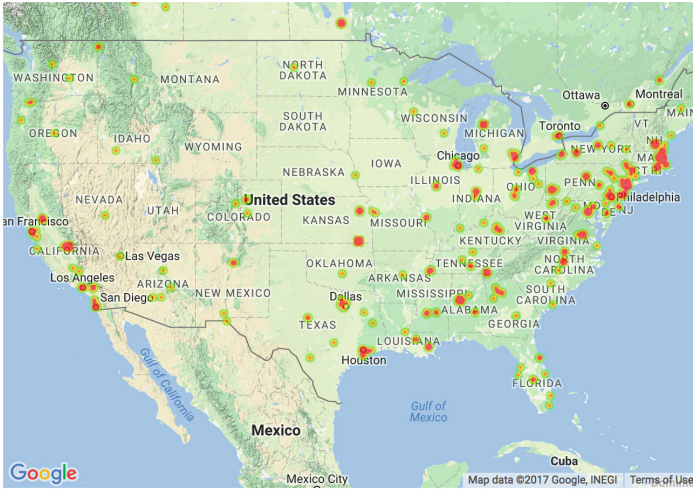


Fig. 5. Example of a map of road infrastructure failures in USA

a word cloud to distill the most popular tweets down to terms that appear with the highest frequency. We remove “pothole” from this list as we already know each tweet will contain it. Figure 6 presents a visualization of the word cloud from the most retweeted tweets that have 2 or more retweets. As expected, the words *road* and *repair* are significant according to this visualization.

Table 2. Examples of the most retweeted tweets on potholes posted on July 21, 2017.

Relevance	Text	Retweets
Irrelevant	Pothole https://t.co/HBxBvHxv5s	3271
Irrelevant	Humour—Mumbai youth drowns in rain induced pothole comes out 2 days later from the Indian Ocean? #FakeItIndia	156
Relevant	Mumbai..BMC pe Bharosa Nai ka ? Officer enroute 2 serve Notice 2 #Malishka gets injured as Bike Falls in Pothole	77
Relevant	There are many ways in which you can travel from one pothole to another	26
Relevant	@KeNHAKenya @lapsset The LAPSSSET Highway between Isiolo and Moyale (500 km) is the longest road without a single pot?	25
Relevant	Mardi gras bead filled pothole in New Orleans	24
Relevant	There’s a pothole left on my street unfilled for almost a year now after repeated complaints. Guaranteed Dengue Br?	14

blue labels indicate relevant users. Observe that both red labels are connected and indicate two celebrities who are collaborators on a song called “Pothole”; whereas blue labels show regular Twitter users acting as citizen sensors.

4 Conclusion and Future Work

In this paper we introduced a framework that detects road infrastructure hotspots based on Social Media reports. We break the framework into 4 components, including data collection, processing of individual messages, processing of groups of messages, spatial hotspot detection, and describe each of them in detail. The presented framework is comprehensive as it analyzes each collected message. We proposed to estimate the sentiment of each detected failure event as a novel perspective on the state of road infrastructure. The proposed sentiment analysis is based on text classification where Word2Vec vectors are used as features. In addition, we provide analysis of the most important terms and users in the collected dataset. Finally, we release both training⁴ and evaluation⁵ datasets as a contribution to research community.

Currently, we use a fixed set of keywords to download Social Media data related to road infrastructure failures, but we are exploring an approach to generate the list of keywords dynamically. Also note, that the current framework only supports English at the moment. We plan to extend support for other languages in order to improve the coverage of detected road issues. Our future work includes support of not only additional social networking platforms, such as YouTube and Facebook, but also integration of social sensor data with the data coming from physical sensors. Physical sensors represent authoritative and relevant sources and may include data coming from various road sensors, traffic lights and weather reports.

References

1. Cha, M., et al.: Measuring user influence in Twitter: The million follower fallacy. In: ICWSM 2010, pp. 10–17, p. 30 (2010)
2. Chainey, S., Ratcliffe, J.: GIS and Crime Mapping. Wiley, Hoboken (2013)
3. Cheng, Z., Caverlee, J., Lee, K.: You are where you tweet: a content-based approach to geo-locating Twitter users. In: 2010 CIKM (2010)
4. Cottrill, C., et al.: Tweeting Transit: an examination of social media strategies for transport information management during a large event. *Transp. Res. Part C Emerg. Technol.* **77**, 421–432 (2017)
5. Crowley, D.N., et al.: Gamification of citizen sensing through mobile social reporting. In: 2012 IEEE International Games Innovation Conference, pp. 1–5, September 2012. <https://doi.org/10.1109/IGIC.2012.6329849>
6. D’Andrea, E., et al.: Real-time detection of traffic from Twitter stream analysis. *IEEE Trans. Intell. Transp. Syst.* **16**(4), 2269–2283 (2015)

⁴ http://aibek.cs.ua.edu/files/train_ids.txt.

⁵ http://aibek.cs.ua.edu/files/eval_ids.txt.

7. Eftelioglu, E., et al.: Ring-shaped hotspot detection: a summary of results. In: 2014 IEEE International Conference on Data Mining, ICDM 2014, Shenzhen, China, 14–17 December 2014, pp. 815–820 (2014)
8. Triantaphyllou, E.: Multi-criteria Decision Making Methods: A Comparative Study. Kluwer Academic Publication, Dordrecht (2000)
9. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL 2005, Ann Arbor, Michigan, pp. 363–370 Association for Computational Linguistics (2005)
10. Fox, V., et al.: Bayesian filtering for location estimation. *IEEE Pervasive Comput.* **2**(3), 24–33 (2003)
11. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In: IJCAI 2007, Hyderabad, India, pp. 1606–1611 (2007)
12. Ge, X.: Address geocoding. US Patent 6,934,634, August 2005. url: <https://www.google.com/patents/US6934634>
13. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, p. 12 (2009)
14. Gu, Y., Qian, Z.S., Chen, F.: From Twitter to detector: real-time traffic incident detection using social media data. *Transp. Res. part C Emerg. Technol.* **67**, 321–342 (2016)
15. Han, E.-H.S., Karypis, G.: Centroid-based document classification: analysis and experimental results. In: Zighed, D.A., Komorowski, J., Żytkow, J. (eds.) PKDD 2000. LNCS (LNAI), vol. 1910, pp. 424–431. Springer, Heidelberg (2000). https://doi.org/10.1007/3-540-45372-5_46
16. Hecht, B., et al.: Tweets from Justin Bieber’s heart: the dynamics of the location field in user profiles. In: CHI (2011)
17. Heimerl, F., et al.: Word cloud explorer: text analytics based on word clouds. In: 2014 47th Hawaii International Conference on System Sciences (HICSS), pp. 1833–1842. IEEE (2014)
18. Huang, A., Gallegos, L., Lerman, K.: Travel analytics: understanding how destination choice and business clusters are connected based on social media data. *Transp. Res. Part C Emerg. Technol.* **77**, 245–256 (2017)
19. Kirk, R.S., Mallett, W.J.: Funding and financing highways and public transportation. Technical report R42877. Congressional Research Service, p. 33, September 2015
20. Kocatepe, A., et al.: The reach and influence of DOT Twitter accounts: a case study in Florida. In: 2015 IEEE 18th International Conference on Intelligent Transportation Systems (ITSC), pp. 330–335. IEEE (2015)
21. Kuffik, T., et al.: Automating a framework to extract and analyse transport related social media content: the potential and the challenges. *Transp. Res. Part C Emerg. Technol.* **77**, 275–291 (2017)
22. Kulldorff, M.: SaTScan user guide for version 9.0 (2011)
23. Mikolov, T., Yih, W.-T., Zweig, G.: Linguistic regularities in continuous space word representations (2013)
24. Musaev, A., Wang, D., Pu, C.: LITMUS: a multi-service composition system for landslide detection. *IEEE Trans. Serv. Comput.* **8**(5), 715–726 (2015). ISSN 1939-1374
25. Musaev, A., et al.: Toward a real-time service for landslide detection: augmented explicit semantic analysis and clustering composition approaches. In: 2015 IEEE International Conference on Web Services (ICWS), pp. 511–518. IEEE (2015)

26. Page, L., et al.: The PageRank citation ranking: bringing order to the web. Technical report, Stanford InfoLab (1999)
27. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: LREc, Vol. 10 (2010)
28. Pang, B., Lee, L., et al.: Opinion mining and sentiment analysis. *Found. Trends® Inf. Retrieval* **2**(1–2), 1–135 (2008)
29. Perrin, A.: Social networking usage: 2005–2015. Technical report, Pew Research Center (2015)
30. Rashidi, T.H., et al.: Exploring the capacity of social media data for modelling travel behaviour: opportunities and challenges. *Transp. Res. Part C Emerg. Technol.* **75**, 197–211 (2017)
31. Read, J.: Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In: Proceedings of the ACL Student Research Workshop, pp. 43–48. Association for Computational Linguistics (2005)
32. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: WWW (2010)
33. Salk, C.F., et al.: Limitations of majority agreement in crowdsourced image interpretation. *Trans. GIS* **21**(2), 207–223 (2017)
34. Sheth, A.: Citizen sensing, social signals, and enriching human experience. *IEEE Internet Computing* **13**(4), 87–92 (2009). <https://doi.org/10.1109/MIC.2009.77>. ISSN 1089–7801
35. Tian, T., Zhu, J.: Max-margin majority voting for learning from crowds. In: Cortes, C., et al. (eds) *Advances in Neural Information Processing Systems*, vol. 28, pp. 1621–1629. Curran Associates Inc. (2015)
36. To, H., et al.: Real-time task assignment in hyperlocal spatial crowd-sourcing under budget constraints. In: 2016 IEEE International Conference on Pervasive Computing and Communications (PerCom), pp. 1–8. IEEE (2016)
37. Vieweg, S., et al.: Microblogging during two natural hazards events: what Twitter may contribute to situational awareness. In: CHI 2010, pp. 1079–1088. ACM, Atlanta (2010). ISBN 978-1-60558-929-9