Applying machine learning and geolocation techniques to social media data (Twitter) to develop a resource for urban planning

Sveta Milusheva[1*], Robert Marty[1], Guadalupe Bedoya[1], Sarah Williams[2], Elizabeth Resor[3], Arianna Legovini[1]

1 Development Impact Evaluation Department, World Bank, Washington DC, United States of America

2 School of Architecture and Planning, Massachusetts Institute of Technology, Cambridge MA, United States of America

3 School of Information, University of California, Berkeley CA, United States of America

* smilusheva@worldbank.org

The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

Abstract

With all the hype around big data, it is easily overlooked that even basic vital statistics remain difficult to obtain in most of the world. What makes this frustrating is that potentially useful data exists in private companies, just not in the hands of people who can use it to track poverty, reduce disease, or build urban infrastructure. In this project, we set out to test whether we can transform an openly available dataset (Twitter) into a resource for urban planning and development. We test our hypothesis by creating Road Traffic Crashes (RTC) location data, scarce in most resource-poor environments, yet essential for addressing the number one cause of mortality for children over five and young adults. We scrape 874,588 traffic related tweets in Nairobi, Kenya, apply a machine learning model to capture the occurrence of a crash, and develop an improved geoparsing algorithm to identify its location. We dispatch a motorcycle delivery service in real-time to verify the crash and its location; the results show 92% accuracy. We geolocate 36,538 crash reports for 2012-2020 and cluster them into 26,791 unique crashes during this period. This is the first geolocated dataset of crashes for the city and allowed us to produce the first crash map for Nairobi. By clustering locations based on concentration of crashes we identify the portion of the road network (<1%) where 50% of the crashes identified occurred, providing urban planners with essential information they can use to plan safety roadways. The methods can be adapted to collect other essential data that can be hard to obtain such safety and natural disasters.

Introduction

The World Bank has declared that data is the next deprivation to end; they argue that the lack of data causes many of the world's poorest populations to be overlooked when resources are allocated to address their essential needs [1]. Data deprivation is a pressing challenge with as many as 74% of the global and 97% of the Sub-Saharan African population living in countries without adequate vital registration [2]; one third of countries lacking any poverty statistics [1]; and only 17% of the estimated road traffic deaths reported in official figures of low-income countries [3]. Without data to inform national and urban policies, the gap between low- and high-income countries will worsen [4]. Yet, while official statistics are poor, data in the hands of private providers is plentiful, populated by the rapid expansion of mobile phones and social media. Globally, phone penetration reached 67% in 2019 [5], and social media penetration is almost 50% [6]. This provides an opportunity for using crowdsourced data to study major urban and development policies [7–11].

In this project we test the hypothesis of whether privately maintained data can be transformed into a resource to better understand development challenges. Private data has been used to characterize populations [12–17]. Here, we use private data to characterize the environment that affects those populations. The events we are interested in are events reported on social media that affect people's lives and their safety such as road traffic crashes, crime or floods. We focus on road traffic crashes (RTC), the number one cause of death for children and young adults aged 5-29 years, for which lack of adequate data is a recognized and unmet challenge [18]. The objective is to improve RTC data for urban planning that can contribute to addressing the high toll of road deaths, estimated globally at 1.35 million a year [3]. Our case study is Kenya, a country with high road mortality, where the official figures are said to underestimate the number of fatalities by a factor of 4.5 [3].

The United Nations' Sustainable Development Goal (SDG) 3 sets a target to halve road mortality by 2020; progress has been slow, and the target moved to 2030. The Stockholm Declaration by the Third Global Ministerial Conference on Road Safety "Achieving Global Goals 2030" reiterated the call for country investments in road safety–from legislation and regulation, safe urban and transport design, safe modes of transport and vehicles, to modern technologies for crash prevention, trauma care, and urban management. However, resource constraints make it unlikely that countries will be able to do it all. Instead, countries should invest smartly where it matters most. This requires knowing where and when crashes happen, so that resources can be targeted to risky locations and times.

Social media data, with all its biases, can contribute to filling some of the data gaps for urban analysis, planning and management [19]. In this study, we create an algorithm that classifies transport-related tweets into geolocated RTCs for Nairobi. This is done by building on existing literature to test two natural language processing algorithms to identify crash reports [20, 21], developing an improved geoparsing algorithm to extract data on crash time and location [22–28], and groundtruthing the Twitter data. The paper also contributes to a broader literature that uses machine learning methods for road safety analysis [29–31].

This study innovates on three fronts and demonstrates the value of using social media to expand data availability. (1) Geospatial Twitter data analysis usually uses the approximately 1% of tweets with geolocation [32–34]; we improve this by using a machine learning geoparsing algorithm to leverage the 99% of tweets that do not contain a geotag. (2) To our knowledge there are no other studies which physically validate the locational accuracy of tweets in real time. 92% of verified tweets were found to be valid crashes, demonstrating the validity of crowdsourced crash data. (3) The work created an essential resource by generating the first real-time map of RTCs in an African city (Nairobi). We identify 58,065 crash reports and geolocate those with enough information provided in the text (36,538 of them). In a context where there is no systematic georeferenced data on crashes to support policy planning, the process outlined here could be used to capture this data for cities all over the world who need this essential resource.

Overall, the method expands the coverage of road crashes that can be used to analyze road safety and to prioritize policy action around the locations where crashes occur more often. Especially in many country contexts where the only data available for analysis are aggregated statistics on total fatalities in the country, with no detailed breakdown of location or time, crowdsourced data can help act as an additional input that can be used by policymakers in better understanding the situation. By using a clustering algorithm to identify and rank crash locations, we find that the top 14% of crash clusters (97 out of 703) account for half of all crashes. Knowing that a small portion (<1%) of the road network hosts 50% of RTCs in the crowdsourced data, can help reduce an intractable problem into a more manageable one. This analysis shows the potential of using this data to complement road safety diagnostics and potentially guide investments and planning in road safety in Kenya and in other contexts, especially those with similar data deficiencies and with sufficient social media density like India and the Philippines [35].

The approach can be extended to other events reported on social media, whether related to disaster relief, crime, personal safety, urban mobility or road maintenance.

The work on disaster relief and response makes prominent use of geoparsing of tweets [36–43]. Geoparsing of tweets that lack geolocation information could enable more comprehensive crime analytics [44–46]. Improved algorithms can lead to faster and better geolocation of events, helping urban planners and policy makers improve response and better target interventions.

Method

The goals of this analysis are to create the missing road crash data with times and locations, understand how crashes cluster in the city and create recommendations for the spatial prioritization for urban investments in road safety. The technical challenges this study addresses are: i) improve the protocols for geolocation, ii) apply applications of AI to classify tweets reporting crashes and identify their location from multiple geographical references, iii) cluster the crashes geographically and identify areas with many crashes. See Supplemental Information (SI) for detailed methodology. The components are as follows:

1. **Scrape data.** We scrape 874,588 tweets posted by Ma3Route, an existing urban mobility platform with 1.1 million followers, since its inception in May 2012 - July 2020 (see SI for examples of tweets and for a figure of the daily number of tweets across time).

2. **Develop and augment a gazetteer.** We build a gazetteer of landmarks for the five counties that constitute the Nairobi metro area using: OpenStreetMap, Geonames and Google Places. The gazetteer includes landmark, geocoordinates and type of landmark (e.g., school, bus stop). We use consecutive combinations of 2 and 3 words (known as n-grams) and skip-grams of landmarks in the gazetteer, alternate spellings and abbreviations and splitting of landmarks with select punctuation (e.g., slashes, parentheses, commas). We innovate by developing alternate names that exclude the landmark type from the name (e.g., excluding "Hotel" from the name).

3. **Develop a truth dataset.** We develop a truth dataset to train the algorithm. Taking all tweets for July 2017 - July 2018, we restrict tweets to the ones most

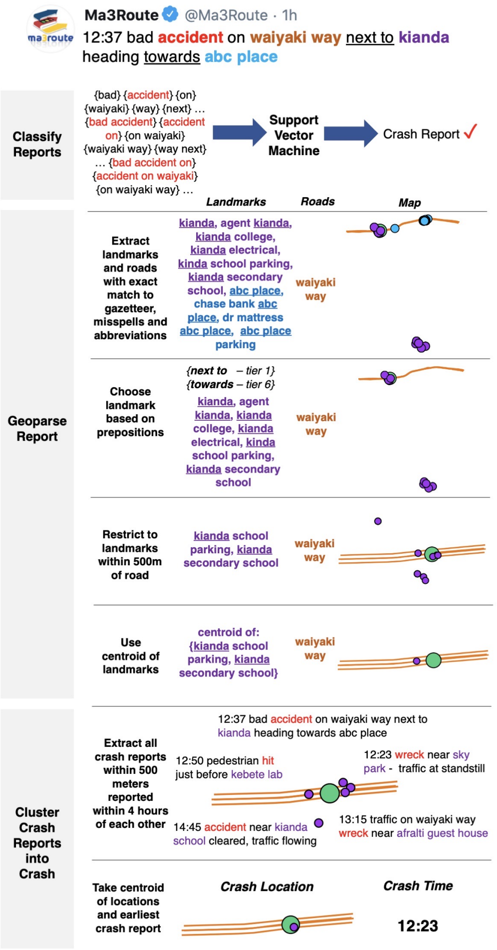
likely related to a crash based on a broad list of words and their variations. Each tweet is manually coded, indicating (1) if the tweet reported a crash and (2) the approximate latitude and longitude of any reported crash whenever enough information is provided. 9,480 tweets were coded, of which 69% (6,602) reported a crash and of these, 63% (4,192) identified an approximate location of the crash. On average, users posted 10 crash reports that could be geolocated to Twitter daily.

4. **Identify RTCs and their location.** We use a three-step process to convert unstructured crowdsourced text into a dataset. The first is to identify relevant reports from hundreds of thousands of reports. The second is to extract needed information from the relevant reports. And the third is to consolidate unique record information from multiple reports of the same event. In Fig 1, we illustrate how the algorithm works to classify and geolocate RTCs. We use the tweet "Bad accident on Waiyaki Way next to Kianda heading towards ABC Place."

   (a) **Classify relevant crowdsourced reports.** We restrict the analysis to tweets that contain one of a broad list of English and Kiswahili road safety keywords such as "accident" or "overturn." This approach follows previous research and allows for misspellings [20]. We use natural language processing to classify and exclude tweets that contain road safety keywords but discuss road safety conditions rather than specific crash events (e.g., "terrible drivers keep causing crashes"). We test two approaches that analyze the combination of words in a tweet – Naive Bayes and support vector machines (SVM). We extract the number of occurrences of words or n-grams in a tweet, removing n-grams that occur in less than 1% and more than 99% of tweets.

   (b) **Geolocate reports.** We extract all landmarks and roads that have an exact match between the gazetteer and the tweet. In Fig 1, "kianda" and "abc way" match several entries in the gazetteer. We extract misspelled matches based on Levenshtein distance varied by length of the n-gram, matches based on the word following a preposition, and matches based on intersections between multiple roads.

   Existing geoparsers extract all possible location references, without



Ma3Route @Ma3Route · 1h
12:37 bad accident on waiyaki way next to kianda heading towards abc place
Classify Reports
{bad} {accident} {on} {waiyaki} {way} {next} … {bad accident} {accident on} {on waiyaki} {waiyaki way} {way next} … {bad accident on} {accident on waiyaki} {on waiyaki way} …
Support Vector Machine
Crash Report ✓
Landmarks
Roads
Map
Geoparse Report
Extract landmarks and roads with exact match to gazetteer, misspells and abbreviations
kianda, agent kianda, kianda college, kianda electrical, kinda school parking, kianda secondary school, abc place, chase bank abc place, dr mattress abc place, abc place parking
waiyaki way
Choose landmark based on prepositions
{next to – tier 1} {towards – tier 6}
kianda, agent kianda, kianda college, kianda electrical, kinda school parking, kianda secondary school
waiyaki way
Restrict to landmarks within 500m of road
kianda school parking, kianda secondary school
waiyaki way
Use centroid of landmarks
centroid of: {kianda school parking, kianda secondary school}
waiyaki way
Cluster Crash Reports into Crash
Extract all crash reports within 500 meters reported within 4 hours of each other
12:37 bad accident on waiyaki way next to kianda heading towards abc place
12:50 pedestrian hit just before kebete lab
12:23 wreck near sky park - traffic at standstill
14:45 accident near kianda school cleared, traffic flowing
13:15 traffic on waiyaki way wreck near afralti guest house
Take centroid of locations and earliest crash report
Crash Location
Crash Time
12:23

Fig 1. Illustration of classification and geolocation algorithm developed for extracting data from crowdsourced information

identifying the unique location that makes the data useful. We resolve two technical challenges to find the location of the crash:

i. When multiple locations are mentioned in the tweets, we use prepositions to sort locations into tiers, based on the probability of a location being correct after a particular preposition. For example, in Fig 1, “next to” is ranked as tier 1 while “toward” is ranked as tier 6, resulting in the correct geolocation for the crash at “kianda” and not “abc place”.

ii. When a name refers to multiple landmarks, we adopt a toponym resolution approach. In Fig 1, more than 6 landmarks across Nairobi have “kianda” in the name. We resolve the toponym in three steps: (1) we search for landmarks that are within 500 m of a road if it is mentioned, (2) we use the centroid of the clustered location if 90% or more of the landmarks are in a 500 m radius, or (3) we rank the landmarks by the probability of being correct using the landmark type in the truth data (see SI for statistics on location type). In the example, we use "Waiyaki Way" to narrow down the landmarks "kianda" in a 500 m radius (from 6 to 3) and then use the centroid as the crash location.

We define a correct geoparse as one located within 500 m of the coordinates in the truth dataset. As a benchmark, we compare our algorithm to the Location Name Extraction tool (LNEx), which was shown to have better accuracy than other geoparsers [40]. As LNEx and other geoparsers are not designed to extract one unique location from text [26, 40, 47], we first judge performance by examining whether any location references are near the true coordinates. Next, we define the crash location as determined by LNEx to be the centroid of all locations it finds in the tweet, and compare this with the unique location identified by our algorithm.

(c) **Identify unique reports.** To avoid over-counting, we develop a clustering algorithm that uses time and location to identify which tweets refer to the same crash. In Fig 1, five tweets report a crash within two hours of each other, referencing different landmarks that are all close together. To develop reasonable parameters for clustering, we manually identify tweets that report the same crash in the truth dataset based on the time, location and crash characteristics. The 4,192 crash reports are clustered into 2,648 unique crashes. For unique crash clusters, 97% of tweets reported landmarks within 500 m and within 4 hours of each other (see additional details in SI for how parameters were chosen).

(d) **Groundtruth.** To ensure that the crowdsourced data is reliable and provides correct information, we conduct a ground-truthing exercise to validate the quality of the data and the performance of the underlying algorithm. We processed tweets in real-time and dispatched a motorcycle delivery service (Sendy) to the site of the crash within minutes. The Sendy

Table 1. Geolocation Algorithm Results

	Any Location Captured by Algorithm Close to True Crash Location		Crash Location Determined by Algorithm Close to True Crash Location	
	Recall	Precision	Recall	Precision
LNEx	0.674	0.686	0.129	0.132
Alg., Raw Gaz	0.695	0.757	0.57	0.761
Alg., Aug Gaz	0.798	0.857	0.656	0.813
Alg., Aug Gaz [Cluster]			0.666	0.777

'N Crashes' refers to the number of correctly identified crashes. 'Raw Gaz' refers to the raw gazetteer (ie, dictionary of landmarks with original names) and 'Aug Gaz' refers to the augmented gazetteer. We use our raw gazetteer as an input into LNEX, which implements its own augmentation process. For LNEx, the crash location is determined by taking the centroid of all locations captured by the algorithm. Locations are considered close if they are within 500 meters of each other.

driver was tasked to verify and report whether the crash actually happened in that location.

Results

The methods laid out here created a georeferenced RTC dataset that was previously unattainable and produced the first real-time map of RTCs in Nairobi. We classify 58,065 tweets as crash-related out of a universe of 874,588 tweets during 2012 - 2020 (Panel A of Fig 2). This is based on the SVM algorithm, which we find performs better than the Naive Bayes algorithm according to the F1 statistic (see Table S4 in the SI). We geolocate 36,538 time-stamped crash tweets from August 2012 to July 2020 and cluster them into 26,791 unique geolocated crashes (panels B and C of Fig 2 show the unique crashes generated by Twitter daily using the algorithm and clustering). In our truth dataset, where we manually coded each crash-related tweet, we found that 63% of tweets contain enough information in order to be geolocated. Assuming the same proportion of tweets contain enough information to be geolocated in the full dataset, we would expect 36,581 tweets with enough location information. This aligns almost perfectly with the number of tweets that the algorithm is able to geolocate.

The ground-truthing exercise confirms the validity of the crowdsourced data. We find that of the 73 crash related tweets physically verified, 92% correctly corresponded to a crash near the estimated location; 32.8% witnessed the crash scene, 57.5% did not see

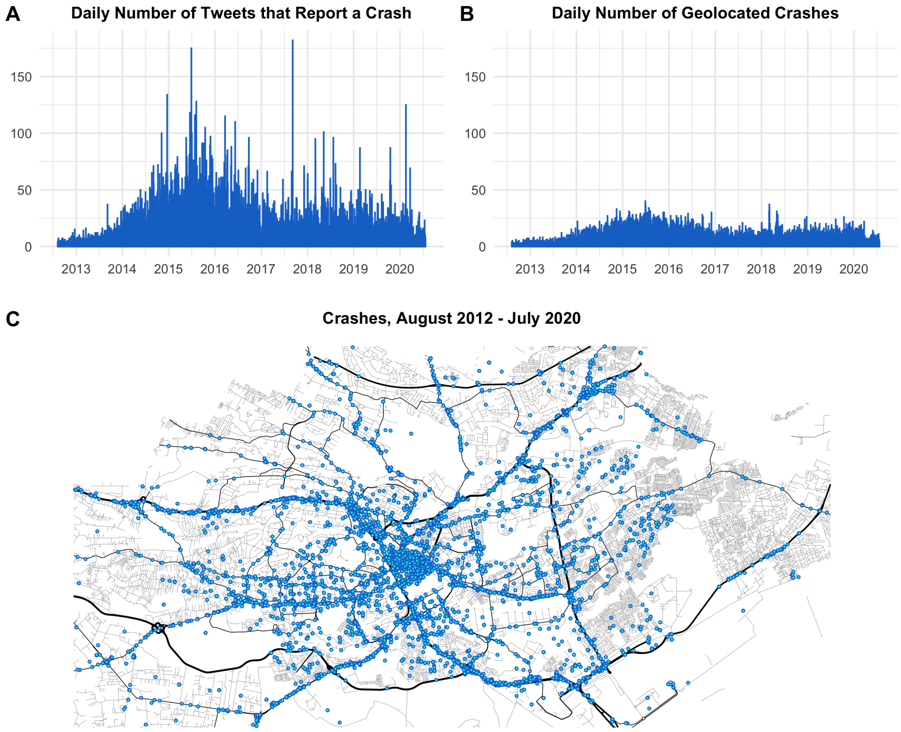


Fig 2. Crowdsourced crash reports from twitter data that the algorithm has geolocated and clustered into unique crashes for the city of Nairobi between 2012 and 2020.

the crash but were told by a bystander that a crash occurred and was recently cleared, and 1.4% reported that the crash did not occur at the specified location but nearby. Furthermore, using our truth dataset to benchmark shows that our algorithm performs significantly better than the current geoparsing standard. Our algorithm's recall rate of 66% is a five-fold improvement in performance compared to the LNEx algorithm (13% recall) in identifying the unique location of a crash (Table 1). This is largely because existing algorithms do not identify a correct location when multiple locations are mentioned. Our algorithm performs 25% better than LNEx even when comparing according to whether any location extracted from the tweet is near the true location.

Analyzing the crash data produced using our algorithm and focusing on the truth dataset within the city limits of Nairobi, we find that all the crashes from July 2017 to July 2018 can be found in 435 clusters, each with a maximum diameter of 300 m. Of these, 67% have two or more crashes and there are 56 clusters with 10 or more crashes. Additionally, 66 crash clusters represent over 50% of all the crashes. When looking at the 7.5 years of crowdsourced data for the city of Nairobi, the number of crash clusters do not grow linearly, implying that the locations where crashes occur and are reported in Twitter are consistent across years. Only 12% of crash locations have only a single crash, and there are 537 crash clusters with 10 or more crashes. We see the concentration of crashes even more when we note that only 9% (151 our of 1541) of crash clusters represent 50% of the crashes reported (Fig 3 shows crash heatmaps when for the truth dataset from July 2017 to July 2018 and for 2012-2020).

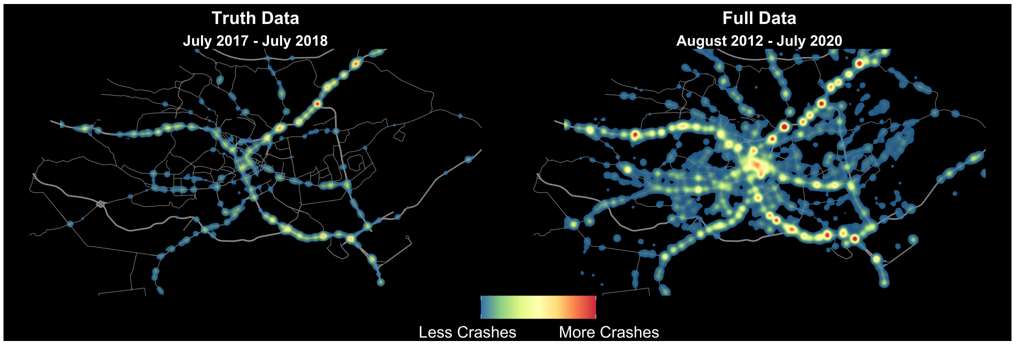


Fig 3. Heatmap of crashes Data in panel a is from July 2017 - July 2018, where we use the manually coded Twitter dataset. Data in panel b is for August 2012 - July 2020.

Discussion

Cities are constantly evolving and understanding urban mobility is critical to designing interventions that leverage the best that cities offer, while managing risks. Many of the policy interventions needed to manage risks face severe data limitations especially in low- and middle-income resource-constrained countries. Closing the data deprivation gap can help avert divergence in socioeconomic conditions between data-poor and -rich countries. By focusing on RTCs–the number one cause of death among young people—we demonstrate that social media could be an inexpensive way to produce non-existent RTC data in resource-poor contexts that can support government analyses of road safety and potentially inform policy. This tool could be especially powerful when combined with investments in building a digital administrative dataset that would provide information on the crashes attended by police. The answer to the seemingly simple question of where and when crashes occur has profound implications for public policy response that can save lives. And while official data deprivation can be an impediment to economic development, data generated by private operators can be transformed and placed in the hands of policy makers as a resource for policy making. By expanding the amount of data, we can generate more input that could help resource-constrained countries prioritize policy action where most needed.

In this example, geolocating for the first time data from crowdsourcing can help to guide infrastructure redesign or enforcement policies to reduce RTCs. The city starts off with an extensive road network of 6,200 km. With limited resources, addressing road safety across the whole network is difficult. By using geolocated data, urban planners and policy makers can narrow down the problem to the areas with the highest number of crashes. This has been proven to work in developed countries where targeting risky locations led to reductions in the concentration of crashes [48]. As shown in the results, crashes reported in Twitter are highly concentrated, with the top 14% of locations spread across 29 km of road having 50% of the crashes reported on Twitter.

There are important limitations to the approach. The data generated is limited by the coverage of the crowdsourced data. Users are more active on social media at particular times, and it is necessary to possess a smartphone and have access to internet to be able to use the service. This can lead to bias in the reports generated via the crowdsourced data. Only 7.5% of tweets are sent between the hours of 9 p.m. and 6 a.m., and as a result only 11% of the crash reports from Twitter are during this time. There could also be geographic bias if there are areas of the city where people with smartphones are more likely to be present or passing by, and therefore more likely to report. Our validation exercise demonstrates the internal validity of the crowdsourced data and the improved algorithm. External validity is more difficult to assess because we do not know what the universe of crashes is. Additionally, we do not know the severity of the crashes reported in Twitter. Therefore, we have no way of knowing if the areas where crashes happen are the most dangerous, which is what policy makers likely would want to target. These caveats should be considered by policy makers when using crowdsourced data to inform policies and targeting. Despite these limitations, the better performing geoparsing algorithm discussed in this paper can begin filling some of the gaps in data in low-capacity and data-scarce settings. And while the crash cluster areas identified by the algorithm may not be the most dangerous or may not represent all crash areas; nevertheless, they highlight problem areas. Even less severe crashes have important economic consequences in terms of property damage and lost time and productivity due to the traffic generated (which is one of the reasons the crash is likely reported on Twitter). Therefore, targeting areas where we are seeing high numbers of crashes consistently, could still lead to benefits. Especially in settings where there are limited or non-existent administrative records and, therefore, lack of any geolocated data, this tool can produce information in real-time for one of the most pressing challenges in developing countries.

Furthermore, by improving the tools that generate time-stamped geolocated data and statistics from crowdsourcing on different "events" that are reported on social media, we can hope to expand data availability across contexts and across issues that affect people's lives. Real-time traffic applications like RIDLR in India can be used to expand data on road safety. These improved tools can also help geolocate victims during a natural disaster or alert disaster management teams to the location of unsafe buildings or areas needing immediate attention. They can support law-enforcement or communities to locate and respond to crimes, cases of violence against women, or police violence. Improved identification of time and location of events can help to automate and accelerate policy response across a wide set of issues, potentially leading to better policy outcomes.

Supporting information

S1 File. Supplementary Information

Acknowledgments

We thank Robert Tenorio and Amy Dolinger for their field coordination and research support. We are also grateful to Andrew Muriithi, Purity Kimuru, Rodgers Avuya, Salome Omondi and Pheliciah Mwachofi for their field support. We appreciate comments from anonymous reviewers and participants at the ACM COMPASS Conference and the Netmob Conference. The research has been funded with UK aid from the UK government through the ieConnect for Impact program and the World Bank's Knowledge for Change program.

References

1. Serajuddin U, Uematsu H, Wieser C, Yoshida N, Dabalen A. Data deprivation: Another deprivation to end. The World Bank. 2015;.

2. Notzon F, Nichols EK. Global Program for Civil Registration and Vital Statistics (CRVS) Improvement; 2015.

3. WHO. Global status report on road safety 2018. World Health Organization. 2018;.

4. IEAG. A World that Counts–Mobilising the Data Revolution for Sustainable Development. Independent Expert Advisory Group on a Data Revolution for Sustainable Development. 2014;.

5. GSMA Intelligence. The Mobile Economy 2020. London: GSM Association. 2020;.

6. Kemp S. Digital 2020: Global Digital Overview. Retrieved from Datareportal: https://datareportalcom/reports/digital-2020-global-digital-overview. 2020;.

7. Batty M. Big data, smart cities and city planning. Dialogues in human geography. 2013;3(3):274–279.

8. Miller G. Social scientists wade into the tweet stream. Science. 2011;333(6051):1814–1815.

9. Kitchin R. The real-time city? Big data and smart urbanism. GeoJournal. 2014;79(1):1–14.

10. Einav L, Levin J. Economics in the age of big data. Science. 2014;346(6210).

11. Hao J, Zhu J, Zhong R. The rise of big data on urban studies and planning practices in China: Review and open research issues. Journal of Urban Management. 2015;4(2):92–124.

12. Blumenstock J, Cadamuro G, On R. Predicting poverty and wealth from mobile phone metadata. Science. 2015;350(6264):1073–1076.

13. Kosinski M, Stillwell D, Graepel T. Private traits and attributes are predictable from digital records of human behavior. Proceedings of the national academy of sciences. 2013;110(15):5802–5805.

14. Resch B, Summa A, Zeile P, Strube M. Citizen-Centric Urban Planning through Extracting Emotion Information from Twitter in an Interdisciplinary Space-Time-Linguistics Algorithm. Urban Planning. 2016;1(2):114–127. doi:https://doi.org/10.17645/up.v1i2.617.

15. Jaidka K, Giorgi S, Schwartz HA, Kern ML, Ungar LH, Eichstaedt JC. Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods. Proceedings of the National Academy of Sciences. 2020;117(19):10165–10171.

16. Steiger E, Westerholt R, Resch B, Zipf A. Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. Computers, Environment and Urban Systems. 2015;54:255 – 265. doi:https://doi.org/10.1016/j.compenvurbsys.2015.09.007.

17. Wang Q, Phillips NE, Small ML, Sampson RJ. Urban mobility and neighborhood isolation in America's 50 largest cities. Proceedings of the National Academy of Sciences. 2018;115(30):7735–7740.

18. WHO. Data systems: A road safety manual for decision-makers and practitioners. World Health Organization. 2010;.

19. Williams S. Data Action: Using Data for Public Good. Cambridge, MA: MIT Press; 2020.

20. Gu Y, Qian ZS, Chen F. From Twitter to detector: Real-time traffic incident detection using social media data. Transportation Research Part C: Emerging Technologies. 2016;67:321 – 342. doi:https://doi.org/10.1016/j.trc.2016.02.011.

21. Zhang Z, He Q, Gao J, Ni M. A deep learning approach for detecting traffic accidents from social media data. Transportation research part C: emerging technologies. 2018;86:580–596.

22. Finkel JR, Grenager T, Manning C. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05); 2005.

23. Bender O, Och FJ, Ney H. Maximum Entropy Models for Named Entity Recognition. USA: Association for Computational Linguistics; 2003.Available from: https://doi.org/10.3115/1119176.1119196.

24. Bhargava R, Zuckerman E, Beck L. CLIFF-CLAVIN: Determining Geographic Focus for News Articles; 2014. NewsKDD: Data Science for News Publishing.

25. Ritter A, Clark S, Mausam, Etzioni O. Named Entity Recognition in Tweets: An Experimental Study. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing; 2011.

26. Gelernter J, Balaji S. An algorithm for local geoparsing of microtext. GeoInformatica. 2013;17(4):635–667. doi:10.1007/s10707-012-0173-8.

27. Malmasi S, Dras M. Location Mention Detection in Tweets and Microblogs. In: Hasida K, Purwarianti A, editors. Computational Linguistics. Singapore: Springer; 2016. p. 123–134.

28. Middleton SE, Middleton L, Modafferi S. Real-Time Crisis Mapping of Natural Disasters Using Social Media. IEEE Intelligent Systems. 2014;29(2):9–17. doi:10.1109/MIS.2013.126.

29. Zeng Q, Huang H, Pei X, Wong S. Modeling nonlinear relationship between crash frequency by severity and contributing factors by neural networks. Analytic methods in accident research. 2016;10:12–25.

30. Zeng Q, Huang H, Pei X, Wong S, Gao M. Rule extraction from an optimized neural network for traffic crash frequency modeling. Accident Analysis & Prevention. 2016;97:87–95.

31. Wahab L, Jiang H. A comparative study on machine learning based algorithms for prediction of motorcycle crash severity. PLOS ONE. 2019;14(4):1–17. doi:10.1371/journal.pone.0214966.

32. Salas A, Georgakis P, Petalas Y. Incident detection using data from social media. In: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC); 2017. p. 751–755.

33. Mai E, Hranac R. Twitter Interactions as a Data Source for Transportation Incidents. In: Transportation Research Board 2013 Annual Meeting; 2013.

34. Sloan L, Morgan J. Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. PloS one. 2015;10(11):e0142209.

35. Gatica-Perez D, Santani D, Isaac-Biel J, Phan TT. Social Multimedia, Diversity, and Global South Cities: A Double Blind Side. In: Proceedings of the 1st International Workshop on Fairness, Accountability, and Transparency in MultiMedia. ACM; 2019. p. 4–10.

36. Meier P. Digital humanitarians: How big data is changing the face of humanitarian response. Routledge; 2015.

37. Dhavase N, Bagade AM. Location identification for crime disaster events by geoparsing Twitter. In: International Conference for Convergence for Technology-2014; 2014. p. 1–3.

38. Aggarwal CC, Zhai CX. Mining Text Data. Boston, MA: Springer; 2012.

39. Yin J, Karimi S, Lampert A, Cameron MA, Robinson B, Power R. Using Social Media to Enhance Emergency Situation Awareness: Extended Abstract. In: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI); 2015.

40. Al-Olimat H, Thirunarayan K, Shalin V, Sheth A. Location Name Extraction from Targeted Text Streams using Gazetteer-based Statistical Language Models. In: Proceedings of the 27th International Conference on Computational Linguistics; 2018.

41. Premamayudu B, Subbarao P, Koduganti VR. Identification of Natural Disaster Affected Area Precise Location Based on Tweets. International Journal of Innovative Technology and Exploring Engineering. 2019;8(6).

42. Sangameswar MV, Nagabhushana Rao M, Satyanarayana S. An algorithm for identification of natural disaster affected area. Journal of Big Data. 2017;4(39).

43. de Bruijn JA, de Moel H, Jongman B, de Ruiter MC, Wagemaker J, Aerts J. A global database of historic and real-time flood events based on social media. Scientific Data. 2019;6(311).

44. Ristea A, Boni MA, Resch B, Gerber MS, Leitner M. Spatial crime distribution and prediction for sporting events using social media. International Journal of Geographical Information Science. 2020;0(0):1–32. doi:10.1080/13658816.2020.1719495.

45. Gerber MS. Predicting crime using Twitter and kernel density estimation. Decision Support Systems. 2014;61:115 – 125. doi:https://doi.org/10.1016/j.dss.2014.02.003.

46. Yang D, Heaney T, Tonon A, Wang L, Cudré-Mauroux P. CrimeTelescope: crime hotspot prediction based on urban and social media data fusion. World Wide Web. 2018;21(5):1323–1347.

47. Karimzadeh M, Pezanowski S, MacEachren AM, Wallgr[U+FFFD]n JO. GeoTxt: A scalable geoparsing system for unstructured text geolocation. Transactions in GIS. 2019;23(1):118–136. doi:10.1111/tgis.12510.

48. Austroads. Guide to roadsafety part 8: Treatment of crash locations; 2015.