

URBAN PLANNING? I SPY CRASHES



Applying machine learning and geolocation techniques to social media data (Twitter) to develop a resource for urban planning¹



"Traffic Jams" by anique, used under CC BY-NC-ND 2.0.

I Spy is a book series that has entertained generations of children searching for clues in busy pictures. Those children, now young adults, die at alarming rates on the roads of our cities. Road traffic crashes (RTC) are the number one cause of death for children and young adults ages 5-29 years.² As we seek to halve mortality on the roads by 2030, in line with United Nations Sustainable Development Goal (SDG) 3, the challenge we face is what the World Bank has declared to be the next deprivation to end: data.

On average, only 17% of road traffic deaths are estimated to be reported in the official figures of low-income countries, and the available reports are often on paper and difficult to analyze.³ At the same time, unprecedented access to social media has democratized reporting and greatly expanded the data. The *I Spy* puzzle we seek to solve is to find and geolocate RTCs. The study summarized here tests whether we can transform an openly available data set (Twitter) into a resource for urban planning and development. The objective is to improve RTC data to help address the high toll of road deaths, which are estimated globally at 1.35 million a year.⁴

Our case study is Kenya, a country with high road mortality and where, according to the World Health Organization, the official figures underestimate the number of fatalities by a factor of 4.5 and no official public data on where RTCs happen exist.⁵ 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, and 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 and across their whole road network. 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.

This brief is based on the World Bank Policy Research Working Paper, "Applying Machine Learning and Geolocation Techniques to Social Media Data (Twitter) to Develop a Resource for Urban Planning," by Sveta Milusheva, Robert Marty, Guadalupe Bedoya, Sarah Williams, Elizabeth Resor, and Arianna Legovini. Policy Research Working Paper, No. 9488. World Bank, Washington, DC; 2020.

Applying Machine Learning and Geolocation Techniques to Social Media Data

The puzzle we set to solve is to scrape massive amounts of tweets to find those that report crashes and then timestamp and geolocate those reports onto a live map. At the start of this work, we realized that there was a missed opportunity with thousands of crashes reported through social media, but less than 1% of the tweets included GPS information, so we lacked a critical piece of needed information. Instead, many people referred to the location of crashes using geographic references (for example, the name of a nearby landmark). We tackle this challenge by developing new machine learning algorithms to classify transport-related tweets into geolocated RTCs, applying them to 874,588 tweets in Nairobi. The algorithms, which are illustrated in figure 1, are basically a set of rules to classify information. The better the rules are, the better the classification is. Training an algorithm to improve its efficiency requires a truth data set to ascertain whether the information is correctly classified and how the rules should be improved. In this case, we manually built a 13-month double-coded crash location truth data set from the Twitter data and used it to train the algorithms. Further, we validated the information on the ground minutes after the crash tweets came in, by dispatching a motorcycle delivery company to the site to verify each crash in a sample of crashes.

As the numbers show, the exercise holds great promise for supplementing the information available to urban planners. We identified 52,228 crash reports, geolocated 32,991 that included enough location information in the tweet, and clustered them into 22,872 unique crashes (figure 2). The dispatch service physically validated the accuracy of the tweets and their geoparsing onsite and in real time in 92% of the cases, giving us confidence in the results. The approach expands the coverage of road crashes that can be used to analyze road safety and prioritize policy actions where crashes occur more often. By using a clustering algorithm, we find that the top 15% of crash clusters (66 of 435) account for half of all crashes. This information is important for prioritizing road safety investments. As it turns out, these 66 clusters represent less than 1% of the road network (figure 3 shows crash heatmaps for the truth data set from July 2017 to July 2018 and for 2012–20). We hope that moving from addressing road safety on the entire road network to focusing on a relatively small portion (<1%) of that network where crashes are concentrated reduces an intractable problem into a more manageable one and can help cities reach their 2030 SDG target.

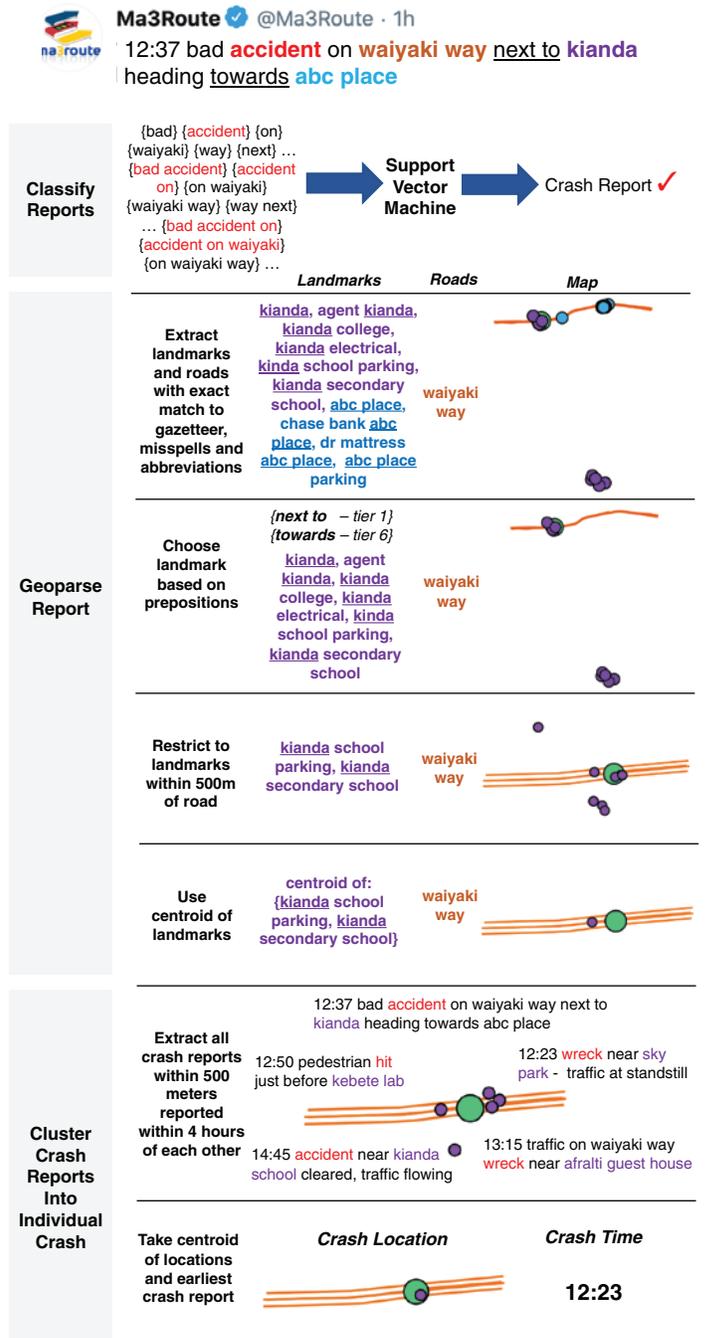


Figure 1. Illustration of the classification and geolocation algorithm developed for extracting data from crowdsourced information.

The wider significance of this work is that we can use the technology to generate time-stamped geolocated data and statistics on different “events” that are reported on social media, and we hope to expand data availability across contexts and issues that affect people’s lives. These improved tools can 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,

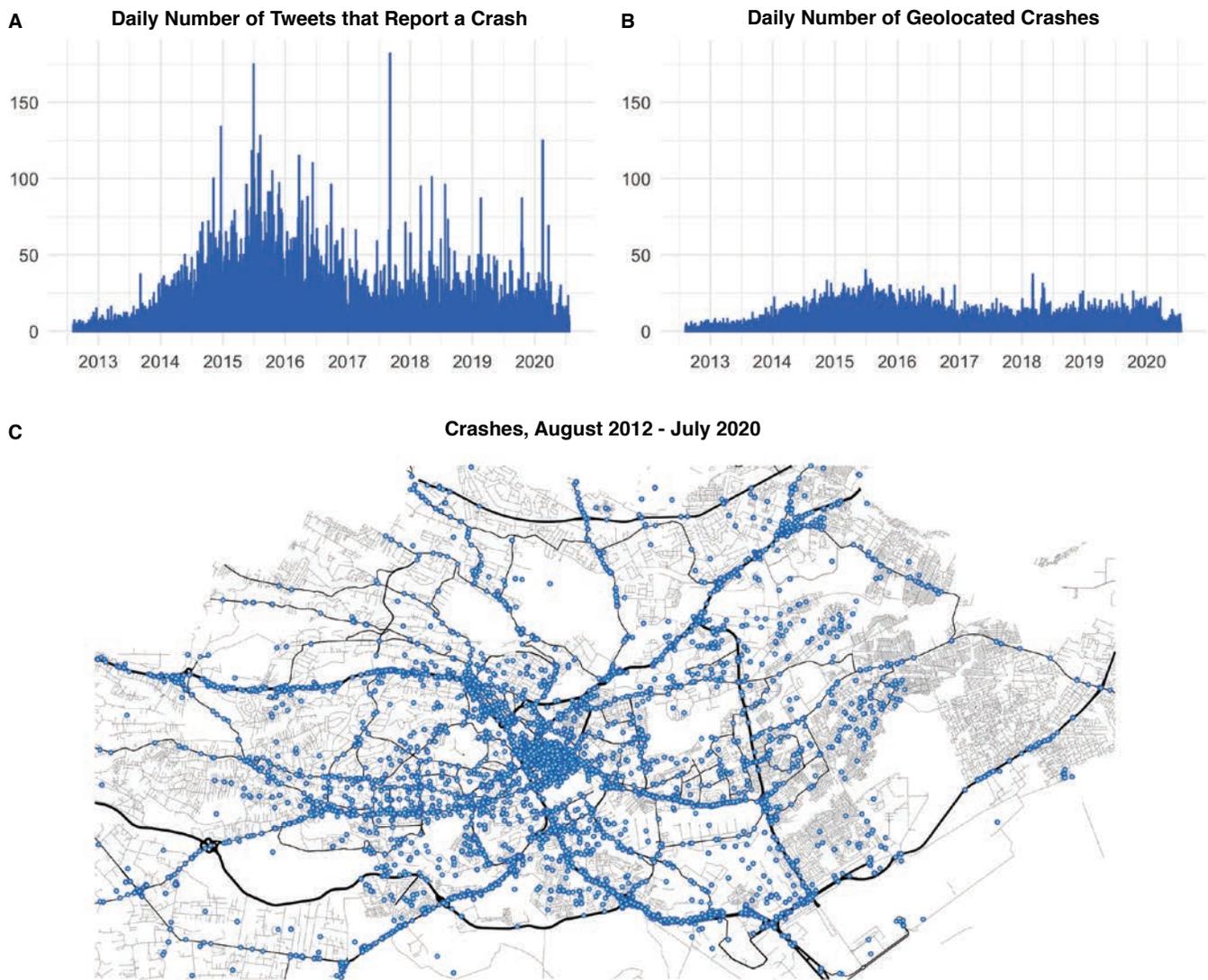


Figure 2. Crowdsourced crash reports from Twitter data, which the algorithm has geolocated and clustered into unique crashes in Nairobi between 2012 and 2020.

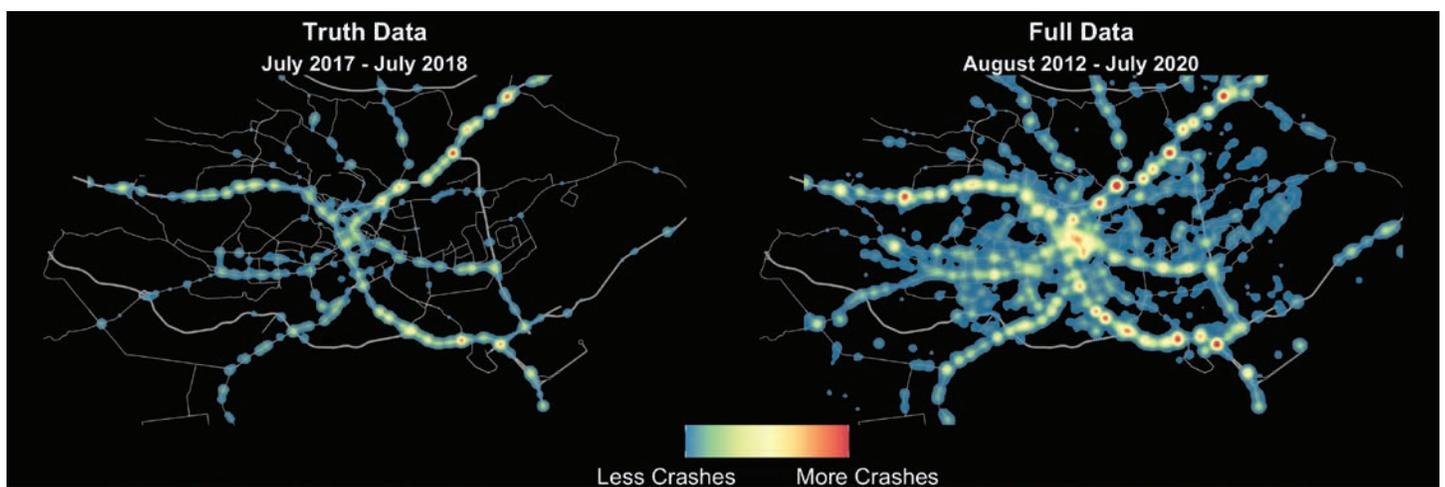


Figure 3. Heatmap of crashes. The data in the left panel are from July 2017 to July 2018, where we use the manually coded Twitter data set. The data in the right panel are for August 2012 to July 2020.



"Social Media apps" by Jason Howie, used under CC BY 2.0.

or police violence. Improved identification of the time and location of events can contribute to automating and accelerating policy response across a wide set of issues, potentially leading to better policy outcomes.

While potentially useful for policy, crowdsourced data have limitations, such as bias in the time and location of reports.

Therefore, it is best to leverage the data by using them to complement traditional data, where and when available, and help authorities improve their administrative data. We are currently working with the National Police Service in Kenya to digitize seven years of police paper records from the 14 police stations in Nairobi, to create a digital administrative data set of georeferenced police records that include crashes with casualties. This collaboration, and the upcoming system of electronic recording on which the National Police Service is working, will place Kenya as a leader in the region in using data and analytics to monitor and design better policies for road safety.

ENDNOTES

¹ This research was funded by the UK government through DIME's ieConnect for Impact program and the World Bank's Knowledge for Change Program.

² WHO (World Health Organization), *Data Systems: A Road Safety Manual for Decision-Makers and Practitioners*, WHO, Geneva, 2010.

³ Calculated using data from: WHO (World Health Organization), *Global Status Report on Road Safety 2018*, WHO, Geneva 2018.

⁴ WHO (World Health Organization), *Global Status Report on Road Safety 2018*, WHO, Geneva 2018.

⁵ WHO (World Health Organization), *Global Status Report on Road Safety 2018*, WHO, Geneva, 2018.

For more information email dimettransport@worldbank.org or visit www.worldbank.org/en/research/dime/brief/transport

IE CONNECT FOR IMPACT

ieConnect has over 30 ongoing impact evaluations across 19 different countries. The IEs focus on urban mobility, transport corridors, road safety, and rural roads sectors with thematic emphasis on gender, female economic empowerment, and fragile situations. From the ieConnect program we will learn how to improve the availability and quality of data that can be used for measuring the impact of transport projects and generate evidence that can be used to improve decision making for transport investments in the long-term. The ieConnect for Impact program is a collaboration between the World Bank's DIME group and the Transport Global Practice. This program has been funded with UK aid from the UK government.