International Cartographic Association
Association Cartographique Internationale

Social Media for Sensing: Do Tweets Represent Events at Geo-Tagged Locations?

Morteza Karimzadeh [a, *]

[a] *Postdoctoral Researcher, Purdue Visualization and Analytics Center (PURVAC), School of Electrical and Computer Engineering, Purdue University, karimzadeh@purdue.edu*

* Corresponding author

Keywords: social media, geoparsing, flow maps, geographic focus, geographic information retrieval, geovisualization

Abstract:

It is difficult to quantify how – and to what extent – the public engages with events in other countries. Twitter users all over the globe post more than 500 million tweets every day[1]. They also discuss places in their tweets. Therefore, Twitter provides a lens through which geographic research can investigate public discourse as it relates to place. Further, many research studies use geo-tagged posts on Twitter (and social media in general) to sense the society in particular locations (according to geo-tags) for various purposes that may need a "local sense" such as sentiment analysis, situational awareness for crisis response, election prediction, or targeted advertising. However, it is unclear to what extent the online discourse by users are about local events versus events in other locations or countries.

In this pilot study, we visualize and characterize relations between places mentioned in Twitter posts and places where users live to identify whether Twitter users in different countries engage more with domestic or international (or transnational) events. We also visualize the extent to which places in other countries are being discussed through online platforms/social media. The results have implications for the design of algorithms in geographic information science attempting to automatically geolocate places mentioned in tweets for use in sentiment or spatial analysis, situational awareness, and advertisement. Additionally, most place names are ambiguous and refer to more than one location. For example, London can refer to London, Texas or London in England. Our analysis gauge whether Twitter users' profile locations can be used to disambiguate places that are mentioned in their tweets.

We used the GeoCorpora dataset (Wallgrün, 2018) to conduct our study using a gold-standard dataset, i.e. a manually geo-annotated dataset in which place names were manually resolved to geolocations, since according to our analysis, the accuracy of automated text geoparsing methods (i.e. place name detection, disambiguation and geolocation) usually does not exceed 80% for tweets. Specifically, we used the Twitter Streaming Application Programming Interface (API) to collect approximately one billion tweets from January 1, 2014 to January 1, 2016. To generate this corpus, crisis-related keywords were used to filter for tweets that were more likely to include place names. From this collection, a random sample of 6000 tweets was drawn by querying for 700 tweets for each of the following 12 terms: earthquake, Ebola, fire, flood, flu, malaria, measles, protest, rebels, riot, tornado, violence. These keywords represent three types of events (a) natural disasters, (b) infectious disease, and (c) human threats/violence, for which one expects to observe a higher number of place-focused tweets (as opposed to an average sample, such as tweets about celebrities). In order to produce a more spatially and temporally distributed sample of tweets, a stratified random selection of tweets was performed, with temporal stratification per month. We ensured only tweets that are primarily in English were selected and that no two selected tweets are the same or very similar (e.g. one is a re-tweet of another). We replaced similar (re)tweets and deleted tweets and, to keep the distribution across keywords, added more tweets from the database. The resulting 6,711 tweets were manually geo-annotated using an annotation software platform, meaning that place names in a tweeter's (user's) profile location and place names in the tweet's text (the content of the tweets) were given geographic locations.

We processed the resulting 2185 tweets that had place names in their textual content to establish a link between every place listed in the profile location (by the user generating the tweet) and every place mentioned in the tweet text. A list of connection triples in the form of [source, target, count of connecting tweets] was formed for geovisualization. We created an interactive flow map (link in footnote[2]) to show these triplets.

[1] http://www.internetlivestats.com/twitter-statistics/
[2] https://www.geovista.psu.edu/GeoCorpora/LinksMap/

29th International Cartographic Conference (ICC 2019), 15–20 July 2019, Tokyo, Japan | https://doi.org/10.5194/ica-abs-1-160-2019

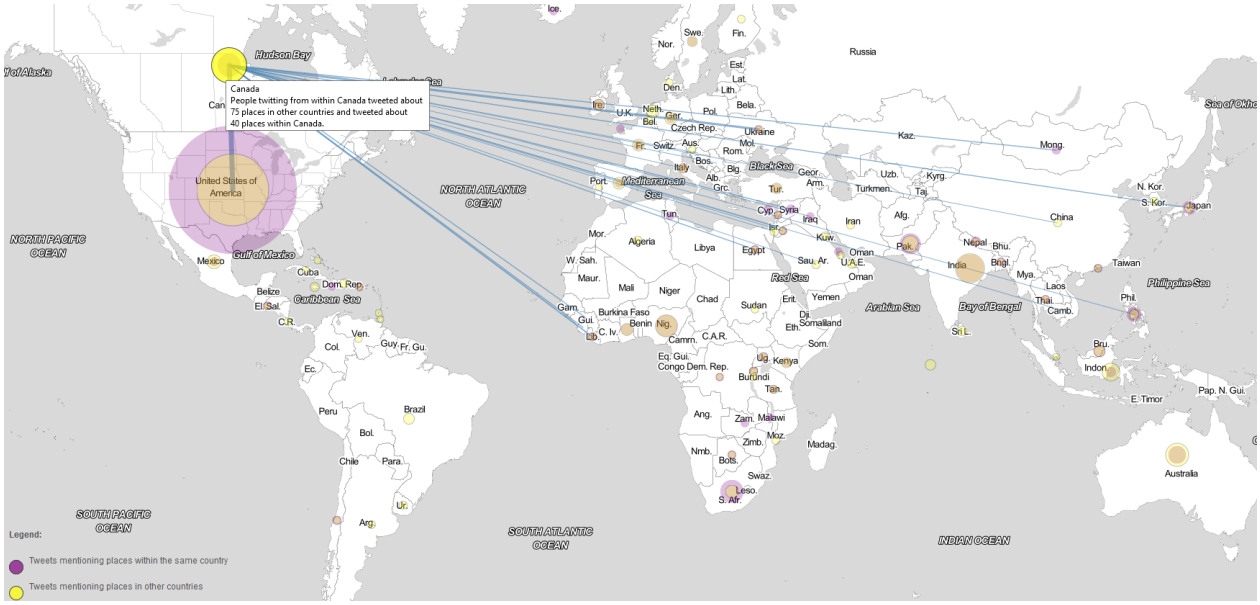

Figure 1. A static view of our interactive flow map[3] that shows the proportion of tweets discussing places within a country and the ones that mention places in other countries, visually linking the other countries that are mentioned from Twitter users in each country. Yellow circle symbols show the proportion of tweets mentioning places within the same country and purple circles indicate the proportion of tweets discussing places located in other countries. Holding the mouse over each country's yellow circle draws links to other countries that are mentioned by users located inside the selected country. Link thickness indicates the proportion of tweets mentioning the linked country in the source country.

Our visual analysis of the ground truth dataset shows that tweets indeed do not focus merely on local events, as many users discuss events in other countries in their tweets, exceeding the number of local tweets for many countries (Figure 1). The extent of local or global focus changes based on the country of origin. For instance, tweets generated in the United States have a more local (within-country) focus, as opposed to tweets generated in Canada, in which users talk about events in other countries more often than they do so about local events. This variation is important for researchers and practitioners who use tweets as sensors on the ground, since – as evident in our results – many of such tweets may not reflect the situation in the immediate vicinity of the originating location. Automatic or human in the loop solutions should be carefully used to filter for posts that represent or characterize the specific location of interest.

We are in the process of expanding our methodology to geo-tagged tweets collected in the Continental United States from the beginning of 2016 to the end of 2018 and use validated automatic geoparsing methods to extract place names in tweets' text. In future research, we will examine finer-scale relationships between tweets originating from different census blocks or tracks and the places that the content of such tweets are about (i.e. the places that tweets mention in text). We will examine different socioeconomic variables as recorded in census as potential predictors for the variability of tweet content locality. Such characterization helps us understand the geographic literacy of the society and its engagement with various kinds of local, regional, or global events, which has important implications for education, policymaking, and any type of research that leverages geo-tagged social media as its data source.

References:

Wallgrün, J. O., Karimzadeh, M., MacEachren, A. M., & Pezanowski, S. (2018). GeoCorpora: building a corpus to test and train microblog geoparsers. *International Journal of Geographical Information Science*, pp. 1–29. https://doi.org/10.1080/13658816.2017.1

[3] https://www.geovista.psu.edu/GeoCorpora/LinksMap/