Data Census of a Geographically-Bounded Tweet Set to Enhance Common Operational Picture Tools

Nathan J. Elrod
University of Cincinnati
elrodnj@ucmail.uc.edu

Howard Hall
University of Cincinnati
hallho@mail.uc.edu

Pranav Mahajan
University of Cincinnati
mahajapp@mail.uc.edu

Rob Grace
Texas Tech University
Rob.Grace@ttu.edu

Jess Kropczynski
University of Cincinnati
jess.kropczynski@uc.edu

Abstract

Location information is of particular importance to crisis informatics. The Twitter API provides several methods to assess a rough location and/or the specific latitude and longitude in which a post originated. This paper offers a comparison of location information provided by Twitter's four geolocation methods. The study aggregates one month of data from the greater Cincinnati, Ohio metropolitan area and assesses the relative contribution that each method can make to common operational picture tools used by crisis informatics researchers. Results show that of 49,744 Tweets, 4% contained geotags, 85.2% contained a location in the users' profile, and 3.5% contained no apparent location data, but were gathered using the bounding box method and would not have been identified using traditional methods of gathering data using geotagged Tweets or user profile information alone. We reflect on these results in light of design implications for common operational picture tools (COPs).

1. Introduction

First responders and 911 dispatchers in Public-Safety Answering Points (PSAPs) serving municipal jurisdictions rely on aging information infrastructures to assess and respond to emergencies [1, 2]. Whereas most U.S. cities rely on citizens' 911 calls and reports from on-scene responders to gather situational awareness about incidents, industry and select government agencies now use multi-channel methods that include social media analytics to monitor citizen-reported information, identify emerging trends, and inform timely decision-making [3–5]. Research in the field of crisis informatics is now positioned to improve information infrastructures in emergency response by addressing first responders' needs for actionable information with social media analytics, such as common operational picture (COP) tools, that can collect and filter social media data to visualize actionable information during emergencies [6–8].

Recent crisis informatics research has emphasized responders' unique needs for actionable information [9]. For first responders and 911 dispatchers, these needs include fine-grained location information associated with social media posts to locate incidents within hyperlocal, municipal-level jurisdictions [4, 10]. This research, in turn, highlights the need to collect precise location information from social media platforms such as Twitter to design COP tools that can provide first responders and dispatchers with actionable information for emergency response.

However, the extent to which adequate location information can be collected for municipal-sized geographic areas using the Twitter API remains largely unknown. While research has examined the availability of location metadata in tweets posted across large geographic areas (e.g., nationally, globally) [11, 12] and the availability of location information present in tweet content [13], our knowledge of the relative amounts of location metadata - including geotags, place tags, and profile locations - that can be collected for hyperlocal contexts remains incomplete. As a result, we cannot connect crisis informatics findings on actionable information (e.g., the granularity of location information required by first responders) with motivations and requirements for the design of COP tools suitable for social media monitoring in municipal jurisdictions.

To address this gap, this paper performs a census of location metadata collected from the Twitter API during two concurrent high-risk events, the outbreak of COVID-19 and Black Lives Matter protests in Cincinnati, Ohio, and introduces PIVOT, a novel COP tool for municipal emergency response.

2. Background

2.1. Social media for crisis response

Seminal works investigating patterns in the use of social media during crisis situations have offered

URI: https://hdl.handle.net/10125/70888
978-0-9981331-4-0

key insights that have shaped crisis informatics research. Early work examined distributed networks in information sharing in Retweet networks [14, 15], while Olteanu et al. [16] recognized the need to understand themes that emerge on social media around various crises and created CrisisLex as a repository of social media emerging around these crises. This type of understanding of social media users' behaviors has contributed to the creation of situational awareness tools for crisis response. A first step in this type of development is an understanding of the information requirements necessary to support insights.

Typical information behaviors have been observed among directly and indirectly impacted social media users during crises. It is widely recognized that citizens who are directly impacted during natural and man-made crises post different types of information and engage differently with other social media users than citizens who are spatio-temporally removed or indirectly impacted [17–19].

When a crisis erupts, directly impacted citizens post information that can provide early warning of crisis events [20, 21]. Later, as crises develop, directly impacted citizens use social media to provide situational information [22], to include descriptions of environmental conditions (e.g. flood levels), harm to people (e.g. injuries), status of critical infrastructure and resources, and current on-the-ground activities among affected people and response personnel [23, 24]. Citizens also routinely use social media to call for assistance, while other social media users respond by offering assistance and needed resources [25, 26]. Indirectly impacted citizens typically use social media during a crisis to express sympathy with those affected and inquire about the condition of potentially affected friends and family [27], while those directly affected update others on their health and personal condition [28]. The first challenge to presenting social media data within crisis informatics tools is to match data with relevant location information.

Common operational picture (COP) software offers a single display of operational information about an area or situation to facilitate shared situational awareness among users [29]. Commonly used by the military [30], software companies such as ESRIs ArcGIS for Emergency Management has helped to promote the mainstream use of these tools in Emergency Operation Centers (EOCs) as a web-based incident response management system [31]. The types of information and logistics that these tools display is evolving at a rapid pace, however, despite literature indicating the potential benefits of social media analytics into COPs [26, 32] there is little documentation indicating the use of such analytics in practice. With an interest in promoting the use of social media data into COPs, we investigate available mechanisms to use location data available in the Twitter API to enhance geographic visualizations.

2.2. Location information in social media data

In the current free version of the Twitter API, there are a number of locations within a Twitter post where geographic information may explicitly reside. Twitter stratifies this data to best ascertain a tweet's geographic origination when an API query attempts to filter on user location. We use this geospatial metadata and the Twitter API query tools as the basis of our comparative analysis, and do not focus on other methods of deducing location from tweet content such as inferential/probabilistic/network models, natural language processing, or gazetteers. Each of these categories of information are described in the following subsections.

2.2.1. Geotags - Coordinates According to Twitter, only 1-2% of all Twitter posts are geotagged [33]; however, prior benchmarking has indicated a range between 1.5-3.2% [12]. A geotag is metadata with specific latitude and longitude coordinates concerning the physical origin of the post that a user may voluntarily include. The Twitter API uses distinct tags to reference this type of information. The 'coordinates' tag in the JSON file (hereafter referred to as a Coordinate) represents geographic data as directly reported by the user or client application [34], and is the most faithful documentation of the location of a Tweet's origin. Not unsurprisingly, this category is also the most rare. All Tweets that are geotagged also contain a reference to a 'place' object, which are described in the following section.

Due to the sparsity of available geo-located social media data [35], alternative approaches to identifying and collecting social media have been employed that may compromise accuracy of location for quantity of data.

2.2.2. Geotags - Places 'Place' tags (hereafter referred to as Places), by contrast, are "specific, named locations with corresponding geo coordinates" [36], each with a corresponding subset of embedded JSON containing (among other information) the place name, a unique identifier, and a place type. The place type also alludes to the geographic size of the Place object, scaling from (in roughly ascending order of area) points of interest (POIs), neighborhoods, cities, admins, and

countries. Since 2010 [37], Twitter has sourced Place POIs from third-party sources such as Foursquare and Yelp [38, 39].

Unlike a Coordinate, which directly affirms a user is tweeting from the GPS location provided, Place data only indicates that a Tweet is *about* a specific place, but not necessarily being issued from that location [34, 36]. As mentioned, all Tweets with a non-null GPS Coordinate value will also contain a Place reference, but not all Tweets with associated Places will be geotagged with a Coordinate.

2.2.3. User Profile Location The User Profile Location is an arbitrarily user-defined text field. As Twitter defines it, the User Profile Location is a "user-defined location for [the] account's profile. Not necessarily a location, nor machine-parseable. This field will occasionally be fuzzily interpreted by the Search service [40]". Previous research has used profile location data as an additional piece of information to affirm hypotheses when layering multiple types of location information [41], but on it's own the User Profile Location is rarely used as a definitive source of location data in crisis informatics research.

2.2.4. Query Search Radius - A Novel Approach The final source of Tweet location is one that is not included in the metadata of a Tweet, but instead implies location through the means by which geo-located Tweets are gathered. The Twitter API offers robust filtering features, even when using only the standard query operators (via a free API key) [42]. Through the API, a query can be made drawing a radius of a distance around a particular GPS point. A Tweet that is returned by this method may or may not have explicit geographic metadata; if it does not, we may infer that it was collected by some mechanism known to Twitter but unknown to the general public. We explore these tweets more closely in section 5.5 as a potential method to increase regional social media data aggregation for crisis informatics research.

3. Research Questions

Based on research indicating the importance of location data in assisting to identify actionable information [10], we investigate location metadata provided by the Twitter API through the following research questions:

RQ1: Using the free Twitter API, what amounts of social media location information can be collected for hyperlocal geographic areas?

RQ2: What data aggregation techniques can improve the amount and quality of hyperlocal location information available to first responders using COP tools?

The following section describes how the query string radius approach was implemented for the case study in this work.

4. Methods

4.1. Sample Dataset: June 2020 Data Aggregation Using the Query String Method

In our own June 2020 data collection, a wide array of techniques were attempted to filter Twitter data to identify information related to emerging local COVID-19 crises in the greater Cincinnati, OH metropolitan area. Early efforts revolved around the cultivation and implementation of relevant keywords, but this presented a number of challenges.

First was that of identifying the correct set of keywords in an emerging event. The query string limits the number of search terms to 10 operators for a standard API key [42]. Relying solely on keywords created a moving target, requiring constant guesswork, the grooming of keywords, and as was often found, the best or most appropriate terms could not be identified until it was too late. Additionally, the fact that a Twitter user might not always adhere to common spelling conventions or use of slang etc., necessitated redundancies to catch misspellings, causing the search term limit to be reached relatively quickly.

Relying exclusively on this *a priori* filtering approach also, in addition to the possibility of exhausting the search term limit, returns a data set that is essentially unbounded. Meaning that, in research terms, it is difficult if not impossible to ascertain the sample population of which Tweets are being collected. In addition to potentially expanding a search beyond explicit geotags, the returned set is now bounded, circumscribed by a new geographic data point (or circle, more correctly). The value of this may not seem readily apparent, but in simplest terms it allows us a researchers to make a claim *"of all Tweets associated with geographic query radius* R *, the following can be said..."*.

Ensuring a set of Tweets that are associated with a location requires using the search operator 'geo' in the query string. The 'geo' operator takes three positional arguments: a latitude, a longitude, and a radius value.

According to the Twitter API documentation, "When conducting geo searches, the search API will first attempt to find Tweets which have lat/long within the queried geo-code [via the 'coordinates' and 'place' tags], and in case of not having success, it will attempt to find Tweets created by users whose profile location can be reverse geo-coded into a lat/long within the queried geo-code, meaning that is possible to receive Tweets which do not include lat/long information" [42].

Initially, our geographic query radius included the extent of the city's farthest suburbs. This proved to be another challenge, however, because of another set of bounding limitations - the 100 Tweet limit of the standard API key [43]. It again became impossible to ascertain the total number of tweets for a particular region as the dataset would reach saturation each collection cycle. By reducing the area of the 'geo' term to focus on particular neighborhoods, we were able to ensure that the 100 Tweet limit was not being reached, and therefore we were able to gain the full set of Tweets for that location (as detailed in the following section).

4.2. Data Aggregation and Analysis

Using Python 3.7.5 and and the open-source Twitter API data collection package *python-twitter*, Tweets were collected using a raw query that filtered on a 3 mile geographic radius of the Cincinnati city center, using standard API key permissions. Additional filtering criteria included the language (English), the maximum number of Tweets per API call (via the 'count' parameter), which was 100, and the 'result_type' parameter with the value of 'recent', which returns all Tweets meeting the above criteria posted within the past 6-9 days [42], ignoring popularity ranking. Also included was the 'since_id' parameter, which restricts each query cycle to gather tweets from a fixed temporal endpoint.

Our Python script collected data on the geography specified as well as a number of other local geographies, and looped to execute data collection on a 15-minute interval. The 'since_id' parameter was updated dynamically each cycle, and due to the nature of the Search API, all tweets posting during the interim would be theoretically collected, as long as this count did not exceed the 100 tweet maximum per call (which for the geography under test, did not).

Of these data, a set of all tweets occurring between the dates of 6/1/2020 to 6/30/2020 were selected. Presumably, this set contains the totality of tweets for that particular radius and date range - none of the JSON files collected reached the value assigned to the *count* parameter. We examine the location metadata in an comparative analysis that follows. The data were parsed, compiled, and evaluated statistically using a variety of python packages, notably *pandas* and *numpy* - visualizations were made using *matplotlib* and *seaborn* packages.

5. Results

The dataset yielded 49,744 unique tweets from 6,902 individual users (Figure 1). Of these individual users the number of posts over the 30-day period ranged from 1-1454 (M=7.2, SD=34.1) while the median number of Tweets per user was 1.

5.1. Total Location Categories

Of the 49,744 tweets in our data set, 42,394 (85.22%) contained User Profile Locations, 32,543 (65.42%) contained Place data, and 1,988 (4%) contained geotag Coordinate data. A comparison of location by category of metadata is shown in Figure 1. Geotag Coordinates are of particular interest because the are evidence of an eyewitness account and are specific in terms of precise location. That this percentage of our current dataset remains slightly higher the previously-cited value of Tweets with geotags (approximately 1-3.2%) [12, 33] is of particular note. In comparison with previous crises, June 2020 was characterized by an impending and particularly divisive election season, an intense period of local protests, and a public health crisis. The fact that this percentage seems to hold may speak to some underlying mechanism or phenomena. Additionally, 1,722 (3.5%) tweets contained no apparent location data and can be attributed to having being aggregated via the query string alone (this is discussed further in section 5.5).

We then examined the tweets of our dataset in terms of total number of metadata location categories present per tweet, specifically, geo-tag Coordinates, Places, and user profile location as shown in Figure 2. 20,924 (42.06%) tweets contained just one metadata location category, 25,291 (50.84%) tweets contained two metadata location categories and 1,806 (3.63%) tweets contained information in all three metadata location categories. The mean categories per tweet was 1.55 (SD=.62) with a median of 2.

Broken down further in Table 1, we find that 15,478 (61.20%) tweets contained only a user's profile location, 5,446 (21.53%) only a place reference, and no tweets contained a GPS Coordinates alone (which was expected, as every tweet containing a GPS coordinate also contains a place reference). Of tweets with 2 locations categories, 182 (.72%) contained only a place reference and GPS coordinate data, while the vast

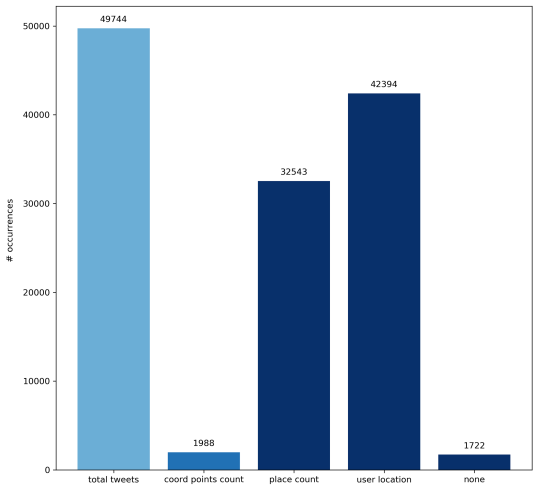

Figure 1. Location Data by Category

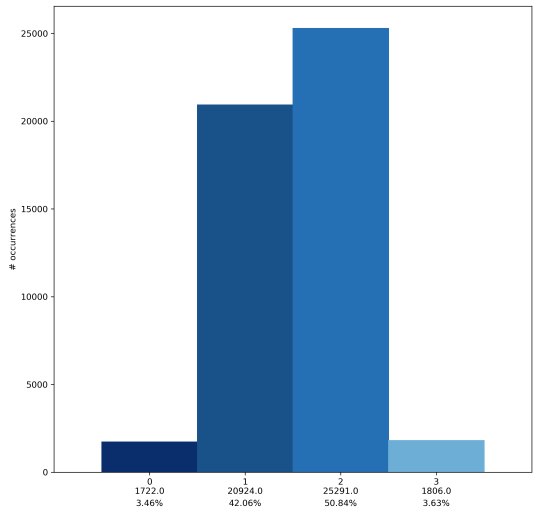

Figure 2. Location Categories per Tweet

majority - 25,109 (99.28%) - contained place reference and a user location. Again, no tweets contained only the paired values of a coordinate and user location, as all tweets with a GPS coordinate are automatically assigned a Place object as well.

5.2. Coordinates

Out of the 1,988 tweets containing GPS Coordinates, our dataset turned up 327 unique coordinate locations, the most referenced occurring 587 times, and the least, only one. This produced an average of 6.08 occurrences per unique coordinate (*SD*=41.9), the mean being 1 (Table 2). As mentioned previously, every tweet with a Coordinate also has a Place reference automatically generated; this will explored in the following subsection.

	1 Category Only		2 Categories Only	
Count	Place	5,446	Place + Coord.	182
% Set	Place	26.03%	Place + Coord.	0.72%
% Total	Place	10.95%	Place + Coord.	0.37%
Count	Loc.	15,478	Place + Loc.	25,109
% Set	Loc.	73.97%	Place + Loc.	99.28%
% Total	Loc.	31.12%	Place + Loc.	50.48%

Table 1. Location Categories per Tweet

	Coordinates	Places	Locations
Count	327	188	2,185
Mean	6.08	173	19.4
SD	41.93	2,280	282.7
Max.	587	31,264	11,843
Min.	1	1	1
Median	1	1	1
Mode	1	1	1

Table 2. Statistics: Unique location data by occurrence

5.3. Places

Of the over 65.42% of tweets containing a Place reference, 188 unique place references were found (Table 2). The average occurrence of a Place in the dataset was extremely varied, with a minimum value of 1 and a max of 31,264, the mean occurrence being 173 times (*SD*=2280). The median unique Place occurrence was 1.

The distribution of Place types was also interesting. The overwhelming majority (32,119, or 98.7%) of place references were to a 'city'; only 422 (1.3%) referred to a 'POI', or 'point of interest', and 2 (.01%) to an 'admin' place type. But even though the majority of place objects referred to a city, only 8 unique city objects were represented. Places of interest, while comprising a small percent of the overall total, were 22.5 times more varied (with 180 unique POI occurrences in the dataset).

All Place objects in our data were stored as coordinate arrays consisting of four point pairs - a square polygon. We found, however, that in many instances these four coordinate pairs were identical, and actually referred to a single point as opposed to an actual area. We found that the total number of point Place Coordinates corresponded directly with the POI place name, the remainder belonging to city and admin place

types. As expected, this correlation extends to unique Coordinates as well.

Excluding points of interest (whose point-value representations have an area of zero), the mean area of all place objects referred to in our dataset was 3.96 mi^2, with a standard deviation of 10.99 mi^2 (see Table 3). We excluded 'admin' place types, which are large bounding boxes roughly equivalent to the size of a U.S. state (the largest in our dataset has an area of 1394 mi^2) and thus skew this number considerably; focusing on only 'city' place types yields a mean area of 3.87 mi^2 (*SD*=.63). These values reflect the mean areas of all duplicated place references as they occur throughout the dataset. Flattening the set to reflect only unduplicated, unique Places we see that the mean area for individual Places changes substantially with the admin place excluded becoming 155.49 mi^2 (*SD*=464.6) and .63 mi^2 (*SD*=1.37) respectively.

5.4. User Profile Locations

The User Profile Location, as stated earlier, is on its own the least reliable source of a tweet's geographic origin. Perhaps unsurprisingly, as also seen earlier in this section, it is by far the most plentiful. Our dataset yielded 2,185 unique User Profile Locations, with a mean occurrence of 19.4 (*SD*=282.7)(see Table 1). These numbers are somewhat misleading, however, as Locations are simple text fields and not curated in the manner of, for instance, a Place object. Thus misspellings of locations and slang or other variations could inflate this count

That said, without using any sort of advanced language processing, we can see that the Location "Cincinnati, OH" was mentioned 11,843 times. Performing a simple substring search of the term "Cincinnati" through our set of unique Locations yielded 140 total instances of this city, and it is conceivable that a number of other local cities and neighborhoods express this redundancy as well.

5.5. 'None' Tweets

One extremely interesting finding is that there exist in our dataset tweets that contain none of Twitter's four location parameters whatsoever. These are the tweets that were returned by the Twitter API as being associated with the supplied geo filter, but contained no explicit geographic references (Coordinate, Place, or User Location), and are identified by the 'none' category in Figure 1. These tweets are of particular interest because they would have escaped any attempt to connect them to a geography using any of the standard methods of geographic identification. We must then infer that Twitter is utilizing additional, more sophisticated geolocation techniques than is at first apparent.

5.5.1. Tweets with No Overt Location Metadata

Performing a census of these tweets lacking geotags, and comparing them to the set as a whole, we found that the 'none' data subset contained 1,722 tweets by 1,224 unique users, with a mean of 1.41 tweets per user (*SD*=2.2)(see Table 1, table 4) An independent t-test verifies this differs significantly from the set as a whole (p=0), which averaged 7.21 Tweets per user (*SD*=34.13).

5.5.2. Tweets with No Overt Location Metadata - Retweets

Interestingly, the 'none' subset is comprised of 99.95% retweets, compared with only 3.46% of the entire dataset; this prompted us to take a deeper look into the structure of retweets themselves.

For retweets, the Twitter API returns the full JSON structure of the original tweet, at the time it was retweeted. While it is conceivable that Twitter could use an algorithm to 'spider' a search into the source tweet to ascertain geographic relevance, Twitter documentation reveals that this embedded information is sanitized of location data, which we evaluated and verified. What we did find, however, is that while all retweets contained no explicit location data, they did contain a user location embedded in the original tweet data. As explained earlier this user location tag is simply saved as a string of text meaning any value can be entered, but an anecdotal look appears to suggest that these embedded user locations reflect the geographic locale for which the API call was made.

6. Discussion

This study investigates the profile of location metadata gathered from one geographic area (*RQ1*). Based on our findings, geographically filtered data from the Twitter API from the region in our sample contained slightly more geotagged tweets (4%) than previously benchmarked studies (1.5-3.2%) [12], was rich with data based on user profile locations (85.2%), and the majority also included place data (65.4%). We also inquire into data returned from a query string radius approach that contains no overt location metadata (*RQ2*), and found that 4% of Tweets gathered using this approach contained no overt metadata.

Additionally, this paper is a dissection of the body of data returned under the umbrella of a geo-radius value. We are by no means the first to contemplate a tiered-approach to tweet geolocation. Layvali et al., for example [44] proposed a "location inference scoring"

	Place areas (all)	Place areas no 'admin')	Place areas (all, undupl.)	Place areas (no 'admin', undupl.)
Place Count	32,121	32,119	9	8
Mean Area	3.96	3.87	155.49	0.63
SD	10.99	0.63	464.61	1.37
Max.	1,394.44	3.98	1,394.44	3.98
Min.	0.03	0.03	0.03	0.03
Median	3.98	3.98	0.093	0.07
Mode	3.98	3.98	0.03	0.03

Table 3. Statistics: Place areas (all)

Unique User Count	1,224
Mean	1.41
SD	2.15
Max.	1,722
Min.	1
Median	1

Table 4. Statistics: Tweets by User (no location)

method that looks at these stratified location positions in hierarchical manner. But to our knowledge we are the first to include the query radius - the underlying object of the Twitter API that unites all the aforementioned data, but which after the API call is performed is typically jettisoned. We choose to instead retain this value, and treat it as a data point no less important than all others referred to above. The greater contribution of this research is that all of this statistical evaluation is made possible by accepting the geo-location query term as the premise for bounding our dataset.

The overarching research question that this research asks is how can social media aggregation techniques be used to improve the amount and quality of information used for crisis informatics? Based on our findings, we propose that COP tools used for crisis informatics clearly label types of metadata and provide filtering features based on the spatial reliability of the data. We propose spatial reliability ranging from that information which may be relatively unreliable (such as profile location) to that with the highest accuracy and reliability (such as geotags). We reflect on these findings in terms of the design of COP tools further in the following section.

6.1. Design Implications

Considering our findings and the data collected, this section reviews their implications when designing common operational picture tools.In visualizing the aforementioned geo-points and polygons of our existing data, several options were explored. In our own experience, what began as a need to simply view our data quickly morphed into a significant software development undertaking, an artifact we came to call PIVOT - the Portal for Intermedia Visualization, Overlay and Triangulation shown in Figure 3. In brief, PIVOT is a web-based geomedia visualization application, utilizing a Python-based Django webserver back-end, and an embedded Google Maps API for front-end visualization. PIVOT allows various types of geographically associated media to be displayed and processed concurrently as well as filtered for specific data sets.

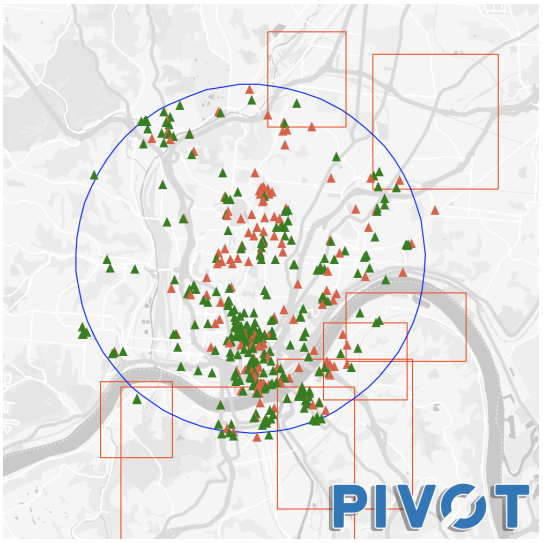

Figure 3. PIVOT prototype screenshot.

6.1.1. Query Radius as a Data Point This preliminary work has already done a great deal to inform the design of our PIVOT tool. One substantial design implication of this research was the decision to treat the API 'geo' operator (what we hereafter call

the Query Radius) as a unique data point unto itself. This perspective informed our decision to modify the PIVOT data collection to store the latitude, longitude, and radius of this operator as a geographic object, referenced to the Tweet data via a foreign key. We anticipate this will enrich the geographic network data of a region and add to overall situational awareness and COP tool performance.

6.1.2. Query Search Term Inclusion In addition to the Query Radius, the PIVOT database will store the remaining operators of the API query string. Using Django equivalents for SQL terms, the search term tool was quickly integrated in order to filter data more efficiently and allows for relevant data to be returned in a more usable and consolidated format, to examine the ways in which users are attempting to collect data, as well as for purposes of data provenance.

6.1.3. Streaming Data The Twitter Search function allows for the 'chunking' of historical tweets based on a geographic reference, but especially when dealing with a live emergency, any lag time between data collection and visualization could be extremely costly. The Twitter API Streaming function allows for instantaneous, real-time collection of tweets, but sacrifices certain other abilities - for instance, if data collection is interrupted, the streaming API has no equivalent to the 'since_id' or 'recent' parameter to fill any gaps in collection. COP tools would benefit enormously from a real-time data gathering option, which we anticipate will be incorporated into PIVOT as well.

6.1.4. Polygon Filtering Given the potential variety in size of Place bounding boxes, and the fact that the Twitter API will potentially return Place objects in the range of thousands of square miles, it is import to include functionality to filter polygons by area, in order to situationally isolate the proper geography.

6.1.5. Distinguishing Place vs. Coordinate points The revelation that Place points of interest were actually single point values stored as polygons prompted internal conversations as to how best store and display this data. While the exact technique has not been refined, it seems a consensus among our researchers that the POI bounding box should be condensed into and stored as a single geographic point, then flagged accordingly as a Place as opposed to a Coordinate. PIVOT would then visually discern the two form of Point data.

6.1.6. Complex polygon generation The shapes utilized by Twitter for Place identification are, as far as our data suggests, simple square polygons, and the Search geographies simple radii. But there are any number of instances where we might desire to associate Tweets with more sophisticated types of polygons; for example, to evaluate tweets originating from a complex geography, to sort by zip code or county, to check for polygon overlap, and so on. We must reserve the ability to evaluate and manipulate more complex polygons, or even generate them via point clouds and advanced machine learning clustering algorithms.

6.2. Limitations

The most evident limitation of our study is that the current statistics have been inferred through data collected over a one month duration, and during a unique time in history which may have impacted users behaviors. Despite the challenge this may present in generalizability of the results, it does offer opportunities to researchers hoping to benchmark results as high or low by comparison to other work.

The data has also been subject to geographical restrictions to a particular region within the United States, and this factor may affect the inferences being drawn. Additionally, the data were collected during a particularly turbulent historical time, in terms of social unrest. This may very well have influenced peoples usage of social media in a myriad of ways.

Finally, the data collected from Twitter has been supplied by the Twitter API, which is subject to change at the whim of the company. While this could be potentially disruptive to data collection and processing efforts, in the event of such change, PIVOT was designed with flexibility in mind and has the advantage of potentially mitigating such compatibility issues.

6.3. Future Work

Through the course of this work, many observations were made and questions answered regarding the anatomy of a geographically retrieved Twitter dataset. However, as seen in the preceding Design Implications and Limitations sections, this gave rise to an even more nuanced set of questions, and numerous potential avenues of research were exposed.

Briefly touched on in the Background section is the idea that network information - that is to say, groups of interconnected and interrelated location data - along with other more sophisticated techniques such as natural language processing (NLP) and image recognition can provide a richer opportunity for tweet analysis than

any one particular piece of information on its own. A wealth of existing research attests to the utility of implementing such models in geospatial visualization, and could certainly be implemented in future iterations of PIVOT.

7. Conclusion

This study finds that among tweets returned by the Twitter API for the area of observation, location information available for collection in hyperlocal geographic areas is mostly course-grained, involving city and place names entered or tagged by users, respectively. It also varies in relative frequency, with a small subset of the dataset including fine-grained geographic coordinates and the majority including course-grained location information, and typically includes multiple types of course-grained location metadata. These hyperlocal observations compliment national and global analyses of geotagging behavior in previous benchmarked studies [11, 12], and fill in gaps on geographic information behavior in crisis informatics literature [13, 45]. Furthermore, this study highlights social media data unaccounted for in prior research: tweets containing no overt location metadata returned via location-based Twitter API queries.

These findings help translate research on actionable social media into design requirements for PIVOT, a novel COP tools suitable for municipal-level emergency response. These requirements include: i) incorporating query radius as a data point to enrich geographic information collected using the Twitter API, ii) storing query search terms to efficiently filter and return data, iii) filtering upwards/downwards by polygon to collect data at appropriate spatial granularities, iv) distinguishing place and coordinate points to visualize incident information for points of interest, and v) generating complex polygons to collect data from irregularly shaped geographic jurisdictions monitored by emergency responders.

References

[1] M. Gardner and D. A. McEntire, "The community dispatch center: An assessment of a neglected component of emergency management," *Journal of Emergency Management*, vol. 1, no. 1, 2003.

[2] J. Van Wagenen, "States seek a 21st-century upgrade to 911 infrastructures." StateTech Magazine https://statetechmagazine.com/article/2017/06/states-seek-21st-century-upgrade-911-infrastructures, 2017. Accessed: 10/8/2020.

[3] K. Barker, J. H. Lambert, C. W. Zobel, J. E. Tapia, A. H.and Ramirez-Marquez, L. Albert, and C. Caragea, "Defining resilience analytics for interdependent cyber-physical-social networks," *Sustainable and Resilient Infrastructure*, vol. 2, no. 2, pp. 59–67, 2017.

[4] R. Grace, S. Halse, A. Kropczynski, J.and Tapia, and F. Fonseca, "Integrating social media in emergency dispatch via distributed sensemaking," in *Proceedings of the 16th ISCRAM Conference*, pp. 734–745, 2019.

[5] A. Wang, "Detecting spam bots in online social networking sites: A machine learning approach," in *24th Annual IFIP WG 11.3 Working Conference*, 2010.

[6] L. Palen and K. M. Anderson, "Crisis informaticsnew data for extraordinary times," *Science*, vol. 353, no. 6296, 2016.

[7] C. Reuter, A. L. Hughes, and M. A. Kaufhold, "Social media in crisis management: An evaluation and analysis of crisis informatics research," *International Journal of Human-Computer Interaction*, vol. 34, no. 4, pp. 280–294, 2018.

[8] J. Wolbers and K. Boersma, "The common operational picture as collective sensemaking," *Journal of Contingencies and Crisis Management*, vol. 21, no. 4, pp. 186–199, 2013.

[9] H. Zade, K. Shah, V. Rangarajan, P. Kshirsagar, M. Imran, and K. Starbird, "From situational awareness to actionability," in *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, pp. 1–18, 2018.

[10] J. Kropczynski, R. Grace, J. Coche, S. Jalse, E. Obeysekare, A. Montarnal, F. Benaben, and A. Tapia, "Identifying actionable information on social media for emergency dispatch," *Proceedings of the ISCRAM Asia Pacific*, 2018.

[11] B. Huang and K. M. Carley, "A large-scale empirical study of geotagging behavior on twitter," 2019.

[12] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, "Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose," in *Seventh international AAAI conference on weblogs and social media*, 2013.

[13] R. Grace, "Hyperlocal toponym usage in storm-related social media," in *Proceedings of the 17th ISCRAM Conference*, pp. 849–859, 2020.

[14] M. Kogan, L. Palen, and K. M. Anderson, "Think local, retweet global: Retweeting by the geographically-vulnerable during hurricane sandy," in *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pp. 981–993, 2015.

[15] K. Starbird and L. Palen, "(how) will the revolution be retweeted? information diffusion and the 2011 egyptian uprising," in *Proceedings of the acm 2012 conference on computer supported cooperative work*, pp. 7–16, 2012.

[16] A. Olteanu, C. Castillo, F. Diaz, and S. Vieweg, "Crisislex: A lexicon for collecting and filtering microblogged communications in crises," in *Eighth international AAAI conference on weblogs and social media*, 2014.

[17] K. A. Lachlan, P. R. Spence, X. Lin, K. M. Najarian, and M. D. Greco, "Twitter use during a weather event: Comparing content associated with localized and nonlocalized hashtags," *Communication Studies*, vol. 65, no. 5, pp. 519–534, 2014.

[18] X. Lin, K. A. Lachlan, and P. R. Spence, "Exploring extreme events on social media: A comparison of user reposting/retweeting behaviors on twitter and weibo," *Computers in human behavior*, vol. 65, pp. 576–581, 2016.

[19] K. Starbird, L. Palen, A. L. Hughes, and S. Vieweg, "Chatter on the red: what hazards threat reveals about the social life of microblogged information," in *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pp. 241–250, 2010.

[20] M. Avvenuti, S. Cresci, A. Marchetti, C. Meletti, and M. Tesconi, "Ears (earthquake alert and report system) a real time decision support system for earthquake crisis management," in *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1749–1758, 2014.

[21] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*, pp. 851–860, 2010.

[22] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, "Microblogging during two natural hazards events: what twitter may contribute to situational awareness," in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1079–1088, 2010.

[23] M. A. Cameron, R. Power, B. Robinson, and J. Yin, "Emergency situation awareness from twitter for crisis management," in *Proceedings of the 21st International Conference on World Wide Web*, pp. 695–698, 2012.

[24] Y. Tim, S. L. Pan, P. Ractham, and L. Kaewkitipong, "Digitally enabled disaster response: the emergence of social media as boundary objects in a flooding disaster," *Information Systems Journal*, vol. 27, no. 2, pp. 197–232, 2017.

[25] H. Purohit, C. Castillo, F. Diaz, A. Sheth, and P. Meier, "Emergency-relief coordination on social media: Automatically matching resource requests and offers," *First Monday*, 2014.

[26] C. M. White, *Social media, crisis communication, and emergency management: Leveraging Web 2.0 technologies*. CRC press, 2011.

[27] Y.-R. Lin and D. Margolin, "The ripple of fear, sympathy and solidarity during the boston bombings," *EPJ Data Science*, vol. 3, no. 1, p. 31, 2014.

[28] L. Palen, S. Vieweg, S. B. Liu, and A. L. Hughes, "Crisis in a networked world: Features of computer-mediated communication in the april 16, 2007, virginia tech event," *Social Science Computer Review*, vol. 27, no. 4, pp. 467–480, 2009.

[29] J. Steenbruggen, P. Nijkamp, J. M. Smits, and M. Grothe, "Traffic incident management, a common operational picture to support situational awareness of sustainable mobility," *International Journal of Transport Economics/Rivista internazionale di economia dei trasporti*, pp. 131–170, 2012.

[30] J. Copeland, "Emergency response: Unity of effort through a common operational picture," tech. rep., ARMY WAR COLL CARLISLE BARRACKS PA, 2008.

[31] "Emergency management operations gallery — arcgis solutions for emergency management." https://solutions.arcgis.com/emergency-management/help/com-operational-picture/. (Accessed on 07/14/2020).

[32] D. Pohl, A. Bouchachia, and H. Hellwagner, "Automatic sub-event detection in emergency management using social media," in *Proceedings of the 21st International Conference on World Wide Web*, pp. 683–686, 2012.

[33] "Tweet geospatial metadata twitter developers." https://developer.twitter.com/en/docs/tutorials/tweet-geo-metadata. (Accessed on 07/14/2020).

[34] "Tweet object." Twitter Developer https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object, 2020. Accessed: 07/09/2020.

[35] R. Lee and K. Sumiya, "Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection," in *Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks*, pp. 1–10, 2010.

[36] "Geo object." Twitter Developer https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/geo-objects, 2020. Accessed: 07/09/2020.

[37] "Twitter launches 'points of interest' pages for locations." Wired https://bit.ly/30eVT8n, 2010. Accessed: 07/13/2020.

[38] "How to add your location to a tweet." Twitter Help Center https://help.twitter.com/en/using-twitter/tweet-location, 2020. Accessed: 07/13/2020.

[39] "Tweet location faqs." Twitter Help Center https://help.twitter.com/en/safety-and-security/tweet-location-settings, 2020. Accessed: 07/13/2020.

[40] "User object." Twitter Developer https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/user-object, 2020. Accessed: 07/09/2020.

[41] R. Grace, J. Kropczynski, S. Pezanowski, S. E. Halse, P. Umar, and A. H. Tapia, "Social triangulation: A new method to identify local citizens using social media and their local information curation behaviors.," in *ISCRAM*, 2017.

[42] "Search tweets." Twitter Developer https://developer.twitter.com/en/docs/tweets/search/guides/standard-operators, 2020. Accessed: 07/09/2020.

[43] "Search tweets." Twitter Developer https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets, 2020. Accessed: 07/15/2020.

[44] F. Laylavi, A. Rajabifard, and M. Kalantari, "A multi-element approach to location inference of twitter: a case for emergency response," *International Journal of Geo-Information*, vol. 5, no. 56, 2016.

[45] T. Shelton, A. Poorthuis, M. Graham, and M. Zook, "Mapping the data shadows of hurricane sandy: Uncovering the sociospatial dimensions of big data," *Geoforum*, vol. 52, pp. 167–179, 2014.