Incorporating Context and Location Into Social Media Analysis: A Scalable, Cloud-Based Approach for More Powerful Data Science

Jennings Anderson
University of Colorado Boulder
jennings.anderson@colorado.edu

Gerard Casas Saez
University of Colorado Boulder
gerard.casassaez@colorado.edu

Kenneth M. Anderson
University of Colorado Boulder
ken.anderson@colorado.edu

Leysia Palen
University of Colorado Boulder
leysia.palen@colorado.edu

Rebecca Morss
National Center for Atmospheric Research
morss@ucar.edu

Abstract

Dominated by quantitative data science techniques, social media data analysis often fails to incorporate the surrounding context, conversation, and metadata that allows for more complete, accurate, and informed analysis. Here we describe the development of a scalable data collection infrastructure to interrogate massive amounts of tweets—including complete user conversations—to perform contextualized social media analysis. Additionally, we discuss the nuances of location metadata and incorporate it when available to situate the user conversations within geographic context through an interactive map. The map also spatially clusters tweets to identify important locations and movement between them, illuminating specific behavior, like evacuating before a hurricane. We share performance details, the promising results of concurrent research utilizing this infrastructure, and discuss the challenges and ethics of using context-rich datasets.

1. Introduction

Social media data analysis is becoming a sensationalized research area typically dominated by machine learning, data mining, and natural language processing. These approaches often fail, however, to incorporate surrounding context to better situate the content and enable more complete, accurate, and informed analysis. This paper describes the development of a scalable Twitter data collection infrastructure using Kubernetes and a location-based analysis dashboard that are designed to collect and interrogate massive amounts of tweets based on user conversations—not keywords—to perform contextualized social media data analysis.[1]

This infrastructure was initially designed to help researchers understand people's perceptions and behaviors related to approaching weather hazards (such as whether or not to evacuate in the face of a hurricane) by analyzing the content of Twitter users' publicly visible tweets as a personal narrative of their experiences [1]. When available, location information is displayed in a dashboard that clusters and displays geo-referenced tweets on an interactive map, allowing analysts to interpret tweets in spatial context while highlighting the movements of users. This can, for example, help researchers identify users who evacuate. Our infrastructure is designed to be scalable and cost-effective while working within the limits of Twitter's APIs. In doing so, we aim to lower the entry barrier to mixed-methods, large-scale social media data analysis. However, analysis of social media narratives and geo-referenced posts also raises challenges—both technical and ethical—that we will discuss.

In the following sections, we explain the value of including the complete conversation and context surrounding social media posts and then review related work. We discuss geo-referenced social media posts and their ability to provide additional context and present common pitfalls that researchers need to protect against when interpreting such data. We then describe the technical implementation of our data collection infrastructure. Finally, we share our results and end with a larger discussion around the challenges and ethics of using context-rich datasets.

1.1. Context Matters

Accounting for the conversational nature of social media is challenging both technically and analytically; it requires collecting significantly more data and then piecing it all together to construct the full narrative [2, 3]. More the fault of the data collection process than the research approach, Twitter data analysis starts with the gathering of publicly posted tweets. This is

[1] Code: github.com/Project-EPIC/context-analysis-infrastructure

URI: https://hdl.handle.net/10125/59666
ISBN: 978-0-9981331-2-6

most commonly done through one of the public (and free) APIs that Twitter makes available for sampling their massive stream of social media posts. Typically, analysts request posts that match a set of relevant keywords through these APIs. These keyword-based collections, however, cannot tell the complete story. Consider the following example (synthetic, based on real tweets):

t1: **User1** There's a hurricane coming, I'm freaking out! #hurricanesandy
t2: **User2** **@User1** Be safe! My cousin's also in the area and just got told to evacuate!
t3: **User1** **@User2** Yeah, I just got the notice, too. It's so scary. I'm packing now.
t4: **User1** The first floor of my building is totally underwater #hurricanesandy
t5: **User2** **@User1** Okay, let me know when you're safe!
t6: **User1** **@User2** just got to my aunt's house in New Jersey. My whole family is here.

Data collections based on storm-related keywords will only collect posts t1 and t4 (”Hurricane” and ”hurricanesandy”). Posts t3 and t6, however, provide critical information about what preparatory actions **User1** is taking and why; yet, these tweets will not be included in a keyword-based dataset. Depending on the analyst's end goal, the lack of this contextual information likely goes unnoticed. Machine classifiers may recognize that these users are talking about the storm, that they are nervous, and that a building is flooding. These pieces of information describe just one more person in a list of potentially millions that are affected by the storm. Far from comprehensive, keyword datasets are still used by many; for example, Kryvasheyeu et al. used keyword-based collections for damage assessment, achieving correlation between location, volume, and damage-levels [4]. These interpretations, however, offer primarily validation that people who tweet, tweet about their current experiences. These tweets are then interpreted as simple proxies for people mentioning the storm, and the rich information contained in the post itself is not utilized. Furthermore, Shelton et al. show that many of these quantitative mappings of tweet densities are incapable of understanding the various contained geographies without more context and qualitative investigation [5]. The problem, then, is that many data science approaches extract the single post (i.e. ”tweet”) for analysis rather than a user's monologue or conversation with other users. This problem has previously been called the *tyranny of the tweet* [2].

Thus, this work comments on the difficulties of social media analysis at scale and argues that contextualized analysis infrastructures help expose a more accurate story, in this case about how users are responding to a disaster, a story more nuanced and descriptive than that provided by the analysis of keyword collections. We stress the concept of *contextualized analysis* over *data science* or *data analysis* because our infrastructure is built to better contextualize data to promote more responsible interpretation and it is built for mixed-methods research approaches to social media data, as advocated by [2, 6].

Referring back to the example conversation, to understand the complete story of whether or not this user evacuated, we must first collect **all** of a user's tweets *and replies* to create a complete narrative and conversation for analysis. This is what our new infrastructure does in a scalable and cost-effective way. Only with this complete record, can we claim that the data represents a record of the user's experience and behaviors. From here, we can begin contextualized analysis of the user's perceptions and behaviors related to the threat, a goal of concurrent work [1, 7].

2. Background and Related Work

Researchers in the crisis informatics and information science disciplines are developing new contextualized knowledge by employing mixed-methods approaches to social media analysis. Using the near real-time personal accounts of a disaster's impact afforded by social media posts, researchers have extracted situational awareness [8], identified local and community-specific relief needs [9], and identified key distinctions between locally-actionable information and general well-being concerns in the global conversation [10]. Within these fields, there has also been multiple attempts at developing methods for the collection, analysis, visualization, and interpretation of social media data based on keywords [11, 12, 13, 14].

Incorporating interviews and more qualitative analysis approaches, researchers have studied topics such as the reunification of lost pets with families in the wake of disasters [15], the self-organization of digital volunteers [16], and the critical information infrastructure that develops through social media during a crisis [17]—as well as the misinformation that travels across these infrastructures [18]. The rich narratives published to social media provide on-lookers with a window to the situation as it is unfolding on the ground. There are major limitations, however, of what can actually be learned from these data collections. These datasets are not always representative of the larger population and often have embedded biases. These data can certainly provide insightful first-hand accounts of the events, and should absolutely be studied in this capacity, but researchers need to be aware of—and honest about—the limitations [6].

2.1. Data Collection Methods

Twitter makes collecting data available through a set of publicly accessible APIs. The most common are the search and filter APIs which allow users to submit specific search rules, typically keywords. This service is available to the public for free and, as such, does not include all possible Twitter data; indeed, only a small percentage of Twitter data is included. In 2018, Twitter adjusted their API access model to offer more options via different subscription levels, such as access to the "decahose," i.e., up to 10% of the current tweets being generated at any time. Paid access to more data has always been available, with some universities and research groups having expanded access through subscriptions [6]. This model of access to Twitter data is unlikely to change. While Twitter may continue to alter rates and provide different levels of access to data, the fundamental access model of searching for a specific keyword, location, or user and getting a rate-limited response is likely to persist. This implies that every researcher is ultimately working with a different dataset, depending entirely on the terms with which they seeded their search and the type of access they could afford. While our data collection infrastructure does not solve this issue entirely, it creates more complete datasets at the user level by collecting all of a user's tweets and conversations.

The design and implementation of data intensive software systems for social media research is an active area of study [19]. This constantly evolving field—also known as data engineering—has existed alongside and pushed the field of crisis informatics by improving data collection and analytics at ever-growing scales. Some examples include infrastructures for collecting and analyzing social media data [12, 20, 21], moving from traditional SQL to NoSQL distributed data stores [21], and moving to real-time social media analysis [22, 23]. In this paper, we look at a new evolution of this type of infrastructure that takes advantage of services known as container orchestration systems that are designed for massive scalability, high availability, and strong reliability.

2.2. Cloud-Based Scalability

Open source tools like Hadoop and Spark have made scalable, distributed computation available to the masses. In addition, systems like Cassandra have greatly simplified hosting internal storage. As a result of companies open-sourcing their distributed orchestration systems, such as Kubernetes or Mesos, data analysis systems can scale up to handle big data loads at a greatly reduced cost. The SMACK architecture is an example of a distributed architecture similar to our system [24]. SMACK distributes workloads using Akka for actor orchestration and Mesos for container orchestration. Our architecture uses Kubernetes and Google Cloud Storage. We chose to use micro-services within Kubernetes instead of Akka because Kubernetes is 1) language agnostic through the use of containers and 2) managed clusters are readily available from many cloud service providers. However, similar to work performed by Hu et al. [25] and Kiran et al. [26], our infrastructure is not tied to any one container orchestration system or cloud vendor. This flexibility provides other researchers with a choice that can help to reduce their costs if they decide to adopt our system.

2.3. Movement Prediction with Geo-Referenced Social Media Data

Knowing the place associated with a tweet provides another dimension for analysis: *location*. Knowing the geographic location and timestamps of two tweets from the same user can provide knowledge of a user's movement during an event. As early as 2008 (before Twitter introduced the geotagging API in 2009), researchers investigated the ability of user-generated content sites to learn about the locality of users. Using data from Flickr, Girardin et al. identified patterns of geo-referenced photos posted by tourists to Rome, creating a proxy for understanding the city's tourism industry [27]. Most notably, this study identified the observational advantages of this form of passive, secondary geo-referenced data over other sources where a user is opting-in to provide location data.

Using data from 200K Twitter users, Krumm et al. built a machine learning classifier to predict whether a user will visit a given location [28]. Relevant to our work, Krumm et al. identified the most likely times for a user to be home (middle-of-the-night) and then identified the location most associated with tweets at these times. We use this same process to predict a user's home location which is critical additional context when investigating, e.g., evacuation behavior. As for human movement patterns, Jurdak et al. found Twitter to be a useful proxy for predicting human movement using geo-referenced tweets posted in Australia [29] while Martín et al. used the amount of geo-referenced social media data, aggregated at the county level in South Carolina as a proxy for identifying evacuation during Hurricane Matthew [30]. Our infrastructure differs from their approach by creating opportunities for user-level, not aggregated, analysis to understand these behaviors.

3. Geo-Referenced Social Media Data

The location from where a tweet was posted provides powerful additional context, especially if there are no other location clues in the post. Most tweets do not contain any geo-referencing, making those that do include this extra metadata all the more valuable. Complicating the ability to learn from this information, however, are the different forms of geo-referencing available to users. In some cases, we think users are likely not even aware they are posting their location information on Twitter. Techniques, norms, and options available for geo-referencing one's social media posts have changed over the years. When discussing these behaviors, we make the following distinction between the terms geo-reference, geo-tag, and geo-location:

Geo-Reference Either form of location information below:

Geo-Tag Includes the actual coordinates (the precise lat/lon) as reported by the GPS of the posting device

Geo-Location Includes a relative location of varying resolution: "Miami" or "USA"

In comparing various disaster-event datasets collected from Twitter over the last six years, we find a decrease in geo-tagged content and an increase in geo-located content. In that time, the Twitter mobile app changed the location-sharing interface to make it easier for users to geo-locate their post than to geo-tag their post. This is achieved by providing the user with a list of suggested locations above the geo-tag (coordinates) when choosing to geo-reference their post. Moreover, many geo-referenced posts are cross-posted from hundreds of other services such as Foursquare or Instagram. Each of these sources then has their own method of geo-locating or geo-tagging. Foursquare, for example, intentionally obscures the exact location of "homes" [31]. Table 1 shows the number of distinct sources per event type for various disaster-related Twitter datasets we have collected since 2012.

Event Type	Geo Tweets	% of total	Sources
Earthquake	2,463,303	2.13	684
Flood	890,409	1.99	701
Hurricane	458,225	0.64	660
Tornado	402,077	1.21	567

Table 1. Geo-referenced tweets collected since 2012 from disaster events and the number of services posting them to Twitter, via the `source` attribute
.

Geo-referenced tweets always hold point-location information in one of two attributes: `coordinates` or `geo` (now deprecated by Twitter). These fields record the point as an exact latitude/longitude. The trouble with this point-coordinate resolution, however, is that those using the data may incorrectly assume that the point represents the exact coordinates from where the tweet was posted (a geo-tag). Given the multitude of sources for this geo-reference, there can be no guarantee that this is an accurate point location. For example, many cross-posts from Instagram contain geo-tags that are really geo-locations. Figure 1 shows geo-referenced tweets posted by a user during Hurricane Irma in Fall 2017. The two points represent "Miami" (red) and "South Beach." In fact, many cross-posts from Instagram for "Miami" appear to be from the red point in Figure 1. It is unlikely that all of these users are standing in that exact location, which happens to be a parking lot for a Yacht club. Since we also find hundreds of Instagram photos cross-posted to Twitter from Hurricane Matthew (2016) at this exact location, it is probably the case that these coordinates are returned by a geo-coding lookup service for "Miami," and are then embedded into the tweet. As an overview, the resolution of these coordinates representing the city is fine, but if looking for neighborhood or block-level resolution, the precision of this geo-reference becomes problematic; users were not at that specific location as one might initially assume from the exact coordinates embedded in the tweet.

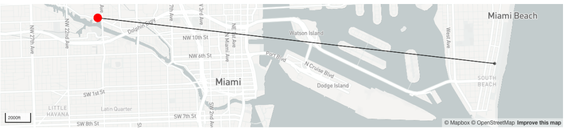


Figure 1. User movement between Miami and South Beach. Red dot represents exact coordinates for "Miami" geo-location.

To address this confusion, geo-referenced tweets include a `place` attribute with additional information about the location the user is geo-referencing. This includes a geographic bounding box, the full name of the location, and a description of the geographic resolution: A city, county, state, etc. The `place` attribute is a powerful feature for Twitter as a data-provider for geographic search and filtering, but the nuances of this attribute seem too often to be ignored by data analysts: people simply produce maps of individual tweets as points from their collected dataset. We also find examples in our datasets of the `place` attribute either missing or containing incorrect descriptions of the geo-reference. With so many different services posting to Twitter (see Table 1), these errors seem inevitable: different services handle location data differently, such as FourSquare intentionally obscuring a private residence. What gets recorded in the tweet, however, is an exact coordinate representing *some* location within *some* distance of the original address, with no additional context to warn future analysts that the user is not actually standing at that location. Similarly, during Hurricane Matthew, hundreds of Instagram cross-posts were embedded with the coordinates for the "Miami" location in Figure 1, however, the tweets do not correctly specify "city" as the `place_type` within the `place` attribute, so they appear to be exact geo-tags. Ultimately, many geo-referenced tweets appear to have better spatial accuracy than they actually do. Differentiating the spatial resolution of a geo-reference is not straightforward without more context about the user's activity.

Moreover, the presence of a geo-reference does not imply that the user is actually at that location. Twitter even notes in the API documentation that "Tweets associated with Places are not necessarily issued from that location but could also potentially be about that location" [32]. We have found tweets from users who appear geographically vulnerable by geo-locating their tweets at casinos in Atlantic City, New Jersey near the time of Hurricane Sandy's landfall when the region was supposed to be evacuated. In reality, they were expressing their concern for the business and chose to geo-reference the location as additional context, making the tweet *about that place*, not *from that place*.

4. Technical Architecture

In this work, our goal was to create a scalable architecture that optimized performance by distributing the workload of data collection between different types of workers. This requires two components: the *user-crawler* and the *user-requester*. We use a container-based deployment system in the cloud—Kubernetes—that enables us to scale the number of workers as necessary. When scaling these components, our primary concern lies with rate-limiting. Twitter implements API rate-limiting to keep their services functional and available; this limit constrains the amount of data available for collection by outside systems. In this context, scalability refers to the ability to work within the bounds of API rate-limiting and still retrieve data at an acceptable rate. An odd trade-off in the big data era is that to be cost-effective, what each crawler spends most of its time doing is simply waiting for an API limit to reset; for Twitter, that interval is typically 15 minutes.

4.1. User crawler

Inspired by web crawlers, this component is a client of Twitter's REST API; it pulls a user's entire timeline (all of a user's tweets, up to limits set by Twitter). The user crawler is a single, robust command-line utility written in Python. Input is either a list of users through `stdin`, a file containing a list of Twitter usernames, or usernames that arrive via a specific Kafka topic (i.e. message queue). The program authenticates with Twitter via credentials stored by Kubernetes as environment variables within the container created for the web crawler.

As of June 2018, the Twitter user timeline API can return a user's most recent 3,200 tweets. Each single request to the API can return 200 tweets. Obtaining all (available) tweets for a user therefore takes 16 requests. This public Twitter API endpoint is rate-limited to 1,500 requests every 15 minutes. This imposes a maximum retrieval of 93 users per 15 minutes (or 6 users per minute). In reality this number is slightly higher because many users have not tweeted 3,200 times, so their entire timelines are retrievable in less than 16 requests. When a user's entire available timeline is downloaded, the results are saved to a cloud storage bucket in a line-delimited JSON file (one tweet per line).

Within each storage bucket, files are stored under an event subdirectory. Events are defined by start and end dates. For each event, there are four subdirectories to further organize the results:

complete Successfully collected tweets from a user that fall within the entire time of an event

incomplete A user has tweeted more than 3,200 times since the start of the event; there may be tweets from the user during the event that we cannot collect.

failed A user has tweeted more than 3,200 times since the end of the event; none of the tweets returned were posted during the time of interest.

not accessible A user's tweets are protected or the account no longer exists

While every event varies drastically in terms of the number of users who are actively talking about it on social media, this number is rarely small. The Twitter rate limit then becomes the primary bottleneck in the collection. Distributing to multiple workers allows us to use more API keys and alleviate this concern, linearly scaling the number of tweets per minute we can obtain at the rate of >6 users per minute per worker.

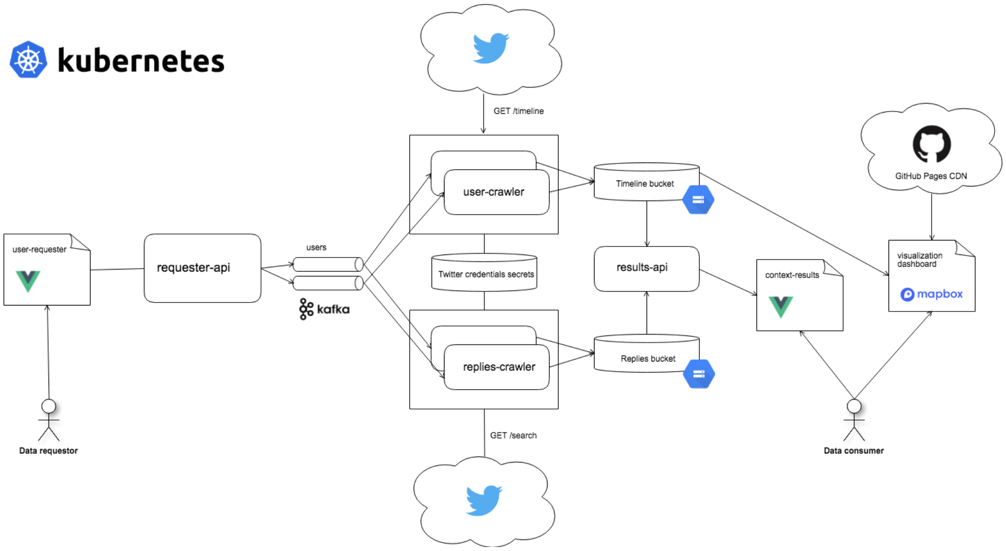


Figure 2. Crawler microservices system architecture overview on Kubernetes

4.1.1. Reply Crawler

Tweets authored by a single user only represent one side of the conversation. To get the other half, we run the *reply crawler*. This microservice uses the search API to pull mentions of a user that are replies. For example, the search string `"@User1 filter:replies"` queries for replies to tweets authored by @User1. We do not use the query operator `"to:"` because it only pulls tweets that directly reply to @User1 which excludes any reply from a third user. Consider the example conversation below. The reply crawler will capture both t2 and t3. Using the `"to:"` operator would only retrieve t2. This is important because t3 is technically a reply to t2/@User2 and therefore the author of t3 may be unaware of t1/@User1, other than the mention of @User1 in t2.

t1: **User1** There's a hurricane coming, I'm freaking out! #hurricanesandy
t2: **User2** **@User1** Be safe! My cousin's also in the area and just got told to evacuate!
t3: **User3** **@User1 @User2** Yeah, I just got the notice, too. It's so scary. I'm packing now.

This component works in a similar way to the user crawler. Inputs can be a list of users via `stdin`, a file, or a Kafka topic. It uses the Standard Search API. Each request returns a maximum of 100 tweets and is rate-limited to 450 requests every 15 minutes. In addition, as of May 2018, Twitter has restricted the access of this endpoint to tweets published in the last seven days. The new Premium APIs allow for searches that go further back in time. It should be noted though, that the new premium APIs do not have the `"filter:replies"` operator. This limits our reply queries to seven days after the event, so analysts must act quickly to identify users of interest.

4.2. Scaling

The primary purpose of distributing this infrastructure on Kubernetes is to increase download speed. For this, we use a publish-subscribe, asynchronous micro-service architecture. We utilize an external micro-service acting as a web interface to easily input the list of usernames to be downloaded. To communicate this to the crawlers we use a distributed Kafka cluster with a topic for the users, partitioned for each worker. Distribution between partitions is done in a round robin fashion which ensures that each of our components will be reading approximately the same amount of users without skipping a user.

4.2.1. Handling Twitter rate-limits

To avoid abuse of its service, Twitter implements strict rate-limits on each API endpoint. Twitter data collection infrastructures are required to work within these requests-per-minute constraints. Our architecture increases performance by using different API credentials for each worker, running all of the crawlers in parallel. In practice, a small number of workers (less than ten) is sufficient to meet the needs of many crisis events.

Given the 15 minute interval to reset the rate limits, each worker sits idle for many minutes at a time. We currently utilize this time to create GeoJSON files (described below) when geographic metadata is present. This is just one example; these idle times could be used for any number of automated user-level analyses. Additional techniques could include sentiment analysis, language detection, location analysis, or even identification of more users to collect.

4.2.2. Making Use of the Cloud To make it easier to deploy and scale, we dockerized the whole infrastructure to allow for distributed deployment in Kubernetes. This container orchestration system allows a system administrator to deploy code without having to worry about languages, operating systems, or dependencies. Kubernetes is quickly becoming a de facto standard to deploy cloud-based infrastructure thanks to its compatibility with various cloud providers. It allows a research team to choose among the major cloud hosts when deploying software infrastructure.

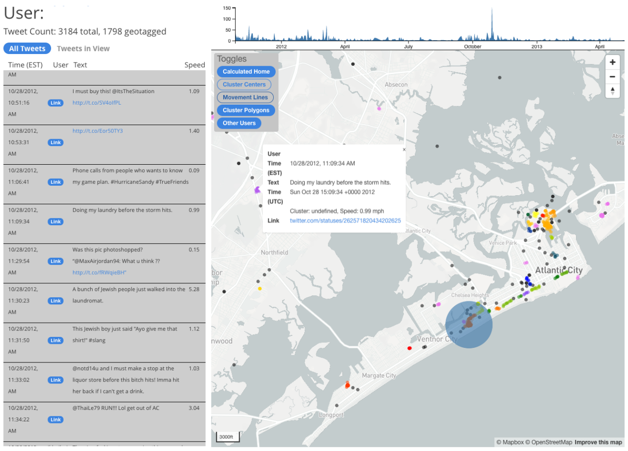
Kubernetes simplifies deployment and makes the task of deployment efficient. With its abstraction, we can deploy several microservices in isolated environments without needing a separate machine for each. This allows us to deploy complex systems with a simpler and reduced underlying infrastructure, keeping infrastructure costs lower than if we had to deploy a machine for each microservice. In our case, our whole testing infrastructure is deployed on only two virtual machines on Google Cloud. For this implementation, we use Google's Kubernetes Engine (GKE). It offers a more advanced interface and is capable of meeting our needs better than current competitors. In addition, taking into account that our infrastructure creates heavy network traffic through the transferring of many tweets in parallel, network performance is a primary concern. Using the latest Google networking systems to route requests, we find GKE provides great performance at low cost (roughly $100/month).

5. Extending with Geo-Location

In addition to saving the JSON representation of a user's tweets, the user crawler script creates a simplified geojson feature representation with the following schema:

```
{"type":"Feature",
 "geometry": <geojson point>,
 "properties": {
   "tweetID": <string>
   "user": <string>
   "date": <string (ISO Timestamp)>
   "text": <string>
 }}
```

If a tweet does not have a `geo` or `coordinates` property, the geometry is set to `null`. If even one tweet has a valid point geometry, a GeoJSON feature collection is created with all of the tweets and saved to a cloud bucket. The GeoJSON format is a lossless, non-compressed, human-readable format for storing geographic data. The limited tweet attributes included here ensure that a user's entire available contextual stream typically remains under 1MB in size for ease of transfer across the web while maintaining enough valuable information for standalone visualization. We include the original tweet id to ensure it is easy to look up the post on Twitter. If there is a reply conversation (@*user*) associated with this user's contextual stream, a separate script will stitch together the entire conversation chronologically. For this reason, the *user* attribute is important to include in every feature.



User:

Tweet Count: 3184 total, 1798 geotagged

Time (EST)	User	Text	Speed
AM			
10/28/2012, 10:51:16 AM	Link	I must buy this! @ItsTheSituation http://t.co/SV4oIfPL	1.09
10/28/2012, 10:53:31 AM	Link	http://t.co/Eor50TY3	1.40
10/28/2012, 11:06:41 AM	Link	Phone calls from people who wants to know my game plan. #HurricaneSandy #TrueFriends	0.09
10/28/2012, 11:09:34 AM	Link	Doing my laundry before the storm hits.	0.99
10/28/2012, 11:29:54 AM	Link	Was this pic photoshopped? "@MaxAirjordan94: What u think ?? http://t.co/fRWqieBH"	0.15
10/28/2012, 11:30:23 AM	Link	A bunch of Jewish people just walked into the laundromat.	5.28
10/28/2012, 11:31:50 AM	Link	This Jewish boy just said "Ayo give me that shirt!" #slang	1.12
10/28/2012, 11:33:02 AM	Link	@notd14u and I must make a stop at the liquor store before this bitch hits! Imma hit her back if I can't get a drink.	1.03
10/28/2012, 11:34:22 AM	Link	@ThaiLe79 RUN!!! Lol get out of AC	3.04

User

Time (EST) 10/28/2012, 11:09:34 AM

Text Doing my laundry before the storm hits.

Time (UTC) Sun Oct 28 15:09:34 +0000 2012

Cluster: undefined, Speed: 0.99 mph

Link twitter.com/statuses/262571820434202625

Figure 3. Visualization Dashboard populated with 1798 geo-tagged tweets.

5.1. Web-Based Interactive Visualization

To view a user's tweet stream with geographic context, we created an interactive web map built on Mapbox-GL (mapbox.com/maps) embedded within a larger javascript-powered dashboard. This visualization tool shows a map alongside a scrollable timeline of a user's entire contextual stream. The dashboard is populated by loading a single GeoJSON file from a location specified in the URL, making it easy to share links for specific users among teams of analysts. As seen in Figure 3, geo-referenced tweets appear as interactive, clickable points on the map, while all of a user's tweets populate the accompanying user timeline on the side. Clicking on geo-referenced tweets in the timeline will pan and zoom the map to the location of the tweet. Above the map is a graph showing the volume of tweets overtime. The user depicted in Figure 3 tweeted more during Hurricane Sandy than any other time that year. Selecting a time range filters the tweets shown on the map to only those posted during that time while simultaneously auto-scrolling the contextual timeline on the side to the beginning of that time period. Analysts may also choose to filter tweets in the timeline to only those visible on the current map. In this way, both the map and the timeline are capable of driving the interaction of the other.

5.2. Clustering, Home Location, and Movement

Using a javascript implementation of DBScan[2], all of a user's tweets (not replies or conversations) are spatially clustered in the browser. This identifies areas of consistent posting. If users are consistently geo-tagging their tweets from a similar location, the noise in the GPS signal will show many points close together, but never in the same exact place. Conversely, geo-located tweets that represent the same location will appear with exactly the same coordinates, directly on top of one another. Clustering these points simplifies the noise and provides valuable additional information at a glance, such as which cluster does the user tweet from most frequently? As shown in Figures 3 and 4, clusters are differentiated by color.

Home location detection provides further context to the geographic footprint of a user's activity. Using only the time and geo-cluster associated with the tweet, we use the same approach as Krumm et al., identifying likely home times and then classifying the clusters with the most recurring Twitter activity during these times as their potential home location [28]. In practice, we have found that these locations may not be homes but instead gyms, work, or school. We found, however, that the accuracy of a *home* location is not as important as the identification of a location that represents normalcy during non-storm times [33]. Figure 3 shows the user's calculated home location as a transparent blue circle. Given the proximity to the shore (and the mandatory evacuation order of that area), we consider this user to be *geographically vulnerable* [10]. Using this dashboard, we found that this user tweeted from this calculated home location before the storm, then from a motel further inland during landfall, and then from the home location again after the storm, a strong signal of evacuation: all learned from the geographic metadata only, and then confirmed later by a single tweet days after the storm near the home location saying "It feels so great to be home. *sigh*". Additionally, the existence of any two geo-referenced tweets may imply movement between these two points. We create lines connecting each of a user's geo-referenced points as a separate feature that analysts can toggle on/off on the map to see a user's path of movement.

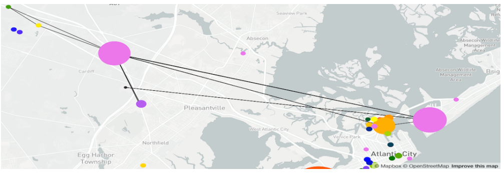


Figure 4. A User's common tweeting location clusters with movement between them.

One complication of working with these types of geo-referenced social media data is the meaning of the location metadata. As mentioned before, a geo-reference does not imply a user is actually at that location. The tweet could simply be *about* that location. These tweets are still important to provide context, but one must understand their intention. For example, many media or storm-tracking accounts geo-reference their tweets to the location of the story or the current path of a storm [34]. This practice ensures that these tweets show up in spatial queries. A simple filtering approach is to identify the speed a user traveled between any two consecutive geo-referenced points and determine if the point is a plausible physical location of the user, or if it's merely *about* the place. We use the approach demonstrated in [29] as reference, but we do not remove the data entirely, instead we show the calculated speed in the tweet timeline within the dashboard to provide context. If users are traveling thousands of miles per hour between their tweets then some of the geo-references must be *about* the place, not *from* the place. We leave this interpretation to the analyst.

6. Results

Our infrastructure was recently used in one of our research studies [33] to help identify evacuation behavior to create a training set for automated detection of evacuation behavior based both on textual clues in a user's tweets as well as their physical movements. The tool is currently deployed for various research efforts associated with [1] and continues to prove itself a stable, scalable social media post collection infrastructure. We currently have two machines running three user-crawlers. This system costs $3.33 per day and can download contextual streams for 30k users per day.

[2] github.com/Turfjs/turf/tree/master/packages/turf-clusters-dbscan

7. Discussion

The ability to use social media posts as a near real-time personal account of those experiencing a crisis or disaster is a powerful affordance of our modern information environment. All researchers in this space should be able to access and perform informed, contextualized analysis of these rich narratives. We created our data collection infrastructure to allow just that, but recognize that it is potentially at odds with official Twitter guidelines. Additionally, this infrastructure is less useful for those with expanded access to Twitter data as they may be able to perform more complex queries and may not be limited to a user's last 3,200 tweets. With that, we offer this infrastructure and discussion of potential difficulties to all researchers currently performing analysis of keyword-based tweet collections obtained from public Twitter APIs.

The distinction made here between geo-tag and geo-location is not common language, but the lack of this distinction has the potential to affect many social media studies. Further, distinctions about tweeting *from a place* versus tweeting *about a place* are difficult to automatically discern.

When offering a technical infrastructure that aims to lower the barriers to big social media data analysis, specifically an infrastructure that focuses on individual users and not the aggregate, we identify an additional responsibility to inform other researchers of the ethical gray areas associated with publicly accessible personal data and the associated privacy concerns. We find this particular conversation timely and appropriate given that the GDPR went into effect in May 2018. While all of our data is publicly accessible, we doubt that the users who appear in these datasets truly consent to their data being used in this manner. Twitter's terms of service clearly describe the potential of public tweets being collected and analyzed by third parties; however, research indicates that users remain unaware of how their data is actually used and likely do not consent [35].

8. Conclusion

We have presented a cloud-based Twitter data collection infrastructure based on collecting all tweets from and to specific users rather than specific keywords to allow for more contextualized analysis, especially in support of mixed-methods research approaches to social media data analysis. The infrastructure scales to maximize the number of users it can collect while working within the artificially imposed rate-limits of Twitter's public API endpoints. For tweets with geo-referencing information available, the infrastructure creates additional files that drive an interactive map-based visualization to provide additional geographic context for analysis. This work addresses concerns put forth by researchers [2, 6] with respect to the lack of contextualization and reliance on privileged access to data streams streams that dominates in social media research, specifically in the fields of hazards research and disaster response.

The infrastructure has been developed, used, and tested by crisis informatics researchers to better identify and understand the decisions made by social media users who are geographically vulnerable to impending hurricanes. Through further use, we anticipate that this infrastructure and the underlying philosophies and approach can support more contextualized, more powerful data science.

References

[1] R. E. Morss, J. L. Demuth, H. Lazrus, L. Palen, C. M. Barton, C. A. Davis, C. Snyder, O. V. Wilhelmi, K. M. Anderson, D. A. Ahijevych, J. Anderson, M. Bica, K. R. Fossell, J. Henderson, M. Kogan, K. Stowe, and J. Watts, "Hazardous weather prediction and communication in the modern information environment," *Bulletin of the American Meteorological Society*, vol. 98, no. 12, pp. 2653–2674, 2017.

[2] L. Palen and K. M. Anderson, "Crisis informatics-new data for extraordinary times: Focus on behaviors, not on fetishizing social media tools," *Science*, vol. 353, no. 6296, pp. 224–225, 2016.

[3] M. Kogan and L. Palen, "Conversations in the Eye of the Storm: At-Scale Features of Conversational Structure in a High-Tempo, High-Stakes Microblogging Environment," in *CHI '18*, 2018.

[4] Y. Kryvasheyeu, H. Chen, N. Obradovich, E. Moro, P. Van Hentenryck, J. Fowler, and M. Cebrian, "Rapid assessment of disaster damage using social media activity," *Science Advances*, vol. 2, no. 3, 2016.

[5] 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.

[6] K. Crawford and M. Finn, "The limits of crisis data: analytical and ethical challenges of using social and mobile data to understand disasters," *GeoJournal*, vol. 80, no. 4, pp. 491–502, 2015.

[7] J. L. Demuth, R. E. Morss, L. Palen, K. M. Anderson, J. Anderson, M. Kogan, K. Stowe, M. Bica, H. Lazrus, O. Wilhelmi, and J. Henderson, "sometimes da #beachlife ain't always da wave: Understanding peoples evolving hurricane risk communication, risk assessments, and responses using twitter narratives," *Weather, Climate, and Society*, vol. 10, no. 3, pp. 537–560, 2018.

[8] 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 28th international conference on Human factors in computing systems - CHI '10*, p. 1079, 2010.

[9] J. Anderson, M. Kogan, M. Bica, K. M. Anderson, K. Stowe, R. Morss, J. Demuth, H. Lazrus, and O. Wilhelmi, “Far Far Away in Far Rockaway: Responses to Risks and Impacts during Hurricane Sandy through First-Person Social Media Narratives,” in *Proceedings of the 13th International Conference on Information Systems for Crisis Response and Management (ISCRAM 16)*, no. May, 2016.

[10] M. Kogan, L. Palen, and K. M. Anderson, “Think Local, Retweet Global,” in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW ’15*, (New York, New York, USA), pp. 981–993, ACM Press, feb 2015.

[11] T. Onorati, P. Díaz, and B. Carrion, “From social networks to emergency operation centers: A semantic visualization approach,” feb 2018.

[12] M. Imran, C. Castillo, J. Lucas, P. Meier, and J. Rogstadius, “Coordinating Human and Machine Intelligence to Classify Microblog Communications in Crises,” in *Proceedings of the 11th International ISCRAM Conference*, (University Park, Pennsylvania), 2014.

[13] M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, and P. Meier, “Practical extraction of disaster-relevant information from social media,” *Proceedings of the 22nd International Conference on World Wide Web*, pp. 1021–1024, may 2013.

[14] F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, and K. Tao, “Twitcident: fighting fire with information from social web streams,” in *Proceedings of the 21st international conference companion on World Wide Web - WWW ’12 Companion*, (New York, New York, USA), pp. 305–208, ACM Press, 2012.

[15] J. I. White, L. Palen, and K. M. Anderson, “Digital mobilization in disaster response: The Work & Self-Organization of On-Line Pet Advocates in Response to Hurricane Sandy,” in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW ’14*, (New York, New York, USA), pp. 866–876, ACM Press, 2014.

[16] K. Starbird and L. Palen, “”Voluntweeters”: self-organizing by digital volunteers in times of crisis,” in *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI ’11*, (New York, New York, USA), pp. 1071–1080, ACM Press, 2011.

[17] D. Dailey and K. Starbird, “Social Media Seamsters: Stitching Platforms & Audiencies into Local Crisis Infrastructure,” in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW ’17*, (New York, New York, USA), pp. 1277–1289, ACM Press, 2017.

[18] Y. L. Huang, K. Starbird, M. Orand, S. A. Stanek, and H. T. Pedersen, “Connected Through Crisis: Emotional Proximity and the Spread of Misinformation Online,” in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW ’15*, (New York, New York, USA), pp. 969–980, ACM Press, 2015.

[19] K. Anderson, “Embrace the challenges: Software engineering in a big data world,” in *First International Workshop on Big Data Software Engineering*, pp. 19–25, May 2015.

[20] K. M. Anderson and A. Schram, “Design and implementation of a data analytics infrastructure in support of crisis informatics research,” *Proceeding of the 33rd international conference on Software engineering - ICSE ’11*, p. 844, 2011.

[21] K. M. Anderson, A. A. Aydin, M. Barrenechea, A. Cardenas, M. Hakeem, and S. Jambi, “Design challenges/solutions for environments supporting the analysis of social media data in crisis informatics research,” in *Proceedings of the Annual Hawaii International Conference on System Sciences*, vol. 2015-March, pp. 163–172, IEEE, jan 2015.

[22] S. Jambi and K. Anderson, “Engineering scalable distributed services for real-time big data analytics,” in *International Conference on Big Data Computing Service and Applications*, pp. 131–140, April 2017.

[23] A. A. Aydin and K. Anderson, “Batch to real-time: Incremental data collection & analytics platform,” in *Hawaii International Conference on System Sciences*, pp. 5911–5920, January 2017.

[24] R. Estrada and I. Ruiz, *Big Data SMACK: A Guide to Apache Spark, Mesos, Akka, Cassandra, and Kafka*. Apress, 2016.

[25] H. Hu, Y. Wen, T.-S. Chua, and X. Li, “Toward scalable systems for big data analytics: A technology tutorial,” *IEEE access*, vol. 2, pp. 652–687, 2014.

[26] M. Kiran, P. Murphy, I. Monga, J. Dugan, and S. S. Baveja, “Lambda architecture for cost-effective batch and speed big data processing,” in *Proceedings of the 2015 IEEE International Conference on Big Data (Big Data)*, BIG DATA ’15, (Washington, DC, USA), pp. 2785–2792, IEEE Computer Society, 2015.

[27] F. Girardin, J. Blat, F. Calabrese, F. Dal Fiore, and C. Ratti, “Digital footprinting: Uncovering tourists with user-generated content,” *IEEE Pervasive Computing*, vol. 7, no. 4, pp. 36–44, 2008.

[28] J. Krumm, R. Caruana, and S. Counts, “Learning likely locations,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7899 LNCS, pp. 64–76, 2013.

[29] R. Jurdak, K. Zhao, J. Liu, M. AbouJaoude, M. Cameron, and D. Newth, “Understanding human mobility from Twitter,” *PLoS ONE*, vol. 10, p. e0131469, jul 2015.

[30] Y. Martín, Z. Li, and S. L. Cutter, “Leveraging Twitter to gauge evacuation compliance: Spatiotemporal analysis of Hurricane Matthew,” *PLoS ONE*, vol. 12, no. 7, pp. 1–22, 2017.

[31] “Types of places on foursquare.” `support.foursquare.com/hc/en-us/articles/201065100-Types-of-places-on-Foursquare`. Accessed: 2018-05-20.

[32] “Tweet location metadata.” `developer.twitter.com/en/docs/tweets/data-dictionary/overview/geo-objects`. Accessed: 2018-05-20.

[33] K. Stowe, J. Anderson, M. Palmer, L. Palen, and K. M. Anderson, “Improving classification of twitter behavior during hurricane events,” 2018.

[34] M. Bica, L. Palen, and C. Bopp, “Visual Representations of Disaster,” in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW ’17*, (New York, New York, USA), pp. 1262–1276, ACM Press, 2017.

[35] C. Fiesler and N. Proferes, “Participant Perceptions of Twitter Research Ethics,” *Social Media + Society*, vol. 4, p. 205630511876336, jan 2018.