Identifying Firsthand Accounts of Emergencies from Social Media

Lucas A. Overbey*, Bryan Williams*, Scott C. Batson‡, LaKeisha Williams*, Benjamin Greco*, Seth Bartynski§, and Ana Carmen González§

*Space and Naval Warfare Systems Center Atlantic
North Charleston, SC, 29419-9022
‡Email: lucas.overbey@navy.mil
†Mathematics Department
North Carolina State University
Raleigh, NC, 27695
§Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA, 19104
¶Department of Mathematics
University of Puerto Rico at Mayagüez
Mayagüez, Puerto Rico, 00680

Abstract—With the recent explosion in availability and use of internet social media, citizens now utilize these resources to transmit information quickly about events as they unfold. However, for responding personnel in emergency situations, it is often difficult to sift through the enormous quantity of data within such sources to find the most pertinent information. The ability to filter messages is critical to, for example, identify firsthand accounts from persons within the direct vicinity of events. Nevertheless, on social media platforms such as Twitter, location-based information is often missing or unreliable. This paper outlines an approach to probabilistically identify the likely locations of individuals on Twitter based on their content, and from socially connected users with more reliable geographic information. We utilize measurements of user content similarity and Gaussian mixture modeling to infer “hotspots” of the likely location of users in emergencies. We are able to achieve upwards of 70% accuracy of Twitter user home cities without using any prior knowledge of geographic boundaries to look within.

Keywords—disaster response; social media; natural language processing; analytics; geographic forecasting

I. INTRODUCTION

The exponential growth and popularity of internet social networking sites has produced an alternative communication medium from which word-of-mouth communication is more easily observable. According to Alexa [1], 10 of the top 20 most popular sites globally can be classified as social networking sites, e.g., those that involve user-generated content such as personalized pages, blogs and micro-blogs, file/video upload services, and collaborative wikis. One such site, Twitter, is a free social networking and micro-blogging service that enables users to write messages of up to 140 characters, displayed on the authors profile page and delivered to the authors subscribers, or followers. Based on web traffic, Twitter is the eighth most popular website globally, with 71.1% of its traffic coming from outside the United States [1]. Since its inception, Twitter has increased dramatically in users and usage, currently seeing upwards of 500 million tweets per day [2]. Recent international events, including natural disasters, the cascade of protests and revolutions in the Arab world, and the Boston marathon bombings [3]–[6], have uncovered the utility of social networking sites like Twitter for dealing with both social unrest and emergency response. Twitter played a prominent role in these events, either through real-time local and global information sharing [7] or through attempts to coordinate or influence activities [6], [8].

Given this increased prominence, techniques need to be cultivated that can filter out pertinent information from amongst such big data by sifting through a large amount of irrelevant and even misleading information [6]. In particular, geographic proximity is a valuable tool for identifying persons with new, relevant, and valuable intelligence in such situations. Recent work has focused on the idea of amalgamating structured organizational data along with open source intelligence (OSINT) including social media data from multiple sources onto a common platform in the form of a crisis map [9]. However, users commonly conceal location-based content from being released on these social media applications [10], and either would not want to (in the case of social unrest) and do not always have the forethought or means to (in the case of natural disasters) change the availability of this location-based information when disaster strikes.

Some recent research has made attempts at identifying hidden geographic characteristics of Twitter users. Sadilek et al. [11] attempt to identify geophysical information on Twitter in New York City using friendship linkages to obtain predicted latitude/longitude coordinates on Twitter. This approach utilizes a combination of metrics, including a collocation score (based on geographic proximity at the same times) and a text similarity score. The collocation score $S$ is:

$$S(u, v) = \sum_{l_u, l_v \in L} \frac{t(l_u, l_v)}{d(l_u, l_v)}.$$  (1)
where \( L \) is the union of all locations from which users \( u \) and \( v \) are known to have been, \( t(l_u, l_v) \) is the amount of time \( u \) spends at location \( l_u \) while \( v \) is at \( l_v \), and \( d(l_u, l_v) \) is the distance between the tweets of persons \( u \) and \( v \).

Their text similarity score \( C_f \) is as follows:

\[
C_f(u, v) = \sum_{w \in W(u) \cap W(v)} f_u(w)f_v(w),
\]

where \( W(u) \) is the set of words that appear in user \( u \)'s tweets and \( W(v) \) is the set of words that appear in user \( v \)'s tweets. \( S \) is a set of stop words. This measure will find similarities between linked users’ text, but does not take into account common phrases or sequences (n-grams) or whether certain terms may contribute more information to a person’s geographic proximity (or other characteristic) relationships than others. Also, as this measure has no normalization, it will tend to be higher among users that are more prolific (tweet more) rather than just writing about the same things. Hidden locations are then predicted using unsupervised clustering. This approach in general leads to effectively noisy location sensors that form a scalable probabilistic model of human activity, with a focus toward exact locations of persons at given times as opposed to a single home location. However, this approach implies that a bounded search region is already known and that everyone outside of this search region can be ignored, that temporal information is treated discretely, and that users have at least some geolocation data available at given times.

Several other approaches have begun to emerge in recent months. Schulz et al. [12] utilize a multi-indicator approach that utilizes limited text content with named entities, dedicated (location) entry, and additional information from a user’s profile. Priedhorsky et al. [13] use word and phrase n-grams and a Gaussian mixture model (GMM). Mahmud et al. [14] aggregate a content-based statistical classifier with content-based heuristics and time zone behavioral information to determine a user’s home location. Still others [15], [16] have used language and dialect characteristics to infer the likelihood of where a person is from. All of these approaches make several assumptions, including that a user has some known tagged locations, users are located within a known set of geographic boundaries, and/or that geographic entities have been pre-established with known locations.

The goal of this work is to describe an approach for automatically inferring geographic information, with accuracy, for anonymous or semi-anonymous persons on social media platforms. We will leverage these existing models, manipulating them to the current application and integrating them into a combined geographic profiling scheme. Our approach uses known geographic information about connected users and the similarities of these users based on textual content to generate a probabilistic “hotspot” model of likely location(s) for the unknown user.

II. OUR APPROACH

In this work, we are interested in developing a measure that expresses likely locations of users without prior known geographic information or an a priori geographic ontology. Therefore, we will incorporate a hybrid approach that utilizes available geographic information from a subset of a user’s connections, measuring the content similarity between the user of interest and these connected users, and then applying that measure as a scaling to a GMM that can effectively determine “hotspots” of the likely home location(s) of the user. We focus on a home location here rather than any location at any given time, as we are assuming we have no information at all about a user or a user’s followers’ movements. We test several different forms of text similarity measures and apply them to a GMM, as outlined in the below subsections.

A. Text Similarity Measures

We describe several alternatives to the text similarity score from Eq. (2), which we outline below. Other metrics exist as well [17], [18] but suffer from similar problems to \( C_f \) described previously. The goal of our text similarity measures are to identify users that post about the same things, and are often stylistically (dialectally) similar as well. We ultimately want to be able to identify users that write locally-relevant information and how similar they are to other users that write about other (or the same) locally-relevant information. Three related text similarity measures are described below.

1) tf-idf Cosine Similarity: We first consider the cosine similarity of vectors formed from the term frequency - inverse document frequency (tf-idf) [19], a common approach in natural language processing information retrieval and text mining. The normalized term frequency for a given document can be expressed as:

\[
tf(i, j) = \frac{f(s_i, j)}{\sum_{k=1}^{K} f(s_k, j)},
\]

where \( f(s_i, j) \) is the number of occurrences of sequence \( s_i \) in document \( j \). \( D \) is the number of total documents that we are considering. Likewise, the inverse document frequency is

\[
\text{idf}(i) = \log \frac{D}{d|\forall s_i \in d|}.
\]

where \( d|\forall s_i \in d \) is the number of documents where the sequence \( s_i \) appears. The tf-idf is simply

\[
\text{tf-idf}(i, j) = tf(i, j) \times \text{idf}(i).
\]

The tf-idf is a numerical statistic that increases proportionally to the number of times a word appears in the document, offset by the frequency of the word appearance in the corpus of all documents. The inverse document frequency diminishes the weight of terms that appear commonly in a large number of documents such that more meaningful terms are emphasized for evaluating differences between them. Here, our documents are represented by the collection of tweets from a specific user.

Given the tf-idfs, the cosine similarity can be computed by:

\[
C_{\text{tf-idf}}(u, v) = \cos(\theta_{uv}) = \frac{\text{tf-idf}(u) \cdot \text{tf-idf}(v)}{||\text{tf-idf}(u)|| \cdot ||\text{tf-idf}(v)||},
\]

where \( \text{tf-idf}(u) \) is the vector (encompassing all considered terms/sequences) of all tf-idf scores for document \( v \) and \( \theta_{uv} \) is
the angle between these vectors. Unlike the measure used by Sadilek et al. [11] (Eq. 2), the cosine similarity of the tf-idf scales importance by uniqueness of terms across the corpus and normalizes the measure to be in the range $[0, 1]$. The tf-idf is usually computed for a set of given sequence lengths $l(s_i)$ ($n$-grams), typically between one and three. In our tests below, we compare the cosine similarities from tf-idf vectors for sequence lengths of one, two, and three individually.

2) $n$-gram tf-idf Cosine Similarity: An alternative approach is to formulate a text similarity measure that incorporates all possible $n$-grams in the corpus. This becomes computationally intractable for very large documents; however, for Twitter user document sets, this is not impossible. Twitter naturally limits the size of messages to 140 characters, and we can consider sequences only within individual messages. This approach has been used previously with character $n$-grams for general text categorization [20] and for Twitter language identification [21].

We can limit the total vector size further by only considering those sequences that appear in at least two different user document sets and are maximal sequences. Thus, Eq. (5) is computed the same way, but using a dictionary such that

$$W_{tot} = W_{a1} \cup \ldots \cup W_{a2} \cup \ldots \cup W_{a2}$$

where $\mathcal{A} = \{a_1, a_2, \ldots, a_{2^n}\}$ is the set that consists of all the two-document sequence combinations of any length $n$ in the $D$ documents. The similarity measure itself, $C_{tf-idf}(u, v)$, is computed from the cosine similarity as in Eq. (6). The term maximal refers to those sequences of terms of length $n$ such that there is not a larger sequence of length $n + 1$ for which that sequence is fully contained. We compare the different measures (including the Sadilek et al. [11] measure from Eq. (2)) in the Results (Section IV).

B. Hotspot Identification

In order to be able to model the probabilities of a geographically anonymous user’s likely locations, we wish to develop a geographic profile or “hotspot map” that describes local contours of likely presence. We wish to incorporate a text similarity measure described in the previous section to weigh other associated user’s known locations to infer the geographically anonymous user’s location.

1) GMM: Like Priedhorsky et al. [13], we employ a multivariate GMM so that we can sum the contributions of multiple distance metrics to obtain an overall geographic probability surface. This prior work used known geographic information about text-based content and limited geotagged information from the user itself to inform their measure. We wish to only use geographic information from other users to generate the hotspot map. A generalized example of a multivariate gaussian function is the following:

$$f(x) = \frac{1}{(2\pi)^{m/2} |\Sigma|^{1/2}} e^{-(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

where $\mu$ is the mean, $\Sigma$ is the covariance matrix, $m$ is the dimensionality of $x$, and $|\cdot|$ is the determinant. We can treat the multivariate gaussian function as the probable geographic location of a Twitter user, where $x$ is a two-dimensional vector representing the longitude and latitude. For a given user with an unknown location, we can treat each user whom that user connected to (through a follower/following relationship) as a noisy location sensor of the unknown user’s location. We can then sum the information gained by these noisy sensors, weighted by how much we believe they contribute to the user’s probable location:

$$f_i(x) = \sum_{j=1}^{L_i} \frac{C_{i,j}}{|\Sigma|^{1/2}} e^{-(x-\mu_j)^T \Sigma^{-1} (x-\mu_j)}$$

where $j$ are the $L_i$ followers of user $i$ and $C_{i,j}$ is a scaling factor based on that follower’s capacity to impact the probable location of user $i$. For this work, we will use the text content similarity measures described in Section II-A for this GMM scaling. The hypothesized implication is that users that are connected and write messages about the same things tend to be located in similar areas. For simplicity’s sake in the experiments conducted in the following sections, we set $\Sigma_{i,j}$ to the identity matrix.

2) Temporal Considerations: An issue which must be considered is that an individual $i$ has multiple tweets/check-ins/communications within a span of time, $t$. To estimate a single home location in such a case, we could calculate the latitude/longitude centroid $(\mu = (\mu_1, \mu_2))$ of a users geotagged tweets as:

$$\mu = \left(\frac{x_{1}^{(1)} + x_{1}^{(2)} + \ldots + x_{1}^{(m)}}{m}, \frac{x_{2}^{(1)} + x_{2}^{(2)} + \ldots + x_{2}^{(m)}}{m}\right)$$

where $x_{1}^{(j)}$ and $x_{2}^{(j)}$ corresponds to the latitude and longitude of person at a given time $j$. The problem with the above centroid is that it is theoretically possible for $\mu$ to not correspond to an actual location for which an individual $i$ ever is actually located. To avoid this, we can calculate $\mu$ from the Eq. (10) but call the actual centroid $\mu_{act}$ to be the point such that $d(\mu, x)$ is minimum for that user. We will use $\mu_{act}$ as a proxy for a single user location (e.g. “home location”) in geographical space.

III. DATA COLLECTION

Given the need to filter pertinent information based on firsthand accounts in emergencies, we begin our data collection approach with a manually chosen set of key Twitter accounts associated with natural disasters, emergency management, and emergency response. We selected a small sample of Twitter these accounts from different regions and organizations in two distinct countries: Australia and Canada. We chose these countries because they are both English-speaking, yet have very different environments and climates, such that their natural disaster response is focused on very different types of events. In particular, Australia is particularly concerned about fire dangers, while Canada is more concerned with winter storm and avalanche scenarios.

We applied snowball sampling [22] by performing a user query on these accounts using the Twitter search API. This query will return tweets from the selected users as well as mentions and retweets of those selected users. These additional
tweets are the one-degree away users in an ego-centered retweet/mention graph around our selected users [23]. We wish to be able to validate whether our predictions about likely geolocations of selected users are correct. Therefore, we downselected the users from this set to those users that have geotagging enabled on their Twitter accounts and have Australia or Canada listed as their profile locations, respectively. While not every tweet that they produce will have a longitude and latitude, these users have at least the possibility of having that information present for a given tweet. From this one-degree snowball sampling approach, we obtained 998 Australian users and 1,228 Canadian users that we will consider our “test users.” We perform our text similarity and hotspot analysis on each of these users to guess their locations.

From each of these test users, we then determined all of their friends (users following them) and followers (users they follow). We have now snowballed one-degree away in an ego-centered follower/following graph from our 2,226 test users. We then downselected from this set of users to those that have geographic profile or geotagged tweet information, and to users with less than 500 total followers (for computations sake). This downselection results in 542,739 and 690,064 users from the original Australian and Canadian users, respectively. We then perform a user query on these accounts with the Twitter search API over an extended period of time. We collected data for a month from each of these two datasets (Table I), and filtered our test and link users down to those users that have at least one geotagged tweet and a minimum of 10 tweets present in the set. The test user near centroid locations ($\mu_{u_{\text{act}}}$) and their linked friends/followers near centroid locations can be found in Figures 1 and 2 respectively. Note that while we restricted our test users to the two countries described, the linked users can be located anywhere in the globe. Thus, when we test whether we can identify one of the test users’ locations, we do not pre-suppose knowledge that they are in one of the two countries we started with.

<table>
<thead>
<tr>
<th>Properties</th>
<th>Australian Dataset</th>
<th>Canadian Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start Date</td>
<td>June 2, 2014</td>
<td>June 2, 2014</td>
</tr>
<tr>
<td>End Date</td>
<td>July 3, 2014</td>
<td>July 3, 2014</td>
</tr>
<tr>
<td>No. Users (min 10 tweets)</td>
<td>60,807</td>
<td>69,888</td>
</tr>
<tr>
<td>No. Geotagged Users (min 10 tweets)</td>
<td>6,443</td>
<td>6,697</td>
</tr>
<tr>
<td>No. Test Users (min 10 tweets)</td>
<td>136</td>
<td>180</td>
</tr>
<tr>
<td>Mean No. Followers / Test User</td>
<td>39.9</td>
<td>24.7</td>
</tr>
<tr>
<td>Median No. Followers / Test User</td>
<td>23</td>
<td>12.5</td>
</tr>
</tbody>
</table>

IV. RESULTS

A. Text Similarity Results

We first evaluate the text similarity measures described in Eq. (2) and Section II-A above. Here, we use the Spearman rank correlation to compare each of the measures for all users in our dataset with geotagged information (Table II). We also evaluate the correlation between each of these measures and the distance between the users nearest tweet to the geographic centroids ($\mu_{u_{\text{act}}}$) for each of these users. Almost all of the measures are negatively correlated with geographic distance, giving credence to the concept that one can scale geographic effect of other users based on text similarity. In this dataset, geographic distance is mostly negatively correlated with $C_{\text{tf-idf}}$ (1-gram), although the 2-gram and 3-gram versions are quite close. $C_f$ is not strongly correlated with geographic distance. The correlation between similarity measures tends to be fairly significant, with the weakest being between $C_f$ and others. $C_{\text{tf-idf}}$ (3-gram) is perhaps the mostly consistently strongly correlated with others, which is logical given that it is in between the lower n-gram and the infinite-gram approaches. Overall, it is difficult to determine which of these measures is optimal for estimating a relationship between text content and geographic distance by looking at the text similarity alone. Of particular note, the majority of the geotaggable users in this set were located in major Australian or Canadian cities and their suburbs. Therefore, the distances tend to cluster in disjoint groups based on city distances. It is likely that text similarity does a better job of distinguishing whether a user is from the same city rather than specific distance away.

B. Hotspot Forecasting Results

We apply Eq. (9) to the test user data from Table I, summed across the users’ friends and followers. We compute $f_i(x)$ for discrete longitudes and latitudes at increments of approximately 0.5 degrees. An example hotspot map is shown in Figure 3 for one of the Canadian test users. In this example, the user’s nearest tweet to their centroid location (in Edmonton, Alberta, Canada) is very near the same location as the largest $f_i(x)$ value in the GMM-based hotspot map. Therefore, the user was correctly identified to be in the city in which they are primarily located. However, the user has another location with increasing contours in Toronto, Ontario, Canada. Thus, the text similarity with users in that area was

![Fig. 1. Nearest tweet locations to the centroids ($\mu_{u_{\text{act}}}$) of each of the test users in our dataset for Australia (for the 136 test users from Table I). Test Users are in red and their connections are in blue.](image1)

![Fig. 2. Nearest tweet locations to the centroids ($\mu_{u_{\text{act}}}$) of each of the test users and their connections in our dataset for Canada (for the 180 test users from Table I). Test Users are in red and their connections are in blue.](image2)
also not insignificant. This secondary hotspot may indicate that
the user has affiliations and/or travels to that other location. Other friend/follower locations outside of these two cities do
not have these same visible increasing contours because the
text similarity was below significance relative to these two
locations.

Although the hotspot map contains additional information
about possible affiliations in multiple cities, we can also reduce
the hotspot map to a single prediction of a user’s likely
location by determining:

$$x_i = \max_x f_i(x) \tag{11}$$

This estimate gives a prediction as to a user’s most likely home
location. We can compare that prediction to the users’ known
nearest tweet to the centroid ($\mu_{act}$) of their geotagged tweets.

Figures 4 and 5 show the accuracy as a function of log distance.
Overall for Australia, our predictions result in guesses that are
within 50 kilometers (km) 57-64% of the time depending on
the text similarity measure employed. These predictions are
accurate to within 100 km about 64-74% of the time, and
with 500 km 74-80% of the time. The inflection points on
this graph are likely to be geography-specific and related to the
disjoint city-separation problem described in Section IV-A. For
Canada, the GMM hotspot predictions result in a guess that is
within 50 km 61-64% of the time, 100 km 68-75% of the time,
and 500 km 84-88% of the time. Because Canada has a lesser
population in between major cities, and these major cities are
fairly spread out, it has only one major inflection point at
around 100 km. This means that we are achieving fairly good

\begin{table}
<table>
<thead>
<tr>
<th>Measure</th>
<th>$d(\mu_{act, u}, \mu_{act, v})$</th>
<th>$C_F(u, v)$</th>
<th>$C_{\text{act}}(u, v)$</th>
<th>$C_{\text{act}}(u, v)$</th>
<th>$C_{\text{act}}(u, v)$</th>
<th>$C_{\text{act}}(u, v)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$d(\mu_{act, u}, \mu_{act, v})$</td>
<td>1.0, 1.00</td>
<td>-0.0129, -0.00366</td>
<td>-0.0594, -0.0518</td>
<td>-0.0431, -0.0194</td>
<td>-0.0234, -0.0146</td>
<td>0.0527, -0.0318</td>
</tr>
<tr>
<td>$C_F(u, v)$</td>
<td>-0.0129, -0.00366</td>
<td>1.0, 1.00</td>
<td>0.271, 0.390</td>
<td>0.254, 0.519</td>
<td>0.137, 0.315</td>
<td>0.0437, 0.148</td>
</tr>
<tr>
<td>$C_{\text{act}}(u, v)$ (1-gram)</td>
<td>-0.0594, -0.0518</td>
<td>0.271, 0.390</td>
<td>1.0, 1.00</td>
<td>0.765, 0.735</td>
<td>0.570, 0.627</td>
<td>0.118, 0.433</td>
</tr>
<tr>
<td>$C_{\text{act}}(u, v)$ (2-gram)</td>
<td>-0.0431, -0.0194</td>
<td>0.254, 0.519</td>
<td>0.765, 0.735</td>
<td>1.0, 1.00</td>
<td>0.826, 0.730</td>
<td>0.101, 0.281</td>
</tr>
<tr>
<td>$C_{\text{act}}(u, v)$ (3-gram)</td>
<td>-0.0234, -0.0146</td>
<td>0.137, 0.315</td>
<td>0.570, 0.627</td>
<td>0.826, 0.730</td>
<td>1.0, 1.00</td>
<td>0.0777, 0.251</td>
</tr>
<tr>
<td>$C_{\text{act}}(u, v)$ (∞-gram)</td>
<td>0.0527, -0.0318</td>
<td>0.0437, 0.148</td>
<td>0.118, 0.433</td>
<td>0.101, 0.281</td>
<td>0.777, 0.251</td>
<td>1.0, 1.00</td>
</tr>
</tbody>
</table>
\end{table}
users as noisy sensors for the user of interest’s predicted location. We were able to predict the city/suburb area of users in Australia and Canada at accuracies of 70-80%. As it stands, this algorithm could be incorporated into a larger platform such as a crisis map to assist with identifying messengers on social media with firsthand accounts of emergencies and crises. These firsthand accounts are much more likely to provide new, real-time information, and allow emergency management professionals to filter out secondhand or potentially misleading information.

Future research should investigate whether an aggregation of text similarity measures could improve results. This may not be the case given that the $\infty$-gram tf-idf-based measure is effectively an aggregation of all possible $n$-grams below it, and did not always outperform the others. Perhaps a filtering approach would instead be better, where a geographic ontology or named entity recognition (NER) is utilized to further limit the considered sequences to those that are most likely to relate to a person’s geography. Users also tend to have linked users that are also linked to each other in various fashions, and this information may provide even greater information into a user’s likely home location. For example, clustering users based on their joint geographic similarity or common social circles may either speed the algorithm or improve discernment into their likely contribution to an unknown, connected user’s location information.

**ACKNOWLEDGMENT**

The authors acknowledge support and funding through the Space and Naval Warfare Systems Center Atlantic NISE Basic and Applied Research (BAR) Program. The third author would like to acknowledge funding through the Department of Defense (DOD) Science, Mathematics and Research for Transformation (SMART) Program. The sixth and seventh authors would also acknowledge funding through the Office of Naval Research (ONR) Summer Research Programs.

**REFERENCES**


