Twitter Geolocation: A Hybrid Approach

JORDAN BAKERMAN, KARL PAZDERNIK, and ALYSON WILSON,
North Carolina State University
GEOFFREY FAIRCHILD and RIAN BAHRAN, Los Alamos National Laboratory

Geotagging Twitter messages is an important tool for event detection and enrichment. Despite the availability of both social media content and user network information, these two features are generally utilized separately in the methodology. In this article, we create a hybrid method that uses Twitter content and network information jointly as model features. We use Gaussian mixture models to map the raw spatial distribution of the model features to a predicted field. This approach is scalable to large datasets and provides a natural representation of model confidence. Our method is tested against other approaches and we achieve greater prediction accuracy. The model also improves both precision and coverage.

CCS Concepts: • **Information systems** → **Data mining**; • **Mathematics of computing** → *Density estimation*;

Additional Key Words and Phrases: Twitter, Geotag, Gaussian mixture model, simple accuracy error, comprehensive accuracy error, prediction region area

ACM Reference format:
Jordan Bakerman, Karl Pazdernik, Alyson Wilson, Geoffrey Fairchild, and Rian Bahran. 2018. Twitter Geolocation: A Hybrid Approach. *ACM Trans. Knowl. Discov. Data.* 12, 3, Article 34 (March 2018), 17 pages.
https://doi.org/10.1145/3178112

1 INTRODUCTION

Twitter was created in 2006 as a free online social networking service. The service allows its users to post messages with a maximum 140 characters called "tweets" without restriction on the number of tweets sent in any given time period. As of 2015, Twitter generates more than 500 million tweets daily, representing in excess of 300 million active monthly users (Ajao et al. 2015). The popularity of Twitter has spawned a variety of research efforts to use the micro-blogging service as a tool to support many applications, including event detection (Korkmaz et al. 2015), event monitoring (Gelernter and Mushegian 2011), and influenza diffusion (Generous et al. 2014). Each of these applications benefits from location information to identify and utilize relevant tweets.

This work is supported by the Consortium for Nonproliferation Enabling Capabilities (CNEC) Grant DE-NA0002576, which is sponsored by the Department of Energy, NNSA Office of Defense Nuclear Nonproliferation, and also National Science Foundation, under grant DGE-1633587.

Authors' addresses: J. Bakerman (corresponding author), K. Pazdernik, and A. Wilson, Statistics Department, North Carolina State University, Raleigh, NC 27695; emails: {jbakerm, ktpazder, agwilso2}@ncsu.edu; G. Fairchild, Analytics, Intelligence, and Technology Division, Los Alamos National Laboratory, Los Alamos, NM 87545; email: gfairchild@lanl.gov; R. Bahran, Nuclear Engineering and Nonproliferation Division, Los Alamos National Laboratory, Los Alamos, NM 87545; email: bahran@lanl.gov.

ACM 1556-4681/2018/03-ART34 $15.00
https://doi.org/10.1145/3178112

Each tweet can be geotagged with a latitude and longitude pair, either according to a user's cell phone geo-location service or the HTML5 geolocation API when a person tweets from their computer. However, Twitter users may remove this geotagging option, and most users do so. A recent study of Twitter showed that less than 1% of tweets are geotagged (Ajao et al. 2015). As a result, recent research has focused on estimating location information in the Twittersphere.

The possible active Twitter features are as follows: (1) *Geotagged coordinates* (latitude × longitude); **(2)** the *tweet* (140 characters maximum), which may include, words, symbols, toponyms, abbreviations, slang, and the like; (3) *Language*—each user chooses a language when joining Twitter, and each tweet is tagged with that chosen language; (4) *Location field*—when joining Twitter, users may specify their location with as much or little specificity as desired; that is, the user may specify *Raleigh, NC* or *my treehouse*, both are acceptable; (5) *Time zone*—the user may also choose their time zone; each tweet is then tagged with a generic time zone stamp; and (6) *Network*—Twitter users may "follow" other members of the community and converse with people directly using the *@username* syntax. It is through the *follower* and *followee* relationships or direct messaging that one can derive a Twitter user's network and use it as a model feature. If a particular tweet is intended for a specific user, the tweet field contains the user name. This direct engagement in an online conversation can be considered the requirement for two users to be "friends" and to create individual friendship networks.

The language, location, and time zone fields can be highly inaccurate because users may set these fields without protocol. Some authors use these fields as model features or as a test of accuracy for the ground truth and other authors only include tweets for model training with a real location field that can be mapped using a gazetteer (Abrol and Khan 2010; Cheng et al. 2010; Davis Jr. et al. 2011; Rout et al. 2013; Schulz et al. 2013). However, an "actual" location does not mean it is the user's true location. Only 66% of users provided accurate location field information, and nearly all at the city or state level (Hecht et al. 2011).

The current state of the Twitter geotagging literature can be described according to the model features. One segment uses the tweet text only, and another uses network information only, but there is little overlap between the two methodologies. The first segment mines the training set of tweets, which are tweets with true geotagged coordinates, to find *n*-grams with little geographic variability. In other words, these methods identify sets of co-occurring words that are geographically narrow and predictive of location. For example, the term *Celtics* is vernacular most used in the Boston area. The network segment, on the other hand, uses the distribution of friends to predict location. This notion is predicated on the idea that Twitter relationships mimic off-line relationships (Rout et al. 2013).

Regardless of the features used for location prediction, Twitter geotagging algorithms have evolved to consider prediction accuracy as the metric for quality. Whether classifying a tweet into a predefined region or predicting the latitude × longitude origin of a tweet, the current approaches can predict a tweet to within a few hundred kilometers. The literature is described more fully in Section 2.

In this article, we propose a hybrid method (Figure 1) that can harness the power of both text and network features. Our method proves to be highly competitive with the current state of the literature according to multiple metrics. We are able to pinpoint the origin of a tweet to within 19km more than 50% of the time. In addition, our approach has several key advantages compared to the literature.

(1) *We use a hybrid model.* We provide a method to incorporate both text and network information to obtain a more accurate location estimate. In addition, increasing the number of model features correlates to predicting a larger number of tweets. The hybrid approach

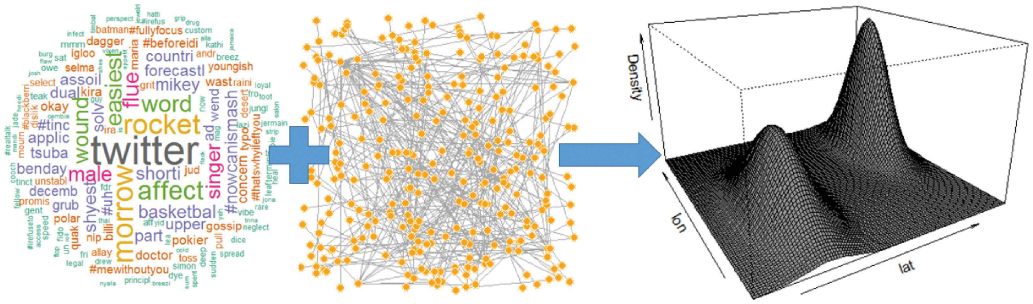

Fig. 1. The general approach to our hybrid method. Combine Twitter text and network information to predict location with a bivariate density estimate.

can still be used if at least one of the features, text or network, is available. If a tweet does not contain any predictive text, the hybrid model will use the network information only, and vice versa.

(2) *The text and network features are used jointly as predictors.* Our model weights the features according to their geographic scope. This contrasts to the only other hybrid model in the literature (Rahimi et al. 2015), which uses text and network features consecutively without regard for the predictive power of each feature.
(3) *We avoid using a gazetteer.* Filtering training tweets by accurate location field information can highly reduce the training set size. We use all tweets (divided into training and test sets) tagged with a latitude × longitude pair.
(4) *We avoid fabricated boundaries.* We treat geotagging as a coordinate prediction problem and not a classification problem. We avoid creating predefined geographic regions that are subject to data sparsity difficulties.
(5) *More complete preprocessing.* We take several text mining preprocessing steps to clean the data. We also remove outliers for the text features that skew predictions.
(6) *We use Gaussian mixture models (GMMs).*
    (a) *GMMs are scalable.* This modeling technique is computationally inexpensive and, as a result, scales to large datasets. It also allows for efficient testing of different component mixture models to find the best fitting model.
    (b) *GMMs are multi-modal.* Model features must be geographically narrow to represent a good predictor, but a single prediction region is not necessary. Multiple locations are estimated for a single tweet with associated probabilities to gauge each location's possibility as the origin. For example, the word *Burlington* is linked to over 20 cities in the United States.
    (c) *GMMs are interpretable probability distributions.* GMMs provide a natural way to interpret results and model error. Location estimates are bivariate probability distributions, visually displaying the most likely origin of a tweet and the model confidence in the estimate.

The article is organized as follows. In Section 2, we discuss the current state of geotagging literature and the need for a hybrid approach. Section 3 describes the necessary preprocessing steps to remove signal noise and Section 4 details the new hybrid algorithm. Section 5 shows the performance of the model and its utility compared to the current literature. Finally, Section 6 describes an example application to use geotagged Twitter data, and Section 7 presents conclusions.

2 RELATED WORK

The current geotagging literature can be grouped into three categories: text, network, and hybrid methods. We briefly describe the state of the literature for each category.

2.1 Text Approaches

In one of the first Twitter geotagging methods, Eisenstein et al. (2010) developed a latent variable model that treats the text and location of tweets as outputs from a generative process. The authors showed that lexical variation, and text in general, can be predictive of location. Furthermore, they created a freely available dataset of nearly 380,000 tweets that we use for model comparison with the literature. The dataset is described more fully in Section 5.

Subsequent text-based approaches regarded geotagging as a classification problem. Several authors (Chandra et al. 2011; Cheng et al. 2010; Hecht et al. 2011; Hulden et al. 2015; Kinsella et al. 2011; Roller et al. 2012; Wing and Baldridge 2011) predict location according to a region of varying granularities, such as country, state, city, zip code, or geographic areas defined by degrees of latitude and longitude. These methods use the frequency of words within regions to train a model and predict location using Bayes theorem. For example, naive Bayes is a simple and scalable method used to estimate the probability of a categorical event, such as the city origin of a tweet. The highest probability event is then regarded as the classification estimate.

The more recent text-based approaches move away from classification techniques to predict the latitude and longitude coordinate pair. The authors in Schulz et al. (2013) use a method called *polygon stacking* to estimate unique geographic regions. The authors use words that can be mapped to a region according to a gazetteer, which is a geographical dictionary. Words with geographical meaning, such as “Okemo Mountain,” are mapped to a physical location representing a geographic polygon. Each polygon for each word in the tweet is then stacked on top of each other and the center of the tallest plateau is considered the location estimate.

In a more intuitive approach, Priedhorsky et al. (2014) use GMMs to estimate geographic probability distributions. The authors show that the technique is scalable, multi-modal, geographically interpretable, and avoids artificial boundaries. For these reasons, the GMM is the basis of our new algorithm.

2.2 Network Approaches

There are sets of network-based methods that also considered geotagging as a classification problem. In the simplest network-based approach, Davis Jr. et al. (2011) classifies a tweet at the city level according to where the largest proportion of their friends are located. Abrol and Khan (2010) use the same method as Davis Jr. et al. (2011), except they consider the variable depth of a user’s network to classify at the city level. For example, at a variable depth of $d = 2$, the algorithm considers the locations of an individual’s friends, $F_1, \ldots, F_n$, and the locations of each of $F_1, \ldots, F_n$’s friends as well. The authors in Rout et al. (2013) use support vector machines to classify at the city level in the United Kingdom using features such as city population and the number of reciprocated tweets. In addition, Rout et al. (2013) scale the population density of each city to avoid weighting the model toward the largest city, i.e., London.

As with the text-based approaches, some methods focus on classification, while others try to predict location according to the latitude and longitude coordinate pair. Although not initially intended for Twitter, Backstrom et al. (2010) establishes one of the first geotagging methodologies for social media, specifically Facebook. The authors show empirically that the probability of friendship is inversely proportional to the geographic distance between two people and use these probabilities within a likelihood approach to estimate the most probable location of a user compared to the locations of a set of other users. The authors in McGee et al. (2013) extend this method to Twitter and use regression trees to first estimate the strength of friendship between users. The strength of friendship determines which probability curve is used within the likelihood.

The most effective network-based approaches use an extension of label propagation, a common algorithm found in network analysis literature. This is a simple iterative algorithm designed to infer labels for unknown items in a network based on the community structure of the network. The unknown item receives the most frequent label from the group of items with which it shares a connection. Jurgens (2013) proposes spatial label propagation to account for the geographic distribution of a network. This method predicts unknown locations as the geometric median of the observed locations from a user's network. In addition, the spatial label propagation algorithm process is repeated in order to expand the set of predictions that can be made. Predictions in previous iterations are used as ground truth for the subsequent iterations. Compton et al. (2014) extend spatial label propagation and consider the frequency of directed tweets as edge weights.

2.3 Hybrid Approach

The authors are aware of only one hybrid geotagging approach, which uses both text and network structure. Rahimi et al. (2015) borrow techniques from the text and network approaches to form a hybrid model. First, the authors discretized the geographic space using the method detailed in Wing and Baldridge (2014). Then, unigrams appearing in at least ten tweets are used as features to train a logistic regression model. The model is then used to predict the most likely cell for an unknown user and the center of the cell is the coordinate location estimate. Next, the authors use the spatial label propagation method proposed by Jurgens (2013). This iteratively adjusts the estimates from the text method by using the geometric median of locations of a user's friends to predict location.

In general, a hybrid approach is advantageous because it includes additional model features. When text features are not predictive, the model can rely on network information and vice versa. A drawback of the hybrid algorithm detailed by Rahimi et al. (2015) is that the algorithm uses text and network features independently and sequentially. In doing so, it uses predictions as ground truth before beginning the spatial label propagation algorithm and does not quantify model uncertainty. Our algorithm uses both model features jointly for prediction while accounting for model uncertainty. The next two sections provide details of the algorithm, and in Section 5, we illustrate that our method provides greater prediction accuracy when applied to a sample dataset.

3 PREPROCESSING

Prior to model creation, we filter the data extensively. These steps are designed to filter noise from Twitter data and allow discovery of relationships between the model features and the geographic coordinates.

3.1 Text Preprocessing

The four text preprocessing steps below, in combination, reduce the number of unigrams present in the data, and increase the rate of sparse words. For example, the words *cool* and *COOOOL* are considered the same unigram. Therefore, we obtain two instances of a single unigram instead of one instance of two separate unigrams. As a result, we are able to supply our subsequent analyses with more data. To illustrate the steps, consider the following tweet as an example.

— Hey @kpazphd OMG! my new car is SWEEEET!!! #TESLAlife watch out Raleigh! livin the dream. . .

First, we remove all special characters, emojis, and punctuation from the text except for the @ symbol, which indicates a direct tweet, and the hashtag (#) symbol. In Twitter, the hashtag is used before keywords or phrases to denote a topic or theme. This effectively lets users categorize tweets and allows for simple queries to monitor trending topics in the Twittersphere. For example, *#blueandblack* and *#whiteandgold* were tweeted more than 4.4 million times over a two day period as people debated the color of a peculiar dress online. The hashtag gives the phrase an entirely different meaning and context than if we considered the phrases *blue and black* and *white and gold* as is. Simultaneously, we also reduce all characters to lower case to avoid case sensitivity. We assume case bears no predictive power.

- hey @kpazphd omg my new car is sweeeet #teslalife watch out raleigh livin the dream

Next, we remove all stop words because we assume there is no association between the most common words used in Twitter and geographic location. Stop words are simply the most common words in a given language. For example *the*, *is*, and *no* are common English stop words. However, there is no universal set of stop words. We combine two dictionaries for this preprocessing step. We combine the SMART stop word list, a list of 571 words developed by Cornell (Feinerer et al. 2008), and also the top 500 most common Twitter words uncovered by TIME magazine (TIME 2009). The second list is used to remove common slang terms such as *lol*, *omg*, and *wtf* from the Twitter feed.

- hey @kpazphd car sweeeet #teslalife watch raleigh livin dream

Next, we perform a spell check on the remaining words that do not start with the @ or # symbol. Thus, we assume the association between misspelled words and geographic location is the same as correctly spelled words and geographic location. Each remaining word is compared to an English dictionary for accuracy. If the word is misspelled, it is replaced with the closest matching word according to the Jaccard coefficient for strings. The Jaccard coefficient simply measures the similarity between two sets, $|A \cap B|/|A \cup B|$. In the context of language processing, A and B are two separate unigrams for which the letters are compared.

- hey @kpazphd car sweet #teslalife watch raleigh living dream

Finally, we apply the Porter stemming algorithm, a widely used English stemming algorithm (Porter 1980). This reduces each remaining word to its root. For example, the words *jumping*, *jumper*, and *jumped* are set to the root word *jump*.

- hey @kpazphd car sweet #teslalife watch raleigh live dream

3.2 Text and Network Preprocessing

The unigrams present after the preprocessing steps above are then used to create unigram variables. Each unigram variable simply represents the locations (in latitude × longitude coordinates) from which it was tweeted.

Likewise, we create each user's network by considering *direct* tweets the barometer for friendship. A directed tweet is equivalent to sending a message to a specific person in a public forum. Directed tweets are performed using the *@username* syntax to refer to a specific user. That is, *userA* sends a direct tweet to *userB* by simply including the syntax *@userB* in the tweet. In one study, Huberman et al. (2009) shows that approximately 25% of posts are directed tweets. To create *userA*'s network, we find every instance when *userA* sends a direct tweet. The users receiving the direct tweet are considered *userA*'s friends. We then search for tweets initiated by these friends that are

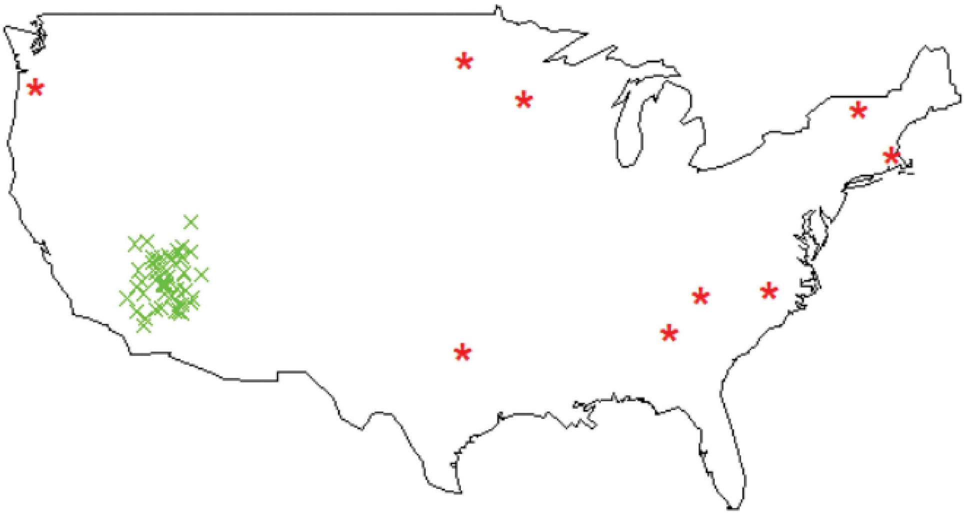

Fig. 2. A set of text or network coordinates. The ★ observations are considered outliers and removed before the Gaussian mixture model is created.

also geotagged and include them in the network. *userA*'s network variable is now associated with a set of coordinates (latitude × longitude). This process is repeated for each user.

A final preprocessing step is applied to both the text and network variables. We cluster the data into subpopulations for outlier prepossessing using the well-known k-means clustering procedure. This algorithm clusters data into k clusters and finds their centers by minimizing the sum of Euclidean distances of all data points to their respective centers. To estimate the number of appropriate clusters for each variable, we use the Gap statistic (Tibshirani et al. 2001). This technique chooses k by finding the largest discrepancy of the pooled within-cluster sum of squares, W_k, for the data and the expectation of W_k under a null distribution,

$$\underset{k \in \{1,\ldots,20\}}{\text{argmax}} \quad Gap(k) = E\{\log(W_k)\} - \log(W_k).$$

For simplicity, we use a uniform distribution over the range of observed data for each variable.

If an observation is farther than the mean distance from the center of its subpopulation, it is considered an "outlier" and removed from the dataset (Aggarwal and Singh 2013). For tightly clustered data, this removes subpopulation anomalies. Figure 2 illustrates this technique.

In addition, we remove all text variables with less than 10 observations to avoid modeling sparse data in the next section. This final technique is widely adopted in the geotagging literature (Eisenstein et al. 2010; Hulden et al. 2015; Priedhorsky et al. 2014; Wing and Baldridge 2011).

4 NEW HYBRID MODEL

Our hybrid model is constructed using a series of GMMs. We first create a bivariate density estimate for each unigram u_i and network n_j. Here, i and j represent the index for the unique unigrams and networks, respectively. Then, we combine the appropriate unigram and network GMMs using an intuitive reweighting algorithm to create a final GMM, $h(\cdot)$. This final model is a geographic probability distribution representing an estimate of the origin location of a non-geotagged tweet. The rest of this section details the creation of $h(\cdot)$.

Let $y \in \mathbb{R}^2$ be a single pair of coordinates, latitude × longitude. For each unigram u_i, we fit a GMM to the observed locations

$$g(y|u_i) = \sum_{k=1}^{c_i} \phi_{ik} N(y|\mu_{ik}, \Sigma_{ik}), \tag{1}$$

where N is the bivariate normal density function with parameters μ_{ik} and Σ_{ik} as the mean vector and covariance matrix, respectively. In addition, ϕ_{ik} are the mixture weights that combine the subpopulation probability distributions into a single density estimate. The subscript k represents the individual mixture components for each GMM and c_i is the number of mixture components for unigram u_i. The parameters ϕ_{ik}, μ_{ik}, and Σ_{ik} are estimated using the expectation maximization algorithm from the *MCLUST* package in *R* (Fraley et al. 2016). The volume, shape, and orientation of each mixture are free to vary, which implies that all three covariance matrix parameters must be estimated.

The authors in Priedhorsky et al. (2014) use a simple heuristic approach to choose the number of components for each GMM, $\min(20, \log(s)/2)$, where s is the sample size. Instead, we choose the number of components, c_i, which yields the best Bayesian Information Criterion (BIC) for the GMM when testing components $1, \ldots, 20$ individually,

$$\underset{c_i}{\arg\min} \left(-2 \sum_{b=1}^{s_i} \log \left[\sum_{k=1}^{c_i} \hat{\phi}_{ik} N(y_b|\hat{\mu}_{ik}, \hat{\Sigma}_{ik}) \right] + p \log(s_i) \right).$$

Here, p represents the number of parameters to be estimated and s_i is the sample size. We use the BIC to balance both model fit and model complexity.

Similarly, define the GMM for a user's network as

$$f(y|n_j) = \sum_{k=1}^{c_j} \theta_{jk} N(y|\mu_{jk}, \Sigma_{jk}). \tag{2}$$

The number of components, c_j, as well as the parameters θ_{jk}, μ_{jk}, and Σ_{jk} are estimated in the same fashion as before. Recall from Section 3 that the data used to fit each network GMM are the locations where users within the network tweeted. That is, the coordinates of the network are targets to fit the GMM and estimate parameters similarly to the unigram GMMs.

Next, we combine the applicable text and network GMMs above to create a final density estimate of the origin of a tweet. Let m be the set of unique unigrams that appear in a single tweet for which individual density estimates, $g(y|u_i)$, were created. Define the hybrid model for a given tweet, m, and the network of the user initiating the tweet, n_j, as

$$h(y|m, n_j) = \delta_0 f(y|n_j) + \sum_{l=1}^{T_m} \delta_l g(y|m_l), \tag{3}$$

where $\{\delta_0, \ldots, \delta_{T_m}\}$ represents the mixture weights. Specifically, δ_0 is the network GMM weight and $\{\delta_1, \ldots, \delta_{T_m}\}$ are the text GMM weights. In addition, T_m is the cardinality of m and m_l represents the lth element of the set m. That is, m_l refers to a unique unigram of a single tweet.

A useful text and network location estimator, from Equations (1) and (2), respectively, should be geographically narrow, i.e., the subpopulations it models should be small in area. Thus, we weight the hybrid density estimate in Equation (3) toward the GMMs with small prediction areas while also balancing the probabilities of the mixture components. We first define the weights for the network GMM, δ_0^*, and unigram GMMs, δ_l^*, as the inverse weighted average of the mixture probabilities (θ_{jk} and ϕ_{ik}) and the area of the highest $100 \times (1 - \alpha)\%$ density of each mixture component. The area of each mixture component is defined as $\pi \chi_2^2(\alpha) det(\Sigma)^{1/2}$ and is simply a function

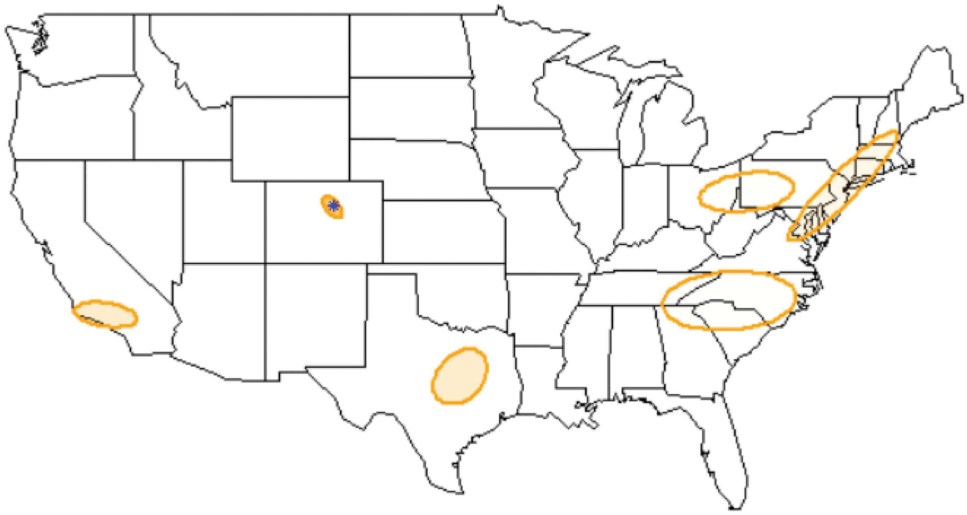

Fig. 3. An example of the hybrid model density estimate, $h(y|m, n_j)$, predicting location of a single tweet. Each ellipse models a separate subpopulation and the darker the transparency the higher the probability. The true origin of the tweet is marked by the ⋆ symbol.

of the cumulative density, $\chi_2^2(\alpha)$, and the shape and magnitude of the ellipse, $det(\Sigma)^{1/2}$, containing $100 \times (1 - \alpha)\%$ of each separate bivariate density. The initial weights δ_0^* and δ_l^* are calculated as

$$\delta_0^* = \frac{1}{\sum_{k=1}^{c_j} \theta_{jk} \pi \chi_2^2(0.05) det(\Sigma_{jk})^{1/2}},$$

$$\delta_l^* = \frac{1}{\sum_{k=1}^{c_i} \phi_{ik} \pi \chi_2^2(0.05) det(\Sigma_{ik})^{1/2}}. \tag{4}$$

To ensure that the hybrid model is also a GMM, we normalize the weights, $\{\delta_0^*, \ldots, \delta_{T_m}^*\}$, for each prediction. Thus, the weights used in Equation (3) are

$$\{\delta_0, \ldots, \delta_{T_m}\} = \left\{ \frac{\delta_0^*}{\sum_{l=0}^{T_m} \delta_l^*}, \ldots, \frac{\delta_{T_m}^*}{\sum_{l=0}^{T_m} \delta_l^*} \right\}.$$

To illustrate the weighting algorithm, consider the location prediction of a tweet with a network and only one unigram. Let each GMM, $f(y|n_j)$ and $g(y|u_i)$, be a two component model with the mixture probabilities and associated mixture component areas $\{(\theta_{j1}, \theta_{j2}) = (0.1, 0.9), (A_{j1}, A_{j2}) = (15\text{km}^2, 5\text{km}^2)\}$ and $\{(\phi_{i1}, \phi_{i2}) = (0.5, 0.5), (A_{i1}, A_{i2}) = (10\text{km}^2, 10\text{km}^2)\}$, respectively. Note, A_{jk} and A_{ik} correspond to the areas calculated as a function of Σ_{jk} and Σ_{ik}, respectively. Using Equation (4) to calculate the weights, $\delta_0^* = 1/6$ and $\delta_1^* = 1/10$ ($\delta_0 = 0.625$ and $\delta_1 = 0.375$ normalized). In this case, we weight the hybrid model toward the network estimate, even though the total area for each GMM is the same, because 90% of the data comes from a subpopulation with a small area, 5km^2.

Figure 3 displays one such estimate of $h(y|m, n_j)$. In this instance, the highest probability subpopulation corresponds to the user's network, which is near Denver, Colorado. Here, the user's network is the best predictor of the origin of the tweet. The text portion of the model relates unigrams to other subpopulations throughout the country with lower probability.

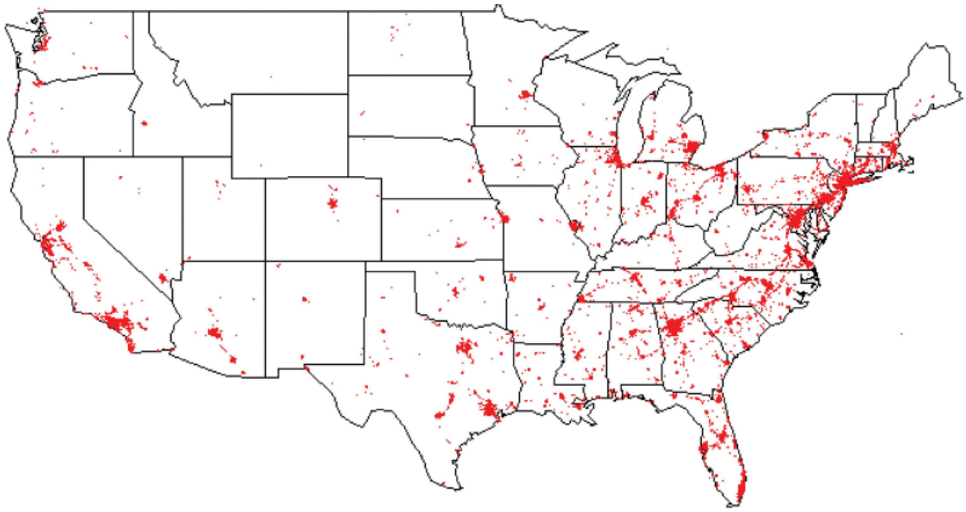

Fig. 4. Locations of all tweets within the Eisenstein dataset.

5 RESULTS

In general, the hybrid model, $h(y|m, n_j)$, estimates the probability of each point on the Earth's surface, y, being the origin of the tweet, given the text contained in the tweet, m, and the user's friendship network, n_j. In this section, we test our hybrid model on a common dataset from the geotagging literature and evaluate the hybrid model according to several metrics. The metrics we use are described in Priedhorsky et al. (2014), who also use GMMs for geotagging.

5.1 Data

The dataset we use was collected by Eisenstein et al. (2010). The data was accumulated in March 2010 from Twitter's "Gardenhose" Streaming API, which was a 15% sample of all daily messages. The authors kept only geotagged data within the contiguous United States. In addition, they filter the remaining data to include users following and followed by less than 1,000 people in an attempt to avoid celebrities. The final dataset consists of approximately 380,000 tweets and 9,500 users. The locations of the tweets from this dataset are displayed in Figure 4.

For subsequent analyses, we randomly split the data into 90% for training and 10% for testing. The training data is preprocessed and a model is fit as described in Sections 3 and 4. The test data is filtered through the same text preprocessing steps to maintain consistency between the two sets.

5.2 Performance Metrics

The *simple accuracy error* (SAE) metric measures the distance from the most probable location to the origin of the tweet. For example, the highest density estimate in Figure 3 coincides with an estimate near the actual location. Letting $d(\cdot)$ be the great-circle distance function and y' be the true origin of the tweet, the SAE is defined as

$$\text{SAE} = d\left(\arg\max_{y} \; h(y|m, n_j), \quad y'\right). \tag{5}$$

This metric is directly comparable to prior work, most of which forego geographic probability distribution estimates. Table 1 contrasts the SAE for our model compared to others in the literature.

Table 1. The Simple Accuracy Error Metric Compared to the Literature Using the Eisenstein Dataset from 2010

Algorithm	Features	Mean SAE	Med SAE
Eisenstein et al. (2010)	Text	845km	501km
Wing and Baldridge (2011)	Text	967km	479km
Roller et al. (2012)	Text	897km	432km
Hulden et al. (2015)	Text	765km	357km
Priedhorsky et al. (2014)	Text	923km	645km
Rahimi et al. (2015)	Hybrid	654km	151km
New hybrid model	Hybrid	593km	19km

Results are reported in kilometers using the great-circle distance function.

We achieve a median prediction error of only 19km on the test dataset, considerably better than most of the literature. Although the hybrid model is multi-modal by default, our algorithm maintains highly accurate single point estimates. In addition, we have a mean SAE of 593km, which also outperforms the other algorithms in Table 1. The difference in the mean and median SAEs indicates that the origin of some tweets are highly difficult to predict, skewing the overall results.

Table 1 also shows that our algorithm outperforms the only other hybrid model in the literature by a factor of nearly 8. Recall, the method presented by Rahimi et al. (2015) uses the text and network features sequentially. The authors use the text features to predict locations without a network and then perform spatial label propagation to finish the algorithm. Our model, on the other hand, uses both features jointly and weights each according to their geographic distribution. Spatial label propagation is competitive with our hybrid algorithm (Compton et al. 2014). However, the authors heavily filter the data and only predict approximately 10% of the data for competitive results. Our algorithm is able to successfully predict nearly 99% of the test data.

To account for the geographic distribution of subpopulation mixtures, the *comprehensive accuracy error* (CAE) is a measure of the expected distance between the true origin of the tweet and a random point generated from our model. There is no requirement for subpopulations to be clustered near each other for $h(y|m, n_j)$ to be a good estimator. However, this metric indicates whether or not the best subpopulation mixture is large in probability, small in area, and near the true location. It measures how accurate the density estimate is entirely, as opposed to a single best guess. The CAE is defined as

$$\text{CAE} = E_h[d(y, y')] = \int_y d(y, y')h(y|m, n_j)dy. \tag{6}$$

To estimate the integral in Equation (6), we use the Monte Carlo method, CAE $\approx \frac{1}{|z|} \sum_{y \in z} d(y, y')$, where z is a sample from $h(y|m, n_j)$ of size 100 in our testing.

The *prediction region area*, denoted PRA_α, is the area encompassed by $100 \times (1 - \alpha)\%$ density of $h(y|m, n_j)$. There are multiple ways to calculate PRA_α. We sum the area covered by the highest $100 \times (1 - \alpha)\%$ density of each mixture contributing to $h(y|m, n_j)$. The area of each mixture is calculated as a function of α and its covariance matrix Σ:

$$\pi \chi_2^2(\alpha) det(\Sigma)^{1/2}. \tag{7}$$

A small prediction region is ideal yet unserviceable if it rarely covers the true location. Therefore, the PRA_α performance is assessed concurrently with the *coverage*, COV_α, or the proportion of times the prediction region covers the true origin of the tweet. To calculate COV_α, we simply measure

Table 2. The Comprehensive Accuracy Error, Prediction Region Area, and Coverage Metrics Compared to Priedhorsky et al. (2014) because Both Algorithms use Gaussian Mixture Models

Metric	Priedhorsky et al. (2014)	New hybrid
Mean *CAE*	1,445km	557km
Median *CAE*	1,205km	117km
Mean $PRA_{0.95}$	3,156,517km^2	251,120km^2
Median $PRA_{0.95}$	2,846,610km^2	183,954km^2
$COV_{0.95}$	0.99	0.94

The weighting scheme applied by Priedhorsky et al. (2014) is the sum of the product of the elements in each covariance matrix of each GMM.

the proportion of times the true origin of the tweet is within the ellipses defined by PRA_α for the test set. A geographic point $y' = (y'_1, y'_2)$ is within the boundary of an ellipse with center μ and covariance matrix Σ if

$$(y' - \mu)^T \Sigma^{-1} (y' - \mu) \leq \chi^2_{2(\alpha)}. \tag{8}$$

The *comprehensive accuracy error*, *prediction region area*, and *coverage* metrics are unique to models with geographic distribution estimates. As a result, Table 2 compares these metrics with Priedhorsky et al. (2014) only.

As expected, the hybrid model heavily outperforms the text-based GMM method. The expected distance between the true origin of the tweet and a random point generated from our model is less than 117km, 50% of the time. Similar to the discrepancy between the mean and median SAE, the difference between the mean and median CAE indicate the predictions are skewed. However, this simply means that some subpopulations in $h(\cdot)$ are a large distance from one another. For example, if a person moves from Atlanta, Georgia to Las Vegas, Nevada, it is likely the hybrid model will estimate a subpopulation for each location, causing the CAE to be large by default. More importantly, the bimodal structure of the model captures both possible locations.

The mean and median $PRA_{0.95}$ are approximately 200,000 km^2. Although this is a large geographic area, it is more than 10 times smaller than the $PRA_{0.95}$ of Priedhorsky et al. (2014). To make it tangible, consider that the median prediction region of Priedhorsky et al. (2014) is more than a third of the contiguous United States. The median prediction region for the hybrid model, on the other hand, is approximately the geographic area of only North and South Carolina combined. Additionally, the coverage for the hybrid model is nearer to the nominal level in our experiments, whereas the previous method has much higher than nominal coverage due to extremely large prediction regions.

Achieving nominal coverage for the hybrid model was challenging. Initial experiments and methodology underestimated this metric by 3–14%. We eventually chose to sum the area covered by the highest $100 \times (1 - \alpha)\%$ density of each mixture contributing to the hybrid model. This technique achieves a coverage nearly equal to the nominal rate and also keeps the *prediction region area* metric small.

Coverage is also closely associated with the outlier preprocessing described in Section 3.2. Failure to remove outliers widens the GMM's prediction regions causing the coverage to inflate. We also examined the effectiveness of removing observations

- farther than 200 miles from at least five other observations,
- within a cluster of less than five observations where the number of subpopulations is determined by the "elbow" phenomenon in the k-means clustering algorithm.

We ultimately chose to remove observations from both text and network variables farther than the mean distance from their respective subpopulation center. Although this technique provides similar results to the aforementioned methods, it is an automated routine for removing outliers.

6 EXAMPLE APPLICATION

One of the more common applications of Twitter data is the monitoring and prediction of influenza and disease outbreaks around the world. In the United States, practitioners use daily geotagged tweets containing influenza relevant keywords, such as flu, fever, cough, and the like, to monitor the health of the country (Lee et al. 2013). The authors in Achrekar et al. (2011) show daily word counts relevant to influenza that can be used as a leading indicator to predict the Center for Disease Control and Prevention's (CDC) influenza-like-illness (ILI) reports. That is, the frequency of influenza keywords are used as a proxy for the true rate of influenza among US citizens.

In Brazil, researchers are using Twitter to monitor the rate and study the diffusion of Dengue fever, a mosquito-borne illness that can lead to death if untreated. In countries without efficient government agencies to monitor the spread of disease, it is important for researchers to develop tools that can identify regions of disease outbreak in order to allocate resources properly. Gomide et al. (2011) use geotagged tweets relevant to Dengue fever to cluster data into regions. The frequency of tweets is used as a barometer for the severity of Dengue fever within each region. In an effort to utilize more information, Davis Jr. et al. (2011) classify non-geotagged tweets at the city level using network information. This allows the authors to use Dengue fever relevant tweets that are not geotagged in their analyses.

In both applications, influenza in the United States and Dengue fever in Brazil, the hybrid method is advantageous. First, the hybrid method allows practitioners to geotag more data because both the text and network features are available to be used as possible predictors. In our experiments in Section 5, the hybrid method was able to geotag 98.2% of test tweets. Using text or network features independently, only 79.7% and 90.1% of test tweets were able to be geotagged, respectively. Furthermore, the use of model uncertainty provides more information regarding the location of disease outbreak. That is, the smaller the prediction regions of the hybrid method, the more confident we are in the estimated location. Additionally, the greater the certainty in the location of a set of geotagged tweets, the greater the evidence for allocating resources to specific regions.

As an example, consider monitoring influenza in the United States using the Eisenstein dataset from Section 5. Recall, this dataset was accumulated in March 2010 from the Twitter API and only geotagged tweets in the contiguous 48 states were considered. From the 376,510 tweets, only 1,031 contain influenza-related keywords. As evidenced by the performance metrics, the hybrid model can be used to accurately predict the location of influenza-related tweets. However, the uncertainty associated with the estimates is ignored if the best predictions are simply binned into regions.

To account for the uncertainty in the predictions, suppose that we stack the geographic density estimates, $h(\cdot)$, for each predicted tweet and reweight to obtain a final geographic probability distribution of the prevalence of influenza throughout the country. Let $h_q(\cdot)$ be the qth hybrid model density estimate of a sample of non-geotagged tweets S. The final distribution that combines all predictions and uses the uncertainty of each is $\sum_{q=1}^{|S|} h_q(\cdot)/|S|$. Figure 5 displays the estimated distribution of influenza in the United States during March of 2010 according to the Eisenstein dataset.

To compare the estimated distribution of influenza to ground truth, we use the weekly CDC influenza report for the week of February 28–March 6 2010 (CDC 2010), the week the Eisenstein data was collected. In general, these reports summarize by region the percentage of medical patients

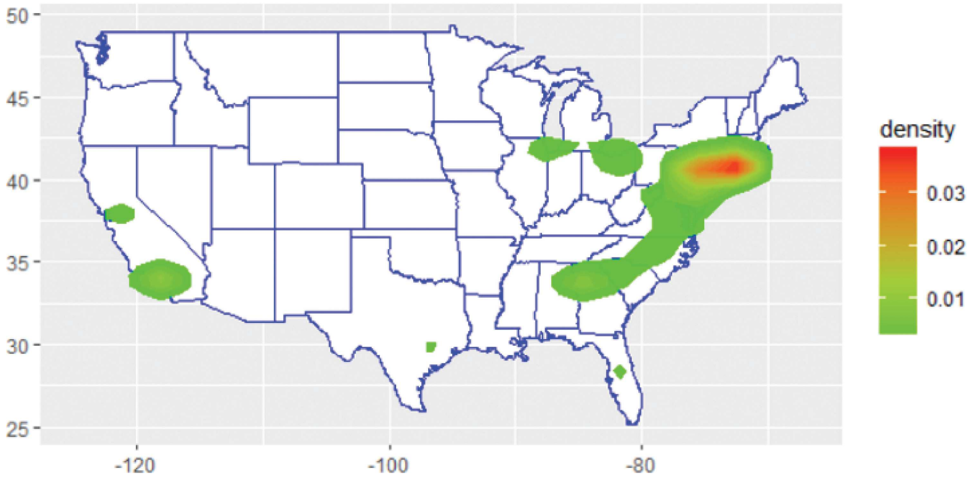

Fig. 5. Geographic probability distribution of influenza in the United States during March of 2010.

with ILI, percentage testing positive for influenza, and the number of jurisdictions reporting influenza activity. For the first week of March, only region 1 (CT, ME, MA, NH, RI, VT) and region 4 (AL, FL, GA, KY, MS, NC, SC, TN) reported widespread influenza activity. In addition, the Midwest and Southwest experienced sporadic influenza activity, and the Northwest reported no influenza activity. The report is similar to the estimated distribution in Figure 5. That is, there is elevated influenza in the northeast and southeast regions. In addition, there is activity in the Midwest and no activity in the Northwest. However, Figure 5 indicates possible elevated influenza activity in southern California; the CDC report does not support this finding.

Note, the results here do not simply imitate the most dense locations of Twitter users from the Eisenstein dataset (Figure 4). Also, Twitter usage has grown significantly since the data was collected in 2010, but there is no reason to suspect the application is no longer suitable.

7 CONCLUSION

In this article, we presented a hybrid geolocation model for Twitter. Our model exploits both text and network features and weights the features according to their geographic scope. Our method outperforms other geotagging algorithms according to four metrics. In particular, the median distance between the most probable location to the true origin of a single tweet is only 19km on average in our experiments.

Additionally, our hybrid model is one of only two geotagging methods that quantifies uncertainty using GMMs. It estimates the probability of each point within a spatial domain of being the true origin of a tweet. This structure allows us to visually interpret model confidence for a single tweet (Figure 3) or combine the uncertainty for a set of geotagged tweets to monitor an event such as influenza (Figure 5).

Any analysis of Twitter data is subject to the biases and limitations of the data itself. The limitations of Twitter data in general are attributable to the demographics of its users, Twitters streaming API sampling scheme, and users turning off their location services when tweeting. First, a 2012 investigation by the Pew Research Center found that age is inversely related to the likelihood of using Twitter (Duggan and Brenner 2013). Internet users in the age group 18–29 are the most likely to use Twitter, and users 65 or older are the least likely. Women are more likely than men and urban residents are more likely than suburban or rural dwellers to use Twitter. Also, different levels of education and levels of household income correspond to similar rates of Twitter use. As a result of the Pew study, we know any data gathered from Twitter does not mimic the population in general.

To obtain a free sample from Twitter, any person can connect to the streaming API and download approximately 1% of all tweets daily. Procuring additional tweets is a substantial cost and therefore, most researchers opt for the free sample. However, the algorithm used to sample from the streaming API is currently unknown and may not be uniformly random. Morstatter et al. (2013) show that tweets collected freely are generally biased according to trending topics and the most used hashtag strings. One final additional point of concern is that to our knowledge there is no study describing the user demographics for those who are more willing to leave the Twitter location services turned on. There may be revealing information simply from tweets with or without geotags.

The aforementioned limitations result in the scenario where the probability of detecting anomalous events on Twitter and being able to accurately impute geotags for those tweets is very low. However, Twitter data is a good source of information to monitor national or regional events. For example, an event such as the spread of influenza, as discussed in Section 6, is assumed to affect the population demographics equally and affects a significant proportion of Twitter users that will overcome the bias of the Streaming API. Also, people with the flu are assumed no more or less likely to turn off their location services. Events like the spread of flu overcome the limitations of Twitter data.

For future work, we seek to understand how the uncertainty in imputed geotags affects post geotagging analyses. We believe our model holds promise for scenarios where geotagged tweets are used for clustering locations for a specific event and when the geotag is used as a barometer to include or exclude the tweet from subsequent models.

REFERENCES

Satyen Abrol and Latifur Khan. 2010. Tweethood: Agglomerative clustering on fuzzy k-closest friends with variable depth for location mining. In *Proceedings of the IEEE 2nd International Conference on Social Computing (SocialCom'10)*. 153–160.

Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. 2011. Predicting flu trends using twitter data. In *Proceedings of the IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS'11)*. 702–707.

Shruti Aggarwal and Janpreet Singh. 2013. Outlier detection using K-mean and hybrid distance technique on multi-dimensional data set. *International Journal of Advanced Research in Computer Engineering and Technology* 2, 9 (2013), 2626–2631.

Oluwaseun Ajao, Jun Hong, and Weiru Liu. 2015. A survey of location inference techniques on Twitter. *Journal of Information Science* 41, 6 (2015), 855–864.

Lars Backstrom, Eric Sun, and Cameron Marlow. 2010. Find me if you can: Improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*. ACM, 61–70.

CDC. 2010. 2009-2010 Influenza Season Week 9 ending March 6, 2010. Accessed December 7, 2016 from http://www.cdc.gov/flu/weekly/weeklyarchives2009-2010/weekly09.htm.

Swarup Chandra, Latifur Khan, and Fahad Bin Muhaya. 2011. Estimating twitter user location using social interactions–A content based approach. In *Proceedings of the IEEE 3rd International Conference on Social Computing (SocialCom'11) and Proceedings of the Privacy, Security, Risk and Trust (PASSAT'11)*. 838–843.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: A content-based approach to geolocating twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM'10)*. ACM, 759–768.

Ryan Compton, David Jurgens, and David Allen. 2014. Geotagging one hundred million twitter accounts with total variation minimization. In *Proceedings of the IEEE International Conference on Big Data (Big Data'14)*. 393–401.

Clodoveu A. Davis Jr., Gisele L. Pappa, Diogo Rennó Rocha de Oliveira, and Filipe de L Arcanjo. 2011. Inferring the location of twitter messages based on user relationships. *Transactions in GIS* 15, 6 (2011), 735–751.

Maeve Duggan and Joanna Brenner. 2013. *The Demographics of Social Media Users,* Vol. 14. Pew Research Center's Internet & American Life Project, Washington, DC.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1277–1287.

Ingo Feinerer, Kurt Hornik, and David Meyer. 2008. Text mining infrastructure in R. *Journal of Statistical Software* 25, 5 (March 2008), 1–54. http://www.jstatsoft.org/v25/i05/.

Chris Fraley, Adrian E. Raftery, Thomas Brendan Murphy, and Luca Scrucca. 2016. mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. *The R Journal* 8, 1 (2016), 205–233.

Judith Gelernter and Nikolai Mushegian. 2011. Geo-parsing messages from microtext. *Transactions in GIS* 15, 6 (2011), 753–773.

Nicholas Generous, Geoffrey Fairchild, Alina Deshpande, Sara Y. Del Valle, and Reid Priedhorsky. 2014. Global disease monitoring and forecasting with Wikipedia. *PLoS Computational Biology* 10, 11 (2014), e1003892.

Janaína Gomide, Adriano Veloso, Wagner Meira Jr., Virgílio Almeida, Fabrício Benevenuto, Fernanda Ferraz, and Mauro Teixeira. 2011. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. In *Proceedings of the 3rd International Web Science Conference*. ACM, 3.

Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. 2011. Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 237–246.

Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. 2009. Social networks that matter: Twitter under the microscope. *First Monday* 14, 1

Mans Hulden, Miikka Silfverberg, and Jerid Francom. 2015. Kernel density estimation for text-based geolocation. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*.

David Jurgens. 2013. That's what friends are for: Inferring location in online social media platforms based on social relationships. In *Proceedings of the International Conference on Web and Social Media (ICWSM'13)*. 273–282.

Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. 2011. I'm eating a sandwich in Glasgow: Modeling locations with tweets. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*. ACM, 61–68.

Gizem Korkmaz, Jose Cadena, Chris J. Kuhlman, Achla Marathe, Anil Vullikanti, and Naren Ramakrishnan. 2015. Combining heterogeneous data sources for civil unrest forecasting. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, 258–265.

Kathy Lee, Ankit Agrawal, and Alok Choudhary. 2013. Real-time disease surveillance using twitter data: Demonstration on flu and cancer. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1474–1477.

Jeffrey McGee, James Caverlee, and Zhiyuan Cheng. 2013. Location prediction in social media based on tie strength. In *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management*. ACM, 459–468.

Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M. Carley. 2013. Is the sample good enough? Comparing data from twitter's streaming API with twitter's firehose. *International Conference on Weblogs and Social Media*. 400–408.

Martin F. Porter. 1980. An algorithm for suffix stripping. *Program* 14, 3 (1980), 130–137.

Reid Priedhorsky, Aron Culotta, and Sara Y. Del Valle. 2014. Inferring the origin locations of tweets with quantitative confidence. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 1523–1536.

Afshin Rahimi, Duy Vu, Trevor Cohn, and Timothy Baldwin. 2015. Exploiting text and network context for geolocation of social media users. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT'15)*. 1362–1367.

Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. 2012. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 1500–1510.

Dominic Rout, Kalina Bontcheva, Daniel Preoţiuc-Pietro, and Trevor Cohn. 2013. Where's@ wally? A classification approach to geolocating users based on their social ties. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. ACM, 11–20.

Axel Schulz, Aristotelis Hadjakos, Heiko Paulheim, Johannes Nachtwey, and Max Mühlhäuser. 2013. A multi-indicator approach for geolocalization of tweets. In *Proceedings of the International Conference on Web and Social Media (ICWSM' 13)*.

Robert Tibshirani, Guenther Walther, and Trevor Hastie. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63, 2 (2001), 411–423.

TIME. 2009. The 500 Most Frequently Used Words on Twitter. Accessed December 7, 2016 from http://techland.time.com/2009/06/08/the-500-most-frequently-used-words-on-twitter/.

Benjamin Wing and Jason Baldridge. 2014. Hierarchical discriminative classification for text-based geolocation. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP'14)*. 336–348.

Benjamin P. Wing and Jason Baldridge. 2011. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 1. Association for Computational Linguistics, 955–964.

Received April 2017; revised September 2017; accepted December 2017