Geocoding Without Geotags: A Text-based Approach for reddit

Keith Harrigian
Warner Media Applied Analytics
Boston, MA
`keith.harrigian@appliedanalytics.net`

Abstract

In this paper, we introduce the first geolocation inference approach for reddit, a social media platform where user pseudonymity has thus far made supervised demographic inference difficult to implement and validate. In particular, we design a text-based heuristic schema to generate ground truth location labels for reddit users in the absence of explicitly geotagged data. After evaluating the accuracy of our labeling procedure, we train and test several geolocation inference models across our reddit data set and three benchmark Twitter geolocation data sets. Ultimately, we show that geolocation models trained and applied on the same domain substantially outperform models attempting to transfer training data across domains, even more so on reddit where platform-specific interest-group metadata can be used to improve inferences.

1 Introduction

The rise of social media over the past decade has brought with it the capability to positively influence and deeply understand demographic groups at a scale unachievable within a controlled lab environment. For example, despite only having access to sparse demographic metadata, social media researchers have successfully engineered systems to promote targeted responses to public health issues (Yamaguchi et al., 2014; Huang et al., 2017) and to characterize complex human behaviors (Mellon and Prosser, 2017).

However, recent studies have demonstrated that social media data sets often contain strong population biases, especially those which are filtered down to users who have opted to share sensitive attributes such as name, age, and location (Malik et al., 2015; Sloan and Morgan, 2015; Lippincott and Carrell, 2018). These existing biases are likely to be compounded by new data privacy legislation

that will require more thorough informed consent processes (Kho et al., 2009; European Commission, 2018).

While some social platforms have previously approached the challenge of balancing data access and privacy by offering users the ability to share and explicitly control public access to sensitive attributes, others have opted not to collect sensitive attribute data altogether. The social news website reddit is perhaps the largest platform in the latter group; as of January 2018, it was the 5th most visited website in the United States and 6th most visited website globally (Alexa, 2018). Unlike real-name social media platforms such as Facebook, reddit operates as a pseudonymous website, with the only requirement for participation being a screen name.

Fortunately, there has been significant progress made using statistical models to infer user demographics based on text and other features when self-attribution data is sparse (Han et al., 2014; Ajao et al., 2015). On reddit in particular, Harrigian et al. (2016) has used self-attributed "flair" as labels for training a text-based gender inference model. However, to the best of our knowledge, user geolocation inference has not yet been attempted on reddit, where a complete lack of location-based features (e.g. geotags, profiles with a location field) has made it difficult to train and validate a supervised model.

The ability to geolocate users on reddit has substantial implications for conversation mining and high-level property modeling, especially since pseudonymity tends to encourage disinhibition (Gagnon, 2013). For instance, geolocation could be used to segment users discussing a movie trailer into US and international audiences to estimate a film's global appeal or to inform advertising strategy within different global markets. Alternatively, user geolocation may be used in conjunction with

sentiment analysis of political discussions to predict future voting outcomes.

Moving toward these goals, we introduce a text-based heuristic schema to generate ground truth location labels for reddit users in the absence of explicitly geotagged data. After evaluating the accuracy of our labeling procedure, we train and test several geolocation inference models across our reddit data set and three benchmark Twitter geolocation data sets. Ultimately, we show that geolocation models trained and applied on the same domain substantially outperform models attempting to transfer training data across domains, even more so on reddit where platform-specific interest-group metadata can be used to improve inferences.

2 reddit: the front page of the internet

Founded originally as a social news platform in 2005, reddit has since become one of the most visited websites in the world, offering an expansive suite of features designed to "[bridge] communities and individuals with ideas, the latest digital trends, and breaking news" (reddit, 2018). Although the so-called *front page of the internet* still boasts an impressive amount of link-sharing, reddit has gradually evolved into a self-referential community where original thoughts and new content prevail (Singer et al., 2014).

reddit is structured much like a traditional online forum, where over one-hundred thousand topical categories known as subreddits separate user communities and conversation. Subreddits may cover topics as general as humor and movies (e.g. *r/funny*, *r/movies*) or as specific as an unusual fitness goal (e.g. *r/100pushups*). Within each subreddit, users can post thematically relevant submissions in the form of an image, text blurb, or external link. Users are then able to post comments on the submission, responding to either the content in the original submission or to comments made by other users.

reddit users tend to feel protected by the site's pseudonymity and consequently eschew their offline persona in favor of a more genuine online persona (Gagnon, 2013; Shelton et al., 2015). Thus, there is much potential in reddit as a data source for conversation mining. Unfortunately, the same policies which promote this favorable behavior currently preclude the segmentation of conversation based on demographic dimensions and serve as a fundamental motivation to our research.

3 Geocoding reddit Users

Previous research on geolocation inference for social media has primarily used three Twitter data sets for model training and validation (Eisenstein et al., 2010; Roller et al., 2012; Han et al., 2012). Although Twitter's topical diversity typically supports generalization to other platforms (Mejova and Srinivasan, 2012), it has not been tested thoroughly in the geolocation context or on reddit at all.

There is substantial reason to believe that geolocation models trained on Twitter data will not perform optimally when applied to reddit. One of the most glaring concerns is the variation in user demographics between the platforms. Namely, Twitter tends to skew female while reddit tends to skew male (Barthel et al., 2016; Smith and Anderson, 2018). Coates and Pichler (2011) have shown that language varies between genders, which suggests that language-model priors may vary based on social platform.

Additionally, the lack of geolocation ground truth for reddit necessarily excludes several promising inference models from being applied to the platform. For example, network-based geolocation approaches exploit the empirical relationship between physical user distance and connectedness on a social graph to propagate known user locations to unlabeled users (Backstrom et al., 2010). Rahimi et al. (2015) argue that network-based approaches are generally superior to content-based methods, especially when the social graph is well-connected. However, these models require within-domain grounding for the propagation algorithms to be useful.

Finally, while the reddit platform lacks certain features useful for geolocation on Twitter (e.g. timezones, profile location fields), it possesses its own unique assets which we hypothesize can prove useful for user geolocation. In this paper, we quantify the predictive value of subreddit metadata, leaving other features such as user flair and the platform's hierarchical comment structure for future research.

3.1 Labeling Procedure

Since reddit does not offer geotagging capabilities, nor do reddit user profiles include location fields, we design a free-text geocoding procedure to associate reddit users with a home location. While we focus on reddit for this paper, we believe that

Figure 1: Comments from one of the seed submissions. Replies (gray background) are filtered out because they often contain location names not representative of user location.

the following labeling method should generalize to other pseudonymous platforms that have a question and answer-based submission structure.

Data. We begin by querying the reddit API for submissions with a title similar to "Where are you from?", "Where do you live?", or "Where are you living?" This query yields 3,600 English-language submissions, from which we manually curate a subset of 1,200 seed submissions where we expect users to self-identify their home location. Examples of submissions within the final set include "Where do you support Liverpool from?" and "How much is your rent, and where do you live?" Examples of submissions in the original query which were excluded from the final set include "Where are you banned from?" and "Without naming the location, where are you from?"

We then query comment data from the 1,200 seed submissions using the Python Reddit API Wrapper (PRAW)[1]. We filter out comments which mention "move," "moving," "born," or "raised" to ensure only current home locations are identified. We also remove comment replies, finding that they often include location mentions from a perspective of discussion as opposed to self-identification. Representative comments from one of the seed submissions[2] are displayed in Figure 1. Ultimately, we keep 96,071 comments from 89,697 unique users.

Entity Extraction. We employ a string matching approach to identify locations mentioned within the comment text. As a gazetteer, we use a subset of the GEONAMES data set (Wick and Vatant, 2012) with city populations greater than 15,000. To reduce the false positive rate of location recognition, 90 location names found within the 5,000 most common English words from the

CORPUS OF CONTEMPORARY ENGLISH (Davies, 2009) are removed from the gazetteer. After applying this filter, our gazetteer is left with 23,018 cities and their associated location hierarchies (i.e. city, state, country).

To aid in the disambiguation of common location names, we create a small dictionary of 57 frequently occurring abbreviations found during a preliminary exploration of the data. This dictionary includes abbreviations for each state and territory of the United States, in addition to the following: USA (United States), UK (United Kingdom), BC (British Columbia), and OT (Ontario). Abbreviations are only extracted if they directly follow an n-gram found within the location gazetteer.

Each comment is tokenized into an ordered list using standard processing techniques—contraction and case normalization, number and hyperlink removal, and whitespace splitting. To support identification of multi-token locations, we then chunk each list into all possible n-grams for $n \in [1, 4]$. We remove any n-gram which does not have an exact match to a location in the gazetteer or the abbreviation dictionary. When two or more of the matched n-grams occur in an order of a known location hierarchy from GEONAMES, they are concatenated together. Any remaining n-gram which is a substring of a larger n-gram in the list of matches is removed.

One issue with this approach is that cities under the population threshold, or cities missing from our gazetteer, are ignored entirely. However, we note that users often reference their home location in the form *City, State, Country*. To improve our labeling procedure's recall, we devise an additional rule that captures location mentions which match this syntactic pattern and contain at least one of our gazetteer or abbreviation dictionary entries as a substring. The inclusion of this rule expands the set of users with estimated city-level locations from 25k to 38k.

Geocoding. To associate each n-gram with a coordinate position, we employ Google's Geocoding API, which ingests a string and returns the most probable coordinate pair and nominal location hierarchy information. To help disambiguate common location names, we manually map region biases to a subset of location-related subreddits that house the seed submissions. We feed these region biases into the Geocoding API as a

---

[1]https://praw.readthedocs.io/en/latest/
[2]www.reddit.com/r/LiverpoolFC/comments/3wmjd4/
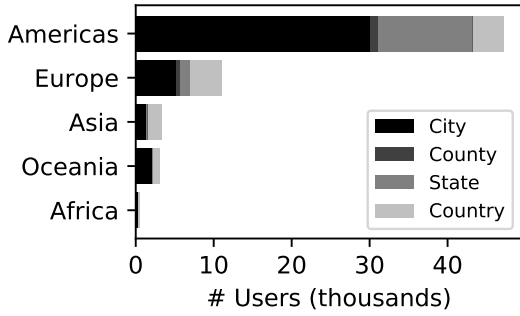
Country	Alexa Traffic	Labeled Users
United States	58.7%	60.1% (n=39,236)
United Kingdom	7.4%	5.4% (n=3,544)
Canada	6.0%	9.4% (n=6,163)
Australia	3.1%	3.5% (n=2,344)
Germany	2.1%	1.7% (n=1,097)

Figure 2 & Table 1: The figure (left) shows the geographic distribution of labeled users. Users outside the Americas are generally more likely to self-identify using a state-level resolution and above. The table (right) compares relative reddit traffic estimated using the proprietary Alexa (2018) panel to the distribution within our labeled data set.

query parameter for comments which were made in any of the mapped subreddits. Thus, a comment that mentions Scarborough in *r/ontario* will be properly mapped to Canada, while a comment that mentions Scarborough in *r/CasualUK*, will be mapped to England.

In the final stage of our procedure, we aggregate location extractions across each user, taking the maximum overlap within their identified location hierarchies as the ground truth. For each location above a city-level resolution, we use the geodesic median (Vardi and Zhang, 2000) of labeled users at the given resolution and below as the true coordinate pair.

3.2 Geographic Representation

Ultimately, we associate 65,245 users with geolocation labels at varying maximum resolutions—38,773 at a city level; 1,541 at a county level; 13,774 at a state level; and 11,157 at a country level. The label distribution and resolution breakdown over continents can be seen in Figure 2.

While the underlying geographic distribution of reddit users is not publicly available, Alexa (2018) publishes an estimate of country-level activity for the top-5 sources of reddit traffic using a proprietary data panel. We find that the same countries make up the top-5 most represented nations in our labeled data set, albeit having slightly different proportional breakdowns (see Table 1). In particular, we note that our data set slightly over-indexes on North American reddit users and suffers from a low sample size for most countries outside of the top-5.

3.3 Label Accuracy

To estimate precision of our labeling procedure across the larger data set, we randomly sample 500 of the labeled users and the comment from which their labeled location was extracted. For each of the 500 comments, we hand label whether the user's home location was properly extracted and, if correct, whether the location was extracted at the appropriate resolution.

Formally, we score any extracted location which falls within the true location hierarchy as a correct label. For example, if the true location is Boston, Massachusetts, but our procedure extracts Massachusetts (the state), we score the label as correct, but at an incorrect resolution. We find that our labeling procedure correctly extracted and geocoded 96.6% of the location names from the raw comments. Of these correct labels, 92.55% were labeled at the correct resolution.

Of the false positives in our random sample, 8 were due to disambiguation errors on part of the Google Geocoding API, 7 were due to users referencing locations that were not their actual home location, and 2 were due to usage of a common word not filtered out using the CORPUS OF CONTEMPORARY ENGLISH. Future work will look into ways to effectively leverage more granular subreddit-location mappings during the geocoding procedure to aid in disambiguation. This exploration may prove most helpful for ambiguous locations which are housed within the same national region (e.g. Kansas City, MO and Kansas City, KS).

For labels extracted within the appropriate hierarchy, but at the incorrect resolution, there were two main sources of error. First, over half of the incorrect resolution labels were due to our system (as designed) aggregating multiple location mentions within a comment up to the maximum overlap present. Second, several of the true cities did not exist within the gazetteer and were not captured by our syntactic rule. More extensive named

entity resolution systems may be useful in addressing the issue of missing gazetteer entries. However, correctly handling multiple location names and location names not representative of a user's home location will likely require a more robust natural language understanding system.

4 Geolocation Inference

In the remainder of the paper, we evaluate whether our "imperfect" geolocation labels are still useful in the context of a common task—user geolocation inference. We begin this analysis with a series of experiments and qualitative diagnostics to understand geolocation inference performance when training and applying models within the reddit domain. To quantify the effect of domain transfer, we conclude with a comprehensive comparison of inference performance across three benchmark Twitter data sets and our new reddit data set.

4.1 Related Work

Accounting for variations in feature selection and model architecture, existing geolocation inference approaches broadly fall into the following three categories (and their hybridizations): network-based, content-based, and metadata-based. We refer the reader to Han et al. (2014) and Ajao et al. (2015) for a comprehensive overview of the literature, but will highlight research pertinent to this particular study below. We ignore network-based models because they are not relevant to our chosen inference architecture.

Content-based. Content-based approaches, in which geographically predictive features are extracted from user-generated multimedia and modeled thereafter, remain some of the most fundamental to user geolocation on social media. Drawing upon well-documented phenomena regarding lexicon usage and its geographical variation (Trudgill, 1974; Vaux and Golder, 2003), text-driven models are particularly suited for application on reddit, where written comments are the lowest level of user behavior data accessible via the platform's public API.

One of the earliest contributors to user geolocation inference for social media, Cheng et al. (2010) introduced a generative model that operates on word usage alone to infer a user's home location, achieving an average prediction error of 535 miles for US Twitter users. Chang et al. (2012) extended this work by replacing frequency-based

word likelihoods with smoothed estimations using Gaussian Mixture Models (GMM). We use this approach within our evaluation due to its ease of implementation and interpretability. However, others have substantially improved inference accuracy using more flexible modeling architectures such as spatial topic models (Eisenstein et al., 2010; Hu and Ester, 2013), stacked denoising autoencoders (Liu and Inkpen, 2015), sparse coding and dictionary learning (Cha et al., 2015), and most recently, neural networks (Rahimi et al., 2017).

Metadata-based. We define metadata as all user behavior not explicitly expressed in multimedia, such as text or image posts, nor directly encoded as a social network connection. While metadata has generally been used to improve performance of content- and network-based methods, there exist some cases where metadata-based models outperform competing approaches outright (Han et al., 2014; Dredze et al., 2016). With reddit only recently introducing user profiles to the platform and their adoption by the larger community still an open question (Shelton et al., 2015), we focus primarily on comment metadata.

First, we suspect that the subreddits a user posts in, which often cover geographically localized topics such as sports and news, will be predictive of user location. While subreddits are specific to the reddit platform, they theoretically align with group membership, which has been shown to correlate with and predict user location in other domains (Zheleva and Getoor, 2009; Chen et al., 2013).

Of additional interest is temporal metadata, which can capture longitudinal variations in cyclical user activity patterns (Gao et al., 2013) or identify geographically-centered and time-dependent events (Yamaguchi et al., 2014). In particular, Dredze et al. (2016) and Do et al. (2017) find self-identified timezone information to be a useful geolocation predictor. Unfortunately, this explicitly encoded geographic indicator is absent from reddit and thus motivates our work to transform and model raw timestamp data in Section 4.2.

4.2 Location Estimation Model

Each user is represented by a concatenation of three feature vectors: word usage $\vec{w}$, subreddit submissions $\vec{s}$, and posting frequency $\vec{\tau}$. $\vec{w}$ is constructed by concatenating all comment text for a user, tokenizing into uni-grams (using the same

process in 3.1), and counting token frequencies. $\vec{s}$ is simply the frequency distribution of subreddits that a user has posted in across their comment history.

We represent the temporal posting habits of a user by $\vec{\tau}$, a 24-dimensional vector where each index contains the comment counts for one hour of the day. reddit comment timestamps are reported in Coordinated Universal Time (UTC) and can therefore be interpreted uniformly across users when constructing $\vec{\tau}$. Moving forward, we let $\vec{u}$ represent the concatenation of $\vec{w}$ and $\vec{s}$, allowing either modality to be turned off within the model. However, we keep $\vec{\tau}$ separate for notational clarity.

Model Architecture. As mentioned briefly above, we use the generative model introduced by Cheng et al. (2010) as the basis for our work, modifying it to enable ingestion of temporal metadata. Formally, given a user with feature set $\vec{u}$ (each u having an occurrence count $\|u\|$) and posting frequency $\vec{\tau}$, we estimate the probability of the user being located at geographic coordinate pair c using the following model:

$$P(c|\vec{u},\vec{\tau}) \propto P(c|\vec{\tau}) \sum_{u \in \vec{u}} \|u\| P(c|u) P(u). \quad (1)$$

To make a singular location prediction, we take the $argmax$ of $P(c|\vec{u},\vec{\tau})$ over an arbitrarily discrete set of coordinate pairs C.

Probability Density Estimation. As a baseline, Cheng et al. (2010) use the count-based frequency of features over cities in their training data to estimate $P(c|u)$. They recognize as a shortcoming to this approach the issue of feature sparsity—features with a low frequency or selective location presence contribute a probability close to zero for several candidate locations.

Drawing upon the relationship of geographic query dispersion quantified by Backstrom (2008), Chang et al. (2012) and Priedhorsky et al. (2014) use a bivariate Gaussian Mixture Model (GMM) to estimate $P(c|u)$ and demonstrate a significant performance improvement over the Cheng et al. (2010) baseline approach. However, the authors highlight as a potential caveat to their work the sensitivity of performance to GMM hyper-parameters, namely the number of mixture components.

To mitigate this issue, we use a Dirichlet Process Mixture Model (DPMM) to estimate $P(c|u)$. Generally, DPMM is able to better describe mixtures with a varying number of components by
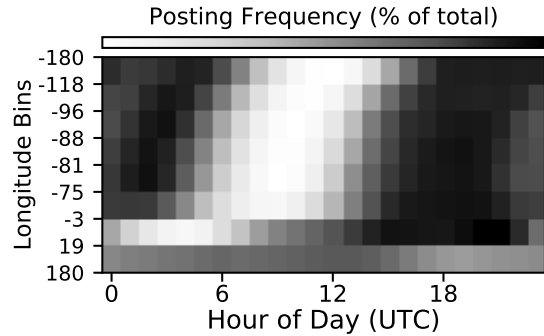

Figure 3: The relative posting frequencies within each longitude bin $\ell \in L$. Users east of 3° are more likely to post between the 6th and 12th hours (UTC) of the day.

constructing an infinite mixture with some components damped to near-zero amplitude; additionally, learned DPMM parameters remain relatively stable regardless of hyper-parameter choice (Attias, 2000; Blei et al., 2006).

We perform a series of experiments to compare performance between GMM and DPMM. Holding constant the hyper-parameter for number of mixture components[3], we find DPMM reduces error relative to GMM on a fixed test set by 5% to 32% depending on the type of covariance matrix used (i.e. diagonal, spherical). Ultimately, we use DPMM with a diagonal covariance matrix and 5 components to optimize inference performance.

Temporal Variation. A significant change to the Chang et al. (2010) model is the addition of a temporal adjustment term $P(c|\vec{\tau})$, designed to re-weight the word and subreddit posterior according to how well a user's posting frequency $\vec{\tau}$ aligns with the posting frequency distribution for users at coordinate pair c.

To make this estimate, we begin by discretizing the longitudes of users within our training data into a set of percentile-based bins L. Then, we fit a Logistic Regression model to map each user's posting frequency vector $\vec{\tau}$ to their associated longitude bin $\ell \in L$, using cross-validation to select hyper-parameters (e.g. L2-regularization) that maximize classification accuracy. Additionally, we use 5-fold cross-validation to select the number of longitude bins $\|L\|$ that minimizes downstream inference error.

To use the trained temporal feature model within our estimator, we first estimate $P(\ell|\vec{\tau})$ for each $\ell \in L$. Then we assign each $c \in C$ to its appropriate longitude bin ℓ and map the predicted

---

[3]For DPMM, we use a truncated distribution with a maximum number of mixture components equal to the chosen number of mixture components for GMM.

probability distribution across C. We visualize the temporal variation across longitude bins in our reddit data set within Figure 3.

Paralleling shifts in timezone, we note that peak user activity occurs earlier (within the UTC normalization) for users in eastern longitude bins than users in western longitude bins. Additionally, users in the eastern hemisphere are significantly more likely to post between the 6th and 12th hours (UTC) of the day.

5 Within-Domain Evaluation

We first examine geolocation inference performance within the domain of our reddit data set. We query all comment data for users within our set of geolocation labels using a publicly available corpus of reddit comments made between December 2005 and May 2018 (Baumgartner, 2018). For each user, we keep a maximum of 1000 comments posted up to one-month after they commented in the set of submissions used for labeling; this date-based filtering is done to mitigate the effect of users moving after posting in the seed set of submissions. Additionally, to ensure our model is not overtly biased by toponym mentions, we remove comments that were used as a part of the labeling procedure.

We separate our reddit data set into two versions—US (restricted to users from the contiguous United States) and GLOBAL (no location restrictions). We require that users within the United States have a minimum city-level resolution, while users outside the United States have been labeled with at least a state-level resolution.

We quantify model performance using three standard metrics from the user geolocation literature: Average Error Distance (*AED*), Median Error Distance (*MED*), and Accuracy at 100 miles (*Acc@100*). *AED* and *MED* are simply the arithmetic mean and median of error between predicted coordinates and true coordinates, respectively. *Acc@100* is the percentage of users whose predicted location is less than 100 miles from their true location.

5.1 Results

Feature Selection. Dimensionality reduction methods have been used to improve geolocation inference performance while reducing computational cost (Cheng et al., 2010; Chang et al., 2012; Han et al., 2012). Of existing approaches, the

	Top Words	Top Subreddits
Massachusetts, USA	allston, mbta, waltham, saugus, brookline, masshole, somerville, alewife, braintree	r/PokemonGoBoston, r/WorcesterMA, r/massachusetts, r/bostonhousing
Ohio, USA	ohioan, westerville, cincinnatis, jenis, clevelander, graeters, cuyahoga, bgsu, cbus	r/uCinci, r/ColumbusSocial, r/columbusclassifieds, r/Columbus
Germany	zeigen, dennoch, wenige, zeigt, solltest, deutlich, wollt, kriegt, stck	r/FragReddit, r/de_IAmA, r/rocketbeans, r/kreiswichs
Belgium	telenet, walloon, vlaams, jupiler, leuven, vlaanderen, ghent, molenbeek, azerty	r/belgium, r/brussels, r/Vivillon, r/ecr_eu

Table 2: Examples of the top features ranked using non-localness. Toponyms and non-English tokens are often the most indicative of location.

non-localness (NL) criteria introduced by Chang et al. (2012) stands out as being both effective at improving inference performance and useful as a means to understand feature alignment. Formally, NL is computed according to Equation 2, where sim_{SKL} is the Symmetric Kullback-Liebler divergence, S is a set of "stopword-like" features which are expected to occur uniformly across locations, and f is a generic feature in a larger feature set F.

$$NL(f) = \sum_{s \in S} sim_{SKL}(f, s) \frac{count(s)}{\sum_{s' \in S'} count(s')} \quad (2)$$

We apply non-localness to $\vec{w}$ and $\vec{s}$ separately to control how many features from each modality are kept. For $\vec{w}$, we let S be a set of 130 English stopwords taken from the Natural Language Toolkit[4]; to apply non-localness to subreddit features $\vec{s}$, we assume the 30 most active subreddits[5] in our data set make up S.

We evaluate NL over a discretized set of "State, Country, Continent" combinations, rolling up locations with less than 50 users in the training data to the next level of the location hierarchy. While Chang et al. (2012) finds modest performance improvements using GMM to estimate feature frequencies in the sim_{SKL} computation, we limit ourselves to frequency-based feature likelihoods to reduce computational expense.

In alignment with previous research, we find that NL produces qualitatively intuitive feature rankings (see Table 2). Furthermore, dimensionality reduction of both words and subreddits using NL significantly reduces error compared to using the full feature set for US and GLOBAL. The most significant source of error for large feature set sizes is noise added by "stopword-like" features, which generally have a large $\|u\|$ and effectively negate the contribution of more geographi-

---

[4]http://www.nltk.org/
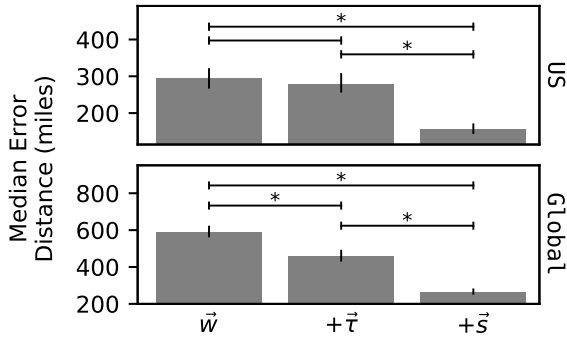[5]Examples include *r/Politics*, *r/AskReddit*, and *r/funny*.

Figure 4: The effect of temporal and subreddit metadata on inference error. Temporal features do not affect model performance on the US data set, but reduce error for the GLOBAL data set; subreddit metadata improves performance on both data sets.

cally predictive features.

Final feature set sizes are selected to minimize *AED* — 40k words and 650 subreddits for US (originally 118k words and 13k subreddits); 50k words and 1.1k subreddits for GLOBAL (originally 120k words and 14k subreddits).

Feature Modalities. Below, we discuss how the addition of the temporal adjustment factor $P(c|\vec{\tau})$ and subreddit features $\vec{s}$ affect model performance. We carry out 5-fold cross validation, with splits varied for US and GLOBAL, but held constant within each data set, to evaluate the modality effects fairly.

As seen in Figure 4, the temporal adjustment factor does not significantly affect model performance on the US data set, but reduces error when included within the GLOBAL data set, where underlying differences in $\vec{\tau}$ are magnified across continents. The most significant reduction in error occurs for European users, whose posting levels tend to peak around the 10th hour (UTC) of the day.

While subreddit features reduce error within both data sets when combined with word features, they do not outperform word features on their own. Rather, models trained using word features alone achieve an *AED* which is 17% and 4% lower for US and GLOBAL, respectively, than models trained using subreddit features alone. This implies that text-based models trained on other domains may perform adequately on reddit, but will likely suffer from the inability to take advantage of subreddit metadata in a supervised manner.

6 Cross-Domain Evaluation

While within-domain experiments suggest that reddit-specific metadata offers substantial predictive value, we wish to compare the highest degree

of performance achieved within-domain to performance achieved using models trained outside the reddit domain. To do so, we use three benchmark Twitter geolocation data sets—GEOTEXT (Eisenstein et al., 2010), TWITTER-US (Roller et al., 2012), and TWITTER-WORLD (Han et al., 2012).

All three data sets were created by monitoring Twitter's streaming API over a discrete period of time and caching comments for users who enabled geotagging features on the Twitter platform. Due to Twitter's terms of service, TWITTER-US and TWITTER-WORLD must be compiled from scratch using tweet IDs. Unfortunately, several users within the original data sets have since either deleted their accounts or restricted access to their tweet history. As such, we were not able to perfectly recreate the original data sets. Ultimately, our compilation of TWITTER-US contains 246k out of the original 440k users, while TWITTER-WORLD contains 888k out of the original 1.4M users.

To understand the impact that domain transfer has on geolocation inference performance, we set up a systematic model comparison. First, we run 5-fold cross-validation within each of the Twitter data sets using both word and timestamp features. For the TWITTER-WORLD data set, we run two independent cross-validation procedures, evaluating on the subset of US users alone and also on the entire data set. Then, we train models for each of our data sets using all available data and apply them to the data sets not used for training. When evaluating performance on a data set that only contains US users, we train the corresponding model only using US user data. All model hyper-parameters (e.g. feature set sizes, regularization, etc.) were chosen to optimize within-data-set performance.

6.1 Results

Experimental results are summarized in Table 3 and explored in greater detail below. The baseline model assigns each user in the test set to the maximum a posteriori (MAP) of a DPMM fit to the locations of all users in the training data. As an additional reference point, we also include results from two recent approaches for the Twitter data sets.

Transfer Performance. In validation of our labeling procedure, we note that models trained on reddit data outperform within-domain baselines for nearly all Twitter data sets. The only exception

	Test	REDDIT-US			REDDIT-GLOBAL			GEOTEXT			TWITTER-US			TWITTER-WORLD (US)			TWITTER-WORLD (All)		
Train		*Acc@100*	*AED*	*MED*	*Acc@100*	*AED*	*MED*	*Acc@100*	*AED*	*MED*	*Acc@100*	*AED*	*MED*	*Acc@100*	*AED*	*MED*	*Acc@100*	*AED*	*MED*
Rahimi et al. (2017) (Text-based)		-	-	-	-	-	-	0.38	844	389	0.54	554	120	-	-	-	0.34	1456	415
Do et al. (2017) (Text + Network)		-	-	-	-	-	-	0.62	532	32	0.66	433	45	-	-	-	0.53	1044	118
Baseline (MAP Estimate)		0.05	894	750	0.03	2207	1221	0.33	679	424	0.07	1659	1869	0.07	1651	1869	0.04	3893	2463
REDDIT-US	$\vec{w}$	0.36	602	295	-	-	-	0.21	695	479	0.31	609	362	0.14	748	605	-	-	-
	$\vec{w}+\vec{\tau}$	0.36	580	278	-	-	-	0.21	695	480	0.31	603	358	0.13	**747**	592	-	-	-
	$\vec{w}+\vec{\tau}+\vec{s}$	**0.45**	**502**	**157**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
REDDIT-GLOBAL	$\vec{w}$	-	-	-	0.24	1751	590	-	-	-	-	-	-	-	-	-	0.09	2717	1329
	$\vec{w}+\vec{\tau}$	-	-	-	0.25	1475	457	-	-	-	-	-	-	-	-	-	0.09	2708	**1329**
	$\vec{w}+\vec{\tau}+\vec{s}$	-	-	-	**0.36**	**1259**	**266**	-	-	-	-	-	-	-	-	-	-	-	-
GEOTEXT	$\vec{w}$	0.07	1210	1019	-	-	-	0.38	591	280	0.12	982	755	0.13	992	755	-	-	-
	$\vec{w}+\vec{\tau}$	0.07	1209	1019	-	-	-	**0.38**	**575**	**271**	0.12	982	755	0.13	992	755	-	-	-
TWITTER-US	$\vec{w}$	0.36	670	326	-	-	-	0.34	631	301	0.40	547	225	0.24	790	583	-	-	-
	$\vec{w}+\vec{\tau}$	0.36	635	294	-	-	-	0.33	632	304	**0.40**	**536**	**220**	**0.24**	789	**582**	-	-	-
TWITTER-WORLD (US)	$\vec{w}$	0.20	963	729	-	-	-	0.12	635	313	0.23	772	564	0.22	795	589	-	-	-
	$\vec{w}+\vec{\tau}$	0.18	923	717	-	-	-	0.13	628	311	0.22	768	563	0.22	791	584	-	-	-
TWITTER-WORLD (All)	$\vec{w}$	-	-	-	0.15	2737	1829	-	-	-	-	-	-	-	-	-	0.16	2716	1665
	$\vec{w}+\vec{\tau}$	-	-	-	0.16	1793	817	-	-	-	-	-	-	-	-	-	**0.16**	**2610**	1405

Table 3: Summary statistics for the domain-transfer experiment. The best results from our cross-validation procedure are bolded. Models trained on reddit data (third and fourth rows) outperform the baseline for Twitter data sets in nearly all cases (see text for caveats). Within the reddit data sets (first and second columns), models with access to platform-specific metadata outperform all models transferred from the Twitter domain.

occurs for GEOTEXT, where less than 10% of reddit features are also present. Due to the lack of feature overlap, many of the GEOTEXT, users are assigned to the MAP of users in the reddit data by default. While TWITTER-US and TWITTER-WORLD also have low feature overlap with GEOTEXT, their MAP estimates are much closer to majority of users in GEOTEXT.

We also note that there is a significant loss incurred by most domain transfers. This loss is magnified for models trained on Twitter data and applied to reddit data, since Twitter models critically lack access to subreddit metadata during training.

Temporal Features. The effect of our temporal adjustment term $P(c|\vec{\tau})$ varies between each data set. Specifically, the temporal features significantly improve within-domain performance for both of our "international" data sets (TWITTER-WORLD and REDDIT-GLOBAL), but offer no significant gain for data sets with US users only. Additionally, we note that temporal features significantly reduce the loss in performance incurred by domain transfer going from TWITTER-US to REDDIT-US and from TWITTER-WORLD to REDDIT-GLOBAL.

Location Estimator. Based on within-domain performance for each of the Twitter data sets, we recognize that our inference modeling approach is below state of the art. For example, in the space of text-only models, Rahimi et al. (2017) have achieved an *Acc@100* of 0.34 on TWITTER-WORLD using a multilayer perceptron and *k-d* tree discretization over the label set.

The performance gap between our model and state of the art approaches widens when considering multi-modal architectures. Notably, Do et al. (2017) have achieved an *Acc@100* of 0.62 on GEOTEXT using multi-view neural networks that simultaneously leverage text, profile metadata, and social network connections. Thus, we hypothesize that implementing models which are more complex than our current architecture will magnify the performance gain achieved by including subreddit metadata alongside text-based features.

7 Discussion and Future Work

In this paper, we introduced the first user geocoding and geolocation inference approach for reddit, demonstrating that pseudonymity is not an exhaustive barrier to supervised learning. In addition to designing a labeling procedure capable of geocoding user home locations in noisy comment data with a precision of 0.966, we demonstrated that reddit-specific metadata can be used to significantly improve inferences. Ultimately, we trained a multi-modal inference model which achieves a median error of 157 miles and 266 miles for US and international reddit users, respectively.

Moving forward, we plan to thoroughly examine underlying biases that may exist within the users identified by our labeling procedure. Specifically, we will build on the work of Sloan and Morgan (2015) and Lippincott and Carrell (2018) to understand differences in user activity, interests, and conversation topicality, relative to the general reddit population. We also plan to explore different seed-submission sampling methods to improve the representation of non-North American users.

References

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

Amazon Alexa. 2018. Amazon top 500 global sites.

Hagai Attias. 2000. A variational baysian framework for graphical models. In *Advances in neural information processing systems*, pages 209–215.

Lars Backstrom, Jon Kleinberg, Ravi Kumar, and Jasmine Novak. 2008. Spatial variation in search engine queries. In *Proceedings of the 17th international conference on World Wide Web*, pages 357–366. ACM.

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*, pages 61–70. ACM.

Michael Barthel, Galen Stocking, Jesse Holcomb, and Amy Mitchell. 2016. Seven-in-ten reddit users get news on the site. [Online; posted 26-May-2016].

Jason Baumgartner. 2018. pushshift.io.

David M Blei, Michael I Jordan, et al. 2006. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143.

Miriam Cha, Youngjune Gwon, and HT Kung. 2015. Twitter geolocation and regional classification via sparse coding. In *ICWSM*, pages 582–585.

Hau-wen Chang, Dongwon Lee, Mohammed Eltaher, and Jeongkyu Lee. 2012. @ phillies tweeting from philly? predicting twitter user locations with spatial word usage. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 111–118. IEEE Computer Society.

Yan Chen, Jichang Zhao, Xia Hu, Xiaoming Zhang, Zhoujun Li, and Tat-Seng Chua. 2013. From interest to function: Location estimation in social media. In *AAAI*.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768. ACM.

Jennifer Coates and Pia Pichler. 2011. *Language and gender: A reader*. Wiley-blackwell Chichester.

Mark Davies. 2009. The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics*, 14(2):159–190.

Tien Huu Do, Duc Minh Nguyen, Evaggelia Tsiligianni, Bruno Cornelis, and Nikos Deligiannis. 2017. Multiview deep learning for predicting twitter users' location. *arXiv preprint arXiv:1712.08091*.

Mark Dredze, Miles Osborne, and Prabhanjan Kambadur. 2016. Geolocation for twitter: Timing matters. In *HLT-NAACL*, pages 1064–1069.

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 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics.

European Union European Commission. 2018. General data protection regulation.

Tiffany Gagnon. 2013. The disinhibition of reddit users. *Adele Richardson's Spring*.

Huiji Gao, Jiliang Tang, Xia Hu, and Huan Liu. 2013. Exploring temporal effects for location recommendation on location-based social networks. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 93–100. ACM.

Bo Han, Paul Cook, and Timothy Baldwin. 2012. Geolocation prediction in social media data by finding location indicative words. In *Proceedings of COLING*, pages 1045–1062.

Bo Han, Paul Cook, and Timothy Baldwin. 2014. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49:451–500.

Keith Harrigian, Nathan Sanders, Jonathan Foster, and Arjun Sangvhi. 2016. When anonymity is not anonymous: Gender inference on reddit. In *Northeastern Research, Innovation, and Scholarship Expo*, Boston, MA.

Bo Hu and Martin Ester. 2013. Spatial topic modeling in online social media for location recommendation. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 25–32. ACM.

Xiaolei Huang, Michael C Smith, Michael J Paul, Dmytro Ryzhkov, Sandra C Quinn, David A Broniatowski, and Mark Dredze. 2017. Examining patterns of influenza vaccination in social media. In *Proceedings of the AAAI Joint Workshop on Health Intelligence (W3PHIAI), San Francisco, CA, USA*, pages 4–5.

Michelle E Kho, Mark Duffett, Donald J Willison, Deborah J Cook, and Melissa C Brouwers. 2009. Written informed consent and selection bias in observational studies using medical records: systematic review. *Bmj*, 338:b866.

Tom Lippincott and Annabelle Carrell. 2018. Observational comparison of geo-tagged and randomly-drawn tweets. In *Proceedings of the Second Workshop on Computational Modeling of People?s Opinions, Personality, and Emotions in Social Media*, pages 50–55.

Ji Liu and Diana Inkpen. 2015. Estimating user location in social media with stacked denoising autoencoders. In *VS@ HLT-NAACL*, pages 201–210.

Momin M Malik, Hemank Lamba, Constantine Nakos, and Jurgen Pfeffer. 2015. Population bias in geotagged tweets. *People*, 1(3,759.710):3–759.

Yelena Mejova and Padmini Srinivasan. 2012. Crossing media streams with sentiment: Domain adaptation in blogs, reviews and twitter. In *ICWSM*.

Jonathan Mellon and Christopher Prosser. 2017. Twitter and facebook are not representative of the general population: Political attitudes and demographics of british social media users. *Research & Politics*, 4(3):2053168017720008.

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*, pages 1523–1536. ACM.

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2017. A neural model for user geolocation and lexical dialectology. *arXiv preprint arXiv:1704.04008*.

Afshin Rahimi, Duy Vu, Trevor Cohn, and Timothy Baldwin. 2015. Exploiting text and network context for geolocation of social media users. *arXiv preprint arXiv:1506.04803*.

reddit. 2018. reddit: the front page of the internet.

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 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1500–1510. Association for Computational Linguistics.

Martin Shelton, Katherine Lo, and Bonnie Nardi. 2015. Online media forums as separate social lives: A qualitative study of disclosure within and beyond reddit. *iConference 2015 Proceedings*.

Philipp Singer, Fabian Flöck, Clemens Meinhart, Elias Zeitfogel, and Markus Strohmaier. 2014. Evolution of reddit: from the front page of the internet to a self-referential community? In *Proceedings of the 23rd International Conference on World Wide Web*, pages 517–522. ACM.

Luke Sloan and Jeffrey Morgan. 2015. Who tweets with their location? understanding the relationship between demographic characteristics and the use of geoservices and geotagging on twitter. *PloS one*, 10(11):e0142209.

Aaron Smith and Monica Anderson. 2018. Social media use in 2018. [Online; posted 1-March-2018].

Peter Trudgill. 1974. Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in society*, 3(2):215–246.

Yehuda Vardi and Cun-Hui Zhang. 2000. The multivariate l1-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97(4):1423–1426.

Bert Vaux and Scott Golder. 2003. The harvard dialect survey. *Cambridge, MA: Harvard University Linguistics Department*.

Mark Wick and Bernard Vatant. 2012. The geonames geographical database. *Available from World Wide Web: http://geonames. org*.

Yuto Yamaguchi, Toshiyuki Amagasa, Hiroyuki Kitagawa, and Yohei Ikawa. 2014. Online user location inference exploiting spatiotemporal correlations in social streams. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1139–1148. ACM.

Elena Zheleva and Lise Getoor. 2009. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *Proceedings of the 18th international conference on World wide web*, pages 531–540. ACM.