Geolocation Prediction in Social Media Data by Finding Location Indicative Words

HAN Bo[1,2] *Paul COOK*[1] *Timothy BALDWIN*[1,2]
(1) University of Melbourne
(2) NICTA
hanb@student.unimelb.edu.au, paulcook@unimelb.edu.au, tb@ldwin.net

Abstract

Geolocation prediction is vital to geospatial applications like localised search and local event detection. Predominately, social media geolocation models are based on full text data, including common words with no geospatial dimension (e.g. *today*) and noisy strings (*tmrw*), potentially hampering prediction and leading to slower/more memory-intensive models. In this paper, we focus on finding location indicative words (LIWs) via feature selection, and establishing whether the reduced feature set boosts geolocation accuracy. Our results show that an information gain ratio-based approach surpasses other methods at LIW selection, outperforming state-of-the-art geolocation prediction methods by 10.6% in accuracy and reducing the mean and median of prediction error distance by 45km and 209km, respectively, on a public dataset. We further formulate notions of prediction confidence, and demonstrate that performance is even higher in cases where our model is more confident, striking a trade-off between accuracy and coverage. Finally, the identified LIWs reveal regional language differences, which could be potentially useful for lexicographers.

Keywords: Social Media, Geolocation, Feature Selection.

Han, Bo, Paul Cook and Timothy Baldwin (2012) Geolocation Prediction in Social Media Data by Finding Location Indicative Words, In Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012), Mumbai, India, pp. 1045-1062.

1 Introduction

With the ever-growing popularity of social media, massive volumes of user-generated data are produced everyday, e.g. in the form of Twitter messages (tweets) and Facebook updates.[1] This data provides many new opportunities and challenges for natural language processing. One such challenge is geolocation prediction: predicting the geolocation of a message or user based on their social media posts. In this paper, we focus on user-level geolocation based on the aggregated body of tweets from a user, and estimate the user's location at the city level.

As is well established in previous work (Cheng et al., 2010; Wing and Baldridge, 2011; Kinsella et al., 2011), it is reasonable to assume that user posts in social media reflect their geospatial locum, because lexical priors differ from region to region. For example, a user in London is much more likely to talk about *Piccadilly* and *British* than a user in New York or Beijing. That is not to say that those terms are uniquely associated with London, of course: *British* could be used by a user outside of the UK to discuss something relating to the UK. However, the use of a range of such terms with high relative frequency is strongly indicative of the fact that a user is located in London.

Our objective in this work is to automatically identify "location indicative words" (LIWs), that is words that implicitly or explicitly encode an association with a particular location. To this end, we refine the geolocation task in Section 3, and explore the impact of LIWs on user geolocation.

Our contributions are as follows: (1) we apply feature selection methods for automatically learning LIWs, and show that both accuracy and efficiency in geolocation are vastly improved using the resultant feature set, achieving state-of-the-art performance over an existing dataset; (2) we develop a city-based world division and a new global geolocation dataset, and demonstrate the effectiveness of the proposed method over this dataset;[2] (3) we conduct a pilot study on the correlation between prediction confidence, as measured by a series of heuristic variables, and classifier accuracy (see Section 5.4); and (4) we find that LIWs selected by our methods are both intuitive and have potential utility in lexicographic research on regional language differences.

The remainder of the paper is organised as follows: Section 2 introduces related work and describes the key questions investigated in this paper. Section 3 outlines the task setting, datasets and evaluation metrics used in this research. Section 4 describes different feature selection methods for extracting LIWs. Section 5 compares the results of the different feature selection methods and discusses their impact on geolocation prediction, and proposes several methods for associating beliefs with predictions. Finally, we conclude the paper and outline possible future work.

2 Related Work and Key Questions

While acknowledging potential privacy concerns (Mao et al., 2011; Pontes et al., 2012), accurate user geolocation is a key driver for location-specific services such as localised search, and has been the target of research across different disciplines. The most reliable and straightforward approach to geolocation prediction is IP-based methods (Buyukokkten et al., 1999), but in many contexts, it is not possible to access the IP of the device used to post content, or the IP is relatively uninformative (as is the case with, e.g. mobile devices). As a result, research has focused on the harder task of geolocation prediction via the textual content of a document (or document set). In the information retrieval community, e.g. web pages (Ding et al., 2000; Amitay et al., 2004; Zong et al., 2005; Silva et al., 2006), search query logs (Wang et al., 2005; Backstrom et al., 2008), Wikipedia edit logs

[1] `www.twitter.com`; `www.facebook.com`

[2] The dataset is available from `http://www.csse.unimelb.edu.au/~tim/etc/coling2012-geo.tgz`.

(Lieberman and Lin, 2009) and Flickr image tags (Crandall et al., 2009; Serdyukov et al., 2009; Hauff and Houben, 2012) have been used as the basis for geolocation prediction. These methods are primarily designed for longer or more homogeneous document sets. In contrast, social media data consists of terse noisy texts, presenting a challenge for these approaches. For instance, any reliance on named entity recognition is thwarted by the unedited nature of social media data, where spelling and capitalisation are much more ad hoc than in edited document collections.

The spatial data mining community has tended to approach the task via identifying geographical references in documents (also known as *geoparsing*: Leidner and Lieberman (2011)). Methods range from naive gazetteer matching and rule-based approaches (Bilhaut et al., 2003), to machine learning-based methods (mainly based on named entity recognition: Qin et al. (2010); Gelernter and Mushegian (2011)). The principal drawback of these methods is that they rely on explicit mentions of addresses or formal placenames in the text, rather than words which are more informally associated with a place. In social media data, we can't rely on a given user mentioning an address or formal placename, severely limiting the coverage of such methods.[3]

There has been a limited amount of work on geolocation prediction based on social network analysis (Backstrom et al., 2010), but social networks are dynamic and the data is often hard to obtain. In terms of text-based geolocation prediction, Cheng et al. (2010) estimate the city-level user geolocation for the continental US with a simple probabilistic model, which they complement with strictly local words and smoothing. Compared with their approach, our LIW selection requires no explicit training data and is more flexible. Wing and Baldridge (2011) use KL-divergence (Kullback and Leibler, 1951) to measure the similarity between different geo-grids specified by geospatial coordinates. Recently, Roller et al. (2012) extend this idea using a KD-tree-based adaptive grid and grid centroids, achieving state-of-the-art geolocation prediction results. Li et al. (2011) investigated the prediction of Places of Interest (POIs) based on linear rank combination of content and temporal factors. Kinsella et al. (2011) compare a variety of geolocation prediction classification models at different location granularities. Adams and Janowicz (2012) utilise external geo-reference data to infer locations. Mahmud et al. (2012) combine timezone information and content-based classifiers in a hierarchical model for geolocation. They only consider nouns, hashtags and place names as features. Recently, Li et al. (2012) integrate both friendship and content information in a probabilistic model. In addition, topic modelling has been applied to the study of geospatially-related tasks including user dialect (Eisenstein et al., 2010), topic discovery (Yin et al., 2011), object matching (Dalvi et al., 2012), factorisation of different geospatial features (Hong et al., 2012), and spatial-temporal analysis (Bauer et al., 2012).

Most work uses the full token set from the training document collection, or a relatively rudimentary approach to feature selection. In this paper, we propose a targeted approach to identify LIWs using various feature selection methods (Yang and Pedersen, 1997), focusing on two key questions:

1. What empirical properties do we observe in LIWs, and what feature selection methods best capture those properties?
2. Can we boost the accuracy of geolocation prediction through targeted identification of LIWs?

3 Geolocation Task Scope and Formulation

We approach geolocation as a text classification task. Tweets from each city are employed to represent a class. All tweets from a given user are aggregated and assigned to the city where that user is based. There are four key components to a geolocation prediction system, which we discuss

[3] An exception is automatically-generated posts from services such as FourSquare which explicitly mention an address.

in turn below: (1) the representation of different geolocations, (2) the model, (3) the data, and (4) the feature set. We then discuss evaluation metrics for geolocation prediction.

3.1 Representation: Earth Grid vs. City

Geolocations can be captured as points, or clustered based on grids (Wing and Baldridge, 2011; Roller et al., 2012) or population centres (Cheng et al., 2010; Kinsella et al., 2011). A point-based representation presents computational challenges, and is too fine-grained for our task. We opt for a city-based representation rather than a grid-based representation because there is considerable variability in the shape and size of geographical regions: a coarse-grained grid cell is perhaps appropriate in central Siberia, but for densely-populated and linguistically/culturally diverse regions such as Luxembourg, doesn't lead to a natural representation of the administrative, population-based or language boundaries in the region. A city-based representation is able to capture these boundaries more intuitively. The only downside to a city-based representation is that it is inappropriate for classifying users in rural areas. As we will see, however, the bulk of users on services such as Twitter are, unsurprisingly, based in cities.

We use the publicly-available Geoname dataset as the basis for our city categorisation.[4] Geoname contains city-level metadata, including the full city name, population, latitude and longitude. The city name is associated with hierarchical regional information, like the state and country it is based in, so that London in Britain, e.g. is distinguished from London in Canada. We hence use a city-region-country format to represent each city (e.g. Toronto, Canada is represented as toronto-08-ca, where 08 signifies the province of Ontario and ca signifies Canada). Because region coding schemes vary across different countries, we only employ the first and second level region fields in Geoname as the region. Furthermore, if the second level field is too specific (i.e. longer than 4 letters), we then only incorporate the first level region field (e.g. instead of using melbourne-07-24600-au, we use melbourne-07-au). Moreover, because cities are sometimes complex in structure (e.g. Boston in Massachusetts colloquially refers to the metropolitan area rather than the city, which is made up of cities including the cities of Boston, Revere and Chelsea), we collapse together cities which are adjacent to one another within a single administrative region, as follows:

1. Identify all cities which share the same region code (i.e. are located in the same state, province, county, etc.) in the Geoname dataset.
2. For each region, find the city c with the highest population.
3. Collapse all cities within 50km of c into c.
4. Select the next-largest city c, and repeat.
5. Remove all cities with a population less than 100K. The remaining cities form our city-based representation of geolocations.

As a result of these procedures, Boston ends up as a single city (incorporating Revere and Chelsea), but neighbouring Manchester is a discrete city (incorporating Bedford) because it is in New Hampshire. This algorithm identifies a total of 3,709 cities throughout the world.

3.2 Generative vs. Discriminative models

Generative models (e.g. naive Bayes) are based on estimation of the class priors (i.e. $P(c_i)$) and the probability of observing a given term vector given a class (i.e. $P(w_1, w_2, \ldots, w_n|c_i)$). In contrast, discriminative models are based on estimation of a given class given a term vector (i.e.

[4] http://www.geonames.org.

Name	Cities	Users	Tweets	Types	Tokens	Region
NA	378	500K	38M	4.92M	436M	North America
WORLD	3135	1.39M	12M	0.85M	103M	the world

Table 1: Details of the two datasets used in this research.

$P(c|w_1, w_2, \ldots, w_n)$). The objective of both models is to find a city $c_{max} \in C$ such that the relevant probability is maximised. We experiment with both types of models in our experiments in Section 5.3. In this paper, we choose a generative multinomial naive Bayes (NB) model as our benchmark, for two reasons: (1) it incorporates a class prior, allowing it to classify an instance in the absence of any features shared with the training data; and (2) generative models outperform discriminative models when training data is relatively scarce (Ng and Jordan, 2002).[5]

3.3 Data

In this paper, we employ two geo-tagged datasets: (1) the regional North America geolocation dataset of Roller et al. (2012) (NA hereafter), for benchmarking purposes; and (2) a novel dataset that covers the entire globe (WORLD), collected for the purposes of this research via the Twitter public Streaming API[6] from 2011.09.21 to 2012.02.29.

In building WORLD, we first filter non-English tweets using `langid.py`, an open-source language identification tool (Lui and Baldwin, 2012), and then apply a Twitter tokeniser (adapted from O'Connor et al. (2010)). We restrict WORLD to English tweets in order to create a dataset similar to NA (in which English is the predominant language), but covering the entire world. A further reason for only using English tweets is to control for the influence of language priors on geolocation performance. For example, we expect that a language such as Japanese would tend to be more skewed towards particular cities than English, making the task of text-based geolocation easier.[7] We further eliminate Foursquare check-ins, as they mention the location of the user and are geo-tagged, and duplicate tweets. We also remove tweets from users with less than 10 geo-tagged tweets to reduce feature sparsity. Finally, we eliminate all tweets which aren't close to a city by dividing the earth into $0.5' \times 0.5'$ grids, and discarding any tweet for which no city is found in any of the 8 neighbouring grid cells. We then assign each user to the single city in which the majority of their tweets occur. Note that the processing described in this paragraph applies only to WORLD; NA was left as-is to ensure comparability with previous work.

Analysis of a sample of 26 million tweets (not filtered as above) reveals that 92.1% of tweets are "close" to (in a neighbouring $0.5' \times 0.5'$ grid cell) of one of our 3,709 cities, and that the top 40% of cities contain 90% of the tweets, as shown in Figure 1.

A statistical profile of NA and WORLD is presented in Table 1.[8] We also analyse the spread of WORLD in Figure 2, in terms of: (1) the number of users with a given number of tweets; and (2) the number of users with differing levels of geographical spread in their tweets, measured as the average distance between each of a user's tweets and the centre of the city to which that user is allocated.[9] This analysis shows that most users have a relatively small number of geo-tagged tweets, and most

[5] There is certainly an abundance of Twitter data to train models over, but the number of Twitter users with sufficient amounts of geo-tagged tweets to be able to perform geolocation prediction is small, relative to the number of parameters in the model (the product of the number of features and classes).

[6] `https://dev.twitter.com/docs/streaming-apis`

[7] All of the methods we consider could nevertheless be easily applied to a mixed-language setting in the future.

[8] WORLD has only 3,135 (as opposed to 3,709) cities because some cities have no tweets.

[9] The geographical spread is calculated over a random sub-sample of 10 tweets for a given user, for efficiency reasons.

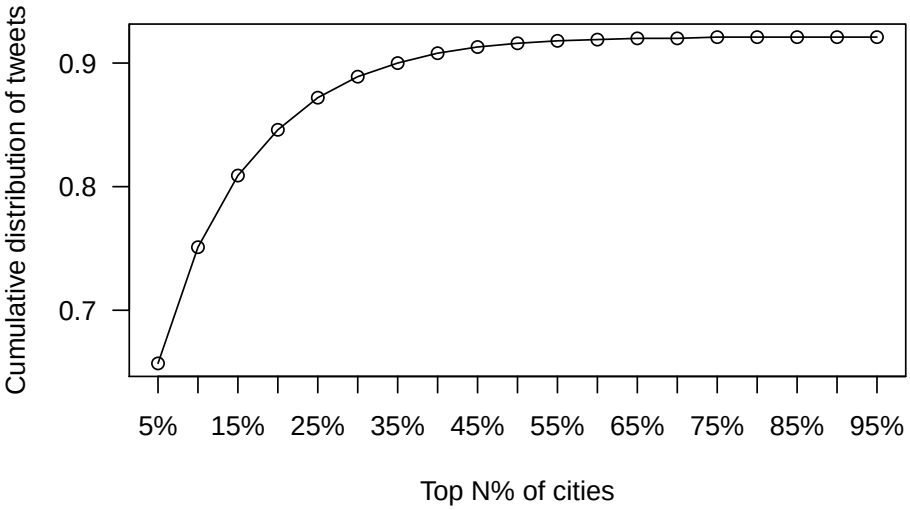

Figure 1: Cumulative coverage of tweets for increasing numbers of cities based on 26 million geo-tagged tweets.

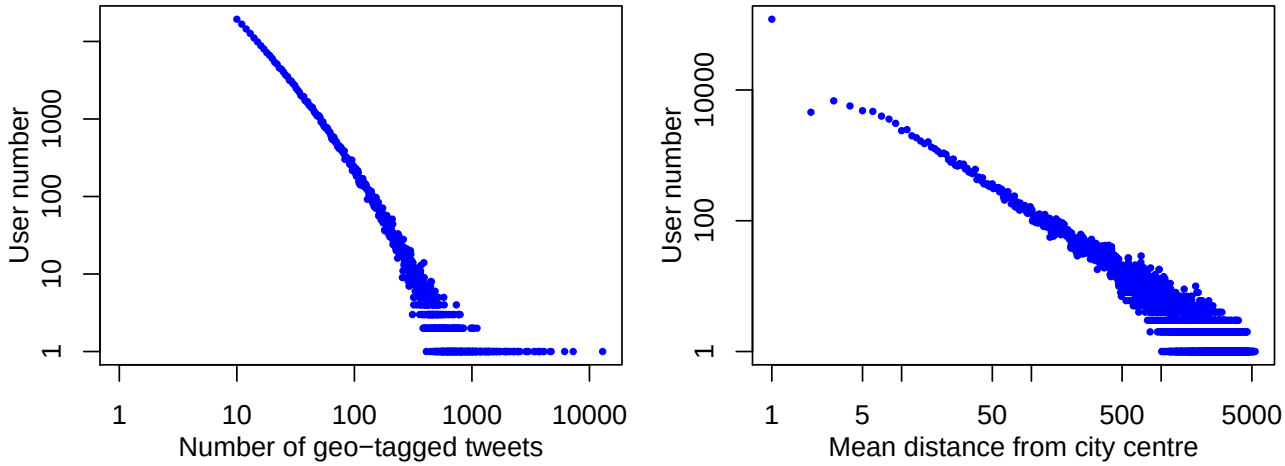

Figure 2: The number of users with different numbers of tweets, and different mean distances from the city center, for WORLD.

users stay near a single city.

3.4 Features: All unigrams vs. Location Indicative Words

Feature selection is a key contribution of this paper, based on the notion of "location indicativeness", as described in Section 4. Our hypothesis is that using only location indicative words (LIWs) as features will be more efficient and effective than using all terms. Rather than engineering new features or attempting to capture named entities or higher-order n-grams, we focus on feature selection over simple term unigrams. This is partly a pragmatic consideration (preliminary results with both named entities and higher order n-grams were disappointing). Partly, however, it is for comparability with past work, in determining whether a strategically selected subset of terms can lead to significant gains in geolocation accuracy.

3.5 Evaluation Metrics

Having reformulated the geolocation prediction task into a discrete class space through the use of our city class set, it is possible to use simple classification accuracy to evaluate our models. However, given that all of our class labels have a location (in the form of latitude–longitude coordinates),

we can also sensitise our evaluation to the distance-based error in predictions. For instance, if the correct location for a user is Seattle, USA, a prediction of Vancouver, Canada is arguably better than a prediction of Los Angeles, USA, on the basis of geospatial proximity. In line with past work (Cheng et al., 2010; Wing and Baldridge, 2011; Roller et al., 2012), we use a number of evaluation metrics which capture spatial proximity, in addition to classification accuracy:

1. **Acc** : the classification accuracy of the highest-probability prediction of the model;
2. **Acc@161** : the classification accuracy of the highest-probability prediction of the model, within a circle of radius of 100 miles (161 kilometres) from the true city centre of the user;
3. **Mean and Median Error**: mean and median prediction error, measured in kilometres between the predicted city centres and the true geolocations.

4 Finding Location Indicative Words

In this section, we experiment with different methods for ranking location indicative words. As a first step, to determine the statistical "signature" of LIWs, we manually pre-identified seed sets of: (1) local words (denoted as 1-local) that are used primarily in a single city, namely *yinz* (used in Pittsburgh to designate locals), *dippy* (used in Pittsburgh to refer to a style of fried egg, or something that can be dipped in coffee, etc.) and *hoagie* (used primarily in Philadelphia, to refer to a kind of sandwich);[10] (2) semi-local words (*n*-local) that refer to some feature of a relatively limited subset of cities, namely *ferry* (found, e.g. in Seattle, New York and Sydney), *Chinatown* (common in many of the largest cities in the USA, Canada and Australia, but much less common in European and Asian cities), and *tram* (found, e.g, in Vienna, Melbourne and Prague); and (3) common words (common) which aren't expected to have substantial regional frequency variation, namely *twitter*, *iphone* and *today*. We use this small set of 9 words to empirically motivate our feature selection approach.

4.1 Decoupling City Frequency and Word Frequency

High-utility LIWs should have both of the following properties:

1. High Term Frequency (TF): there should be a reasonable expectation of observing it for a given user;
2. High Inverse City Frequency (ICF): the term should occur in tweets associated with a relatively small number of cities.

We calculate the ICF of a term i simply as $icf_i = \frac{N}{cf_i}$, where N is the number of cities and cf_i is the number of cities with users who use that term in the training data. Combining the two together, we are seeking words with high "TF-ICF", analogous to seeking terms with high TF-IDF values in information retrieval. As with TF-IDF, however, the exact formulation for calculating the individual values and combining them into a single term weight is not necessarily obvious. A simple TF×ICF product is dominated by the TF component: for example, *twitter* scores as highly as *Jakarta*, because *twitter* has a very high TF. We resolve this issue by decoupling the two factors and applying a radix sort ranking: we first rank features by ICF then by TF, in decreasing order. This procedure has the desired effect of promoting local words and demoting common words. We present evidence for this hypothesis over the pre-identified 1-local, *n*-local and common words in Figure 3.

We can observe that 1-local words have high ICF and relatively low TF, *n*-local words have mid-range ICF and TF values, and common words have low ICF and high TF values, anecdotally justifying our feature ranking method. In order to filter out low-utility words and noise, we only keep words with

[10]These terms are identified with the aid of datasets of regional terms such as DARE `http://dare.wisc.edu/`.

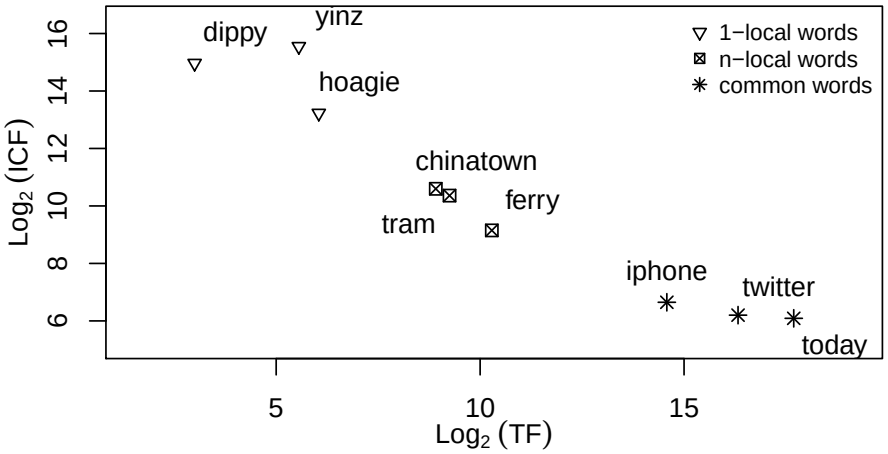

Figure 3: Inverse city frequency vs. term frequency on WORLD

1-local terms	IGR	n-local terms	IGR	common terms	IGR
yinz	0.287	ferry	0.125	iphone	0.012
dippy	0.169	tram	0.188	twitter	0.005
hoagie	0.183	chinatown	0.107	today	0.027

Table 2: Information gain ratio numbers for our sample terms based on WORLD

minimum character length of 3 and $TF \geq 10$ hereafter. As this approach is largely based on the inverse city frequency, we denote it as *ICF* below.

4.2 Information Gain Ratio

In addition to ICF, we also employ Information Gain (IG), an information-theoretic measure of the decrease in entropy a word brings about, where higher values indicate greater predictability on the basis of that feature. Given a set of words $\mathbf{w}$, the IG of a word $w_i \in \mathbf{w}$ across all cities (C) is calculated as follows:

$$IG(w_i) = H(C) - H(C|w_i) \quad (1)$$

$$\propto P(w_i)\sum_{j=1}^{m} P(c_j|w_i)logP(c_j|w_i) + P(\bar{w}_i)\sum_{j=1}^{m} P(c_j|\bar{w}_i)logP(c_j|\bar{w}_i) \quad (2)$$

where $H(C|w_i)$ is the conditional entropy given w_i, which is proportional to $IG(w_i)$.

Words carry varying amounts of "intrinsic entropy", which is defined as $IV(w_i) = -P(w_i)logP(w_i) - P(\bar{w}_i)logP(\bar{w}_i)$. Local/regional words occurring in a small number of cities often have a low intrinsic entropy, where non-local common words have a high intrinsic entropy. For words with comparable IGs, the words with smaller entropies are preferred. Therefore, following Quinlan (1993) we further normalise $IG(w_i)$ using the intrinsic entropy of word $IV(w_i)$, culminating in information gain ratio (IGR): $IGR(w_i) = IG(w_i)/IV(w_i)$. Returning to our earlier sample words, we present IGR scores for WORLD in Table 2.

4.3 Maximum Entropy-based Feature Weights

The previous two feature selection methods optimise across all classes simultaneously. Given that some LIWs may be strongly associated with certain locations, but are less tied to other locations, we also conduct per-class feature selection based on maximum entropy (ME) modelling.[11]

[11] https://github.com/lzhang10/maxent

1-local	Weight	City	*n*-local	Weight	City	common	Weight	City
yinz	6.8e-3	Pittsburgh, US	ferry	1.4e-2	San Francisco, US	iphone	3.2e-2	London, UK
dippy	4.6e-4	Gosport, UK	tram	2.7e-2	Melbourne, AU	twitter	3.7e-2	Bowie, US
hoagie	4.2e-3	Philadelphia, US	chinatown	9.5e-3	Singapore	today	9.3e-2	London, UK

Table 3: ME-based feature weights, and associated cities, for our sample terms based on WORLD.

Given a collection of cities C, the ME model calculates the probability of a user (e.g. represented by word sequence: $w_1, w_2, \ldots, w_n$) assigned to a city c by linearly combining eligible ME feature weights:

$$p(c|w_1, w_2, \ldots, w_n) = \frac{1}{Z} \sum_{k=1}^{m} \lambda_k f_k \tag{3}$$

Here, Z is the normalisation factor, m is the total number of features, and f_k and λ_k are the features and feature weights, respectively. As with other discriminative models, it is possible to incorporate arbitrary features into ME, however, a feature (function) in our task is canonically defined as a word w_i and a city C_j. When w_i occurs in C_j, a feature $f_k(w_i, C_j)$ is denoted as $[class = C_j \wedge w_i \in C_j]$. Each f_k maps to a feature weight denoted as $\lambda_k \in \mathscr{R}$.

Our goal is to estimate feature weights using a ME model. The key idea is that ME features connect a word and a class, with larger λ_k weights indicating stronger word–class associations. Therefore, we should be able to use the learned weights as a means of both ranking features and grouping features by cities. With all features, we generate a weight per term–class pair and rank all weights in decreasing order. Only the first rank position is kept for a given term, and this then forms the final aggregated class-independent feature rank. We don't incorporate a regularizer in our ME model (which can help to avoid over-fitting) because we already removed infrequent words (see Section 4.1), which serves as a basic count-based heuristic regularizer. The comparable results on the development and test sets presented in Sections 5.2 and 5.3 indicate that the feature selection is indeed not overfitting.

We show the ME-based weights, and associated cities, for our sample terms for the WORLD dataset in Table 3. Note that in the case of the 1-local terms *yinz* and *hoagie*, the expected city associations have been learned. Such associations could further be of use to lexicographers in identifying regional usages from social media.

5 Experiments and Analysis

5.1 Comparison of Feature Selection Methods

First we compare the effectiveness of the different feature selection methods in experiments on both NA and WORLD. In total, 214K and 96K features are extracted from the training sections of NA and WORLD, respectively. For each feature selection method, we select the $top\ N\%$ of these features, and then use the selected features in multinomial naive Bayes classification; we compare performance using Acc@161. In this section we consider results only on the development portion (10K held-out users) of each dataset; we use these results to optimise feature selection parameters which we then use in subsequent experiments in the following subsections. Results for NA are shown in Figure 4; results for WORLD are similar and thus omitted from the paper.

For ICF and IGR, Acc@161 rises as the percentage of features selected is increased, and drops dramatically at a very high percentage of selected features. The features which are selected last

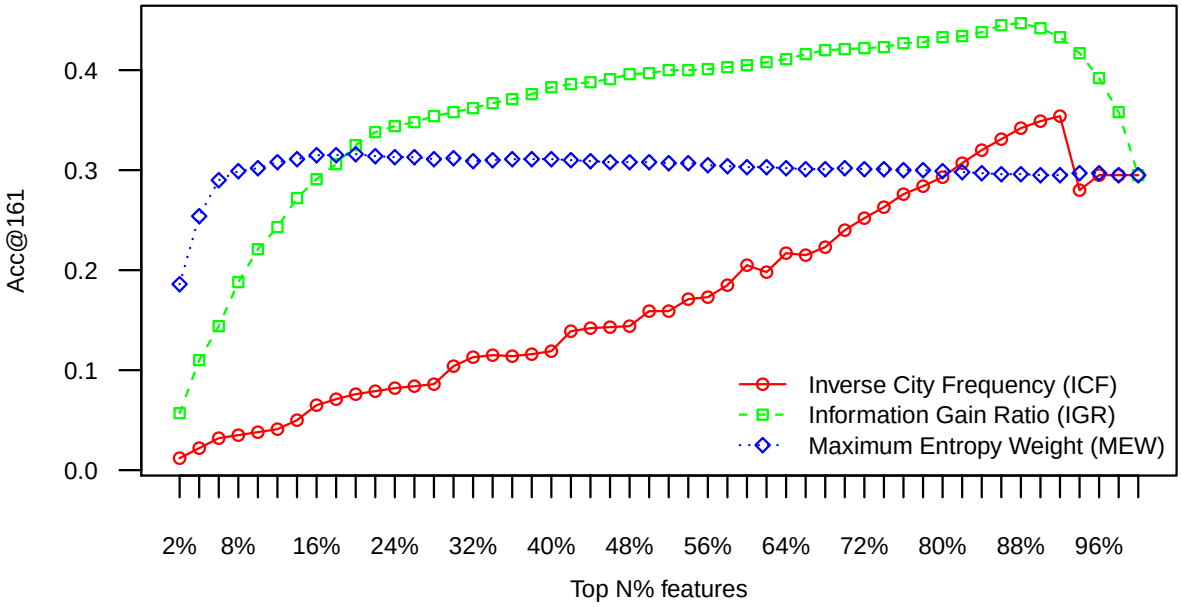

Figure 4: Acc@161 for varying percentages of features selected using the three feature selection methods on the NA dataset. The Optima for ICF, IGR, and MEW are 92%, 88%, and 20%, respectively.

appear to be high-frequency function words (e.g. *the*) and common terms (e.g. *facebook*), which give little indication as to geolocation and lead to prediction errors. By identifying and removing these features, performance can be improved. On the other hand, when insufficient features are used, naive Bayes appears to be under-fitting the training data, and tends to assign classes according to the prior. For instance, when using just the top 2% of features, the most likely class in each case is monterrey-19-mx, because Spanish words are highly location indicative of the small number of Mexican Cities in the NA dataset. For maximum entropy weighting (MEW) we see a very different pattern: we attain the highest Acc@161 using the top 20% features, with performance gradually decreasing as more features are added. The overall poor performance of MEW seems to be due at least in part to the first-occurrence heuristic (see Section 4.3), which causes some non–location indicative words to be ranked higher than local words. Overall, IGR achieves the highest accuracy.

We observe that, as expected, the highly-ranked LIWs for IGR include local dialectal terms (e.g. *yinz*) and place names (e.g. *portland*). We further note the importance of word frequency on location indicativeness. For example, in WORLD the *n*-local term *tram* is roughly an order of magnitude more frequent than the 1-local term *hoagie*, but has higher IGR (see Table 2). Although 1-local words might be useful in geolocation, the impact of infrequent terms on overall accuracy is limited. In particular, frequent words with some geographical association might be more informative for geolocation than words with highly-local distribution but lower frequency. Furthermore, from this analysis it seems that a binary distinction cannot be made between local and non-local words (as in Cheng et al., 2010), but rather that many words carry some geographical indications. As for LIW selection by class (city) with MEW, city names are unsurprisingly strong indicators for locations. For example, *philadelphia* and *philly* are amongst the top-three words associated with philadelphia-pa101-us in NA. Furthermore, upper level administrative regions are also useful for geolocation with, e.g. *georgia* being a strong indicator of atlanta-ga121-us in NA.

Dataset	Features	Mean	Median	Acc@161	Acc
NA	Full	1010	571	0.308	0.171
	ICF	1026	533	0.359	0.209
	IGR	**814**	**260**	**0.450**	**0.260**
	MEW	890	520	0.326	0.183
WORLD	Full	2215	917	0.203	0.081
	ICF	2299	878	0.239	0.107
	IGR	3002	926	**0.259**	**0.124**
	MEW	**1953**	**646**	0.241	0.103

Table 4: Geolocation performance of the full feature set compared to that of each feature selection methodology on both NA and WORLD. The best results for each dataset and accuracy metric are shown in boldface.

5.2 Improved Accuracy with Location Indicative Words

In this subsection we compare the accuracy of classifiers trained using just the optimised LIWs obtained in the previous subsection to that of the full model. The performance is measured on the test data (10K held-out users for both NA and WORLD).

Results on NA show that using LIWs offers an improvement over the full feature set for all evaluation metrics and all feature selection approaches (except for ICF with mean distance). On WORLD the findings are similar in terms of Acc and Acc@161, however, mean distance in particular is substantially higher for IGR. We hypothesize that this is because incorrect world-level predictions can potentially be off by thousands of kilometres, driving up the mean distance more than the other metrics. Overall, these numbers clearly demonstrate that identification of LIWs can improve text-based geolocation. IGR performs best in terms of Acc@161 on both datasets, achieving a 14.2% and 5.6% absolute improvement over the full feature set on NA and WORLD, respectively. Finally, it is worth noting that the raw accuracy on NA is higher than that on WORLD. This is unsurprising because the smaller average number of tweets and larger number of classes for WORLD make it a more challenging dataset.

5.3 Comparison with Benchmarks

We further compare the best-performing method from Section 5.2 to benchmarks and baselines. Here we only consider NA, for which results have been previously reported. We experiment with two partitionings of the Earth's surface: (1) the new city-based division used in the previous experiments, and (2) the KD-tree based partitioning of Roller et al. (2012) which creates grid cells containing roughly even amounts of data, but differing geographical sizes, such that higher-population areas are represented with finer-grained grids. We consider the following methods:

Baseline Because the geographical distribution of tweets is skewed towards higher-population areas (as indicated in Figure 1), we consider a most-frequent class baseline. We assign all users the coordinates of the most-common city centre or KD-tree grid centroid in the training data.

Placemaker Following Kinsella et al. (2011), we obtain results from Yahoo! Placemaker,[12] a publicly-available geolocation service. The first 50K bytes (the maximum query length allowed by Placemaker) from the tweets for each user are passed to Placemaker as queries. The returned city centre predictions are mapped to our collapsed city representations. For queries without results, or with a predicted location outside NA, we back off to the baseline.

KL divergence The previous best published results over NA were achieved using KL divergence

[12] `http://developer.yahoo.com/geo/placemaker/`, accessed in August 2012.

Partition	Method	Mean	Median	Acc@161	Acc
KD-tree	Baseline	1528	1189	0.118	0.003
	KL	859	469	0.344	0.117
	KL+IGR	766	**273**	**0.437**	**0.161**
	NB	835	404	0.367	0.122
	NB+IGR	**763**	280	0.432	0.153
City	Baseline	2707	3089	0.062	0.003
	Placemaker	2188	1857	0.150	0.049
	NB	1010	571	0.308	0.171
	NB+IGR	**814**	**260**	**0.450**	**0.260**
	ME	1336	878	0.232	0.129
	ME+IGR	891	369	0.406	0.229

Table 5: Geolocation performance for baselines, KL divergence (KL), multinomial naive Bayes (NB), and Maximum Entropy (ME). Results using the optimised feature set (+IGR) are also shown. The best-performing method for each evaluation metric and partitioning is shown in boldface.

and the KD-tree grid. Specifically, KL divergence is measured between the distribution of terms in a user's aggregated tweets and that in each grid cell, with the predicted location being the centroid of the most-similar grid cell. We use the same settings as Roller et al..

Multinomial naive Bayes This is the model used in Section 5.2.

Maximum entropy The features from Section 5.2 with a maximum entropy learner.[13]

The results are shown in Table 5. We begin by considering the baseline results. The most-frequent class for the KD-tree grid is New York (the state), while for the city-based partition, it is los angeles-ca037-us. Both baselines perform below the other models, suggesting that geolocation cannot be trivially solved. Looking at the results for Placemaker (which we only consider for the city-based partition) we see very high mean and median scores. This appears to be due to the scope and domain of this service, which predicts locations at the world level, whereas the other methods are restricted to North America.

The KL-divergence method of Roller et al. (KL) and multinomial naive Bayes method (NB) both clearly outperform the baseline. Moreover, approaches incorporating the best features selected in Section 5.2 — KL+IGR and NB+IGR — both outperform KL and NB, demonstrating that for a variety of approaches, identification of LIWs can improve text-based geolocation.[14] From the results on the KD-tree grid it is not decisively clear which of KL or NB is better for our task: in terms of Acc@161, e.g., NB outperforms KL, but KL+IGR outperforms NB+IGR.

Turning to the results for the city-based grid, our best-performing method from Section 5.2 (City, NB+IGR) performs best overall in terms of Acc, Acc@161, and median distance, confirming the effectiveness of LIWs. Compared to the best published results at the time of writing (KD-tree, KL), our method offers a 10.6% absolute improvement in terms of Acc@161, and reduces the mean and median prediction error by 45Km and 209Km, respectively.[15] Finally, although maximum entropy (ME) performs poorly compared to NB, ME+IGR is still a substantial improvement over ME. We

[13] Although there are many other classifiers we could consider, when sufficient training data for each class is available, the performance of different methods is comparable (Yang and Liu, 1999). Moreover, many state-of-the-art classifiers (Wu et al., 2007) are not primarily designed for massively multi-class problems (e.g. support vector machines (Vapnik, 1995)), or are not efficient when applied to such problems (e.g. k-nearest neighbour (Steinbach et al., 2006)).

[14] Note that after LIWs are selected, a small proportion of users end up with no features. These users are not geolocatable in the case of KL, a discriminative model. We turn off feature selection for such users, and backoff to the full feature set, so that the number of test cases is consistent in different settings.

[15] Acc is not comparable for the different partitionings, i.e. KD-tree vs. city, because of the differing numbers of grid cells. High accuracy could trivially be achieved with a very coarse-grained grid.

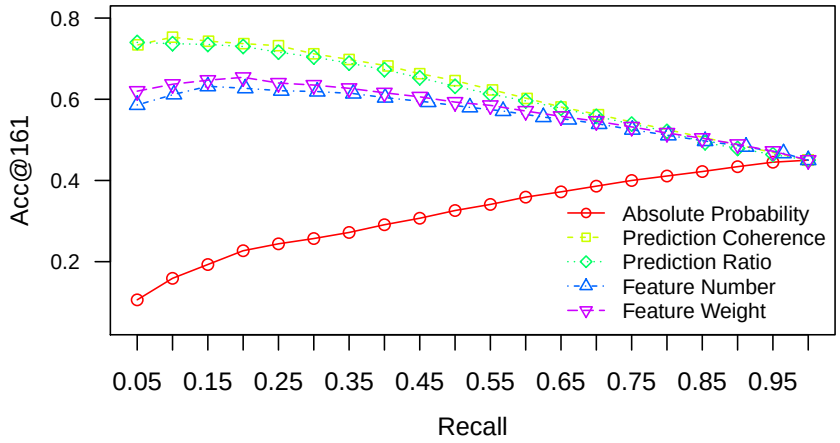

Figure 5: Acc@161 for classification of the top-n% most-confident predictions for each measure of prediction confidence.

plan to further explore the reasons for ME's poor performance in future work.

In addition to improving the accuracy of geolocation, LIW-based feature selection leads to more compact models, which are more efficient in terms of computational processing and memory. Comparing the model based on LIWs selected using IGR with the full model, we find that the prediction time is faster by a factor of roughly five.

5.4 The Confidence of Geolocation Prediction

In the task setup to date, we have forced our models to geolocate all users. In practice, however, many users don't mention any explicitly geolocating terms in their posts, making the task nigh on impossible even for a human oracle. An alternative approach would be to predict a user geolocation only when the model is confident of its prediction. Here, we take our best-performing method (city-based grid, multinomial naive Bayes classifier with LIWs selected using IGR) and consider different methods for selecting users where the model is sufficiently confident of its prediction:

Absolute probability (AP) Only consider predictions with probability above a specified threshold.

Probability ratio (PR) If the model is confident in its prediction, the first prediction will tend to be much more probable than other predictions. We formulate this intuition as PR, the ratio of the probability of the first and second predictions.

Prediction coherence (PC) We hypothesize that for reliable predictions, the top-ranked locations will tend to be geographically close. In this preliminary exploration of coherence, we formulate PC as the sum of the reciprocal ranks of the predictions corresponding to the second-level administrative region in our class representation (i.e. state or province) of the top-ranking prediction, calculated over the top-10 predictions. For example, suppose the top-10 second-level predictions were in the following states: TX, FL, TX, TX, CA, TX, TX, FL, CA, NY. The top-ranking state-level prediction is therefore TX, which also occurs at ranks 3, 4, 6 and 7. In this case, PC would be $\frac{1}{1} + \frac{1}{3} + \frac{1}{4} + \frac{1}{6} + \frac{1}{7}$.

Feature number (FN) We take the number of features found in a user's posts as the prediction accuracy. The intuition here is that a geolocation prediction based on more features is more reliable than a prediction based on less features.

Feature weight (FW) Similar to FN, but in this case we use the sum of IGR of all features, rather than just the number of features.

For this analysis we use the NA dataset. We sort the predictions by confidence (independently for each measure of prediction confidence) and measure Acc@161 amongst the top-n% of predictions for the following values of n: $\{0.0, 0.05, ..., 1.0\}$, akin to a precision–recall curve. Results are shown in Figure 5. The naive AP method is least reliable with, surprisingly, accuracy increasing as AP decreases. It appears that the raw probabilities are not an accurate reflection of prediction confidence. In comparison, PR — which focuses on relative, as opposed to raw, probabilities — performs much better, with higher PR generally corresponding to higher accuracy. Nevertheless, the best-performing method is PC, which only uses the probabilities to rank the class predictions, and roughly captures the geographical proximity of the top predictions, confirming our hypothesis that accuracy will tend to be higher when the top-ranked predictions are relatively near each other. For PC, Acc@161 is about 75% when only the 10% most-confident predictions are considered, which is well above the 45% Acc@161 for the full dataset. FN and FW show similar trends to PC and PR, but don't perform as well. These experiments suggest that there is indeed a trade-off between coverage and accuracy, which could be exploited to obtain higher-accuracy predictions by applications that do not require all the data to be classified.

Discussion and Conclusion

We have investigated various methods for applying feature selection to identify LIWs (location indicative words) for the task of text-based geolocation. Our results on two different datasets demonstrate that using LIWs leads to an improvement over using a full feature set for a variety of evaluation metrics. Furthermore, our best method using LIWs outperforms the previous state-of-the-art on a standardised dataset, and is much faster. These results demonstrate the potential for improving text-based geolocation through feature selection. The LIWs identified by our method, and their associations with particular locations, may also be useful for lexicographers in describing regional usage and variation. We further considered prediction confidence, and showed that it is possible to strike a trade-off between coverage and accuracy; given the very large amount of Twitter data available, a system which gives more-accurate predictions, but only for a subset of the data, may be useful in some applications. Finally, the proposed LIW selection methods, although developed and evaluated on English datasets, could be easily applied in a multilingual setting.

This paper (as well as previous work on this topic) only considered tweets with gold-standard geo-tags, but in an applied setting we envision these models being applied to non-geotagged tweets to infer their locations. However, it might not be the case that geo-tagged tweets (typically sent from a GPS-enabled device such as a smart phone) have the same properties as those which are not geo-tagged (and are sent from a variety of devices, including desktop computers). In future work, we intend to investigate the relationship between these two sources of data. Although the aim of this paper was to examine the relationship between text and location, there are nevertheless further sources of information available on Twitter, such as user profile and social network information, that could be leveraged in a method for geolocation. In future work we intend to consider the incorporation of such information into our methods. Finally, this paper proposed a new city-based representation of locations. We plan to continue in this direction in future work to explore alternative regional partitionings, as well as hierarchical classification methods.

Acknowledgements

NICTA is funded by the Australian government as represented by Department of Broadband, Communication and Digital Economy, and the Australian Research Council through the ICT Centre of Excellence programme.

References

Adams, B. and Janowicz, K. (2012). On the geo-indicativeness of non-georeferenced text. In *Proceedings of Sixth International AAAI Conference on Weblogs and Social Media*, ICWSM '12, Dublin, Ireland.

Amitay, E., Har'El, N., Sivan, R., and Soffer, A. (2004). Web-a-where: geotagging web content. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 273–280, Sheffield, United Kingdom. ACM.

Backstrom, L., Kleinberg, J., Kumar, R., and Novak, J. (2008). Spatial variation in search engine queries. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, pages 357–366, Beijing, China. ACM.

Backstrom, L., Sun, E., and Marlow, C. (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, pages 61–70, Raleigh, North Carolina, USA. ACM.

Bauer, S., Noulas, A., Seaghdha, D. O., Clark, S., and Mascolo, C. (2012). Talking places: Modelling and analysing linguistic content in foursquare. In *Proceedings of The ASE/IEEE International Conference on Social Computing*, SocialCom 2012, Amsterdam, Netherlands.

Bilhaut, F., Charnois, T., Enjalbert, P., and Mathet, Y. (2003). Geographic reference analysis for geographic document querying. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references - Volume 1*, HLT-NAACL-GEOREF '03, pages 55–62. Association for Computational Linguistics.

Buyukokkten, O., Cho, J., Garcia-Molina, H., Gravano, L., and Shivakumar, N. (1999). Exploiting geographical location information of web pages. In *ACM SIGMOD Workshop on The Web and Databases (WebDB'99)*, pages 91–96.

Cheng, Z., Caverlee, J., and Lee, K. (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*, CIKM '10, pages 759–768, Toronto, ON, Canada. ACM.

Crandall, D. J., Backstrom, L., Huttenlocher, D., and Kleinberg, J. (2009). Mapping the world's photos. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 761–770, Madrid, Spain. ACM.

Dalvi, N., Kumar, R., and Pang, B. (2012). Object matching in tweets with spatial models. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 43–52, Seattle, Washington, USA. ACM.

Ding, J., Gravano, L., and Shivakumar, N. (2000). Computing geographical scopes of web resources. In *Proceedings of the 26th International Conference on Very Large Data Bases*, VLDB '00, pages 545–556, Cairo, Egypt.

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2010). A latent variable model for geographic lexical variation. In *Empirical Methods in Natural Language Processing*, pages 1277–1287, Cambridge, MA, USA.

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

Hauff, C. and Houben, G.-J. (2012). Geo-location estimation of flickr images: social web based enrichment. In *Proceedings of the 34th European conference on Advances in Information Retrieval*, ECIR'12, pages 85–96, Barcelona, Spain. Springer-Verlag.

Hong, L., Ahmed, A., Gurumurthy, S., Smola, A. J., and Tsioutsiouliklis, K. (2012). Discovering geographical topics in the twitter stream. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 769–778, Lyon, France. ACM.

Kinsella, S., Murdock, V., and O'Hare, N. (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*, SMUC '11, pages 61–68, Glasgow, Scotland, UK. ACM.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22:49–86.

Leidner, J. L. and Lieberman, M. D. (2011). Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special*, 3(2):5–11.

Li, R., Wang, S., Deng, H., Wang, R., and Chang, K. C.-C. (2012). Towards social user profiling: unified and discriminative influence model for inferring home locations. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, pages 1023–1031, Beijing, China. ACM.

Li, W., Serdyukov, P., de Vries, A. P., Eickhoff, C., and Larson, M. (2011). The where in the tweet. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 2473–2476, Glasgow, Scotland, UK. ACM.

Lieberman, M. D. and Lin, J. (2009). You are where you edit: Locating wikipedia contributors through edit histories. In *ICWSM*.

Lui, M. and Baldwin, T. (2012). langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea.

Mahmud, J., Nichols, J., and Drews, C. (2012). Where is this tweet from? inferring home locations of twitter users. In *Proceedings of Sixth International AAAI Conference on Weblogs and Social Media*, ICWSM '12, Dublin, Ireland.

Mao, H., Shuai, X., and Kapadia, A. (2011). Loose tweets: an analysis of privacy leaks on twitter. In *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society*, WPES '11, pages 1–12, Chicago, Illinois, USA. ACM.

Ng, A. Y. and Jordan, M. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In Thomas G. Dietterich, S. B. and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*. MIT Press.

O'Connor, B., Krieger, M., and Ahn, D. (2010). TweetMotif: Exploratory search and topic summarization for Twitter. In *Proceedings of Fourth International AAAI Conference on Weblogs and Social Media*, pages 384–385, Washington, USA.

Pontes, T., Vasconcelos, M., Almeida, J., Kumaraguru, P., and Almeida, V. (2012). We know where you live: Privacy characterization of foursquare behavior. *4th International Workshop on Location-Based Social Networks (LBSN 2012)*.

Qin, T., Xiao, R., Fang, L., Xie, X., and Zhang, L. (2010). An efficient location extraction algorithm by leveraging web contextual information. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, pages 53–60, San Jose, California. ACM.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, USA.

Roller, S., Speriosu, M., Rallapalli, S., Wing, B., and Baldridge, J. (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, Jeju Island, Korea.

Serdyukov, P., Murdock, V., and van Zwol, R. (2009). Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 484–491, Boston, MA, USA. ACM.

Silva, M. J., Martins, B., Chaves, M. S., Afonso, A. P., and Cardoso, N. (2006). Adding geographic scopes to web resources. *Computers, Environment and Urban Systems*, 30:378–399.

Steinbach, P., Kumar, M., and Tan, V. (2006). Introduction to data mining. *International Edition.–NY.: Addison Wesley*.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA.

Wang, L., Wang, C., Xie, X., Forman, J., Lu, Y., Ma, W.-Y., and Li, Y. (2005). Detecting dominant locations from search queries. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 424–431, Salvador, Brazil. ACM.

Wing, B. P. and Baldridge, J. (2011). Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 955–964, Portland, Oregon, USA. Association for Computational Linguistics.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., and Steinberg, D. (2007). Top 10 algorithms in data mining. *Knowl. Inf. Syst.*, 14(1):1–37.

Yang, Y. and Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 42–49, Berkeley, California, United States. ACM.

Yang, Y. and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 412–420, San Francisco, CA, USA.

Yin, Z., Cao, L., Han, J., Zhai, C., and Huang, T. (2011). Geographical topic discovery and comparison. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 247–256, Hyderabad, India.

Zong, W., Wu, D., Sun, A., Lim, E.-P., and Goh, D. H.-L. (2005). On assigning place names to geography related web pages. In *ACM/IEEE Joint Conference on Digital Libraries*, pages 354–362.