Geolocation Prediction in Twitter Using Location Indicative Words and Textual Features

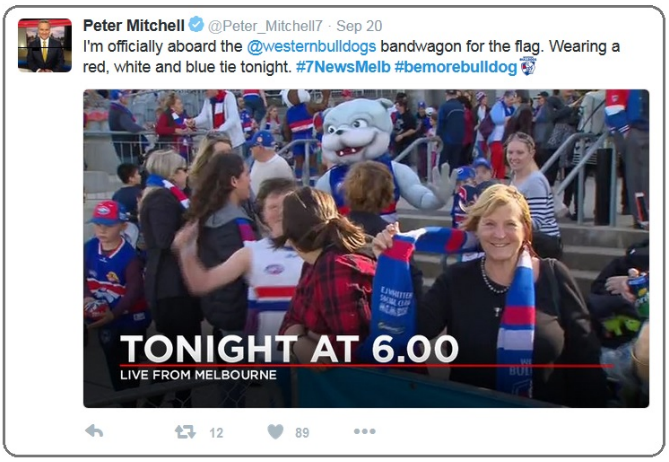
Lianhua Chi[†], **Kwan Hui Lim**[‡†], **Nebula Alam**[†] and **Christopher J. Butler**[†]
[†]IBM Research - Australia
[‡]The University of Melbourne, Australia
lianhuac@au1.ibm.com, limk2@student.unimelb.edu.au,
anebula@au1.ibm.com, chris.butler@au1.ibm.com

Abstract

Knowing the location of a social media user and their posts is important for various purposes, such as the recommendation of location-based items/services, and locality detection of crisis/disasters. This paper describes our submission to the shared task "Geolocation Prediction in Twitter" of the 2nd Workshop on Noisy User-generated Text. In this shared task, we propose an algorithm to predict the location of Twitter users and tweets using a multinomial Naive Bayes classifier trained on Location Indicative Words and various textual features (such as city/country names, #hashtags and @mentions). We compared our approach against various baselines based on Location Indicative Words, city/country names, #hashtags and @mentions as individual feature sets, and experimental results show that our approach outperforms these baselines in terms of classification accuracy, mean and median error distance.

1 Introduction

Determining the location of a social media user and where a message is posted from is important for location-based recommendation (Ye et al., 2010), crisis detection and management (Sakaki et al., 2010), detecting location-centric communities (Lim et al., 2015), demographics analysis (Sloan et al., 2013) and targeted advertising (Tuten, 2008). This work aims to assign a geographical location (most probable location from a list of pre-defined locations, such as cities or countries) to a piece of text. For this textual content, we focus on the Twitter social networking site, which boosts more than 500 million tweets posted on a daily basis (Internet Live Statistics, 2016). In Twitter, tweets are short messages of 140 characters or less, and can also include #hashtags to indicate the topic of the tweet and @mentions to refer to another user. Fig. 1 shows an example of a tweet that contains a text message with a mention of @westernbulldogs, two hashtags of #7NewsMelb and #bemorebulldog, along with an embedded image.

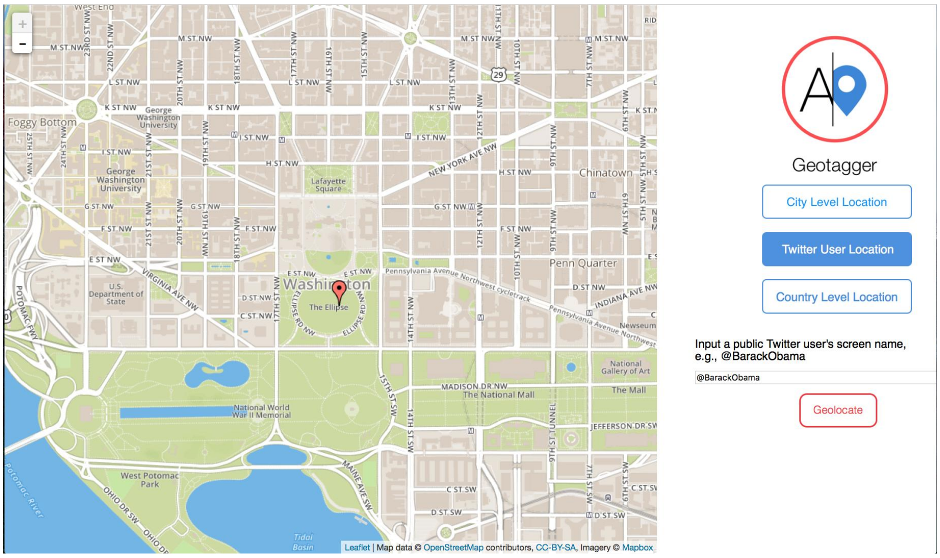


Peter Mitchell @Peter_Mitchell7 · Sep 20
I'm officially aboard the @westernbulldogs bandwagon for the flag. Wearing a red, white and blue tie tonight. #7NewsMelb #bemorebulldog

Figure 1: Example of a tweet message containing a mention (@westernbulldogs), two hashtags (#7NewsMelb and #bemorebulldog) and an attached picture.

Despite the popularity of Twitter and a large volume of tweets, only a small amount of tweets (les than 1%) are geotagged with the location that they were posted from (Sloan et al., 2013), thus restricting the usability of many tweets for location-based services and studies. Due to this motivating factor, the geolocation prediction of tweets and Twitter users have garnered immense interest in recent years. In this paper, we describe our submitted geolocation prediction algorithm for the shared task "Geolocation Prediction in Twitter" in the 2nd Workshop on Noisy User-generated Text (Han et al., 2016). Our algorithm utilizes a multinomial Naive Bayes classifier, using a textual feature set that includes a combination of location indicative words, city/country names, #hashtags and @mentions, which are automatically learnt from a large collection of Twitter data.

Here, we use two examples respectively from tweet level and user level to describe the basic application of our method. From tweet level, for instance, "I plan to take a tram to the federation square this arvo to watch the cricket..." is assigned to Melbourne, Victoria, Australia. The text doesn't contain gazetted terms such as "Melbourne", but our geotagger is able to geolocate this text on basis of location indicative words like "tram", "federation square" and "arvo" and other textual features. From the user level, we can see an example from Figure 2. In Figure 2, the input is the twitter ID of Barack Obama (the president of United States), and the predicted location is Washington.



Input a public Twitter user's screen name, e.g., @BarackObama

@BarackObama

Leaflet | Map data © OpenStreetMap contributors, CC-BY-SA, Imagery © Mapbox

Figure 2: A simple demo from the Twitter user level prediction task.

1.1 Main Contributions

Our main contributions include:

1. We present a geolocation prediction algorithm for Twitter based on a multinomial Naive Bayes classifier, using text-based features that are automatically learnt from tweets.

2. We study the effects of using various feature sets including location indicative words, city/country names, #hashtags, @mentions and a combination of all of the above.

3. We experiment on a Twitter dataset comprising 9.05 million tweets generated by 778K users and show that our approach out-performs a state-of-the-art text-based classifier that solely uses location indicative words.

1.2 Structure and Organization

The rest of the paper is structured as follows. Section 2 describes our proposed approach, including data pre-processing, feature set selection, model training, and evaluation. Section 3 presents the experimental results of our proposed approach and various baselines. Section 4 summarizes our paper and highlights some future directions for geolocation prediction.

2 Proposed Approach

In this section, we describe our proposed approach to the geolocation prediction of Twitter users and tweets. Our proposed approach comprises three main phases, namely: (i) data pre-processing to identify the set of textual features; (ii) model training to train our prediction algorithm based on multinomial Naive Bayes classifier; and (iii) evaluating our prediction algorithm on the development and testing sets.

2.1 Data Preprocessing and Feature Set Selection

Our pre-processing of tweets in the training set comprises the following: (i) converting all text in the tweets to lowercase; (ii) removing all punctuation characters; (iii) tokenizing tweets into individual words based on whitespaces. These processed tweets are then used as input to our multinomial Naive Bayes classifier, where the usage frequency count of a set of feature words is derived from these processed tweets. We now describe the various feature sets used in our experiments, which are:

1. **Location Indicative Words (LIW)**. The set of words that are indicative of a specific city, as generated in (Han et al., 2014; Han et al., 2012). These LIWs are uni-grams that are used in only one or a small subset of all cities in the world and are selected from the set of all unigrams based on their information gain ratio.

2. **City/Country Names (CC)**. The set of city names and country names as listed in the GeoNames database (GeoNames, 2016) and U.S. Department of State list of countries (U.S. DoS, 2016). The basic intuition is that a Twitter user is more likely to mention a city or country that this user is residing in, compared to a city or country that is not.

3. **#Hashtags (HASH)**. The set of #hashtags used in our training set of tweets, we select the top 10,000 #hashtags based on their usage frequency. Twitter #hashtags are typically used to indicate the topic associated with the tweet and this choice of feature set allows us to capture any location-based topics that are indicative of a specific city or country, e.g., #bemorebulldog is used to support the Western Bulldogs, an Australian Football League team that is based in Victoria, Australia.

4. **@Mentions (MENT)**. The set of @mentions used in our training set of tweets, we select the top 10,000 @mentions based on their usage frequency. Twitter @mentions are used as a reference to another user and appears on the wall/notification page of that user. Similar to our use of the HASH feature set, the MENT feature set allows us to infer the location of a user based on who he/she is mentioning, e.g., @7NewsMelbourne is the twitter username of a Melbourne-based news broadcaster.

5. **Combination of all the above (ALL)**. A combination of the LIW, CC, HASH and MENT feature sets as a single feature set.

Using these feature sets, we then proceed to train our multinomial Naive Bayes classifier, which we describe in more details in the next section.

2.2 Training of multinomial Naive Bayes classifier

The multinomial Naive Bayes classifier has been frequently used for various text classification tasks such as sentiment analysis (Melville et al., 2009), news categorization (Kibriya et al., 2004), among others. Using the feature sets described in Section 2.1, we apply a multinomial Naive Bayes classifier to our five different set of features using a bag-of-features approach, which we now describe in greater detail.

Given that C is the set of all cities (i.e., our labels) and T is the set of all tweets in our training set, our aim is to geotag each tweet $t \in T$ with a city $c \in C$ such that the probability $P(c|t)$ is maximized. We utilize a bag-of-features approach and represent each tweet t as a set of features $f_i \in t$ (out of N total features), where each feature f_i indicates the number of times (frequency count) that a feature word f_i is used in a tweet t. Thus, we have:

$$\arg\max_{c \in C} P(c|t) = \arg\max_{c \in C} P(c) \prod_{1<i<N} P(f_i|c) \tag{1}$$

Given that t_c is the set of all tweets that are posted in a specific city c and T is the set of all tweets, we can calculate the prior probability based on:

$$P(c) = \frac{|t_c|}{|T|} \tag{2}$$

To cater for feature words that may not appear in our training set, we apply Laplace smoothing by adding 1 to the frequency count of each feature word. Let $Freq_f, c$ be the frequency count of a feature word f in tweets from a specific city c, we have the conditional probability:

$$P(f_i|c) = \frac{Freq_f, c + 1}{\sum_{f_x=1}^{N} Freq_{f_x,c} + N} \tag{3}$$

Our geolocation prediction task is targetted at two levels, namely: (i) at tweet-level for individual tweets; and (ii) at user-level for an individual user based on his/her collection of posted tweets. For the tweet-level prediction task, Equation 1 is defined based on individual tweets and can be utilized for this task. For the user-level prediction task, there are two possible approaches to apply Equation 1 for a single user, namely:

1. **Tweet Aggregation As One**. In this approach, we aggregate all tweets posted by a user as a single document and apply Equation 1 on this combined document. As our multinomial Naive Bayes classifier accounts for the frequency of feature word usage, this approach is suitable for the users who are likely to mention words associated with his/her home city multiple times in their tweets, compared to cities that they do not reside in.

2. **Most Frequent Tweet City**. Alternatively, we can apply Equation 1 to the collection of tweets posted by a user, then perform a frequency count of each city that is labelled to each tweet. Thereafter, the home location of a user is assigned based on the city that has the highest frequency count.

We performed some preliminary experiments to evaluate both approaches and found that "Tweet Aggregation As One" (Approach 1) out-performs "Most Frequent Tweet City" (Approach 2). As such, we selected Approach 1 for the user-level geolocation task and use it in for the remaining of this paper.

2.3 Algorithms and Baselines

There are different variants of our geolocation prediction algorithms, with each algorithm differing based on the different set of features (described in Section 2.1) that it was trained with. The five algorithms used in our paper are:

1. **MNB-LIW**: A multinomial Naive Bayes classifier trained using the LIW feature set, i.e., the set of location indicative words. This algorithm also serves as the baseline for the current state-of-the-art on text-based geolocation prediction on Twitter (Han et al., 2012).

2. **MNB-CC**: A multinomial Naive Bayes classifier trained using the CC feature set, i.e., the list of city and country names.

3. **MNB-HASH**: A multinomial Naive Bayes classifier trained using the HASH feature set, i.e., the top 10,000 #hashtags used in our training set.

4. **MNB-MENT**: A multinomial Naive Bayes classifier trained using the MENT feature set, i.e., the top 10,000 @mentions used in our training set.

5. **MNB-ALL**: A multinomial Naive Bayes classifier trained using a combination of location indicative words, city/country names, #hashtags and @mentions as a single feature set.

6. **MNB-PART**: Same as MNB-ALL, except that we select a subset of features from the combined feature set, using a collection frequency-based feature selection strategy (Manning et al., 2008).

2.4 Evaluation Metrics

To evaluate these six algorithms, we use the following evaluation metrics:

- **Accuracy**. The proportion of tweets (and users) that is correctly classified to their home location (city), out of all tweets (and users). This metric allows us to measure the correctness of our prediction algorithm in terms of percentage of correct labelled cities.

- **Mean Error Distance**. The average error, in terms of distance, between the predicted cities and the ground truth cities of the tweets (and users). Even for mislabelled cities, a mislabelled city that is nearer to the ground truth city is deemed better, e.g., New York mislabelled as Chicago, is considered better than New York mislabelled as London. This metric aims to measure this aspect.

- **Median Error Distance**. The median error, in terms of distance, between the predicted cities and the ground truth cities of the tweets (and users). Similar to the Mean Error Distance, except that we are measuring the error distance in terms of median values

3 Experiments and Results

In this section, we describe the dataset used in our experiment, highlight our experimental setup and discuss the key results of our proposed algorithm and various baselines.

3.1 Dataset Description

As part of the shared task, a total of 12 million tweets were made available, which the participants have to retrieve via the Twitter API. However, due to inactive Twitter accounts, deleted tweets and time constraint, we were only able to crawl a total of 9.05 million tweets. Similarly, for the validation dataset, we were only able to crawl 7,215 users and 7,789 tweets out of the 10,000 users and 10,000 tweets. As the testing dataset was directly provided by the organizers, we were able to experiment on all 10,000 users and tweets. Table 1 shows the summary statistics of our training, validation, and testing dataset.

Table 1: Dataset description

Set	Prediction Task	No. of Users	No. of Tweets
Training	N.A.	778,383	9,053,573
Validation	User-level	7,215	-
Validation	Tweet-level	-	7,789
Testing	User-level	10,000	-
Testing	Tweet-level	-	10,000

3.2 Experimental Setup

Our experimental setup is aligned to the setup of the shared task and comprises three main phrases. In the first phrase, we use the training dataset to extract our various feature sets and train our geolocation predictors (MNB-LIW, MNB-CC, MNB-HASH, MNB-MENT, MNB-ALL, MNB-PART). In the second

Table 2: Results for Tweet-level Geolocation Prediction. The bolded results indicate the best performing statistics (highest value for accuracy and lowest value for mean/median error) for the validation and testing set.

Algorithm	**Set**	**Accuracy**	**Mean Error**	**Median Error**
MNB-LIW	Validation	0.1013	8751.9379	8153.5067
MNB-CC	Validation	0.0608	12438.9015	10790.024
MNB-HASH	Validation	0.0815	5293.6749	6216.3957
MNB-MENT	Validation	0.0449	11576.5707	9552.6721
MNB-ALL	Validation	0.1163	3314.5639	4993.7693
MNB-PART	Validation	**0.1221**	**3129.8084**	**4933.3203**
MNB-LIW	Testing	0.125	7778.6698	7453.0552
MNB-CC	Testing	0.0862	11395.3637	9439.0232
MNB-HASH	Testing	0.0991	5532.4149	6379.0754
MNB-MENT	Testing	0.0461	9458.3402	9016.1264
MNB-ALL	Testing	0.1376	3582.5483	5457.1922
MNB-PART	Testing	**0.1455**	**3424.6398**	**5338.8984**

phase, we evaluate our trained predictors (models) on the validation dataset. For the final phrase, we re-run our various predictors on the testing dataset. As the ground truth labels for the testing dataset was made available only after the shared task, we selected the best performing predictor (MNB-PART) from the second phrase and submitted it to the shared task for the third phrase.

3.3 Results

Table 2 shows the geolocation prediction results for the tweet-level, in terms of accuracy, mean and median error distances. The results show that our proposed MNB-PART algorithm outperforms all baselines for the validation and testing datasets, in terms of all three evaluation metrics. In particular, MNB-PART shows a relative improvement of more than 16% compared to MNB-LIW in terms of accuracy, and offers predictions with less than half the mean error distances than that of MNB-LIW.

Similarly, Table 3 shows the user-level geolocation prediction results, in terms of the same three evaluation metrics. In terms of accuracy, MNB-PART shows a relative improvement of 7.1% over MNB-LIW for the testing dataset, while MNB-LIW out-performs MNB-PART by 4.5% for the validation dataset. In terms of mean and median error distances, MNB-PART incurs less than half the mean error distances and less than two-thirds of the median error distances for the testing dataset.

4 Conclusion and Future Work

In this paper, we proposed an approach for geolocation prediction on Twitter based on a multinomial Naive Bayes classifier using a feature set derived from the textual features of tweets. Specifically, our feature set is based on a combination of location indicative words, city/country names, #hashtags and @mentions as a combined feature set. Instead of using the entire feature set, our approach uses a subset of the features, which are selected based on a frequency-based feature selection strategy. We compared our proposed approach MNB-PART against various baselines that uses location indicative words, city/country names, #hashtags and @mentions separately as independent feature sets, i.e., MNB-LIW, MNB-CC, MNB-HASH and MNB-MENT, respectively. The experimental results show that MNB-PART outperforms all baselines in most cases, including against its counterpart MNB-ALL that utilizes all features. These results show the effectiveness of using a combined feature set of location indicative words, city/country names, #hashtags and @mentions, then employing a feature selection strategy to use a sub-

Table 3: Results for User-level Geolocation Prediction. The bolded results indicate the best performing statistics (highest value for accuracy and lowest value for mean/median error) for the validation and testing set.

Algorithm	Set	Accuracy	Mean Error	Median Error
MNB-LIW	Validation	**0.2046**	662.3987	2970.8039
MNB-CC	Validation	0.1476	8605.2476	7994.2556
MNB-HASH	Validation	0.1609	872.7039	2706.0175
MNB-MENT	Validation	0.0973	3377.6792	5292.4654
MNB-ALL	Validation	0.1767	928.2935	2599.0432
MNB-PART	Validation	0.1953	**629.8692**	**2290.3172**
MNB-LIW	Testing	0.21	1373.1077	4533.2978
MNB-CC	Testing	0.1976	5207.3125	7037.8555
MNB-HASH	Testing	0.1812	961.3378	3352.8145
MNB-MENT	Testing	0.0944	4650.7975	5933.8057
MNB-ALL	Testing	0.1996	962.144	3234.4641
MNB-PART	Testing	**0.225**	**630.2436**	**2860.1745**

set of these features for training and testing on a multinomial Naive Bayes classifier.

Our work focuses on the use of textual features in tweets for geolocation prediction. Apart from textual features, there are interesting future directions for utilizing non-textual features of a user for this geolocation prediction task, such as:

1. Using friendship (bi-directional) and following (uni-directional) links to infer the location of a user based on his/her friends and followings.

2. Using demographics information listed in a user's profile, such as the user description, selected timezone, user-entered location field.

3. Using the temporal information in posted tweets, use the time distribution of a user's tweeting activities to determine possible time zones and thus cities that a user is in.

References

GeoNames. 2016. Geonames database. Internet. http://www.geonames.org/.

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

Bo Han, Afshin Rahimi, Leon Derczynski, and Timothy Baldwin. 2016. Twitter Geolocation Prediction Shared Task of the 2016 Workshop on Noisy User-generated Text. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (W-NUT)*. ACL.

Internet Live Statistics. 2016. Twitter usage statistics. Internet. http://www.internetlivestats.com/twitter-statistics/.

Ashraf M. Kibriya, Bernhard Pfahringer Eibe Frank, and Geoffrey Holmes. 2004. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 17th Australasian Joint Conference on Artificial Intelligence (AI'04)*, pages 488–499.

Kwan Hui Lim, Jeffrey Chan, Christopher Leckie, and Shanika Karunasekera. 2015. Detecting location-centric communities using social-spatial links with temporal constraints. In *Proceedings of the 37th European Conference on Information Retrieval (ECIR'15)*, pages 489–494.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Introduction to information retrieval. Cambridge University Press.

Prem Melville, Wojciech Gryc, and Richard D. Lawrence. 2009. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'09)*, pages 1275–1284.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web (WWW'10)*, pages 851–860.

Luke Sloan, Jeffrey Morgan, William Housley, Matthew Williams, Adam Edwards, Pete Burnap, and Omer Rana. 2013. Knowing the tweeters: Deriving sociologically relevant demographics from twitter. *Sociological Research Online*, 18(3):7.

Tracy L Tuten. 2008. *Advertising 2.0: Social Media Marketing in a Web 2.0 World: Social Media Marketing in a Web 2.0 World*. ABC-CLIO.

U.S. DoS. 2016. A-z list of country. Internet. http://www.state.gov/misc/list/.

Mao Ye, Peifeng Yin, and Wang-Chien Lee. 2010. Location recommendation for location-based social networks. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems (SIGSPATIAL'10)*, pages 458–461.