Geolocation of Tweets with a BiLSTM Regression Model

Piyush Mishra
University of Colorado
first.last@colorado.edu

Abstract

Identifying a user's location can be useful for recommendation systems, demographic analyses, and disaster outbreak monitoring. Although Twitter allows users to voluntarily reveal their location, such information isn't universally available. Analyzing a tweet can provide a general estimation of a tweet location while giving insight into the dialect of the user and other linguistic markers. Such linguistic attributes can be used to provide a regional approximation of tweet origins. In this paper, we present a neural regression model that can identify the linguistic intricacies of a tweet to predict the location of the user. The final model identifies the dialect embedded in the tweet and predicts the location of the tweet.

1 Introduction

Social media platforms are useful sources for doing data analysis on linguistic patterns. Twitter is one such platform and allows access to tweets posted by the users via APIs. While the metadata of a tweet is accessible, only 3% of tweets include geotagging information (Jurgens et al., 2015). This severely limits the use of a metadata-based approach for location identification. Even if a geolocation model performs well with such metadata, it will fail in the absence of such information. Since tweets are rich in information, a model that relies only on the textual content of tweets will have a better chance of performing consistently.

The location prediction task can be addressed either as a classification problem or a regression problem. The former can be implemented in two ways; by treating known locations as classes or by dividing the entire region at hand into a grid and using the center of the sub-regions as classes. A regression approach can be seen as a simple double regression task where a model predicts the latitude and longitude of a tweet separately, based on its features. Previous studies have shown that tweets, along with their metadata, can be used to predict a user's location (Han et al., 2014). Neural models have been used to better utilize the metadata, along with the tweets (Thomas and Hennig, 2017).

This paper presents an approach to predict the location of a user using a neural model trained solely on the tweets' text content, without any external knowledge sources. This allows the model to generalize more easily to new domains and languages. This paper intends to get justifiable results at a low cost and with limited resources.

The remainder of this paper is organized as follows: The first section provides an overview of related works on location prediction. Section 3 described the datasets used. Section 4 describes the details of the neural network architecture. Results on the validation set are shown in Section 5. Finally, Section 6 concludes the paper with possible future work.

2 Related Work

One of the earliest works on text-based geotagging (Ding et al., 2000) used named entities such as cities, states, etc. mentioned in the text to identify the geographic scope of the web pages in a classification setup. Later, rule-based methods were employed to identify the locations (Bilhaut et al., 2003), while

Proceedings of the 7th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects, pages 283–289
Barcelona, Spain (Online), December 13, 2020

others used named entity recognition in combination with various machine learning techniques (Qin et al., 2010). All these approaches rely on the fact that there is a mention of some location in the text which may not always be the case. Other approaches include supervised classification (Kinsella et al., 2011; Wing and Baldridge, 2011) using words as features and unsupervised learning, where clustering techniques are employed based on topic modeling (Ahmed et al., 2013; Hong et al., 2012).

Some of the previous research has focused on leveraging the metadata associated with the tweet to improve the performance. Many results submitted in the Workshop on Noisy User-generated Text (WNUT2016) (Han et al., 2014) used such metadata heavily. Models were built that focused on user-declared locations, timezone values, and user self-descriptions in addition to the content of the tweet itself (Miura et al., 2016). As part of feature pre-processing, tweet-level geo coordinates were repalced with geo coordinates of the corresponding tweet cities using several mapping services such as GeoNames and time zone boundaries. Finally, neural networks were trained using the FastText n-gram model (Joulin et al., 2017) on posted text, user location, user descriptions, and user timezones.

Ensemble methods are also common, often relying on several information resources like text, user location text, user time zone information, messenger source, and reverse country look-ups for URL mentions. Jayasinghe et al. (2016) relies on specific URL mentions and screened website metadata for geographic coordinates.

Multinomial naïve Bayes methods have been employed that focused on the use of textual features (i.e., location indicative words, GeoNames gazetteers, user mentions, and hashtags) (Chi et al., 2016). Since the authors are using location indicative words, city/country names, hashtags and @mentions as a combined feature set, even though it relies on textual data more than the metadata, dialect related information isn't covered.

Related work to date covers a wide range of languages and dialects including Dutch (Wieling et al., 2011), British (Szmrecsanyi, 2008), American English (Huang et al., 2015; Eisenstein et al., 2010) and African American Vernacular English (Jones, 2015).

3 Data Set

We use the data provided for SMG sub-task (Ljubešić et al., 2016) (Hovy and Purschke, 2018) of the VarDial 2020 shared task is used for the training and testing purpose of the model (Găman et al., 2020). Using data from social media platforms like Twitter and Jodel, three language areas are covered. Each language has a training, a validation, and a test file. Each file contains entries separated by newlines and each entry contains a latitude, a longitude, and a text tuple, separated by tabs.

- Standard German Jodels (DE-AT): Jodel conversations initiated in Germany and Austria, which are written in standard German but commonly contain regional and dialectal forms.

- Swiss German Jodels (CH): Jodel conversations from Switzerland, in Swiss German dialects.

- BCMS Tweets (BCMS): Geo-located tweets published in the area of Croatia, Bosnia and Herzegovina, Montenegro, and Serbia.

The distribution of all three datasets across latitudes and longitudes are shown in Figure 1, Figure 2, and Figure 3 below.

Language	Training Set	Validation Set	Test Set
DE-AT	336983	46582	48239
CH	22600	3068	3097
BCMS	320042	39750	39723

Table 1: Number of data points per set per language

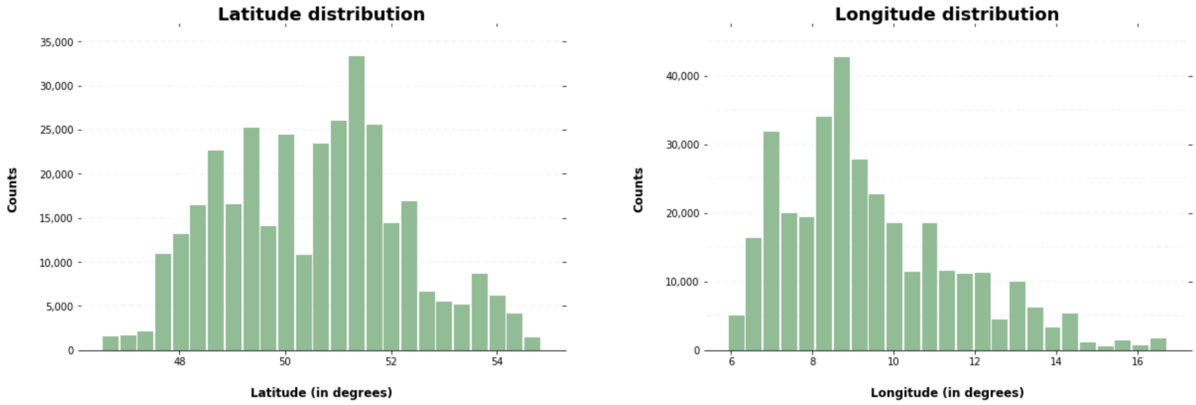


Figure 1: Dataset distribution of DE-AT Jodels across latitudes and longitudes

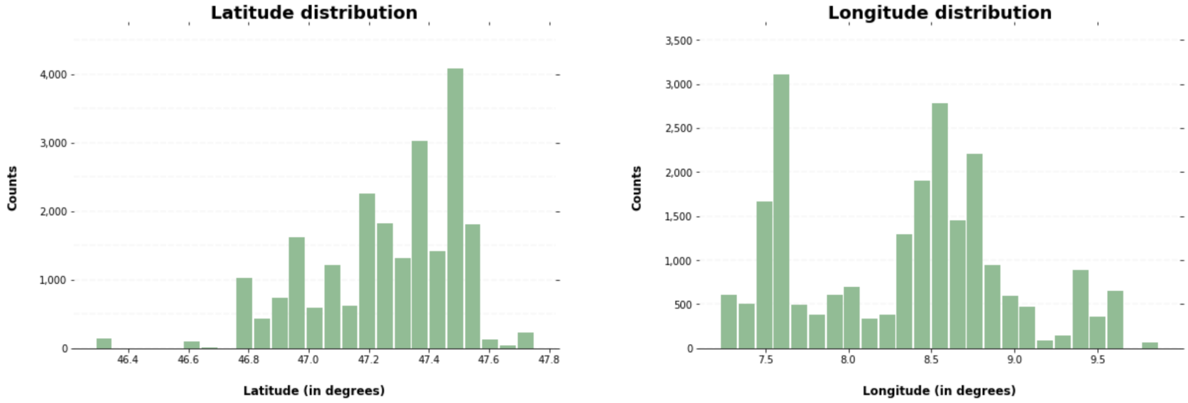


Figure 2: Dataset distribution of CH Jodels across latitudes and longitudes

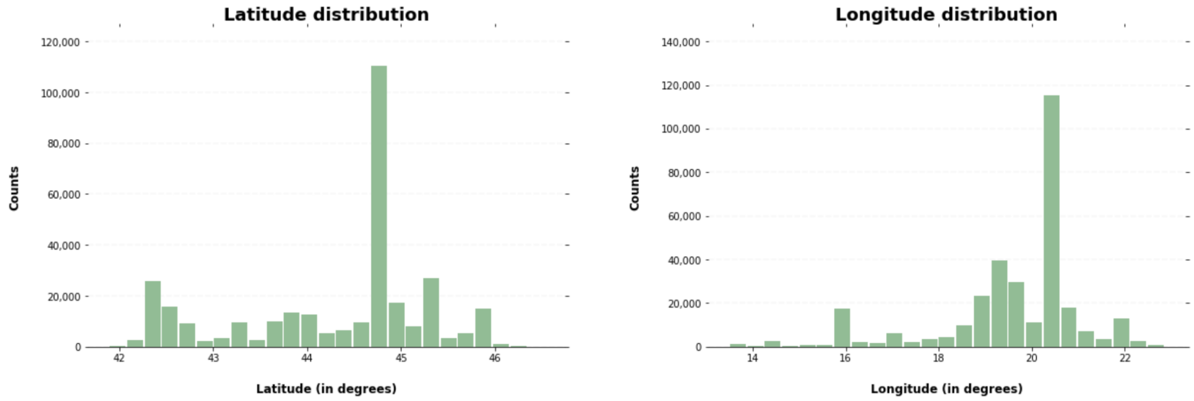


Figure 3: Dataset distribution of BCMS Tweets across latitudes and longitudes

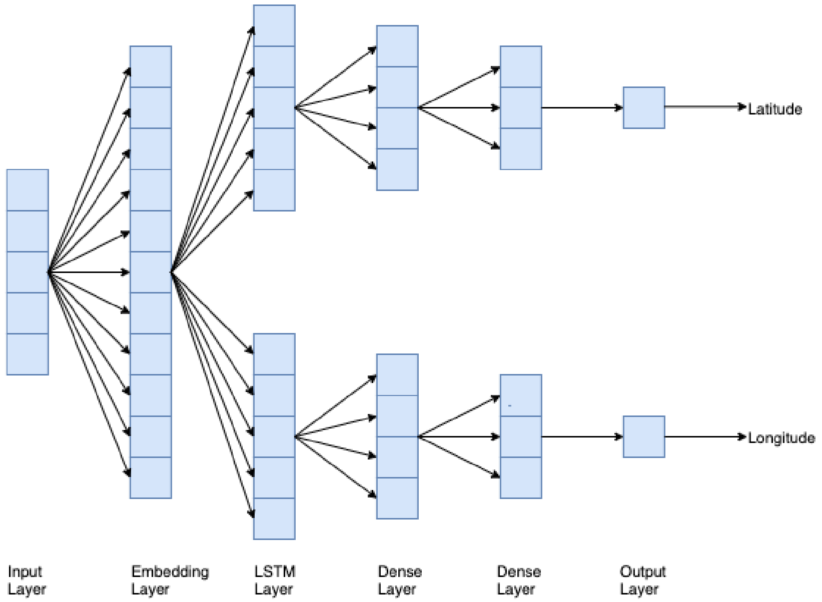
4 Architecture

To identify the location of a given tweet, a double regression approach has been taken. The model will identify the latitude and longitude separately for a given text and the distance between the actual and predicted location will be used to measure its performance. For this purpose, we are using two bidirectional LSTM models.

For text pre-processing, we use a tweet tokenizer as part of the cleanup, to remove unwanted punctuation and lower-case conversion. We use a list of German stopwords to further clean the DE-AT dataset. We decided to keep the stopwords for Swiss German due to the smaller size of the dataset. For BCMS, stopwords from multiple languages are needed which are not as consistent as German, so those are included in the dataset as well. Using TF-IDF weighting allows us to focus on the highly relevant tokens in the text. Using this new set of relevant tokens, the tweets are updated so that irrelevant tokens are removed. A FastText model is trained on this new, modified training set, unsupervised. This trained FastText model is used to produce an embedding matrix. The unsupervised training allows the overall model to be implemented in a similar fashion for all 3 languages.

A neural model is created with a trainable embedding layer, an embedding matrix from the FastText model is used for obtaining embeddings, and two bidirectional LSTM layers are used, one for latitude training and one for longitude. Using a trainable embedding layer with the embedding matrix reduces the training epochs required for the model to converge. Since both LSTM layers share a common embedding layer, both have to rely on the same set of embeddings for learning. Two dense layers are added before the output layer. This allows the model to adjust the weights better for a gradual change. A visual representation of the architecture is presented in Fig. 4.

Figure 4: Neural model architecture



Because the evaluation metric for this task is the median distance loss over the test data, Quantile function with 0.5 as alpha is used as the loss function. While a neural model with MSE or MAE as the loss function predicts better for individual tweets, the distance loss over the entire set becomes greater than the quantile function. On the other hand quantile function takes into account the distribution of the dataset while training and, with alpha set to 0.5, it estimates the median instead of mean. For regularization, since L1 brings weights of unimportant features to zero, only contributing features remain. This helps in better feature selection. We use ReLU as activation functions and Adagrad as the optimizer for adaptive updates of frequently and infrequently occurring feature parameters.

5 Results

The performance of the model on validation and test data is shown in Table 2. Median loss (distance in km) has been used as the metric.

Language	Validation Set	Test Set
DE-AT	174.049	183.99
CH	27.39	27.31
BCMS	95.318	85.70

Table 2: Median distance (in kilometers) over the dataset.

Language	Min Latitude	Max Latitude	Min Longitude	Max Longitude
DE-AT	46.53	54.84	5.92	16.72
CH	46.29	47.75	7.23	9.87
BCMS	41.88730403	46.52582591	13.49204277	22.8606

Table 3: Ranges of latitudes and longitudes with respect to the dataset

Rank	Team	Median distance	Mean distance
1	helsinki-ljubljana	159.59	183.97
2	**This work**	**183.99**	**204.93**
3	CUBoulder-UBC	198.27	218.51
4	ZHAW	205.81	230.78
5	SUKI	243.12	266.85

Table 4: The best results of each team participating on the SMG 2020 shared task DEAT track.

Rank	Team	Median distance	Mean distance
1	ZHAW	15.93	25.06
2	helsinki-ljubljana	17.66	26.21
3	CUBoulder-UBC	19.49	27.63
4	SUKI	23.96	34.59
5	UnibucKernel	25.57	30.52
6	The lingustadors	26.70	31.21
7	**This work**	**27.31**	**33.20**

Table 5: The best results of each team participating on the SMG 2020 shared task Swiss German track.

Rank	Team	Median distance	Mean distance
1	helsinki-ljubljana	48.99	86.83
2	ZHAW	57.24	100.42
3	SUKI	61.01	105.11
4	CUBoulder-UBC	64.76	106.67
5	**This work**	**85.70**	**112.65**
6	The lingustadors	97.16	141.88

Table 6: The best results of each team participating on the SMG 2020 shared task BCMS track.

It can be observed that the deviation of validation and test result is not more than 10km for DE-AT and BCMS dataset and much smaller for the CH dataset. The marginal deviation on the CH dataset can be attributed to the small size of the dataset. Another factor responsible for the scale of deviation for DE-AT and BCMS vs CH is the range of latitudes and longitudes. Below is the range of latitudes and longitudes for all 3 datasets.

From Table 2, it can be observed that, with minimal configuration and constrained training, the model is performing equally well or even better for test data than the validation data. In constrained submissions, our model gave the second-best result for the DE-AT dataset. The performance of all the participating teams for the SMG shared task is shown in Table 4, Table 5, and Table 6.

6 Conclusion and Future Work

We have presented a neural network architecture that addresses the tweet location prediction as a dual regression task. No custom processing of the text was done besides a generic tweet tokenization. As part of the cleanup, stopwords were removed from the Standard German dataset only. No pre-trained model or embedding was used which made the model generic enough to be used for any dataset. The model was trained on Google colab with limited GPU access with a maximum training time was around 8hrs.

In the future, we aim to use the distance between the predicted and actual locations as the main loss parameter to train the model. This should give a better insight into the usage of quantile function since this task focused on median distance loss as the performance metric. We will also analyze the performance of our model against a classification model where the known tweet-emission locations are used as a constrained set of labels.

References

Amr Ahmed, Liangjie Hong, and Alexander Smola. 2013. Hierarchical geographical modeling of user locations from social media posts. pages 25–36, 05.

Frédérik Bilhaut, Thierry Charnois, Patrice Enjalbert, and Yann Mathet. 2003. Geographic reference analysis for geographic document querying. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references*, pages 55–62.

Lianhua Chi, K. Lim, N. Alam, and C. Butler. 2016. Geolocation prediction in twitter using location indicative words and textual features. In *NUT@COLING*.

Junyan Ding, Luis Gravano, and Narayanan Shivakumar. 2000. Computing geographical scopes of web resources. *Proceedings of the 26th VLDB Conference*, 12.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 conference on empirical methods in natural language processing (EMNLP)*, pages 1277–1287.

Mihaela Găman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, and Marcos Zampieri. 2020. A Report on the VarDial Evaluation Campaign 2020. In *Proceedings of the Seventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Bo Han, Paul Cook, and Timothy Baldwin. 2014. Text-based twitter user geolocation prediction. *J. Artif. Intell. Res.*, 49:451–500.

Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander Smola, and Kostas Tsioutsiouliklis. 2012. Discovering geographical topics from twitter streams. In *Proceedings of The 21st International World Wide Web conference (WWW)*, 04.

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394, Brussels, Belgium. Association for Computational Linguistics.

Yuan Huang, Diansheng Guo, Alice Kasakoff, and Jack Grieve. 2015. Understanding U.S. regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, 54, 12.

Gaya Jayasinghe, Brian Jin, J. McHugh, B. Robinson, and Stephen Wan. 2016. CSIRO Data61 at the WNUT Geo Shared Task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 218–226.

Taylor Jones. 2015. Toward a description of African American Vernacular English dialect regions using "Black Twitter". *American Speech*, 90:403–440, 11.

Armand Joulin, E. Grave, P. Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. *ArXiv*, abs/1607.01759.

David Jurgens, T. Finethy, James McCorriston, Yi Tian Xu, and D. Ruths. 2015. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. In *ICWSM*.

Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. 2011. "i'm eating a sandwich in Glasgow": modeling locations with tweets. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 61–68.

Nikola Ljubešić, Tanja Samardžić, and Curdin Derungs. 2016. TweetGeo - a tool for collecting, processing and analysing geo-encoded linguistic data. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3412–3421, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Yasuhide Miura, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma. 2016. A simple scalable neural networks based model for geolocation prediction in Twitter. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*.

Teng Qin, Rong Xiao, Lei Fang, Xing Xie, and Lei Zhang. 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*, pages 53–60.

Benedikt Szmrecsanyi. 2008. Corpus-based dialectometry: Aggregate morphosyntactic variability in British English dialects. *International Journal of Humanities and Arts Computing*, 2:279–296, 10.

Philippe Thomas and Leonhard Hennig. 2017. Twitter geolocation prediction using neural networks. In *International Conference of the German Society for Computational Linguistics and Language Technology*, pages 248–255. Springer.

Martijn Wieling, John Nerbonne, and R. Harald Baayen. 2011. Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PloS one*, 6(9).

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