

Predicting the Geolocation of Tweets Using BERT-Based Models Trained on Customized Data

Kateryna Lutsai and Christoph H. Lampert

Institute of Science and Technology Austria (ISTA), Austria;
National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Ukraine

Received: March 14, 2023; returned: ; .

Abstract: This research is aimed to solve the tweet/user geolocation prediction task and provide a flexible methodology for the geotagging of textual big data. The suggested approach implements neural networks for natural language processing (NLP) to estimate the location as coordinate pairs (longitude, latitude) and two-dimensional Gaussian Mixture Models (GMMs). The scope of proposed models has been finetuned on a Twitter dataset using pretrained Bidirectional Encoder Representations from Transformers (BERT) as base models. Performance metrics show a median error of fewer than 30 km on a worldwide-level, and fewer than 15 km on the US-level datasets for the models trained and evaluated on text features of tweets’ content and metadata context.

Keywords: geolocation prediction, transformers, twitter dataset, machine learning, regression task, Gaussian mixture model, multitask learning

1 Introduction

Much research in recent years was dedicated to processing big data of short text corpora, such as social media posts, to extract geolocation. Location analysis provides data for personalization, making it possible to understand how social media users feel about a particular topic or issue in a specific location. Consequently, it supports social science in identifying patterns of social dynamics in specific areas or regions. It includes public health-related issues, such as vaccination or the impacts of pandemics [37], and possible insights into the demographic characteristics of a candidate’s supporters [1] that are valuable during Presidential elections [41]. Besides that location analysis could be useful for

governmental purposes in terms of natural disasters and crisis management, since it can improve response times, and help to better allocate resources. Furthermore, in a variety of business settings, including retail, real estate, and marketing [18], geolocation information provides a better understanding of people's opinions about a particular brand, product, or service in a specific location.

Twitter, being a widely used online social network, accumulates a large volume of diverse data at a high velocity, this includes short and disordered tweets, a vast network of users and rich contextual information for both users and tweets. This data serves as input for studying three common geolocation problems such as user home location, tweet location and mentioned location prediction [44]. Although, much research dedicated to predicting users' home location utilizes the user network of connections (e. g. followings, replies, mentions) to achieve higher accuracy [46] [45] [43], this work focuses primarily on textual data analysis. Therefore, only tweet content (the text itself) and tweet context (tweet and user meta-data) are adopted to solve the problem of tweet location prediction. Furthermore, grouped by user sets of per-tweet predictions in a form of two-dimensional distributions were leveraged for solving the task of the user's home location prediction.

The solving of the tweet location prediction problem requires the genuine truth location which is commonly extracted from the tweet's meta-data. In terms of Twitter, only 1-2% of the tweets are geotagged with longitude-latitude coordinates [26] which is also followed by the accurate place description (e. g. country, city, place name). Therefore, there are two types of location definition - a pair of numerical coordinates and a textual class label - that imply the type of a task to be either a numerical regression or a textual label classification. The latter is commonly applied for relatively small-scale problems such as location prediction on a city level [20]. However, this study focuses on the country and global worldwide-level scale problems, thus the former method using longitude-latitude coordinates as a genuine truth was chosen as more appropriate. Since Twitter geolocation data is stored in a standard format of a coordinate system called the World Geodetic System 1984 (WGS84), the same geodetic datum was utilized to represent the tweets' genuine location. Although the WGS84 is a continuous, global reference system, when data is displayed on a two-dimensional map, such as a Mercator projection, numerical values of coordinates stay in strict ranges:

$$Y = (y_{lon}, y_{lat}); \quad y_{lon} \in [-180, 180]; \quad y_{lat} \in [-90, 90]$$

The decision to solve the numerical regression rather than the label classification task using BERT-based models was also supported by the experiment results presented in Scherrer's work [33] showing that regression outperformed classification in the geostapital accuracy (mean and median error distance) metrics for the limited Twitter datasets of BosnianCroatian-Montenegrin-Serbian and Deutch languages.

2 Related works

There are many research projects focusing on information retrieval from plain text to estimate geolocation by the application of neural networks. In terms of the previously mentioned geolocation problems that are commonly associated with Twitter data, this section covers works related to the tweet location and the user's home location prediction tasks.



Some solutions are purely text-based and some have a hybrid approach employing the user’s network of friends and mentions in addition to NLP.

Generally, there are three categories of home location granularity: administrative regions, geographical grids, and geographical coordinates. As for the tweet location, point-of-interests (POIs) or coordinates are broadly adopted as representations of tweet locations, instead of administrative regions or grids.

The most straightforward approach is to analyze the natural languages present in social media posts. There are multiple projects processing documents or the summary of user profile tweets to predict the author geolocation [38] [30] [39] [15] [23]. These works apply geographical grids to divide the Earth’s surface into different sizes regions such that highly populated areas are split into smaller pieces while more sparse population areas end up as large grid cells. Early efforts [12], [30] were mainly focused on mining indicative information from users’ posting content relying on location indicative words (LIWs) that can link users to their home locations, based on various NLP techniques (e.g., topic models and statistic models) [46]. For example, Rahimi extracts bag-of-words features from user posts [28]; Wing and Baldridge estimate the word distributions for different regions [38]. Other word-centric works [31] [3] [21] [34] [25] also focus on filtering LIWs from the text and using the gazetteers including not only city/country names but also dialect terms to resolve the geographical indexes. While such methodologies suppose the predefined set of LIWs and their geographical coordinates, Rahimi in the work [27] proposes a based on neural network approach to encoding such words and phrases to the continuous two-dimensional space. In the oldest of relevant works, Eisenstein presents similar unsupervised models which are topic-based [9] [10].

Another way to predict geolocation for social media posts is based on the user network that is covered in many previous works [40] [19] [17] [5] [22] [32]. Consequently, some works suggest a hybrid approach using both content and network to solve the user home geolocation prediction task [29] [8] [46] [45] [43]. Other researchers also define a metadata feature in addition to network and textual input for their models [7] [24] [14]. The usage of context input (such as geo-tags and other meta-data associated with tweets/users) is faintly explored in the previous works due to the deficient data records obtained from publicly available sources of Twitter geotagged datasets.

In terms of single-tweet geolocation prediction, some authors came up with probabilistic models solving the classification task on the country/city levels [11] [16]. Moreover, Yuan in the work [42] also covered the temporal aspect and users’ mobility to estimate each tweet’s location source.

Importantly, Friedhorsky’s methodology was aligned with the present work as it utilized Gaussian Mixture Models (GMM) for geolocation prediction instead of simple coordinates prediction [26]. Another subsequent study, which employed a hybrid approach, also estimated geolocation through GMMs rather than coordinates [2]. This study used the text and network features jointly as predictors for the final estimation. The performance metrics used in both works were adopted to evaluate the proposed probabilistic models as described in Section 5.2.

Most of the previously mentioned works use various approaches based on neural networks for NLP, but none of them applied BERT for the estimation of geolocation. The relatively new BERT model (released in 2018 [6]) is aimed at classification tasks and some works have already used it for country/city names predictions. For instance, Villegas in the work [36] predicts place type or point of interest (POI) using lowercase English BERT.

Furthermore, Li in the work [20] focused on city-scale prediction for the Twitter datasets of Melbourne, and Singapore. The model presented by Li took the text and metadata input and gave the result in a form of probabilities for the set of POI classes.

In the study conducted by Simanjuntak [35], the task of predicting Twitter users' home location was explored using Long-Short Term Memory (LSTM) and BERT models. The dataset used in the experiment comprised text content and user metadata context of tweets in the Indonesian language. However, the input limit of 512 tokens imposed by BERT models posed a challenge in accurately classifying a user's home location, as concatenating all of the user's tweets into a single text would result in an input exceeding the limit. As a result, users' location prediction was performed on a per-tweet basis, and it was observed that the majority vote approach frequently (42%) led to misclassification of the user's home location.

In another work, Scherrer uses BERT models for both regression and classification tasks on limited text datasets of BosnianCroatian-Montenegrin-Serbian and Deutch languages [33]. Evaluation of their regression model on the bound to German-speaking Switzerland dataset shows 21.20 and 30.60 km median and mean distance errors. The authors state that using a smaller tokenization vocabulary and converting the geolocation task to a classification task, in general, yielded worse results. They also conclude that hyperparameter tuning did not yield any consistent improvements, and simply selecting the optimal epoch number on development data showed to be the best approach to the problem of geolocation prediction.

3 Our approach

The main goal of this work was location prediction in a form of geographic points and two-dimensional distributions based on short texts processed by the modified BERT models. In terms of the input data used for model finetuning, this study put into service only textual information such as the tweet content (raw user's text) and tweet context (metadata like information from user's profile and geo-tag textual descriptions). Architecture modifications for the best of proposed models involved the scope of wrapper layers implementing linear regression to the output of parameters defining a GMM (means coordinates, weights, and covariance parameters). Hence, custom loss function computation procedures and complementary realization of multitask learning techniques to handle multiple text features.

Given that the base model has a limited input capacity of 512 tokens (words), it is not feasible to process large text corpora that comprise all user tweets. Although as reported in [35] only the task of tweet geolocation prediction could be efficiently solved using the BERT base models, Simanjuntak was estimating user's home location by the majority vote approach for a set of points predicted for user's tweets. In contrast, this work focuses on processing smaller (less than 300 words) text samples and aggregating a set of probabilistic predictions in a form of GMMs to estimate the set of most probable user's home location points as described in Section 4.3. To achieve this, the surface of the Probability Density Function (PDF) was iteratively calculated for all predicted GMMs on the grid formed of all per-tweet prediction points (GMM peaks), then the average of all PDF values per grid point was computed to get a total per-user score, finally, the local maxima of the evaluated grid were obtained. This approach enabled the significant reduction of the mean spatial error

for the task of user's home location prediction in comparison to the task of tweet location prediction as shown in Table 5.

Moreover, the results presented by [33] indicate that regression outperformed classification in geolocation prediction using modified BERT models. Therefore, this work aims to estimate the geolocation of single tweets in the form of multiple possible location points represented by geographical coordinates (longitude, latitude, weight) or two-dimensional GMMs (longitude, latitude, weight, covariance). As for the task of user's home location prediction, the output of probabilistic models involving the covariance parameter is reduced to a set of weighted coordinate pairs (longitude, latitude, weight).

In this work, the geographic granularity level was defined as a worldwide area covering the whole scope of countries and languages present in the Twitter feed. However, the results of Scherrer's study [33] demonstrated that language-specific BERT models clearly outperformed their multilingual counterpart on most of the used datasets. In light of these findings, the country-level model for the United States was finetuned on the US dataset utilizing the primarily English BERT base model, technically referred to as "bert-base-cased". Nevertheless, the global dataset necessitated the use of the multilingual BERT base model ("bert-base-multilingual-cased"), which resulted in higher error distance metrics for the per-tweet and per-user geolocation prediction, as expected. A critical challenge addressed is the representation of multiple languages, since the ability to process multiple languages is a key factor in dividing text inputs into geospatial regions. To tackle this issue, a multilingual BERT model, pretrained on the 104 largest Wikipedia languages, was utilized as a starting point for finetuning on a global dataset with geolocation labels.

The base BERT model is intended to be fine-tuned on a downstream task which, in this case, was a regression to geographical coordinates (and GMM parameters such as weights and covariance) output as a type of sentence classification task. In general, BERT models have been pretrained using Masked Language Modeling (MLM), which enables the model to learn bidirectional representations of sentences, and Next Sentence Prediction (NSP) to capture the relationships between sentences. Since BERT's layers are hierarchical, early BERT layers learn more generic linguistic patterns, such as differences between multiple languages and their general sentence structures (in the case of the BERT multilingual model). While the later layers learn more task-specific patterns, in this case - the associations between points on the world map and specific terms referred to as LIWs, as well as text constructions and linguistic patterns commonly used in certain geographical areas.

The challenge in this work was to find a substitute for the gazetteers utilized in prior word-centric studies. The gazetteer typically includes entries for cities, towns, villages, landmarks, natural features, and other geographic entities, along with information such as their location, coordinates, elevation, etc. Similarly, Twitter metadata stored in the "place" field of geo-tagged tweets provides automatically generated information about the country, country code, place type, location name, and location full name. Since typical model input shouldn't include "place" context which refers exclusively to geo-tagged tweets, such metadata shouldn't be present in training sequences as well. Therefore multitask learning has been implemented in order to preserve the input format and feed the model with accurate geographic terms obtained from the "place" metadata associated with a genuine truth location.

The multitask learning approach implies separate wrapper layers for each of the textual input features and a common for all features per-batch loss during the process of model finetuning. The key input feature for the BEST of the proposed models included the tweet

text content combined with context obtained from the author profile (username, description, and location), while the minor input feature was dedicated solely to the context data obtained from the "place" field. Considering that evaluation was performed on the typical model input ("text" and "user"), models trained with the described key and minor feature setup showed lower spatial errors in comparison to the models trained without the minor feature or the models trained on a combination of all data in a single text feature ("text", "user", and "place").

The objective was to provide a solution that allows users to fine-tune BERT models of the proposed setup on their own datasets without the need for external knowledge bases such as gazetteers. Despite the fact that only 26% of users provide location information in their profiles according to [4] and that user-generated data can be noisy, the proposed multitask learning procedures enhanced the geospatial accuracy and resulted in the median error of fewer than 30km.

4 Methods

The common approach to solving the regression task for BERT models is adding the dense linear layer on top of the classification output tokens. To account for the multitask learning, individual wrapper layers were used for key and minor textual features. Considering that and the two-dimensional distributions used as the output format of the probabilistic models, this work proposes the custom procedures of loss function computation for the four model types determined in Table 3 based on the model output format. The training procedure for the best of the proposed models is demonstrated on the flowchart diagram in Figure 1.

The regression layer was used to convert the vector of final hidden states made up of 768 floats to the specified number of numerical outputs. According to the task of geolocation prediction, the essential number of outputs was defined as 2 standing for the geographic point (longitude, latitude) in the WGS84 coordinate system. However, the best of the proposed model had an output of 20 numerical values utilized as parameters for the definition of predicted GMM. Note that the models predicting a single coordinate pair could be trained using the standard Mean Squared Error (MSE) loss function. However, more complex output forms such as GMMs required a custom loss function described in Section 4.2. The Probabilistic Single/Multiple Outcome Prediction (PSOP/PMOP) models have slightly outperformed the straightforward approach of Geospatial Single/Multiple Outcome Prediction (GSOP/GMOP) models on the median error distance metrics as shown in Table 6.

The hyper-parameters tuning stage has covered such parameters as the type of the scheduler, minimal and maximal learning rates for the chosen scheduler, and the number of epochs. For the current task, the cosine type was chosen from the set of linear, cyclic, step, and plateau schedulers. Experiments have shown that the optimal learning rate reduction range starts at 1e-5 and ends at 1e-6 at the end of the last training epoch. The entering number of epochs was reduced from 5 to 3 for the dataset consisting of 2.7 million training samples, which are grouped into batches of 10 to 16 tweets in the model data loader. This decision was made as a result of the observed lack of significant error distance reduction after the third epoch for both training and test metrics. By default, the test set of 300,000 samples was evaluated without a gradient propagation at the end of each epoch. The de-



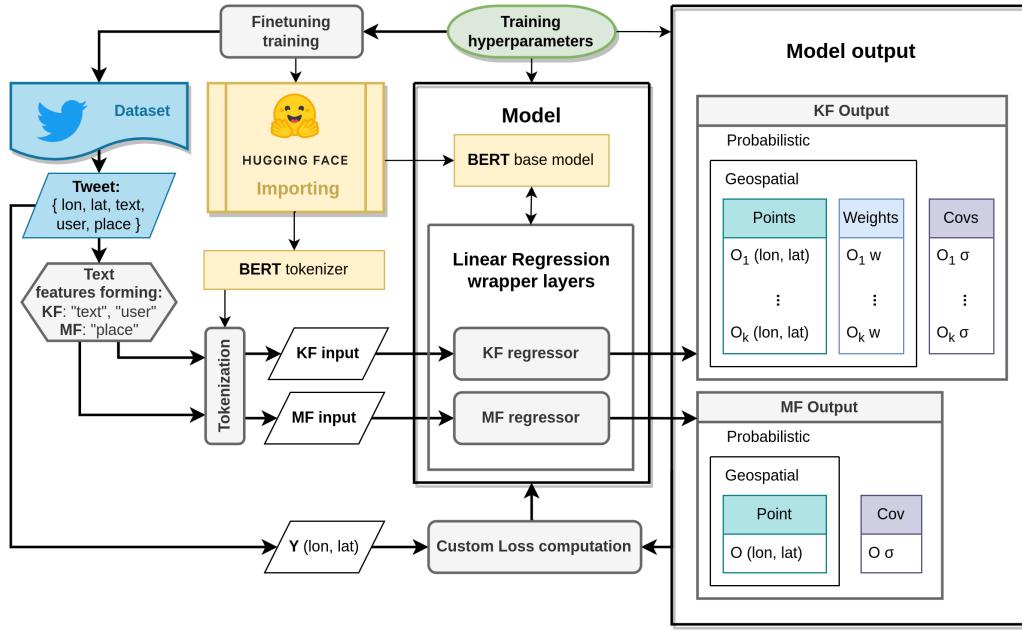


Figure 1: Model finetuning procedure flowchart for the best of the proposed models which has the Probabilistic Multiple Prediction Outcomes (PMOP) output type and individual wrapper layers for Key Feature (KF) and Minor Feature (MF)

scribed parameters were held constant across all proposed models, with variations only occurring in the loss function type, the number of outcomes (prediction points), covariance type, and the combination of text features.

4.1 Data preprocessing

The datasets used for finetuning the models consisted of tweets collected between the years 2020-2022 which have defined "coordinates" and "place" objects in the tweet JSON object. The datasets have undergone several preprocessing stages prior to being uploaded into the model, including text strings filtering, rearrangement of the columns to form text features, tokenization, and split into tensor data loaders of preset batch size. The model was operating solely with input IDs and attention masks representing text features of the dataset and its numerical labels which corresponded to geolocation coordinates.

Generally, only 5% of all tweets in the Twitter archive database collected from 2020 to 2022 have geolocation coordinates, such that the number of available geotagged samples was approximately 150 million. It has been found that up to 20% of the geotagged tweets were posted by bot users, which are automated accounts e. g., newspapers, weather forecasts, airplane trackers, etc. While the finetuning datasets contained tweets of both real and bot users, the evaluation datasets were filtered out based on the assumption that real users don't post more than 20 messages per day. During the preprocessing stage, the orig-

inal tweet objects were condensed into the text content, relevant metadata context, and geolocation coordinates representing the genuine truth.

The genuine truth for this study was defined as the pair of longitude-latitude coordinates associated with each geotagged tweet. There are two inner JSON objects that describe the location information of a tweet: "coordinates" and "place". The former is present only when the exact location was assigned, therefore geolocation coordinate pair in a longitude-latitude format has been collected by parsing only the "coordinates" object of a tweet. In contrast, the "place" field provides an accurate automatically generated textual description of the location, such as country-city naming, which was utilized as a Minor Feature during the finetuning.

Feature name					Dataset columns	JSON object fields
ALL	TEXT-ONLY	USER-ONLY	GEO-ONLY	NON-GEO		
+	+			+	text	text
		+		+	user	location, description, name, screen_name
	+		+		place	country, place_type, location, name, full_name

Table 1: Text feature contents formed from the dataset of parsed tweet JSON object fields

For a purpose of training the model on geospatial terms associated with a tweet, apart from the original tweet text, additional metadata has been collected as well. Such metadata provided Context for the Content of the tweet and has been shown to improve overall model accuracy, as demonstrated in Table 6. Since tweet JSON root-level objects "place" and "user" have multiple fields containing potentially relevant context data, these text fields are concatenated into corresponding string-type columns of the dataset described in Table 1. In general, each tweet always has associated "text" and "user" information, the combination of which was used as the Key training Feature and evaluation input. Furthermore, "place" field is present only in the geotagged tweets, thus it has been passed as a Minor training Feature which is ignored during the evaluation.

There was a limited number of text features combinations explored in this study as described in Table 2. In practice, each model had either one (KF only) or two (KF and MF) wrapper layers which were used according to the input type during the finetuning stage. Note that each training sample had one (KF only) or more (KF and MFs) tokenized sequences assigned to it in the model data loader, such that per-batch loss of each training step depended on one (KF only) or more (KF and MFs) prediction components and their loss values. The Key Feature (KF) was used for validation at the end of each training epoch, while Minor Features (MF) affected primarily the training loss. In order to achieve higher accuracy, test dataset should be formed pursuant to the model KF but without taking into account the model MFs, since MF wrapper layer should be utilized only during the finetuning. The best combination of features, according to the performance metrics shown in Table 6, was NON-GEO KF and GEO-ONLY MF which outperformed all other models of the same output type.



Data type	Training Features		Accuracy
	Key	Minor	
Content, Context	TEXT-ONLY	-	low
	NON-GEO		medium
	ALL		medium
	TEXT-ONLY	USER-ONLY, GEO-ONLY	low
	NON-GEO	GEO-ONLY	high

Table 2: Text features combinations used in text-based (Content) and hybrid (Content and Context) approaches

The final stage of preparing data for model uploading involved converting the dataset’s text features into a data loader of numeric tensors with corresponding IDs, attention masks, and labels. This process started with filtering out URLs and messy punctuation from the text to eliminate irrelevant information. The purified text is then encoded using the default BERT tokenizer, transformed into data loaders with a specified batch size, and made ready for use by the model. Each sample in the batch of data loader represented a single tweet input encoded to IDs and attention masks of its text features, and labeled by its geographic coordinate pair. In practice, the number of words in the tokenized text corpus remained within the limit of 300, which was suitable for the base BERT model’s input size of 512 tokens (words).

4.2 Model architecture

In this work, it is noted that the BERT multilingual base model was tasked with performing regression to predict numerical values like coordinate values, weights, and covariances. This downstream task is similar to text classification and requires modification of the final layer of the model. The base BERT model has a hidden size of 768, which is also the size of the hidden-state token for sequence classification. The proposed wrapper layer operates using only the BERT model pooler output, which consists of processed classification tokens, each representing a separate tweet in the batch. The wrapper layer implements a common linear regression logic, transforming the vector of size 768 to an output vector of a specified size. The exact number of outputs depends on the type of model being used. Furthermore, all models could be grouped into 4 types by the difference in their output structure as shown in Table 3.

The wrapper layer implemented linear regression with a dynamic number of outputs which depended on the feature type (Key or Minor), model type (Geospatial or Probabilistic), and the number of prediction outcomes (Single or Multiple). In the simplest case of the Single Outcome Prediction (SOP) key difference between Geospatial and Probabilistic models was the form of output which could be a two-dimensional point or a Bivariate Normal Distribution.

$$\begin{aligned}\hat{\mathbf{Y}}_{\text{spat}} &= (\hat{y}_{lon}, \hat{y}_{lat}); \quad \hat{y}_{lon} \in \mathbb{R}; \quad \hat{y}_{lat} \in \mathbb{R} \\ \hat{Y}_{prob} &= N(\hat{\mu}, \Sigma); \quad \hat{\mu} = \hat{\mathbf{Y}}_{\text{spat}}; \quad \Sigma = \begin{bmatrix} \sigma_{\hat{c}} & 0 \\ 0 & \sigma_{\hat{c}} \end{bmatrix}; \quad \sigma_{\hat{c}} > 0\end{aligned}$$

The probabilistic output of the model included not only the spatial component of the predicted coordinate pair $\hat{\mu}$ but also a measure of the model’s confidence in its prediction.

Type	Prediction Outcome	Key Feature			Minor Features		Code
		Point	Weight	Cov	Point	Cov	
Spat	1	2	-	-	2	-	GSOP
	M>1	M*2	M	-			GMOP
Prob	1	2	-	1	2	1	PSOP
	M>1	M*2	M	M			PMOP

Table 3: Model types by the number of outputs grouped by Spatial and Probabilistic output forms; Prediction Outcomes determines the number of predicted geographic points; Key Feature and Minor Features stand for the wrapper layers of the proposed models; Code shows the corresponding model abbreviations.

Point - coordinate pair \hat{Y} ; Weight - \hat{w} converted to the weights W ; Cov - \hat{c} converted to the $\sigma_{\hat{c}}$ parameter of the spherical covariance matrix Σ

The Gaussian covariance matrix, which represents the uncertainty of the model, was of spherical type and could be defined by a single positive nonzero numerical value σ . To ensure that $\sigma_{\hat{c}}$ remains positive, the SoftPlus function described in Eq. (SoftPlus) could be applied to the output variable \hat{c} .

$$\sigma_{\hat{c}} = \log(1 + e^{\hat{c}}); \quad \hat{c} \in \mathbb{R}; \quad \sigma_{\hat{c}} > 0 \quad (\text{SoftPlus})$$

However, a lower bound for $\sigma_{\hat{c}}$ was established at $\frac{1}{2\pi}$ to preserve values of the predicted Gaussian's Probability Density Function described in Eq. (PDF) in the range of [0, 1].

$$\sigma_{\hat{c}} = \log(1 + e^{\hat{c}}) + \frac{1}{2\pi}; \quad \hat{c} \in \mathbb{R}; \quad \sigma_{\hat{c}} \in (\frac{1}{2\pi}, +\infty) \quad (\text{LBSP})$$

The selection of the spherical covariance matrix was based on empirical evidence revealing lower geospatial errors compared to models utilizing diagonal, full, and tied covariance matrices. While more complex matrices such as diagonal and full provide more freedom in the shape of the distribution, they necessitate the production of more outputs. Opting for the minimal number of values to determine the shape of the Gaussian distribution eliminated the risk of the model prioritizing optimization of the probabilistic loss component described in Eq. (NLLH) over the spatial component of the loss described in Eq. (SED).

Moreover, the revised lower bound $\frac{1}{2\pi}$ for $\sigma_{\hat{c}}$ curtailed the sharpness of the Gaussian peaks and ensured that their height never exceeds 1, thus avoiding negative values in the probabilistic loss as shown in Figure 3. As a result of applying the lower-bounded SoftPlus described in Eq. (LBSP) on the outputs associated with the covariance parameter, both L_{spat} and L_{prob} remained in the positive domain and were approaching 0 during the finetuning. Hence it eliminated the loss momentum difference and made it possible to include both geospatial and probabilistic errors as tantamount loss components during the total loss computation.

4.2.1 Single and Multiple Prediction Outcomes models

To account for the variations in the output variables, each model type required an individual method for computing its loss. Such that models of the Geospatial type were utilizing solely L_{spat} , while models of the Probabilistic type were utilizing a combination of both L_{spat} and L_{prob} . In terms of KF output evaluation, loss computation procedures for the

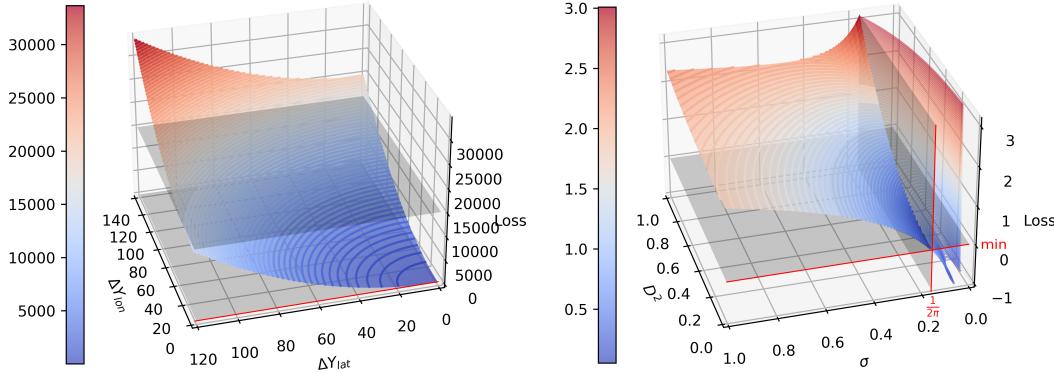


Figure 2: Squared Euclidean Distance (SED) function surface on the axes of ΔY_{lon} and ΔY_{lat} as the error distances per longitude and latitude axes; upper horizontal gray surface indicates the empirical maximum of L_{spat} ; red line indicates the strict minimum of 0 implied by the nature of Eq. (SED)

Figure 3: Negative Log-LikeliHood (NLLH) function surface on the axes of D^2 as the error distance and $\sigma_{\hat{\mu}}$ as the uncertainty in $\hat{\mu}$ of the Gaussian; red lines and gray surfaces indicate the reduction of L_{prob} domain as a result of Eq. (LBSP) application

SOP-type models are visualized in Figure 4, while computational graphs for the MOP-type models can be found in Figure 5. The complete framework of the best of the proposed models, which is PMOP utilizing multiple text features, is also shown as the loss computational graph in Figure 6.

The spatial loss for SOP-type models was calculated as the squared Euclidean distance between the true location specified by the user and the location predicted by the model on the two-dimensional Mercator projection of the worldwide map. Note that the L_{spat} by its nature always remains greater than or equal to zero as shown in Figure 2.

$$L_{GSOP} = (\mathbf{Y} - \hat{\mathbf{Y}})^2 = (y_{lon} - \hat{y}_{lon})^2 + (y_{lat} - \hat{y}_{lat})^2 = D^2; \quad D \geq 0 \quad (\text{SED})$$

The probabilistic component of PSOP models loss was calculated as the negative log-likelihood for the original point \mathbf{Y} to fit in the predicted Gaussian distribution $N(\hat{\mu}, \Sigma)$ as described in Eq. (NLLH).

$$PDF = N(\mathbf{Y} | \hat{\mu}, \Sigma) = \frac{e^{-\frac{D^2}{2\sigma}}}{2\pi\sigma}; \quad PDF \in [0, 1] \quad (\text{PDF})$$

$$L_{PSOP} = -\log(PDF) = \frac{D^2}{2\sigma} + \log(2\pi\sigma); \quad L_{PSOP} \geq 0 \quad (\text{NLLH})$$

Importantly, the definition of the Gaussian covariance matrix implies σ to be greater than 0 which could be achieved by the application of Eq. (SoftPlus) with a lower bound of 0. However, assuming that the error distance is approaching 0, model uncertainty in a predicted point would approach 0 as well. As a consequence, Eq. (PDF) would exceed 1 resulting in negative values of Eq. (NLLH):

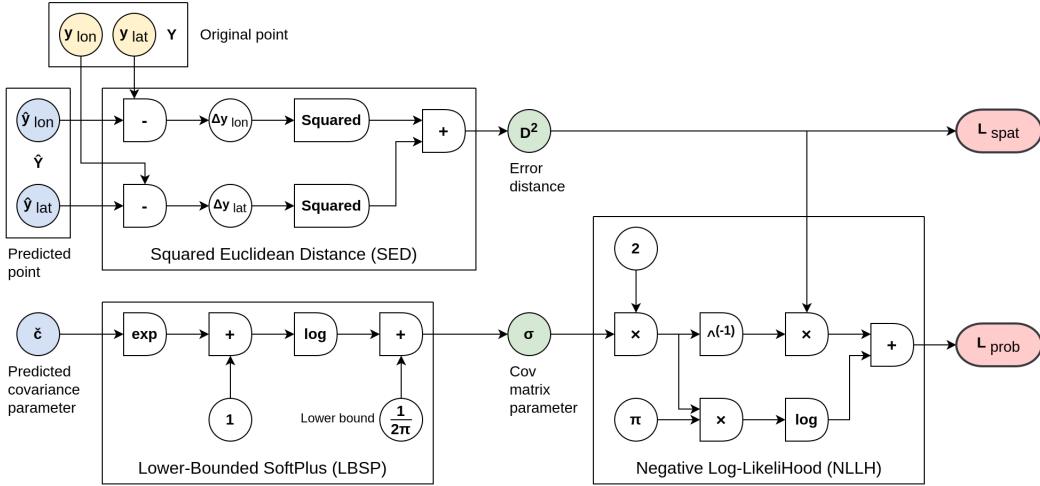


Figure 4: Single Outcome Prediction (SOP) model loss functions computational graph including visualization of Squared Euclidean Distance (SED), Lower-Bounded SoftPlus (LBSP), and Negative Log-LikeliHood (NLLH) components

$$\lim_{(\sigma, D) \rightarrow (0^+, 0^+)} PDF = +\infty; \quad \lim_{(\sigma, D) \rightarrow (0^+, 0^+)} -\log(PDF) = -\infty$$

This study suggests the application of Eq. (LBSP) to limit the covariance parameter as shown in Figure 3. Hence, when the measure of uncertainty σ is approaching its minimum at $\frac{1}{2\pi}$, L_{prob} relies mainly on its spatial component D^2 :

$$\lim_{\sigma \rightarrow \frac{1}{2\pi}} PDF = e^{-\pi D^2}; \quad \lim_{\sigma \rightarrow \frac{1}{2\pi}} L_{\text{PSOP}} = \pi D^2;$$

In the case of MOP-type models, the output included a weight \hat{w}_i for each of M outcomes which indicated its significance among other outcomes. To ensure that all predicted weights sum up to 1, the SoftMax function described in Eq. (SoftMax) was applied to \hat{w}_i .

$$W_i = \frac{e^{\hat{w}_i}}{\sum_{j=1}^M e^{\hat{w}_j}}; \quad \hat{w} \in \mathbb{R}; \quad \sum_{i=1}^M W_i = 1; \quad W_i \in [0, 1] \quad (\text{SoftMax})$$

Therefore, the total of MOP geospatial and probabilistic loss was calculated as the weighted linear combination of all M outcomes errors described in Eq. (WLC).

$$L_{\text{GMOP}} = \sum_{i=1}^M W_i D_i^2; \quad L_{\text{PMOP}} = \sum_{i=1}^M W_i \left(\frac{D_i^2}{2\sigma_i} + \log(2\pi\sigma_i) \right) \quad (\text{WLC})$$

The total loss per text feature f had an essential geospatial component, such that $L_f = L_{\text{spat}}$, and optionally the probabilistic loss added to it. The per-feature loss for the probabilistic models were calculated in three ways: average or sum of L_{spat} and L_{prob} , or

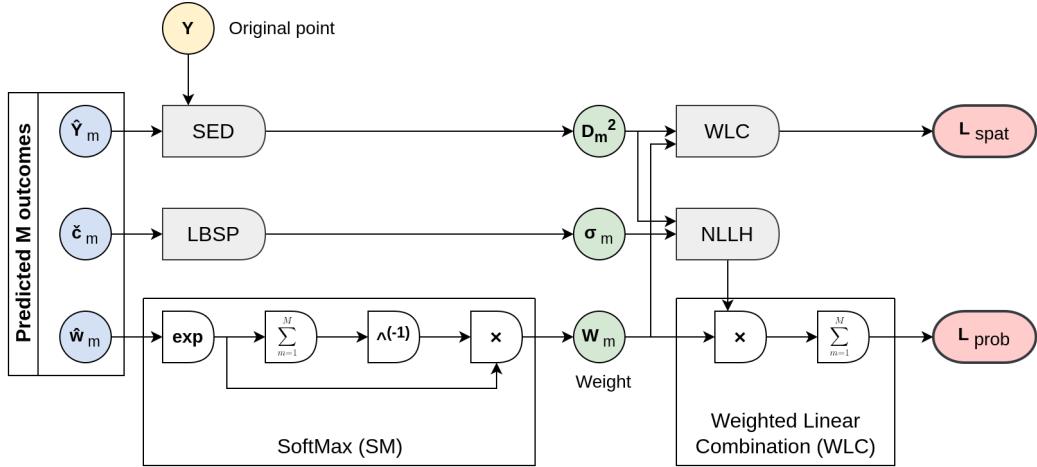


Figure 5: Multiple Outcomes Prediction (MOP) model loss functions computational graph including visualization of Weighted Linear Combination (WLC) and SoftMax (SM) components

L_{prob} with no regard for L_{spat} . The first option has slightly outperformed the second option and has significantly outperformed the last option, thus per-feature loss was calculated as follows:

$$L_f = \frac{L_{spat} + L_{prob}}{2}$$

Finally, the total loss as the average of a single KF and K MFs was calculated to handle multiple textual features F of a single tweet. Note that the number of outcomes M was related solely to the KF output, while the MFs retained SOP-type outputs. However, the number of MFs K was equal to $F - 1$ and could be any integer greater or equal to zero. In addition,

$$L_{total} = \frac{\sum_{i=1}^F L_i}{F}; \quad F \geq 1$$

In practice, the final step was computing the mean of total loss per tweet in a single data loader tensor to back-propagate a float value representing the total per batch loss. The experiments also revealed a computational time growth of 12% in the case of probabilistic models compared to the geospatial analogs. Our equipment (NVIDIA GeForce GTX 1080 Ti) was able to process geospatial loss calculations of 16,7 tweets per second (t/s) in comparison to the 14,9 t/s of the combined geospatial and probabilistic loss computations needed in the case of probabilistic models.

To sum up, Section 4.2 covers a scope of all models reviewed in Table 6, but only the best of the proposed approaches was used to train the models compared to the related works in Table 5. To be more specific, it was the Probabilistic Multiple Outcomes Prediction (PMOP) model of 5 outcomes utilizing NON-GEO as the Key Feature (KF) and GEO-ONLY

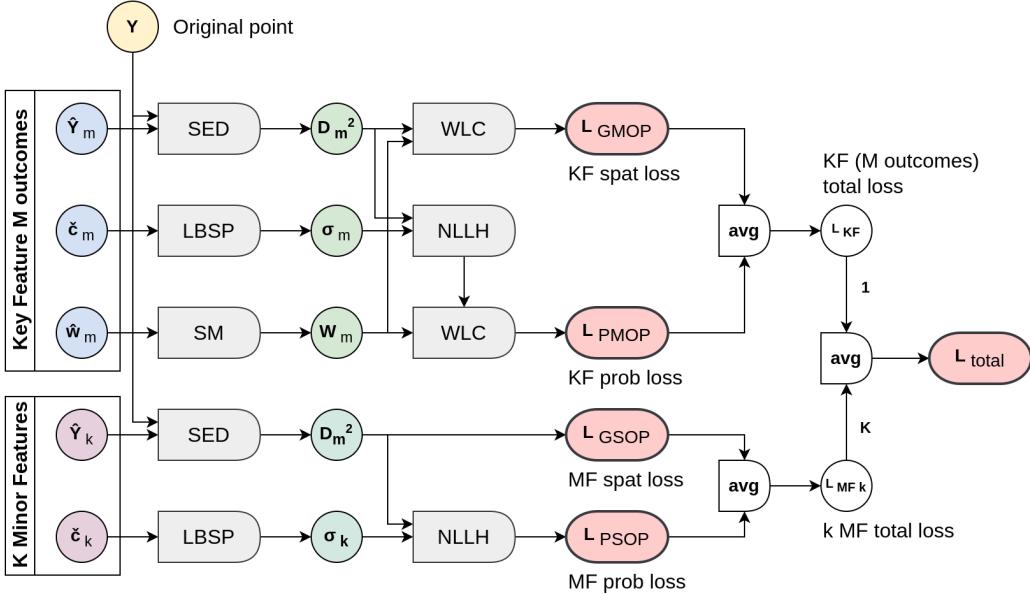


Figure 6: Total loss functions computational graph for the MOP-type Key Feature (KF) with M outcomes, and K Minor Features (MF) of the SOP-type

as the Minor Feature (MF), which was finetuned using custom loss computation procedures described in Figure 6.

4.3 Per-user geolocation estimation

In this work, the focus was primarily on solving the problem of tweet geolocation prediction, thus there is no direct way to predict user's home location using the proposed approaches. However, the PMOP-type models were leveraged to obtain a set of per-tweet predictions for a single user which was then used to estimate the most probable user's home locations in a form of GMOP-type output. Note that in this case the number of selected location points could vary among users, and their weights were set manually based on the summary scores calculated for each outcome.

The summarizing of PMOP-type output of M outcomes was calculated on the grid of $S \cdot M$ GMM peaks gathered from the user's per-tweet predictions put among points of the ground grid G generated on a Mercator projection map with step 10:

$$\mathbf{T} = \{\hat{\mu}_{i,j} \mid 1 \leq i \leq S, 1 \leq j \leq M\}; \quad \hat{\mu}_{i,j} = (\hat{y}_{lon}, \hat{y}_{lat}); \quad \mathbf{T} \in \mathbb{R}^{(S \cdot M) \times (S \cdot M)}$$

$$\mathbf{C} = \mathbf{G} \cup \mathbf{T}; \quad \mathbf{C}_{i,j} = (y_{lon,i}, y_{lat,j}); \quad \mathbf{G} \in \mathbb{R}^{36 \times 19}$$

The union \mathbf{C} provided a multi-set of all GMM peaks \mathbf{T} merged with a background of grid points \mathbf{G} . The average of all S per-tweet scores was calculated for each grid point

$\mathbf{C}_{i,j}$ as its likelihood to fit into a predicted GMM of M peaks defined by their weights W , two-dimensional means μ , and covariance matrices Σ :

$$\text{summary}(\mathbf{C}_{i,j}) = \frac{1}{S} \sum_{s=1}^S \sum_{m=1}^M W_{s,m} \cdot N(\mathbf{C}_{i,j} | \hat{\mu}_{s,m}, \Sigma_{s,m}); \quad \mathbf{C}_{i,j} \in \mathbf{C}$$

Thus forming a two-dimensional matrix \mathbf{Z} containing the average of S probabilities as values of the summarizing function for all grid points $\mathbf{C}_{i,j}$ such that:

$$Z(i, j) = \text{summary}(\mathbf{C}_{i,j}); \quad \text{summary} : \mathbf{R}^2 \rightarrow \mathbf{R}^2$$

Therefore, $Z(i, j)$ was the function value (summary score) of a two-dimensional matrix at point (i, j) , and $M_f Z(i, j)$ was the filtered function value of $Z(i, j)$ using a 10 by 10 maxima filter footprint. The formula took the maximum value of \mathbf{Z} within 10×10 window and assigned it to $M_f Z(i, j)$. This operation could be repeated for every point (i, j) in the matrix \mathbf{Z} to obtain the filtered matrix $M_f Z$:

$$M_f Z(i, j) = \max_{p=-4}^5 \max_{q=-4}^5 Z(i + p, j + q)$$

Then the set of local maxima of \mathbf{Z} was defined as:

$$\mathbf{L}^{user} = \{(i, j) \mid (\mathbf{Z}(i, j) = M_f Z(i, j)) \wedge (i, j) \in \mathbf{C}\}$$

where \mathbf{C} was the multi-set of grid points, and $(i, j) \in \mathbf{C}$ means that (i, j) was a point in the grid \mathbf{C} .

Note that the bag of non-unique points \mathbf{C} was reduced to the set of unique local maxima points \mathbf{L}^{user} that could consist of a single coordinate pair or multiple unweighted ones. Assuming that \mathbf{L}^{user} contained multiple location points, the top K most probable user's home locations were obtained as the first K elements of the sorted in descending order summary scores for all points in \mathbf{L}^{user} which is referenced as \mathbf{Z}^{top-K} :

$$\mathbf{L}^{top-K} = \{(i, j) \mid (\mathbf{Z}(i, j) \in \mathbf{Z}^{top-K} \wedge (i, j) \in \mathbf{L}^{user})\}; \quad \mathbf{L}^{top-K} \in \mathbb{R}^{K \times 2}$$

Then, the coordinate pairs assigned to location indices (i, j) in the set of estimated locations \mathbf{L}^{top-K} were obtained from the initial grid \mathbf{C} as follows:

$$\mathbf{Y}_k^{user} = \mathbf{C}(\mathbf{L}_k^{top-K}) = (y_{lon,i}, y_{lat,j}); \quad \mathbf{Y}^{user} \in \mathbb{R}^{K \times 2}$$

Moreover, weights of the selected location points in \mathbf{Y}^{user} were estimated by the application of Eq. (SoftMax) to their corresponding summary scores in \mathbf{Z}^{top-K} as follows:

$$W_k^{user} = \frac{e^{\mathbf{Z}_k^{top-K}}}{\sum_{j=1}^K e^{\mathbf{Z}_j^{top-K}}}; \quad \sum_{k=1}^K W_k^{user} = 1; \quad W_k^{user} \in [0, 1]$$

Although the initial model output format had to be of PMOP-type, the estimated points \mathbf{Y}^{user} and their weights W^{user} formed a result similar to the GMOP-type output. Note that the set of multiple prediction outcomes for the task of user's home location prediction could be formed only in the case \mathbf{L}^{user} had more than one unique location. Otherwise, the estimation would result in a GSOP-type output containing a single coordinate pair Y^{user} which excludes the calculation of W^{user} .

5 Results

In this section, the proposed models are evaluated by geospatial and probabilistic performance metrics on the test datasets comparable to related works. In the case of probabilistic models with multiple outcomes, the raw model outputs have undergone post-processing according to Eq. (SoftMax) and Eq. (LBSP) applied to the weights and covariance parameters, respectively. The visualization of the post-processed prediction example for the best of the proposed models is shown in Figure 7.

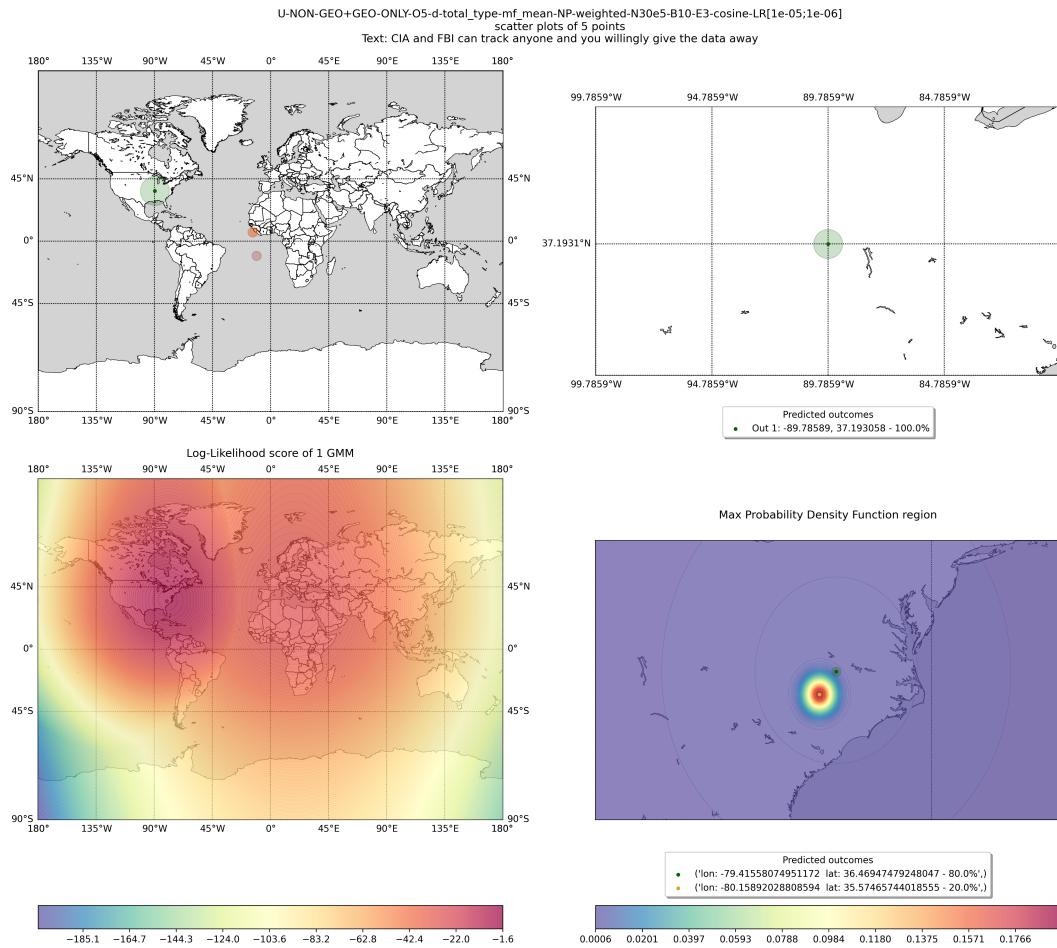


Figure 7: Prediction examples of two models for the same text "CIA and FBI can track anyone, and you willingly give the data away"; both models were trained on the same worldwide dataset with key NON-GEO and minor GEO-ONLY text features, the number of prediction outcomes is 5;
above - GMOP significant (by weight) points as scatter plots; below - PMOP significant (by weight) Gaussian peaks as LLH and PDF plots

5.1 Data

There are three real-world Twitter datasets widely used in previous works for evaluating the geolocation models:

- GeoText [10] is a relatively small Twitter dataset consisting of 9.5K users from US-only. This dataset is designed for evaluation of models solving the problem of user's home location prediction and consists of tweets without user or place meta-data suitable for the hybrid (Content and Context) approach proposed in this work.
- Twitter-US [30] is a larger dataset consisting of 449K users from the U.S., this dataset is also referred to as UTGeo2011 in some papers [7], [30]. A similar dataset of geotagged US-only tweets was collected from our Twitter database archive as a comparable alternative which has both text and meta-data.
- Twitter-World [12] is a much larger dataset that consists of 1.3M users from different countries around the world. The primary locations of users are mapped to the geographic center of the city from where the majority of their tweets are posted. Similarly, a worldwide dataset of geotagged tweets was collected from our archive to perform tweet-oriented geolocation prediction by text-based and hybrid (Content and Context) approaches.

Dataset	Tweets		Users		Scope
	Train	Test Tweet/User	Train	Test Tweet/User	
GeoText	377,616 300K	76K/73K	9,475 7,563	1,900/1,800	US-only
Twitter-US	390M		449,649		
Tw-US Local	16.7M 3M	300K/353K	816,226 379K	48K/5K	
Twitter-World	12M		1.39M		Worldwide
Tw-World Local	24M 3M	300K/394K	1,37M 540K	100K/5K	

Table 4: Original and locally reproduced datasets, split to train and test subsets in numbers of tweets and unique users; *Test* column is divided into per-tweet and per-user evaluation datasets characteristics

Statistics for each dataset are shown in Table 4. There were several reasons to compose alternative Twitter-US and Twitter-World test datasets for the model evaluation. Firstly, there was no direct access to the original datasets at the moment. Secondly, we had an access to the local Twitter archive that contained all tweets from 2020 to 2022 which was more convenient for collecting and parsing the big data. Finally, it allowed us to compose task-oriented datasets that have specific text features like "user" and "place" context meta-data. Since the original datasets are mainly designed for the task of user geolocation prediction, comparison with metrics of the previous works in Table 5 is approximate and serves only as a visible benchmark for the best of proposed BERT model modifications.

Importantly, the test evaluation for locally reproduced Twitter-US and Twitter-World was performed on the datasets filtered from bot users (posting more than 20 messages per day) and users utilized in the train datasets to ensure that the evaluation was unbiased and provided an accurate assessment of the models' generalization capabilities.

Evaluation results of the worldwide models in Table 6 show the difference between geospatial and probabilistic models with a number of prediction outcomes ranging from 1 to 100. The test dataset consisted of 300K tweets from 143K users covering 63 languages and 226 countries. Note that both train and test datasets were the same for all models and contained tweets from both real and bot users. The evaluation was performed on *NON-GEO* ("text" + "user") and *TEXT-ONLY* ("text") features, while the finetuning was performed on the various text feature combinations but primarily on *NON-GEO* as KF and *GEO-ONLY* ("place") as MF.

5.2 Performance metrics

Performance metrics are divided into geospatial and probabilistic, the former is universal for all models and the latter applies to probabilistic models only.

5.2.1 Geospatial metrics

- SAE - *simple accuracy error* metric is utilized to measure the geographic distance between two points on the Earth's surface. This is achieved through the application of the Haversine formula for calculating the great-circle distance in kilometers. The formula is as follows:

$$SAE_{SOP} = Hav(\mathbf{Y}, \hat{\mathbf{Y}}) = D_H$$

where \mathbf{Y} and $\hat{\mathbf{Y}}$ represent the original and predicted points, respectively. When calculating the MOP-type model metrics, the SAE is computed as the weighted linear combination of all M outcomes:

$$SAE_{MOP} = \sum_{i=1}^M W_i Hav(\mathbf{Y}, \hat{\mathbf{Y}}_i) = D_H$$

The mean and median of the SAE for the validation dataset are utilized as key performance indicators. It is worth noting that during the finetuning process, the calculation of the spatial error is simplified to the Euclidean distance formula.

- Acc@161 - the percentage of predicted locations that are within a 161km (100 miles) radius of the actual location.

$$Acc@161 = \frac{|i : D_{H,i} \leq 161|}{N} \cdot 100$$

where D_H is a set of N elements representing the great-circle distance between the predicted locations and the genuine truth locations, and $|i : D_{H,i} \leq 161|$ represents the number of elements in D_H that are less than or equal to 161 km.

Note that this metric based on the imperial units was used as one of the most popular in the area of location prediction due to the great contribution of researchers from the United States.

5.2.2 Probabilistic metrics

- CAE - *comprehensive accuracy error* metric measures the expected distance between the true origin of the tweet and a random point generated from the GMM predicted by

the model. The Haversine formula is used to calculate distances from the original point to each Gaussian sub-population of 100 samples. Bakerman in his work [2] suggests the Monte Carlo method to compute CAE as the mean of the sub-population distance errors:

$$CAE_{SOP} \approx \frac{1}{|z|} \cdot \sum_{\hat{\mathbf{Y}} \in z} Hav(\mathbf{Y}, \hat{\mathbf{Y}})$$

where z is a sample from the Gaussian sub-population. The total CAE of the GMM is calculated as the weighted linear combination of all M Gaussian peaks:

$$CAE_{MOP} \approx \sum_{i=1}^M \frac{W_i}{|z_i|} \cdot \sum_{\hat{\mathbf{Y}} \in z_i} Hav(\mathbf{Y}, \hat{\mathbf{Y}})$$

- PRA_α - *prediction region area* metric is the area covered by $\alpha \cdot 100\%$ density of the predicted GMM:

$$PRA_{\alpha,SOP} = \pi \chi_2^2(1 - \alpha) \sigma_c^2$$

where σ is a determinant of the spherical covariance matrix Σ from the predicted GMM peak $N(\hat{\mu}, \Sigma)$. In the case of MOP, It's calculated as the weighted linear combination of such area for each of M Gaussian peaks, which depends mainly on their covariance parameters and weights:

$$PRA_{\alpha,MOP} = \pi \chi_2^2(1 - \alpha) \cdot \sum_{i=1}^M W_i \sigma_i^2$$

- COV_α - *coverage* metric measures the proportion of times the prediction region PRA_α covers the true origin of the tweet. Following Bakerman's approach, the original geographic point \mathbf{Y} is within the boundary of an ellipse with center $\hat{\mu}$ and covariance matrix Σ if:

$$\frac{D}{\sigma} \leq \chi_2^2(1 - \alpha); \quad D = Euclidean(\mathbf{Y}, \hat{\mu})$$

Note that in the MOP case, COV_α is calculated as the overall mean of N dataset samples with M outcomes.

5.3 Comparing metrics with related works

In general, only Geospatial metrics, such as Acc@161, mean and median SAE, are utilized for the comparison with previous works in Table 5. Note that both Twitter-US and Twitter-World datasets were reproduced locally from the archived database of Twitter posts from the year 2021.

The evaluation of results for the subsidiary task of user's home location prediction was performed on the GSOP-type estimation obtained from the best of the proposed models which were of PMOP-type. In particular, the most probable location was picked from the set of unique local maxima \mathbf{L}^{user} as a point with the highest summary score $\mathbf{Z}(i, j)$:

$$\mathbf{Y}_{user} = \text{argmax}_{(i,j) \in \mathbf{L}^{user}} \mathbf{Z}(i, j) = \mathbf{C}_{i,j} = (y_{lon,i}, y_{lat,j})$$

where \mathbf{C} is the grid of points described in Section 4.3.

Furthermore, the task of the user's home location prediction required a ground truth point to evaluate the summarized prediction results. In this study, this location was picked up from a set of unique genuine truth locations associated with a single user as the most frequent (or the earliest if all locations are equally repeated).

5.4 Worldwide evaluation metrics by model type

The models compared in Table 6 were trained on the local Twitter-World dataset composed of different Train Features (TF) and tested on both *TEXT-ONLY* and *NON-GEO* eValuation Features (VF). The most prevalent *NG+GO* training setup was using the hybrid approach with *NON-GEO* as the Key and *GEO-ONLY* as the Minor Features respectively. Metrics results for the MOP-type models were calculated as the weighted linear combination of all outcomes.

Following the best of the proposed approaches, the total per-tweet loss was composed of two features containing both spatial and probabilistic components. To account for the equivalent of both L_{spat} and L_{prob} loss momentums during finetuning, Eq. (LBSP) set $\frac{1}{2\pi}$ as the lower bound of σ_c covariance parameters thereby maintaining the values of Eq. (NLLH) in the positive domain. The performance results in Table 6 confirm that the PMOP model of 5 outcomes noted by u^* (stands for the obtained from Eq. (SoftPlus) unlimited σ_c) showed downscale scores in all metrics in comparison to the lower-bounded model of 5 outcomes.

As for the Training Feature options, the last three models in Table 6 were trained on *ALL* ("text" + "user" + "place"), *NON-GEO* ("text" + "user"), and *TEXT-ONLY* ("text") features had higher SAE scores compared to the similar PMOP model of 5 outcomes trained on the proposed combination of *NON-GEO* and *GEO-ONLY* features, indicating underperformance.

Moreover, the increasing number of outcomes resulted in interchangeable spatial metrics scores, while revealing a modest growth of CAE and PRA metrics, as well as an undesirable decrease in the COV criteria. While the MOP-type architecture outperformed SOP-type in both Probabilistic and Geospatial variations, the difference was relatively low.

6 Discussion and Conclusion

This work was aimed at the examination of the machine learning techniques for solving the short text geolocation prediction task with NLP techniques employing a scope of BERT-based neural networks. The proposed framework wins in the simplicity of the setup needed for the models finetuning to the user-specific downstream task. The models are flexible to any level of the geospatial granularity (worldwide, country, city, etc) by changing the finetuning dataset and language-specific pretrained base BERT model.

One of the problems mentioned earlier was an alternative for the gazetteers - dictionaries of geographical indexes - used in the previous works containing accurate terms (location naming and LIWs) for geographical objects and their correct geographical coordinates. Considering the noisy user-generated data of the typical model inputs, the idea of feeding a consistent knowledge base of geographic objects into the model appeared practical and has been successfully implemented. The highest performance models were trained on multitask solving that offered an option of parallel learning with a total per-batch training

Model	VF	GeoText			Twitter-US			Twitter-World		
		Mean SAE	Med SAE	Acc @161	Mean SAE	Med SAE	Acc @161	Mean SAE	Med SAE	Acc @161
<i>User's home location task</i>										
Eisenstein [10]	T	845	501	-	-	-	-	-	-	-
Eisenstein [9]	T	900	494	-	-	-	-	-	-	-
Wing [38]	T	967	479	-	-	-	-	-	-	-
Wing [39]	T	808	317	41	704	171	49	1715	490	33
Roller [30]	T	897	432	36	860	463	35	-	-	-
Han [13]	TM	-	-	-	814	260	45	1953	646	24
Cha [3]	T	581	425	-	-	-	-	-	-	-
Melo [23]	T	-	-	-	702	208	-	-	-	-
Rahimi [29]	T	880	397	38	687	159	50	1724	530	32
Rahimi [28]	H	654	151	50	620	157	50	-	-	-
	H	578	61	59	515	77	61	1280	104	53
Rahimi [28]	H	581	57	59	529	78	60	1403	111	53
	T	844	389	38	554	120	54	1456	415	34
Miura [24]	H	-	-	-	336	42	70	780	-	72
Do [7]	H	570	58	59	474	157	51	-	-	-
Ebrahimi [8]	H	476	32	64	438	56	66	1216	95	54
Miyazaki [25]	T	821	325	44	-	-	-	-	-	-
Huang [14]	T	-	-	-	455	64	61	762	86	56
	H	-	-	-	323	28	73	610	6	68
Zhong [45]	T	834	403	41	544	120	54	1456	415	34
	H	514	38	62	452	67	63	1107	102	55
Zheng [43]	H	516	30	64	359	31	72	818	49	62
Zhou [46]	H	316	23	-	263	37	-	636	62	-
Proposed	T	1433	743	0	431	15	78	892	31	74
	TM	1059	741	1	375	13	83	567	26	82
<i>Tweet location task</i>										
Priedhorsky [26]	T	870	534	-	-	-	-	-	-	-
Hulden [15]	T	954	493	-	-	-	-	-	-	-
	T	765	357	-	-	-	-	-	-	-
Liu [21]	T	856	-	-	733	377	-	-	-	-
Bakerman [2]	H	593	19	-	-	-	-	-	-	-
Proposed	T	1216	787	1	1163	599	41	1588	50	61
	TM	1094	770	1	802	25	64	800	25	80

Table 5: Comparison of geospatial metrics results in related works and proposed models trained and evaluated on the datasets GeoText, Twitter-US and Twitter-World using the hybrid approach (Content + Context) and PMOP-type output of 5 prediction outcomes for the estimation of the tweet location and user's home location. Mean and Median SAE in km, Acc@161 in percents.

VF column defines eValuation Features: T - text content, M - metadata context, H - hybrid of text and users network

loss. The best strategy was split into key NON-GEO (always present "text" and "user") and minor GEO-ONLY ('place' present only in geo-tagged tweets) text features. In practice,

TF	Model	VF	OUT	Spatial			Probabilistic				
				Mean SAE	Med SAE	Acc @161	Mean CAE	Med CAE	Mean PRA _{0.95}	Med PRA _{0.95}	COV _{0.95}
NG+GO	PSOP	TO	1	1881.2	153	50.5	1954	352.8	174	60.8	12.5
		NG		568.3	32.1	78.3	639.2	70.4	60.9	3.7	19.6
	PMOP	TO	3	1876.2	134.9	51.5	1911.8	219.9	24.1	16.2	22
		NG		561.7	29.5	78.7	601.1	63.3	13.9	7.9	31.2
		TO	5	1845	135.9	51.5	1896.3	214.7	27.9	15.2	15.9
		NG		551.4	29.4	79	600.4	79.1	15.3	9.6	23.4
		TO	5u*	2036.7	234.8	45.1	2080.9	357.5	27.9	28.8	6.2
		NG		626.7	70	70.7	691	158.5	21.3	18.5	9.7
	GSOP	TO	10	1845	143.2	51.1	1901.4	214.2	32.4	13.9	7.8
		NG		553.2	33.2	78.9	599.8	77.9	15.6	9.4	11
	GMOP	TO	50	1855.9	136.6	51.4	1905	214.7	27.8	14	1.3
		NG		556.9	29.4	79	604.4	77.1	14.7	9.5	2.1
		TO	100	1882.4	149.8	50.6	1939	199.6	193.2	4.5	13.9
		NG		568.9	28.1	78.6	618.3	71.1	15.2	9.1	1.2
NG	TO	TO	1	1872.2	140.3	51.3					
		NG		559.6	36.6	78.4					
	PMOP	TO	3	1859.7	142.8	51.1					
		NG		556.3	36.5	78.5					
	A	TO	5	1986.6	151.3	50.6					
		NG		577.8	35.6	78.6					
		TO	5	3203	585.9	41.2	3225.9	623.4	29.6	8.5	7.4
		NG		1266.2	37.7	67	1292.6	62.6	14.7	7.5	11.7
	PMOP	TO	5	1875.4	172.9	49.3	1927.3	235.2	24.6	13.8	7.5
		NG		581	46.9	76.4	635.2	107	15.9	11.8	12.6
		TO	5	1547.3	176.5	48.7	1708.8	442.1	58.8	38.2	8.8
		NG		782	87.2	64.9	913.1	267.1	36.2	26.4	11.8

Table 6: Worldwide dataset (300,000 tweets) - results of models performance metrics; TF - Training Feature; VF - eValuation Feature;

NG+GO : NON-GEO + GEO-ONLY, A : ALL, NG : NON-GEO, TO : TEXT-ONLY;

all text features abbreviations are described in Table 1

u* - unlimited covariance output, $\sigma_{\hat{c}}$ were obtained by the application of Eq. (SoftPlus) to the covariance parameter output \hat{c}

every feature had its individual linear regression layer wrapping the base model that differed in the number of prediction outcomes, but matched in the per-feature loss function type (Geospatial/Probabilistic). The evaluation was performed using solely the key feature wrapper layer with no regard for the minor feature wrapper layer utilized only during the finetuning. This approach allowed for maintaining the format of the common model input and mapping the official place descriptions to the locations associated with the primary input texts. In terms of the described setup, the best performance results were achieved using the SOP-type minor loss and MOP-type key loss for 3-5 prediction outcomes.

Experiments have shown that the geospatial loss of the minor feature GEO-ONLY was declining much faster than the error distance of key feature NON-GEO since the minor feature had less noisy and more persistent text data associated with the specific geolocations. The straightforward concatenation of the text content and metadata context of the "place" field would require undesirable randomization of all words in the string before tokenization such that models would not learn to only pay attention to the optional part ("place" metadata) of the input sequence. This statement is supported by the results in Table 6

which show that the model trained on *ALL* (union of "text", "user", and "place"), as the key feature has underperformed the model trained on the data split to key and minor features. Moreover, the model trained solely on key *NON-GEO* feature with no regard to the place context demonstrated even higher spatial errors during the evaluation. Similarly, the Geo-Text dataset has no available "place" metadata which resulted in higher spatial errors in comparison to the other datasets reproduced locally and contained all necessary data as shown in Table 5. Thus the proposed methodology of multitask learning has demonstrated a significant improvement in the accuracy of the predictions and should be utilized if it's possible.

Another proposed novelty was a lower-bound limitation of the covariance parameter σ_c in the GMM output of the probabilistic models. Since the probabilistic loss function computed the negative log-likelihood of genuine truth to fit in the predicted Gaussian distributions, it depended on the Probability Density Function (PDF). In the case, σ_c was approaching its minimum at 0, PDF exceeded 1 driving the loss function to the negative values as shown in Fig 3. In terms of probabilistic models, the total loss depended on both spatial and probabilistic components, therefore the latter gained a bigger momentum. As a result, the model paid less attention to the squared Euclidean error distance described in Eq. (SED) which negatively affected the geospatial accuracy. The suggested solution limited PDF to $[0, 1]$ interval by setting the minimum of σ_c to $\frac{1}{2\pi}$. Since the covariance parameter measures model uncertainty about the predicted location point, in the best-case scenarios, the probabilistic loss would depend mainly on the distance error as described in Section 4.2.1.

Considering the main goal of geotagging separate tweets, the problem of the user's home location prediction was explored as one of the possible estimates based on the PMOP-type model predictions. Although this wasn't the foremost performance measure, the results in Table 5 reveal a higher spatial accuracy on the per-user task than on the per-tweet measurements. Hence PMOP-type models are suggested for the user's home location prediction as the comparable BERT-based solution that outperforms most of the previous works.

As for the real-time user location monitoring, Yuan in the work [42] focused on tracking the movement of individuals or groups over time. Monitoring user geolocation in time can provide valuable insights for a variety of applications such as tracking disease outbreaks, analyzing urban mobility patterns, and providing location-based services. Such an analytical framework could be built on top of the proposed models, yet, in this study, only the commonly researched task of the user's home location prediction was properly explored so far.

In addition, there are several different world models that can be utilized for prediction output. While the Mercator projection is commonly used due to the Twitter geolocation stored in the WGS84 format, other projections such as the Robinson projection, conic projection, and Winkel-Tripel projection should be considered as possible alternatives in future studies.

Similarly, the scope of proposed machine learning techniques could be utilized for any base model apart from the BERT variations. This would necessitate an adaptation of the wrapper layer implementation according to the shape of the pooler output vector of the chosen model, since the current approach is set up for the linear regression logic, transforming the vector of size 768 common for all BERT-based models into the predefined number of continuous numerical values.

Overall, location-based sentiment analysis is an important tool for understanding public opinion, social dynamics and patterns, helping decision-makers and researchers to make data-driven decisions, and support the work of various sectors, from marketing to governance. This study provides the NLP-based approach for the estimation of geolocation by processing short text corpora such as social media posts on Twitter. The proposed solution utilizes multitask learning (key and minor features), context data (user and place metadata), and probabilistic output (GMM) to achieve higher spatial accuracy on the tasks of the tweet and the user's home location prediction. Thus contributing to the field of big data analysis with a flexible geographical granularity setup of the custom BERT-based models.

7 Acknowledgments

The authors acknowledge the Institute of Science and Technology (ISTA) for their material support and for granting access to the Twitter database archive, which was essential for the research. We also extend our appreciation to ChatGPT, the language model developed by OpenAI, for its assistance in manuscript editing, coding documentation, and revising parts of the manuscript.

References

- [1] ARAFAT, T. A., BUDI, I., MAHENDRA, R., AND SALEHAH, D. A. Demographic analysis of candidates supporter in twitter during indonesian presidential election 2019. In *2020 International Conference on ICT for Smart Society (ICISS)* (2020), IEEE, pp. 1–6.
- [2] BAKERMAN, J., PAZDERNIK, K., WILSON, A., FAIRCHILD, G., AND BAHRAN, R. Twitter geolocation: A hybrid approach. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 12, 3 (2018), 1–17.
- [3] CHA, M., GWON, Y., AND KUNG, H. Twitter geolocation and regional classification via sparse coding. In *Proceedings of the International AAAI Conference on Web and Social Media* (2015), vol. 9, pp. 582–585.
- [4] CHENG, Z., CAVERLEE, J., AND LEE, K. 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* (2010), pp. 759–768.
- [5] COMPTON, R., JURGENS, D., AND ALLEN, D. Geotagging one hundred million twitter accounts with total variation minimization. In *2014 IEEE international conference on Big data (big data)* (2014), IEEE, pp. 393–401.
- [6] DEVLIN, J., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [7] DO, T. H., NGUYEN, D. M., TSILIGIANNI, E., CORNELIS, B., AND DELIGIANNIS, N. Twitter user geolocation using deep multiview learning. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018), IEEE, pp. 6304–6308.

- [8] EBRAHIMI, M., SHAFIEIBAVANI, E., WONG, R., AND CHEN, F. Twitter user geolocation by filtering of highly mentioned users. *Journal of the Association for Information Science and Technology* 69, 7 (2018), 879–889.
- [9] EISENSTEIN, J., AHMED, A., AND XING, E. P. Sparse additive generative models of text. In *Proceedings of the 28th international conference on machine learning (ICML-11)* (2011), pp. 1041–1048.
- [10] EISENSTEIN, J., O'CONNOR, B., SMITH, N. A., AND XING, E. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (2010), pp. 1277–1287.
- [11] FLATOW, D., NAAMAN, M., XIE, K. E., VOLKOVICH, Y., AND KANZA, Y. On the accuracy of hyper-local geotagging of social media content. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (2015), pp. 127–136.
- [12] HAN, B., COOK, P., AND BALDWIN, T. Geolocation prediction in social media data by finding location indicative words. In *Proceedings of COLING 2012* (2012), pp. 1045–1062.
- [13] HAN, B., COOK, P., AND BALDWIN, T. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research* 49 (2014), 451–500.
- [14] HUANG, B., AND CARLEY, K. M. A hierarchical location prediction neural network for twitter user geolocation. *arXiv preprint arXiv:1910.12941* (2019).
- [15] HULDEN, M., SILFVERBERG, M., AND FRANCOM, J. Kernel density estimation for text-based geolocation. In *Proceedings of the AAAI conference on artificial intelligence* (2015), vol. 29.
- [16] ISO, H., WAKAMIYA, S., AND ARAMAKI, E. Density estimation for geolocation via convolutional mixture density network. *arXiv preprint arXiv:1705.02750* (2017).
- [17] JURGENS, D. That's what friends are for: Inferring location in online social media platforms based on social relationships. In *Proceedings of the International AAAI Conference on Web and Social Media* (2013), vol. 7, pp. 273–282.
- [18] KINSELLA, S., MURDOCK, V., AND O'HARE, N. "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* (2011), pp. 61–68.
- [19] KONG, L., LIU, Z., AND HUANG, Y. Spot: Locating social media users based on social network context. *Proceedings of the VLDB Endowment* 7, 13 (2014), 1681–1684.
- [20] LI, M., LIM, K. H., GUO, T., AND LIU, J. A transformer-based framework for poi-level social post geolocation. *arXiv preprint arXiv:2211.01336* (2022).
- [21] LIU, J., AND INKPEN, D. Estimating user location in social media with stacked denoising auto-encoders. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing* (2015), pp. 201–210.

- [22] MCGEE, J., CAVERLEE, J., AND CHENG, Z. Location prediction in social media based on tie strength. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (2013), pp. 459–468.
- [23] MELO, F., AND MARTINS, B. Geocoding textual documents through the usage of hierarchical classifiers. In *Proceedings of the 9th Workshop on Geographic Information Retrieval* (2015), pp. 1–9.
- [24] MIURA, Y., TANIGUCHI, M., TANIGUCHI, T., AND OHKUMA, T. Unifying text, metadata, and user network representations with a neural network for geolocation prediction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2017), pp. 1260–1272.
- [25] MIYAZAKI, T., RAHIMI, A., COHN, T., AND BALDWIN, T. Twitter geolocation using knowledge-based methods. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text* (2018), pp. 7–16.
- [26] PRIEDHORSKY, R., CULOTTA, A., AND DEL VALLE, S. Y. Inferring the origin locations of tweets with quantitative confidence. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (2014), pp. 1523–1536.
- [27] RAHIMI, A., BALDWIN, T., AND COHN, T. Continuous representation of location for geolocation and lexical dialectology using mixture density networks. *arXiv preprint arXiv:1708.04358* (2017).
- [28] RAHIMI, A., COHN, T., AND BALDWIN, T. A neural model for user geolocation and lexical dialectology. *arXiv preprint arXiv:1704.04008* (2017).
- [29] RAHIMI, A., VU, D., COHN, T., AND BALDWIN, T. Exploiting text and network context for geolocation of social media users. *arXiv preprint arXiv:1506.04803* (2015).
- [30] ROLLER, S., SPERIOSU, M., RALLAPALLI, S., WING, B., AND BALDRIDGE, J. 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* (2012), pp. 1500–1510.
- [31] RYOO, K., AND MOON, S. Inferring twitter user locations with 10 km accuracy. In *Proceedings of the 23rd International Conference on World Wide Web* (2014), pp. 643–648.
- [32] SADILEK, A., KAFTZ, H., AND BIGHAM, J. P. Finding your friends and following them to where you are. In *Proceedings of the fifth ACM international conference on Web search and data mining* (2012), pp. 723–732.
- [33] SCHERRER, Y., AND LJUBEŠIĆ, N. Social media variety geolocation with geobert. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects* (2021), The Association for Computational Linguistics.
- [34] SCHULZ, A., HADJAKOS, A., PAULHEIM, H., NACHTWEY, J., AND MÜHLHÄUSER, M. A multi-indicator approach for geolocalization of tweets. In *Proceedings of the International AAAI Conference on Web and Social Media* (2013), vol. 7, pp. 573–582.



- [35] SIMANJUNTAK, L. F., MAHENDRA, R., AND YULIANTI, E. We know you are living in bali: Location prediction of twitter users using bert language model. *Big Data and Cognitive Computing* 6, 3 (2022), 77.
- [36] VILLEGAS, D. S., PREOTIUC-PIETRO, D., AND ALETRAS, N. Point-of-interest type inference from social media text. *arXiv preprint arXiv:2009.14734* (2020).
- [37] WAKAMIYA, S., KAWAI, Y., ARAMAKI, E., ET AL. Twitter-based influenza detection after flu peak via tweets with indirect information: text mining study. *JMIR public health and surveillance* 4, 3 (2018), e8627.
- [38] WING, B., AND BALDRIDGE, J. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (2011), pp. 955–964.
- [39] WING, B., AND BALDRIDGE, J. Hierarchical discriminative classification for text-based geolocation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (2014), pp. 336–348.
- [40] YAMAGUCHI, Y., AMAGASA, T., AND KITAGAWA, H. Landmark-based user location inference in social media. In *Proceedings of the first ACM conference on Online social networks* (2013), pp. 223–234.
- [41] YAQUB, U., SHARMA, N., PABREJA, R., CHUN, S. A., ATLURI, V., AND VAIDYA, J. Location-based sentiment analyses and visualization of twitter election data. *Digital Government: Research and Practice* 1, 2 (2020), 1–19.
- [42] YUAN, Q., CONG, G., MA, Z., SUN, A., AND THALMANN, N. M. Who, where, when and what: discover spatio-temporal topics for twitter users. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (2013), pp. 605–613.
- [43] ZHENG, C., JIANG, J.-Y., ZHOU, Y., YOUNG, S. D., AND WANG, W. Social media user geolocation via hybrid attention. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020), pp. 1641–1644.
- [44] ZHENG, X., HAN, J., AND SUN, A. A survey of location prediction on twitter. *IEEE Transactions on Knowledge and Data Engineering* 30, 9 (2018), 1652–1671.
- [45] ZHONG, T., WANG, T., WANG, J., WU, J., AND ZHOU, F. Multiple-aspect attentional graph neural networks for online social network user localization. *IEEE Access* 8 (2020), 95223–95234.
- [46] ZHOU, F., WANG, T., ZHONG, T., AND TRAJCEVSKI, G. Identifying user geolocation with hierarchical graph neural networks and explainable fusion. *Information Fusion* 81 (2022), 1–13. <https://doi.org/10.1016/j.inffus.2021.11.004>.