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

Kateryna Lutsai, Christoph Lampert

February 24, 2023

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 an average 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.

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 [WKA+18], and possible insights into the demographic characteristics of a candidate's supporters [ABMS20] that are valuable during Presidential elections [YSP+20]. 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 [KMO11], 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 [ZHS18]. The latter problem is unrelated to this work since it involves the recognition of text fragments referring to location names and the mapping of these fragments to entries in a predefined location database. Likewise, much research dedicated to predicting users' home location utilizes the user network of connections (e. g. followings, replies, mentions) to achieve higher accuracy [ZWZT22] [ZWW+20] [ZJZ+20], however, 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. In addition, users' home locations were estimated by summarizing probabilistic location predictions for a set of their tweets.

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 [PCDV14] 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 [LLGL22]. 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. Furthermore, the Mercator projection map type was chosen as a global world model due to being the simplest in an output understanding and matching the genuine truth location format of the Twitter database per-tweet "coordinates" field.

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

The decision to solve the numerical regression rather than the label classification task was also supported by the experiment results presented in Scherrer's work [SL21] showing that regression outperformed classification in the geostaptial accuracy (mean and median error distance) metrics for the limited Twitter datasets of BosnianCroatian-Montenegrin-Serbian and Deutch languages.

The results of Scherrer's study [SL21] demonstrated that language-specific BERT models clearly outperformed their multilingual counterpart on most of the used datasets. In light of these findings, this study utilizes a custom model based on the primarily English BERT base model, technically referred to as "bert-base-cased", for finetuning and evaluation on datasets sourced solely from the United States. 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 estimation, as expected.

This work focuses on the tweet location prediction based on the tweet content (raw user's text) and tweet context (metadata like information from user's profile and geo-tag textual descriptions). The model finetuning stage covered suggested algorithms that encompass multitask learning (multiple wrapper layers), custom loss function for the predicted GMM, and setup flexibility to the user-defined geospatial granularity and language(s) choice for the proposed custom BERT models.

Furthermore, the tweet's predicted location was utilized to estimate the home location of the user by aggregating the per-tweet probabilistic predictions (GMMs) in the form of the Probability Density Function (PDF) evaluated at the grid of predicted GMM peaks, to identify the highest local maxima on the resulted surface of their total per-user scores.

1.1 Related works

There are many research projects focusing on information retrieval from plain text to estimate geolocation by the application of neural networks. 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. In terms of Twitter data, the raw text corpus could be either multiple tweets of one user or a single tweet, such that the task goal could be a prediction of user home geolocation, tweet geolocation, and mentioned geolocation [ZHS18]. The key difference is model inputs: for home location prediction - all tweets from a user, while for tweet location prediction - only one tweet is given.

Generally, there are three categories of home location granularity: administrative regions, geographical grids, and geographical coordinates. As for the tweet location and mentioned 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 [WB11] [RSR+12] [WB14] [HSF15] [MM15]. 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 [HCB12], [RSR+12] 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)[ZWZT22]. For example, Rahimi extracts bag-of-words features from user posts [RCB17]; Wing and Baldridge estimate the word distributions for different regions [WB11]. Other word-centric works [RM14] [CGK15] [LI15] [SHP+13] [MRCB18] 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 [RBC17] 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 [EAX11] [EOSX10].

Another way to predict geolocation for social media posts is based on the user network that is covered in many previous works [YAK13] [KLH14] [Jur13] [CJA14] [MCC13] [SKB12]. Consequently, some works suggest a hybrid approach using both content and network to solve the user home geolocation prediction task [RVCB15] [ESWC18] [ZWZT22] [ZWW+20] [ZJZ+20]. Other researchers also define a metadata feature in addition to network and textual input for their models [DNT+18] [MTTO17] [HC19]. 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 [FNX+15] [IWA17]. Moreover, Yuan in the work [YCM+13] also covered the temporal aspect and users' mobility to estimate each tweet's location source.

Importantly, Priedhorsky's methodology was aligned with the present work as it utilized Gaussian Mixture Models (GMM) for geolocation prediction instead of simple coordinate prediction [PCDV14]. Another subsequent study, which employed a hybrid approach, also estimated geolocation through GMMs rather than coordinates [BPW+18]. This study used the text and network features jointly as predictors for the final estimation.

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 [DCLT18]) is aimed at classification tasks and some works have already used it for country/city names predictions. For example, Villegas in the work [VPPA20] predicts place type or point of interest (POI) using lowercase English BERT.

In terms of POI estimation, Li in the work [LLGL22] focused on city-scale prediction for the Twitter datasets of Melbourne, and Singapore. They presented a transformer-based model named transTagger which takes text and metadata input and gives the result in a form of probabilities for the set of POI classes. The results of the evaluation show a median error below 4 km for both Melbourne and Singapore datasets.

In the study conducted by Simanjuntak [SMY22], 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 [SL21]. 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.

1.2 Our approach

The main goal of this work was location prediction of short texts using a modified BERT model finetuned on the Twitter dataset. Although most of the previous works focus on the user's home location prediction, as reported in [SMY22] only the task of tweet geolocation prediction could be efficiently solved using the BERT base models. 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. Instead, this work focuses on processing smaller (less than 300 words) text samples and summarization of probabilistic predictions for separate users to estimate the set of most probable user's home location points. Moreover, the results presented by [SL21] 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).

In this study, the main focus was on the global geographic area, with a country-level model (the United States) also included for comparison with previous works. 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. In general, BERT base 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.

The base 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. 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 objective was to provide a solution that allows users to finetune proposed BERT models on their own datasets, regardless of size, and without the need for external knowledge bases. To accomplish this, metadata collected from the "place" field of tweet objects was utilized as a minor training feature, serving as a mapping from location names to geographical coordinates during the finetuning process. The main focus of the finetuning of the BERT base models was to learn the text patterns associated with two-dimensional regions on the Mercator world map projection. The key input feature for the proposed models included the tweet text content combined with context obtained from the author profile (username, description, and location). Despite the fact that only 26% of users provide location information in their profiles according

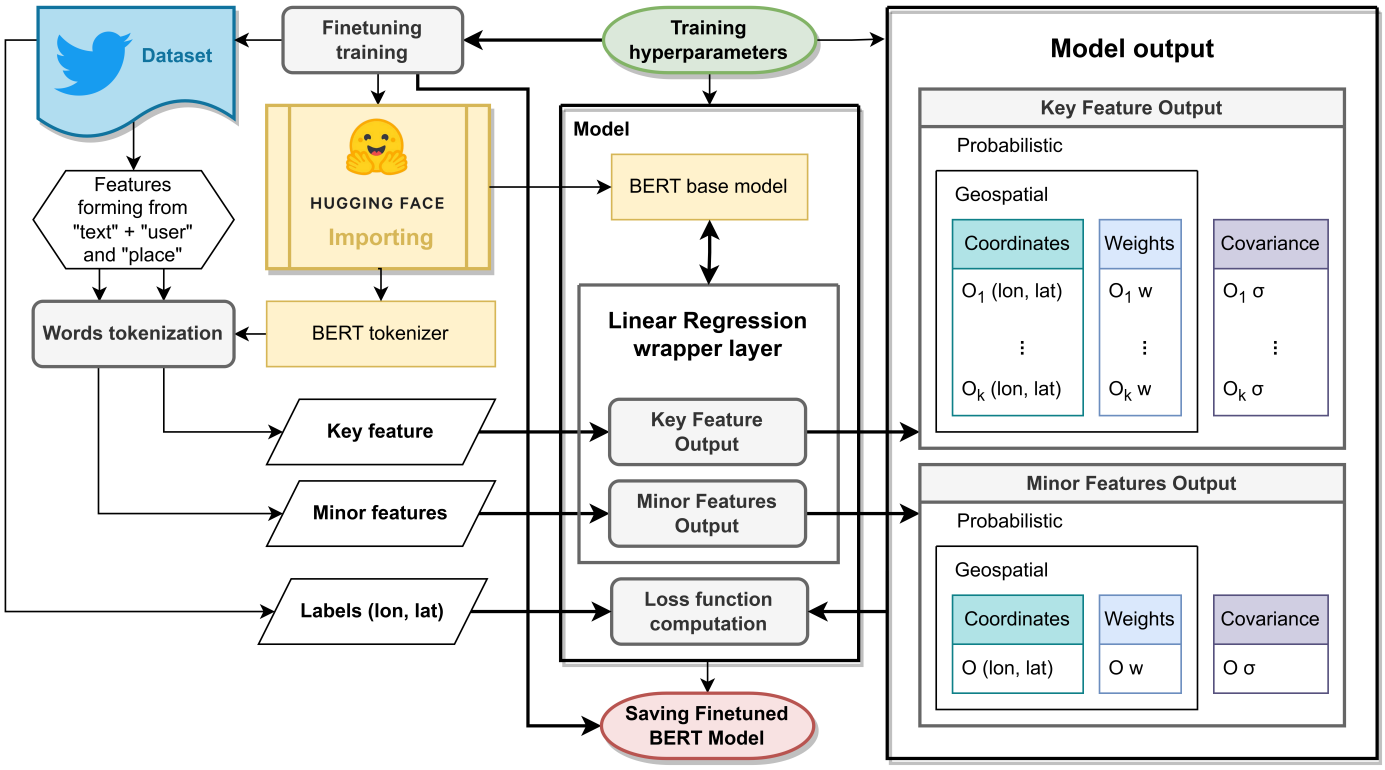


Figure 1: Model finetuning training flowchart

to [CCL10] and that user-generated data can be noisy, the proposed training procedures were able to effectively teach the model to associate correct geospatial object names with their corresponding ground truth locations.

2 Methodology

The common approach to solving the regression task for BERT models is adding the dense linear layer on top of the classification output tokens. The regression layer is used to convert the vector of final hidden states made up of 768 floats to the specified number of outputs. Since the model is used to predict geolocation, the simplest form of output should include a pair of coordinates (longitude, latitude) defining 2 as a minimal number of outputs for linear regression in the last layer. In this case, the standard MSE loss function could be used to compare original and predicted coordinates during finetuning of the model. However, more complex output form such as GMM requires a custom loss function described in 2.2 and outperforms the straightforward approach of training models to predict single point coordinates on the error distance metrics as shown in 6.

The hyper-parameters tuning stage 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 learning 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 dataloader. This decision was made as a result of the observed lack of significant loss reduction after the third epoch for both training and test error distance metrics. By default, the test set of 300,000 samples was evaluated without a gradient propagation at the end of each epoch. The described 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 combination of text features.

2.1 Data preprocessing

The datasets used for finetuning the models consist of tweets collected between the years 2020-2022 which have defined "coordinates" and "place" objects in the tweet JSON object. The datasets undergo several preprocessing stages prior

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

to being uploaded into the model, including text strings filtering, rearrangement of the columns to form text features, tokenization, and split into tensor dataloaders of preset batch size. The model operates solely with input IDs and attention masks representing text features of the dataset and its numerical labels which correspond to geolocation coordinates.

During the preprocessing stage, the original tweet objects are effectively condensed into geolocation coordinates representing the labels for the data loader, the text content, and the most critical metadata properties that serve as context inputs for the model. Null value fields in the objects that should contain a string are replaced with an empty string due to the absence of data. Generally, only 5% of all tweets in the Twitter archive database collected from 2020 to 2022 have geolocation coordinates, which significantly reduces the number of geotagged samples to approximately 150 million. The parsing performed on the filtered samples resulted in a reduction of the size of a single tweet by a factor of 7, from an average of 3.5 kb to 0.5 kb per tweet in the resulting dataset file.

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 process. These geotags would serve as POI class labels, unless the problem scope in this study was on a global scale, encompassing almost the entire world and all country territories.

For a purpose of training the model on geospatial terms associated with a tweet, apart from the original tweet text, additional meta-data has been collected into a training dataset file 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" contents have been passed only as a Minor training Feature which is ignored during evaluation.

The use of different combinations of text features is described in Table 2. The division into Key and Minor Features had a significant impact on training but not on evaluation. The Key Feature (KF) was used for tests at the end of each training epoch, while Minor Features (MF) primarily affected the training loss. Preferably the validation dataset should have the same text feature as the KF used by the model, but it is not a strict requirement. The best combination of features, according to the performance metrics shown in Table 6, was KF *NON-GEO* and MF *GEO-ONLY* which outperformed models trained on *TEXT-ONLY*, *NON-GEO*, and *ALL* text features.

the BERT tokenizer is responsible for both encoding the input text and decoding the output predictions. During the encoding phase, the tokenizer processes the input text and converts it into a sequence of numerical tokens that can be fed into the BERT model. The tokenizer also adds special tokens, such as the [CLS] token at the beginning of the input sequence, which is used to represent the start of the input and [SEP] tokens to separate the sentences.

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.

Data type	Training Features		Accuracy
	Key	**Minor**	
Content	TEXT-ONLY	-	low
Content, Context	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

Type	**Prediction Outcomes**	**Key Feature**			**Minor Features**		**Code**
		Coordinates	Weights	Covariance	Coordinates	Covariance	
Geospatial	1	**2**	-	-	**2**	-	GSOP
	M>1	**M*2**	**M**	-			GMOP
Probabilistic	1	**2**	-	**1**	**2**	**1**	PSOP
	M>1	**M*2**	**M**	**M**			PMOP

Table 3: Model types by number of outputs

Each sample in the batch of data loader represented a single tweet input encoded to IDs and attention masks of the 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 wass suitable for the base BERT model's input size of 512 tokens (words).

2.2 Model architecture

In this work, it is noted that the BERT multilingual base model is tasked with performing regression to predict numerical values like coordinate values, weights, and covariances. This downstream task is similar to text classification and only 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 classification tokens are processed by a linear layer, where the weights are trained from the classification objective during pretraining. 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 implements linear regression with a dynamic number of outputs which depends on the feature type, model type, and the number of prediction outcomes.

In the simplest case of the Single Outcome Prediction (SOP) key difference between geospatial and probabilistic models is the form of output which could be a two-dimensional point or a Bivariate Normal Distribution.

$$\widehat{\mathbf{Y}}_{\mathbf{spat}} = [\widehat{y}_{lon}, \widehat{y}_{lat}];\ \widehat{y}_{lon} \in \mathbf{R};\ \widehat{y}_{lat} \in \mathbf{R}$$

$$\widehat{Y}_{prob} = N(\widehat{\boldsymbol{\mu}}, \boldsymbol{\Sigma});\ \widehat{\boldsymbol{\mu}} = \widehat{\mathbf{Y}}_{\mathbf{spat}};\ \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{\widehat{c}} & 0 \\ 0 & \sigma_{\widehat{c}} \end{bmatrix};\ \sigma_{\widehat{c}} > 0$$

The probabilistic output of the model included not only the spatial component of the predicted coordinate pair $\widehat{\mu}$ but also a measure of the model's confidence in its prediction. The covariance matrix, which represents the uncertainty of the model, was of spherical type and could be defined by a single positive numerical value, $\sigma_{\widehat{c}}$. To ensure that $\sigma_{\widehat{c}}$ remains positive, the SoftPlus function is applied to the output variable $\widehat{c}$. Moreover, a lower bound for $\sigma_{\widehat{c}}$ is established at $\frac{1}{2\pi}$ to preserve the predicted Gaussian's Probability Density Function (PDF) values in the range of [0, 1].

$$\sigma = log(1 + e^{\widehat{c}}) + \frac{1}{2\pi};\ \widehat{c} \in \mathbf{R};\ \sigma \in (\frac{1}{2\pi}, +\infty) \tag{LBSP}$$

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

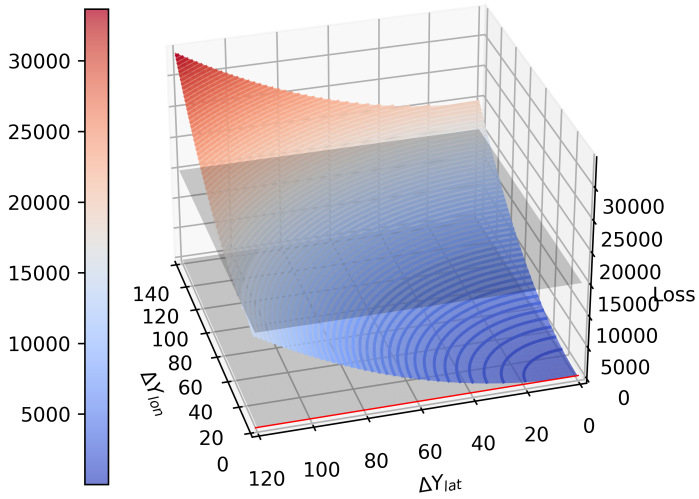


Figure 2: Squared Euclidean Distance (SED) function surface on the axes of ΔY_{lon} and ΔY_{lat} as the error distances per longitude and latitude axes; upper horizontal gray surface indicates the empirical maximum of L_{spat}; red line indicates the strict minimum of 0 implied by SED

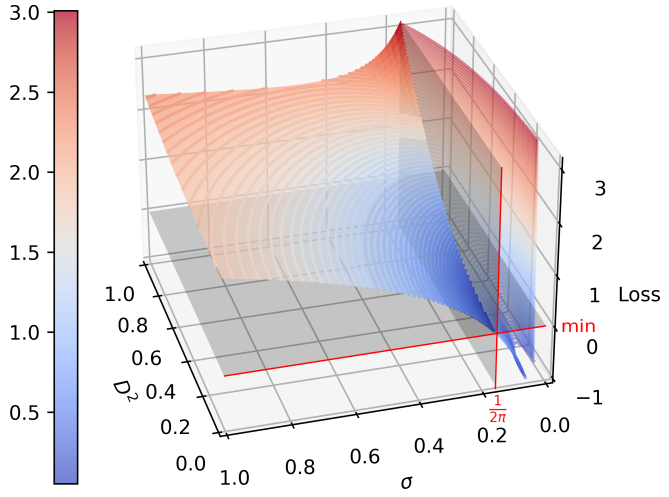


Figure 3: Negative Log-LikeliHood (NLLH) function surface on the axes of D^2 as the error distance and $\sigma_{\widehat{c}}$ as the uncertainty in $\widehat{\mu}$ of the Gaussian; red lines and gray surfaces indicate the reduction of L_{prob} domain as a result of LBSP application on the $\widehat{c}$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

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 Negative Log-LikeliHood (NLLH) optimization over the spatial component of the loss referenced as Squared Euclidean Distance (SED).

Moreover, the revised lower bound $\frac{1}{2\pi}$ for $\sigma_{\widehat{c}}$ curtails the sharpness of the Gaussian peak and ensures the density never exceeds 1, thus avoiding negative values in the probabilistic loss. As a result of applying the Lower-Bounded Soft Plus (LBSP) on the outputs associated with the covariance parameter, the errors of both L_{spat} applied to the coordinate pair outputs and L_{prob} additive, referenced as NLLH, stay in the positive domain and approach 0 during training.

2.2.1 Single and Multiple Prediction Outcomes models

To account for the variations in the output variables, each model type requires a unique method for computing its loss. The spatial loss for SOP is determined by calculating the SED 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 error distance is always greater than or equal to zero.

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

The probabilistic SOP loss is the NLLH for the original point to fit in the predicted Gaussian distribution.

$$PDF = N(\mathbf{Y} \mid \widehat{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) = \frac{e^{-\frac{D^2}{2\sigma}}}{2\pi\sigma};\ \lim_{\sigma \to \frac{1}{2\pi}} PDF = e^{-\pi D^2};\ PDF \in [0,1] \tag{PDF}$$

$$L_{PSOP} = -log(PDF) = \frac{D^2}{2\sigma} + log(2\pi\sigma);\ \lim_{\sigma \to \frac{1}{2\pi}} L_{PSOP} = \pi D^2;\ L_{PSOP} \geq 0 \tag{NLLH}$$

n the case of Multiple Outcomes Prediction (MOP), the model output included a weight $\widehat{w}_i$ for each of M outcomes which indicated its significance among other outcomes. Note that the number of outcomes was related solely to the KF, while the MFs were retaining SOP-type outputs.

$$W_i = \frac{e^{\widehat{w}_i}}{\sum_{j=1}^{M} e^{\widehat{w}_j}};\ \widehat{w} \in \mathbf{R};\ \sum_{i=1}^{M} W_i = 1; W_i \in [0,1] \tag{SoftMax}$$

Therefore, the KF total of MOP geospatial and probabilistic loss is a Weighted Linear Combination (WLC) of all M outcomes errors.

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

The total loss per text feature F always has a geospatial component $L_F = L_{spat}$ and optionally the probabilistic loss added to it:

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

Finally, the average of KF and MF loss is calculated to handle multiple textual features F of a single tweet.

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

In practice, the final step is 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.

3 Results

In this section, the proposed models are evaluated by geospatial and probabilistic performance metrics on the datasets comparable to related works.

3.1 Data

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

- GeoText [EOSX10] is a relatively small Twitter dataset consisting of 9.5K users from US-only. This dataset is designed for user geolocation model evaluation and consists of tweets without user or place meta-data suitable for the hybrid features approach proposed in this work.
- Twitter-US [RSR+12] is a larger dataset consisting of 449K users from the U.S., this dataset is also referred to as UTGeo2011 in some papers [DNT+18], [RSR+12]. 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 [HCB12] 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 by text-based and hybrid (text and meta-data) approaches.

Statistics for each dataset are shown in Table 4. There were several reasons to compose alternative Twitter-US and Twitter-World datasets for the model evaluation. Firstly, there was no direct access to the original datasets at the moment. Secondly, we had a local Twitter archive that contains 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 Twitter-US and Twitter-World was performed on the datasets filtered from bots (posting more than 50 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 validation 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 could contain bot users and a little number of tweets from the same users. The evaluation was performed on both *NON-GEO* ("text" + "user") and *TEXT-ONLY* ("text") features.

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

3.2 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.

3.2.1 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:

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

where $\mathbf{Y}$ and $\widehat{\mathbf{Y}}$ represent the original and predicted points, respectively. When calculating the MOP model metrics, the SAE result is computed as the Weighted Linear Combination (WLC) of all M outcomes:

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

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 each predicted location and the actual location, 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 (US) and the UK.

3.2.2 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 [BPW+18] suggests the Monte Carlo method to compute CAE as the mean of the sub-population distance errors:

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

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

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

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

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

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

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

- 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 $\widehat{\boldsymbol{\mu}}$ and covariance matrix $\mathbf{\Sigma}$ if:

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

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

3.3 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 related works in Table 5. Note that both Twitter-US and Twitter-World datasets were reproduced locally from the whole archived database of Twitter posts from the years 2020-22.

3.3.1 Per-user evaluation

To get the most probable user home locations based on the set of S tweets, only the probabilistic models could be utilized. Although PMOP models are used as the most accurate, the performance metrics are calculated for the GSOP-type location obtained from the user's predictions summary. Note that apart from the test dataset evaluation, the per-user summary of GMOP-type is available for the visualization of up to 5 points manually weighted according to their total scores.

The summarizing of PMOP output of M outcomes is 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 as a Mercator projection with step 10:

$$\mathbf{G} = \{(y_{lon}, y_{lat}) \mid y_{lon} = -180, -170, \ldots, 170, 180, y_{lat} = -90, -80, \ldots, 80, 90\};\ \mathbf{G} \in \mathbf{R}^{37 \times 19}$$

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

$$\mathbf{C} = \mathbf{G} \cup \mathbf{T};\ \mathbf{C}_{i,j} = (y_{lon,i}, y_{lat,j});\ \mathbf{C} \in \mathbf{R}^{(37+S \cdot M) \times (19+S \cdot M)}$$

The union $\mathbf{C}$ provides 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 is 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 $\boldsymbol{\mu}$, and covariance matrices $\boldsymbol{\Sigma}$:

$$summary(\mathbf{C}_{i,j}) = \frac{1}{S}\sum_{s=1}^{S}\sum_{m=1}^{M} W_{s,m} \cdot N(\mathbf{C}_{i,j} \mid \widehat{\boldsymbol{\mu}}_{s,m}, \boldsymbol{\Sigma}_{s,m});\ \mathbf{C}_{i,j} \in \mathbf{C}$$

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

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

Therefore, $Z(i,j)$ is the function value of a two-dimensional matrix at point (i,j), and $M_f Z(i,j)$ is the filtered function value of $Z(i,j)$ using a 10 by 10 maxima filter footprint.

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

The formula takes the maximum value of $\mathbf{Z}$ within 10×10 window and assigns it to $M_f Z(i,j)$. This operation can be repeated for every point (i,j) in the matrix $\mathbf{Z}$ to obtain the filtered matrix $M_f Z$.

Then the set of local maxima of $\mathbf{Z}$ can be defined as:

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

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

Note that the bag of non-unique points $\mathbf{C}$ is reduced to the set of unique local maxima points $\mathbf{L}_{user}$ that could consist of a single coordinate pair or multiple unweighted ones. Moreover, weights of multiple outcomes could be estimated by applying the $SoftMax(Z(i,j))$ function to the summary scores of several points in $\mathbf{L}_{user}$. However, the number of unique locations is inconsistent and varies among users, hence using WLC for the accurate GMOP-type dataset evaluation is impossible. For the purpose of performance evaluation, the output is reduced to GSOP-type by picking a single point that has the highest $summary(Z(i,j))$ score among all in $\mathbf{L}_{user}$:

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

Finally, the task of the user's home location prediction requires a ground truth point to evaluate the summarized prediction results. In this study, such location is picked up from a set of unique locations associated with a single user as the most frequent (or the earliest if all locations are equally repeated). The computation of probabilistic per-user metrics is inapplicable, and geospatial per-user performance metrics imply the usage of SOP-type SAE regardless of the MOP-type model.

3.4 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 is using the hybrid approach with *NON-GEO* as the key and *GEO-ONLY* as the minor features respectively. Models for Probabilistic Single Outcome Prediction (PSOP) and Geospatial SOP have only 1 outcome; Geospatial Multiple Outcome Prediction (GMOP) - 3 and 5; PMOP ranges from 3 to 100 outcomes. Metrics results for all MOP models are calculated as WLC of all outcomes measurements.

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, the LBSP function set $\frac{1}{2\pi}$ as the lower bound of the σ_c covariance parameters thereby maintaining the NLLH in the positive domain. For example, performance results in Table 6 confirm that PMOP model of 5 outcomes noted by $u*$ (stands for the standard 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.

Model	Feature	GeoText			Twitter-US			Twitter-World		
		Mean SAE	**Med SAE**	**Acc @161**	**Mean SAE**	**Med SAE**	**Acc @161**	**Mean SAE**	**Med SAE**	**Acc @161**
User's home location task										
Eisenstein [EOSX10]	Content	845	501	-		-			-	
Eisenstein [EAX11]	Content	900	494	-		-			-	
Wing [WB11]	Content	967	479	-		-			-	
Wing [WB14]	Content	808	317	41	704	171	49	1715	490	33
Roller [RSR+12]	Content	897	432	36	860	463	35		-	
Han [HCB14]	Conent, Context		-		814	260	45	1953	646	24
Cha [CGK15]	Content	581	425	-		-			-	
Melo [MM15]	Content		-		702	208	-		-	
Rahimi [RVCB15]	Content	880	397	38	687	159	50	1724	530	32
	Hybrid	654	151	50	620	157	50		-	
Rahimi [RCB17]	Hybrid	578	61	59	515	77	61	1280	104	53
		581	57	59	529	78	60	1403	111	53
	Content	844	389	38	554	120	54	1456	415	34
		856	360	40	581	91	55	1417	373	36
Miura [MTTO17]	Hybrid		-		336	42	70	780	-	72
Do [DNT+18]	Hybrid	570	58	59	474	157	51		-	
Ebrahimi [ESWC18]	Hybrid	476	32	**64**	438	56	66	1216	95	54
Miyazaki [MRCB18]	Content	821	325	44		-			-	
Huang [HC19]	Content		-		455	64	61	762	86	56
	Hybrid		-		323	28	73	610	**6**	68
Zhong [ZWW+20]	Content	834	403	41	544	120	54	1456	415	34
	Hybrid	514	38	62	452	67	63	1107	102	55
Zheng [ZJZ+20]	Hybrid	516	30	64	359	31	72	818	49	62
Zhou [ZWZT22]	Hybrid	**316**	**23**	-	**263**	37	-	636	62	-
Proposed	Content	1433	743	0	431	15	78	892	31	74
	Content, Context	1059	741	1	375	**13**	**83**	**567**	26	**82**
Tweet location task										
Priedhorsky [PCDV14]	Content	870	534	-		-			-	
		954	493							
Hulden [HSF15]	Content	765	357	-		-			-	
Liu [LI15]	Content	856		-	733	377	-		-	
Bakerman [BPW+18]	Hybrid	**593**	**19**	-		-			-	
Proposed	Content	1216	787	1	1163	599	41	1588	50	61
	Content, Context	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 3 prediction outcomes for the estimation of the tweet location and user's home location. Mean and Median SAE in km, Acc@161 in percents.
Feature column defines evaluation features: Content (text), Context (metadata), Hybrid (text + network)

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
		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
		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
	GSOP	TO	1	1872.2	140.3	51.3	-				
		NG		559.6	36.6	78.4					
	GMOP	TO	3	1859.7	142.8	51.1					
		NG		556.3	36.5	78.5					
		TO	5	1986.6	151.3	50.6					
		NG		577.8	35.6	78.6					
A	PMOP	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
NG		TO		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		TO		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, the Lower-Bounded SoftPlus function of the covariance parameter output $\widehat{c}$ had a standard lower bound of 0 for the yielded σ_c variable

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 MOP-type architecture outperformed SOP-type in both Probabilistic and Geospatial variations, the difference was relatively low.

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.

4 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 which were usually applied at the next step in NER-focused research for mining the text to get the familiar names of geospatial objects. However, such methods are very time-consuming, and searching for matches in a large vocabulary on a global level scale would be very slow, especially in terms of big data streams like a Twitter feed. Nevertheless, the idea of implementing a knowledge base into the BERT model itself is valuable and has been successfully applied before in the models of other architecture types. The base BERT model should be finetuned on the sets of geographical indexes either before or during training on the task-specific text inputs, in this case, tweets and metadata of their author's profile.

The highest performance models were trained on multitask solving that offers an option of parallel learning with a total per-batch training 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. The evaluation was performed only on the key feature, but the minor loss component (*GEO-ONLY* containing accurate and persistent automatically generated geotags) has shown a more rapid drop to small numbers than the key loss (*NON-GEO* containing noisy user-generated content) in the total training average. The proposal is that geotagged metadata is less noisy than users' texts, hence easier to map on the two-dimensional label space of coordinate pairs. 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 wouldn't learn to only pay attention to the optional part ("place" metadata) of the input sequence. In practice, every feature separately has its linear regression layer wrapping the base model that usually differs in the number of prediction outcomes, but matches in the per-feature loss function type (Geospatial/Probabilistic). In the described setup, the best performance results were achieved using the SOP minor loss and MOP key loss for 3-5 prediction outcomes.

Another proposed novelty was a lower-bound limitation of the covariance parameter σ in the GMM output of the probabilistic models. Since the probabilistic loss function computes the Negative Log-LikeliHood of genuine truth to fit in the predicted Gaussian distributions, it depends on the Probability Density Function (PDF). In the case, σ is approaching its minimum at 0, PDF exceeds 1 driving the loss function to the negative values as shown in Fig 4. In terms of probabilistic models, the total loss depends on both spatial and probabilistic components, therefore the latter gained a bigger momentum. As the result, the model paid less attention to the squared Euclidean error distance which negatively affects geospatial accuracy. The suggested solution limits PDF to [0, 1] interval by setting the minimum of σ to $\frac{1}{2\pi}$. The covariance measures the uncertainty of the model about the corresponding point, consequently, in the best-case scenarios probabilistic loss would depend mainly on the distance error:

$$PDF = \lim_{\sigma \to \frac{1}{2\pi}} \frac{e^{-\frac{D^2}{2\sigma}}}{2\pi\sigma} = e^{-\pi D^2};\ PDF \in [0, 1]$$

$$L_{prob} = \lim_{\sigma \to \frac{1}{2\pi}} \frac{D^2}{2\sigma} + log(2\pi\sigma) = \pi D^2;\ L_{prob} \geq 0$$

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 multiple probabilistic 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 [YCM+13] 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, 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"), 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.

References

[ABMS20] Thaufan Ardi Arafat, Indra Budi, Rahmad Mahendra, and Dena Aurum Salehah. Demographic analysis of candidates supporter in twitter during indonesian presidential election 2019. In *2020 International Conference on ICT for Smart Society (ICISS)*, pages 1–6. IEEE, 2020.

[BPW+18] Jordan Bakerman, Karl Pazdernik, Alyson Wilson, Geoffrey Fairchild, and Rian Bahran. Twitter geolocation: A hybrid approach. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 12(3):1–17, 2018.

[CCL10] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768, 2010.

[CGK15] Miriam Cha, Youngjune Gwon, and H Kung. Twitter geolocation and regional classification via sparse coding. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, pages 582–585, 2015.

[CJA14] Ryan Compton, David Jurgens, and David Allen. Geotagging one hundred million twitter accounts with total variation minimization. In *2014 IEEE international conference on Big data (big data)*, pages 393–401. IEEE, 2014.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[DNT+18] Tien Huu Do, Duc Minh Nguyen, Evaggelia Tsiligianni, Bruno Cornelis, and Nikos Deligiannis. Twitter user geolocation using deep multiview learning. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6304–6308. IEEE, 2018.

[EAX11] Jacob Eisenstein, Amr Ahmed, and Eric P Xing. Sparse additive generative models of text. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1041–1048, 2011.

[EOSX10] Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1277–1287, 2010.

[ESWC18] Mohammad Ebrahimi, Elaheh ShafieiBavani, Raymond Wong, and Fang Chen. Twitter user geolocation by filtering of highly mentioned users. *Journal of the Association for Information Science and Technology*, 69(7):879–889, 2018.

[FNX+15] David Flatow, Mor Naaman, Ke Eddie Xie, Yana Volkovich, and Yaron Kanza. 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*, pages 127–136, 2015.

[HC19] Binxuan Huang and Kathleen M Carley. A hierarchical location prediction neural network for twitter user geolocation. *arXiv preprint arXiv:1910.12941*, 2019.

[HCB12] Bo Han, Paul Cook, and Timothy Baldwin. Geolocation prediction in social media data by finding location indicative words. In *Proceedings of COLING 2012*, pages 1045–1062, 2012.

[HCB14] Bo Han, Paul Cook, and Timothy Baldwin. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49:451–500, 2014.

[HSF15] Mans Hulden, Miikka Silfverberg, and Jerid Francom. Kernel density estimation for text-based geolocation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[IWA17] Hayate Iso, Shoko Wakamiya, and Eiji Aramaki. Density estimation for geolocation via convolutional mixture density network. *arXiv preprint arXiv:1705.02750*, 2017.

[Jur13] David Jurgens. 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*, volume 7, pages 273–282, 2013.

[KLH14] Longbo Kong, Zhi Liu, and Yan Huang. Spot: Locating social media users based on social network context. *Proceedings of the VLDB Endowment*, 7(13):1681–1684, 2014.

[KMO11] Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. " 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, 2011.

[LI15] Ji Liu and Diana Inkpen. 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*, pages 201–210, 2015.

[LLGL22] Menglin Li, Kwan Hui Lim, Teng Guo, and Junhua Liu. A transformer-based framework for poi-level social post geolocation. *arXiv preprint arXiv:2211.01336*, 2022.

[MCC13] Jeffrey McGee, James Caverlee, and Zhiyuan Cheng. Location prediction in social media based on tie strength. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 459–468, 2013.

[MM15] Fernando Melo and Bruno Martins. Geocoding textual documents through the usage of hierarchical classifiers. In *Proceedings of the 9th Workshop on Geographic Information Retrieval*, pages 1–9, 2015.

[MRCB18] Taro Miyazaki, Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. Twitter geolocation using knowledge-based methods. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 7–16, 2018.

[MTTO17] Yasuhide Miura, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma. 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)*, pages 1260–1272, 2017.

[PCDV14] Reid Priedhorsky, Aron Culotta, and Sara Y Del Valle. Inferring the origin locations of tweets with quantitative confidence. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 1523–1536, 2014.

[RBC17] Afshin Rahimi, Timothy Baldwin, and Trevor Cohn. Continuous representation of location for geolocation and lexical dialectology using mixture density networks. *arXiv preprint arXiv:1708.04358*, 2017.

[RCB17] Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. A neural model for user geolocation and lexical dialectology. *arXiv preprint arXiv:1704.04008*, 2017.

[RM14] KyoungMin Ryoo and Sue Moon. Inferring twitter user locations with 10 km accuracy. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 643–648, 2014.

[RSR+12] Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 1500–1510, 2012.

[RVCB15] Afshin Rahimi, Duy Vu, Trevor Cohn, and Timothy Baldwin. Exploiting text and network context for geolocation of social media users. *arXiv preprint arXiv:1506.04803*, 2015.

[SHP+13] Axel Schulz, Aristotelis Hadjakos, Heiko Paulheim, Johannes Nachtwey, and Max Mühlhäuser. A multi-indicator approach for geolocalization of tweets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, pages 573–582, 2013.

[SKB12] Adam Sadilek, Henry Kautz, and Jeffrey P Bigham. Finding your friends and following them to where you are. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 723–732, 2012.

[SL21] Yves Scherrer and Nikola Ljubešić. Social media variety geolocation with geobert. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*. The Association for Computational Linguistics, 2021.

[SMY22] Lihardo Faisal Simanjuntak, Rahmad Mahendra, and Evi Yulianti. We know you are living in bali: Location prediction of twitter users using bert language model. *Big Data and Cognitive Computing*, 6(3):77, 2022.

[VPPA20] Danae Sánchez Villegas, Daniel Preoţiuc-Pietro, and Nikolaos Aletras. Point-of-interest type inference from social media text. *arXiv preprint arXiv:2009.14734*, 2020.

[WB11] Benjamin Wing and Jason Baldridge. 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, 2011.

[WB14] Benjamin Wing and Jason Baldridge. Hierarchical discriminative classification for text-based geolocation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 336–348, 2014.

[WKA+18] Shoko Wakamiya, Yukiko Kawai, Eiji Aramaki, et al. Twitter-based influenza detection after flu peak via tweets with indirect information: text mining study. *JMIR public health and surveillance*, 4(3):e8627, 2018.

[YAK13] Yuto Yamaguchi, Toshiyuki Amagasa, and Hiroyuki Kitagawa. Landmark-based user location inference in social media. In *Proceedings of the first ACM conference on Online social networks*, pages 223–234, 2013.

[YCM+13] Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat Thalmann. 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*, pages 605–613, 2013.

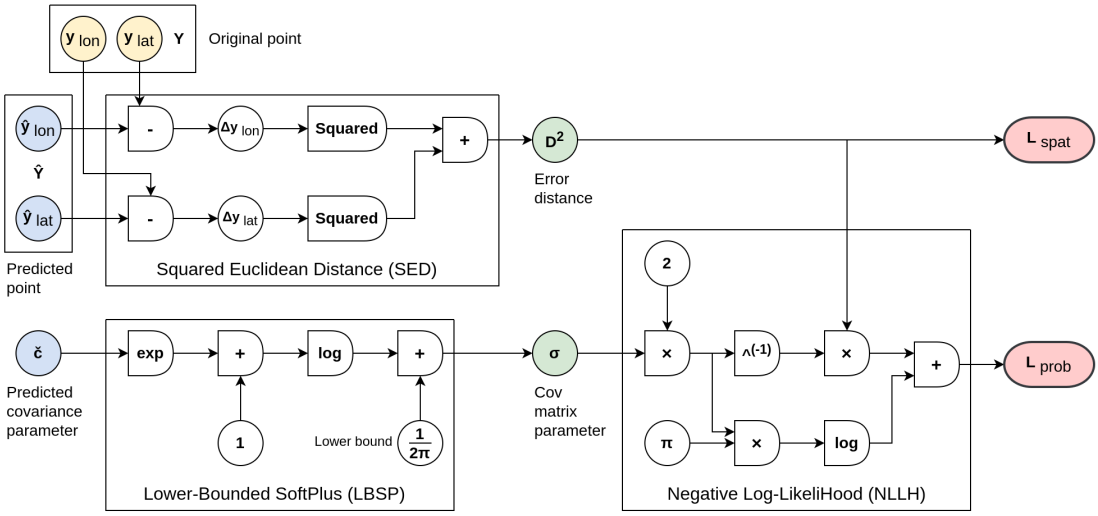
[YSP+20] Ussama Yaqub, Nitesh Sharma, Rachit Pabreja, Soon Ae Chun, Vijayalakshmi Atluri, and Jaideep Vaidya. Location-based sentiment analyses and visualization of twitter election data. *Digital Government: Research and Practice*, 1(2):1–19, 2020.

[ZHS18] Xin Zheng, Jialong Han, and Aixin Sun. A survey of location prediction on twitter. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1652–1671, 2018.

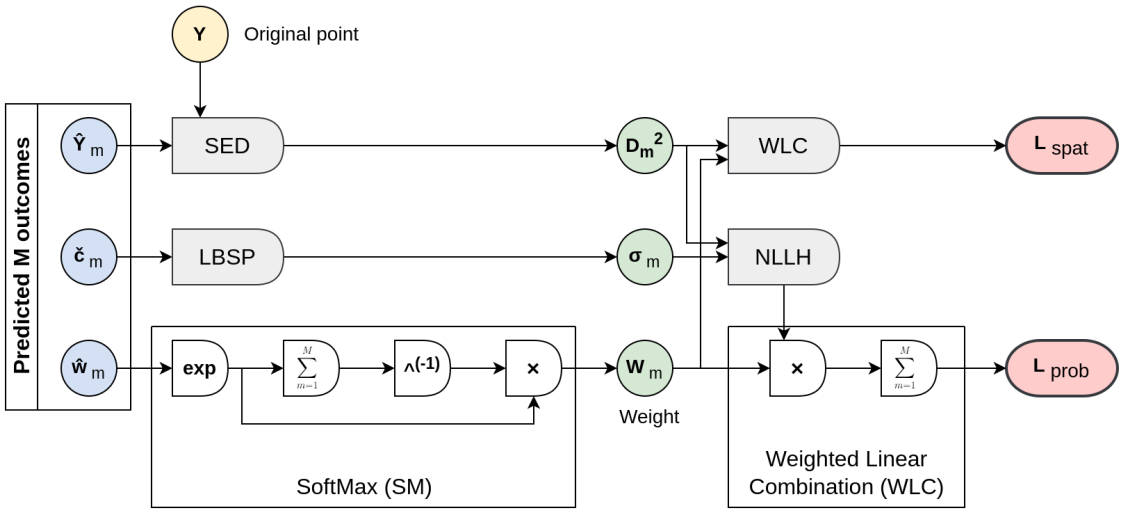
[ZJZ+20] Cheng Zheng, Jyun-Yu Jiang, Yichao Zhou, Sean D Young, and Wei Wang. Social media user geolocation via hybrid attention. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1641–1644, 2020.

[ZWW+20] Ting Zhong, Tianliang Wang, Jiahao Wang, Jin Wu, and Fan Zhou. Multiple-aspect attentional graph neural networks for online social network user localization. *IEEE Access*, 8:95223–95234, 2020.

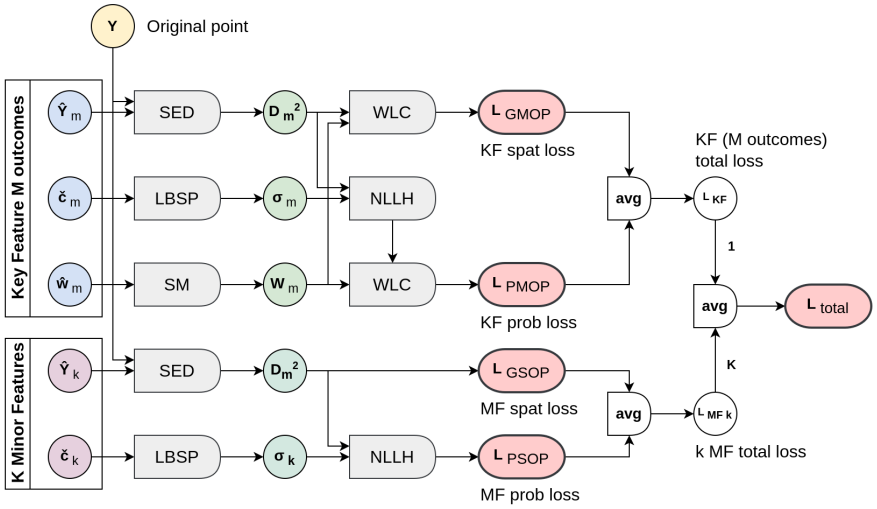
[ZWZT22] Fan Zhou, Tianliang Wang, Ting Zhong, and Goce Trajcevski. Identifying user geolocation with hierarchical graph neural networks and explainable fusion. *Information Fusion*, 81:1–13, 2022.



(a) 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



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



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

Figure 4: Proposed loss functions visualized in the computational graphs

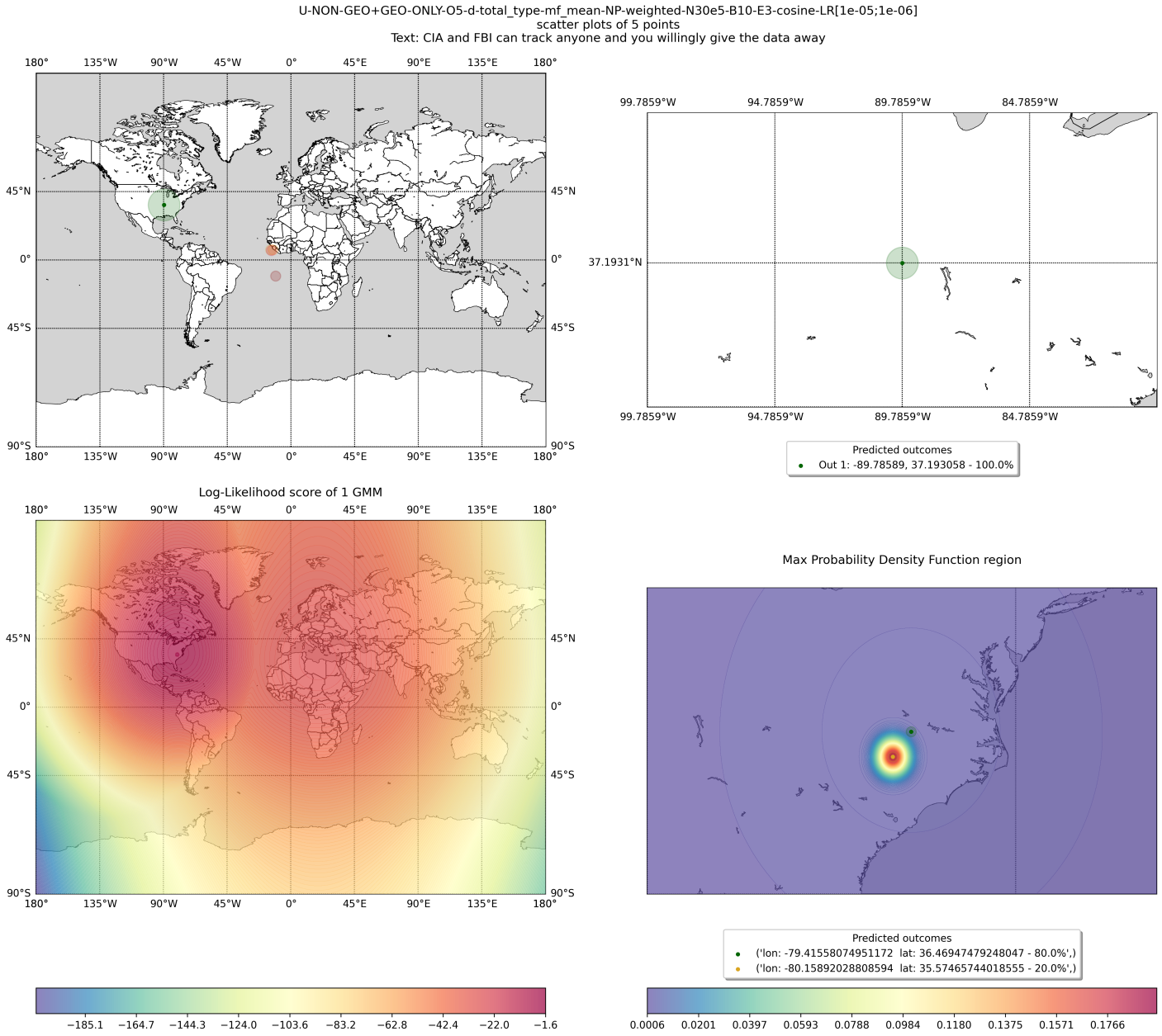


Figure 5: Prediction examples of two models for the same text *"CIA and FBI can track anyone, and you willingly give the data away"*; both models were trained on the same worldwide dataset with key *NON-GEO* and minor *GEO-ONLY* text features, the number of prediction outcomes is 5;
above - GMOP significant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