Predicting the geolocation of tweets using transformer models on customized data

Kateryna Lutsai[1] and Christoph H. Lampert[2]

[1]National Technical University of Ukraine – Igor Sikorsky Kyiv Polytechnic Institute, Ukraine
[2]Institute of Science and Technology Austria, Austria

Abstract: This research is aimed to solve the tweet/user geolocation prediction task and provide a flexible methodology for the geo-tagging 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. Our source code and data are available at github.com/geo-twitter

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 [34], and possible

insights into the demographic characteristics of a candidate's supporters [1] that are valuable during Presidential elections [38]. 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 allocate resources better. Furthermore, in a variety of business settings, including retail, real estate, and marketing [13], 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 [43]. Although, much research dedicated to predicting users' home location utilizes the user network of connections (e. g. followings, replies, mentions) to achieve higher accuracy [42,44,45], this work focuses primarily on textual data analysis. Therefore, only tweet content (the text itself) and tweet context (tweet and user metadata) are adopted to solve the problem of tweet location prediction.

Although the user level geolocation is certainly effective for some applications, message level geolocation supports fine-grained analyses. Thus, grouped by user sets of per-tweet predictions in the form of two-dimensional distributions were leveraged for solving the task of the user's home location prediction. This methodology of using predicted from tweets GMMs for approximating user location can be found in Appendix A.

The solving of the tweet location prediction problem requires the labeled location coordinates which are commonly extracted from the tweet's metadata. In terms of Twitter, only 1–2% of the tweets are geo-tagged with longitude-latitude coordinates [23] which is by default 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 [15].

However, this study focuses on the country and global worldwide-level scale problems, thus the former method using longitude-latitude coordinates as a labeled location 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' labeled location. Although the WGS84 is a continuous, global reference system, when data is displayed on a two-dimensional map, such as a simple Plate carrÃľe projection, where numerical values of longitude and latitude stay in strict ranges of 360 and 180 degrees respectively.

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 [30] showing that regression outperformed classification in the geospatial accuracy (mean and median error distance similar to SAE described in Section B.1) metrics for the limited Twitter datasets of BosnianCroatian-Montenegrin-Serbian and Deutch languages. Detailed descriptions of the proposed regression wrapper layers, outputs post-processing, and loss computation procedures can be found in Section 3.

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 tasks associated with Twitter data, this Section covers works related to the tweet location and the user's home location prediction. 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.

According to [43], the granularity of user's home location prediction task has three levels: administrative regions, geographical grids, and geographical coordinates. As for the separate tweet locations, coordinates or point-of-interests (POIs) are used instead of administrative regions and grids.

The most straightforward approach would be to process 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 [10, 27, 36]. These works apply geographical grids dividing the Earth's globe surface into different-sized regions such that highly populated areas are split into smaller pieces while more sparse population areas end up as large grid cells. Early efforts 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) [45]. For example, Rahimi extracts bag-of-words features from user posts [25]; Wing and Baldridge estimate the word distributions for different regions [35]. Other word-centric works [22, 31, 41] 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 [24] proposes a neural network encoding of such words and phrases to the continuous two-dimensional space. In this light, we aimed to develop another semi-supervised method based on the state-of-the-art models for NLP with no need for LIWs datasets or geographic gazetteers.

Another source of location-indicative information is based on the user interconnection graphs, also referenced as Twitter network in Twitter-related studies. [43]. When tasked with the prediction of users' home locations through an analysis of the Twitter network, we can assume that the user's home is close to their friends' home locations. Likewise, mentions and conversations allude to a closer relationship or shared interests between the two users, also referenced as social closeness.

Furthermore, in the realm of social science, the principle of homophily [20] postulates that individuals who share similarities are more inclined to establish connections at a higher rate compared to those with dissimilar attributes. Research into Twitter network utilization reveals that, in addition to mutual friendships established through the bidirectional following, users' activities such as mentions and active conversations with one another are indicative of their geographic proximity in terms of home locations [19].

Overall, studies in the previous works show that the prediction of user home locations draws equal significance from both Twitter network structures and tweet content [26, 42, 44, 45]. However, it's noteworthy that only a limited number of studies [4, 28, 40] leverage the Twitter network of user connections for predicting tweet locations.

Despite short in-length textual content, a tweet is accompanied by rich context like timestamps and geo-tags associated with it, and multiple attributes of its author profile.

For instance, when sending out a tweet, it is automatically accompanied by a posting timestamp. Moreover, GPS-enabled devices such as smartphones and tablets have become widespread allowing users to easily append geo-tags to their tweets, indicating their current locations [43]. In the case of Twitter, the spatial context is represented as machine-generated geographic object descriptions with countries, cities, place types, and full names attributed by Twitter to the user-defined geolocation coordinates.

Additionally, users can enrich their profiles with details like hometowns, timezones, personal websites, and bios. However, common methods mainly focus on content information due to the deficient data records obtained from publicly available sources of Twitter geo-tagged datasets. Since the usage of context is not well explored for tweet location prediction [43], there was an opportunity for data feature engineering of tweets' metadata contained in our raw datasets of archived Twitter posts.

Some researchers also define metadata features as model inputs, usually in addition to network and textual inputs, in order to predict users' home locations [9, 21]. Among utilized metadata varieties, temporal information like tweet timestamps and user-declared timezones are shown as effective in implying tweet and home locations at coarse-grained granularity [7]. For instance, Yuan in the work [39] managed to cover the temporal aspect and users' mobility to estimate each tweet's location source. However, the inputs of numerical type, such as timestamps and timezone are more ambiguous for the text-oriented transformer models explored in this study, therefore, time-related features were mainly ignored on account of more relevant input options. It's important to note that the user's timezone metadata is no longer used in Twitter, but would only provide insight into the geographic region of the user's home country or its capital city.

In terms of the tweet's author profile metadata, Schulz in the work [31], incorporates users' self-declared home locations, personal websites, and currently outdated timezone attributes, as potential tweet location indicators. For instance, the personal websites were processed to obtain the spatial indicators such as country code and IP address. While the extracted features offer valuable insights into a user's home country, they also lead to a noticeable increase in tweet processing time. Given the imperative of maintaining a swift processing pace, especially when dealing with an overall Twitter feed generation rate of 6,000 tweets per second, the research focus shifted to utilizing readily available fields of the tweet's author profile. In particular, textual fields such as users' home locations, profile bios, names, and screen names were used as model input features for the tweet location prediction.

In the realm of user home location prediction, some authors came up with probabilistic models solving the classification task on the country/city levels [8, 18]. However, employing such classification-based approaches becomes impractical for tweet location prediction, given the potential presence of hundreds of thousands of Points of Interest (POIs) within a city. While classification-based methodologies are commonly employed in home location prediction, the same does not hold true for coordinate-oriented tweet location prediction. Thus, there are relatively few studies that adopt a classification-based approach to predict the location of individual tweets in the form of specific POIs, as exemplified by [4] and [10].

Importantly, Priedhorsky's methodology was aligned with the present work as it utilized GMMs for geolocation prediction instead of simple coordinates prediction [23]. Another subsequent study, which employed a hybrid approach, also estimated geolocation through GMMs rather than coordinates [2]. This study's final estimation was based on the

text and network features as joint predictors. The performance metrics used in both works were adopted to evaluate the proposed probabilistic models as described in Section B.2.

Continuing the exploration of probabilistic models used in the relevant studies, Iso in the work [11] proposed the convolutional mixture density network which is applied to the tweets' content to estimate the parameters of the GMM and employ the mode value of estimated distribution density as the predicted coordinates for tweets. The authors also claim that the choice of a loss function affects the model performance and describe a basic logic of the probabilistic model loss computation using negative log-likelihood as shown in Eq. (NLLH). Similarly, Li in the work [16] utilizes the negative log-likelihood of the GMM over all training examples. Notably, the weights of the Gaussian components were obtained directly from the model's attention scores. In contrast, we propose the prediction of all GMM parameters (weights, covariance) by the model itself rather than the application of available attention weights.

Most of the previously mentioned works use various approaches based on neural networks for NLP, but none of them has applied transformer models for the estimation of geolocation. The relatively new BERT model (released in 2018) is aimed at classification tasks and some works have already used it for country/city names predictions.

The base BERT model is intended to be finetuned 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) to improve the model's understanding of bidirectional sentence representations, along with Next Sentence Prediction (NSP) to capture the inter-sentence relationships. 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.

For instance, Villegas in the work [33] predicts place type or point of interest (POI) using lowercase English BERT. Furthermore, Li in the work [15] focused on city-scale prediction for the Twitter datasets of Melbourne, and Singapore. However, 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 [32], the task of predicting Twitter users' home location was explored using Long-Short Term Memory (LSTM) and BERT models. The authors used a dataset in the Indonesian language containing text content and user metadata context of tweets. 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.

Such constraints were challenging for 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. Authors state that the majority vote approach for per-tweet predictions in 42% cases resulted in misclassification of user home location. Thus only the task of tweet geolocation prediction could be efficiently solved using the BERT-base models. 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 A.

In another work, Scherrer uses BERT models for both regression and classification tasks on limited text datasets of BosnianCroatian-Montenegrin-Serbian and Deutch languages [30]. 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.

Moreover, the results presented by [30] 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 [30] demonstrated that language-specific BERT models clearly outperforms their multilingual counterpart on most of the used datasets. In light of these findings, the country-level model for the English and French-speaking countries was fine-tuned on a limited subset of French, Canadian, and British tweets as described in Section 3.3.2.

Nevertheless, the global dataset was multilingual so the corresponding BERT base model ('bert-base-multilingual-cased') was chosen, 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.

3 Materials and Methods

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. 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. Architecture modifications for the best of proposed models involved only the scope of wrapper layers implementing linear regression to the output of parameters defining a GMM (means coordinates, weights, and covariance parameters).

Considering that the two-dimensional distributions were used as the output format of the probabilistic models, this work proposes the custom procedures of loss function computation for the four model types different in output formats. The training procedure for the best of the proposed models is demonstrated on the flowchart diagram in Figure 1. In terms of the input data used for model fine-tuning, 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).
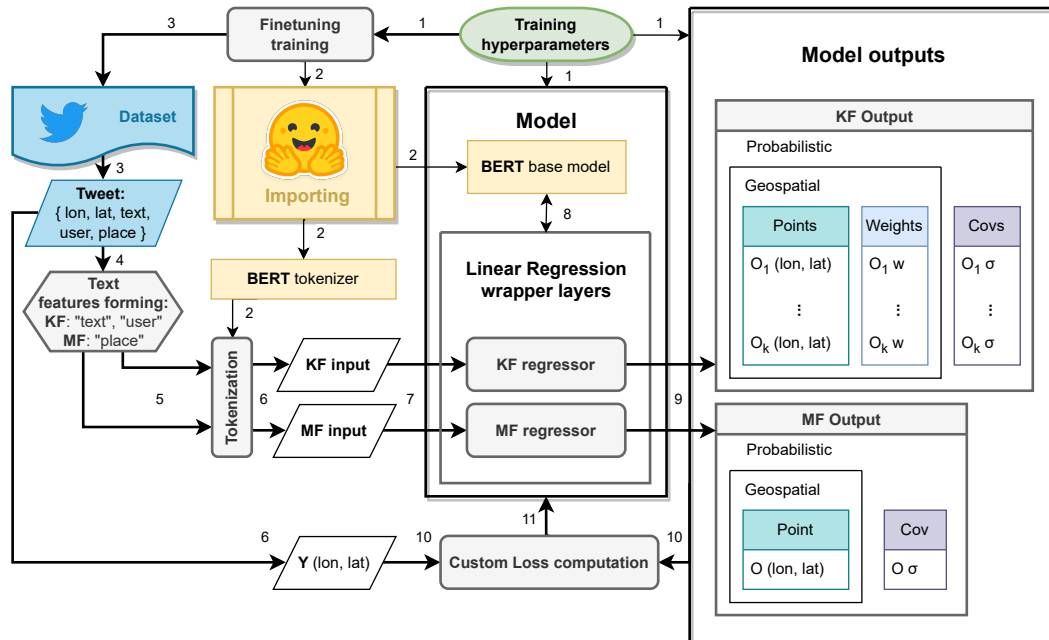

Figure 1: Model finetuning procedure flowchart for the best of the proposed models which has the Probabilistic Multiple Prediction Outcomes (PMOP) output type and individual wrapper layers for the necessary Key Feature (KF) and optional Minor Feature (MF)
Step 1: parameterization of the model's architecture, outputs, and finetuning;
Step 2: importing default tokenizer and base BERT [?] model;
Step 3: reading dataset files from the Twitter archive into a virtual sheet;
Step 4: forming inputs *NON-GEO* as KF and *GEO-ONLY* as MF according to Section 3.1;
Step 5: tokenization of the text inputs through the standard BERT tokenizer;
Step 6: forming of the model data loaders by selection of batch subsets from data;
Step 7: loading input features to the model, each to separate wrapper layer;
Step 8: text processing by the local base BERT model before applying wrapper layers;
Step 9: returning defined in size model outputs according to Table 3;
Step 10: comparing processed predictions and labels according to Section C;
Step 11: backpropagation of the total per-batch loss to the local base BERT version.

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 3.2. The intention was to test whether the Probabilistic Single/Multiple Outcome Prediction (PSOP/PMOP) models outperform the straightforward approach of Geospatial Single/Multiple Outcome Prediction (GSOP/GMOP) models.

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 (development set) metrics. After each epoch, the models were evaluated on the development set, which consists of 300,000 samples. The described hyper-parameters were used for 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.

3.1   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. 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 our Twitter archive database collected from 2020 to 2022 have geolocation coordinates, such that the number of available geo-tagged samples was approximately 150 million. It has been found that up to 20% of the geo-tagged tweets in the collected datasets 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.

The training data we collected for this study includes a comprehensive set of tweets and their metadata (author profiles, location geo-tags) serving as inputs, and labeled location coordinates playing role of the models's 'ground truth' through the training and evaluation. During the preprocessing stage, the original tweet objects were condensed into the text content ('text'), relevant metadata context ('user' and 'place'), and geolocation coordinates ('lon', 'lat') representing the labeled location.

Worth mentioning that the word-centric techniques, which use geocoders to recognize the referred place based on the geographic gazetteers [31]. 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. The challenge in this work was to find a substitute for the accurate knowledge base of geographical references (e.g., gazetteer terms).

Fortunately, 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 tweet JSON root-level objects 'place' and 'user' had multiple fields containing potentially relevant context data, some of them were used in forming the input features as shown in Table 1. Despite the fact that only 26% of users provide location information in their profiles according to [5] and that user-generated data can be noisy, the proposed multitask learning procedures ought to enhance the geospatial accuracy.

Feature name					Dataset columns	JSON 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 fields

Since typical model input shouldn't include 'place' context which refers exclusively to geo-tagged tweets, multitask learning has been implemented in order to both preserve the input format and utilize 'place' metadata during training. This approach implies separate wrapper layers for each of the textual input features and a per-batch loss common for all features during finetuning. The main 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 supplemental input feature was dedicated solely to the Context data obtained from the 'place' field.

Considering that evaluation ought to be performed on the typical model input ('text' and 'user'), models trained with the described Key and Minor Feature setup were expected to show 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'). Moreover, the tweet metadata providing context for the Content of the tweet was expected to improve overall model accuracy by comparison with the 'TO' (*TEXT-ONLY*) model trained only on the 'text' contents without metadata. More details on the comparison of merged and split input features can be found in Section 4 and specifically in Table 4 containing all performance measurements.

Our research primarily focused on using text information to predict tweets and users' home locations, as it is easily accessible from tweet contents and user profile data. Since all tweets are linked to their authors, user profiles were utilized as a major part of the model input sequences. Although profile background images could provide additional insights, incorporating them would require a more resource-intensive hybrid text+image model, which was beyond the scope of our work. Furthermore, the utilization of portrait images would rise more serious privacy issue of sensitive data.

One of the limitations encountered is the unavailability of appropriate datasets that include profile background images. The Twitter datasets commonly used for location predic-

tion do not have these images, and the Twitter streaming API does not provide immediate access to them, requiring manual retrieval of each image individually. This process would introduce significant overhead, especially for a continuously running system.

Considering the challenges of computational complexity and data availability, we chose to concentrate solely on user profile textual information (screen name, name, description, and location of the Twitter user profile), as shown in Table 1. This allowed us to effectively achieve our research objectives of predicting geolocation for the arbitrary tweets within our available resources and datasets. Worth mentioning, that the user's home location wasn't retrieved from the profile but rather sampled from the user's tweets subset during evaluation.

The labeled location for this study was defined as the pair of longitude-latitude coordinates associated with each geo-tagged tweet. By definition, there are two JSON children entities that describe the location information of a tweet: 'coordinates' and 'place'. The former is present only when the exact location was assigned by the user, therefore data labels in a longitude-latitude format have been collected by parsing only the 'coordinates' object of tweets.

In contrast to the tweet numerical geo-tag, the 'place' field provides an accurate automatically generated by Twitter textual description of the location, such as country-city naming, which was utilized as a supplementary input during the finetuning of the best models. As shown in Table 1, the fields of the 'place' object utilized in forming the training dataset included country, place type, name, location, and full name of the geo-tagged location.

The decision to split the training input into Key and Minor Features was dictated by the assumption about the model's attention shift. The model trained on the combination of all input features was expected to learn to pay attention only to the geo-tags. To avoid memorizing of positions of the geo-tag tokens in the input sequence during training, the randomization of words could not be applied due to the semantics distortion.

The comparison of 'A' (*ALL*) and 'NG+GO' (*NON-GEO+GEO-ONLY*) models was intended to prove that the identical model setups trained on merged and split input features would differ in accuracy. In case the separation of *GEO-ONLY* from the always available *NON-GEO* enhanced the *NON-GEO* accuracy, it would support the idea of applying multitask learning techniques in this work despite the slower in total processing of multiple inputs and outputs.

There was a limited number of text feature combinations explored in this study as described in Table 2. Note that each training sample had separate tokenized sequences for all input features assigned to it in the model data loader, such that per-batch loss was computed as a linear combination of per-feature errors. The visualization of the multitask loss computation procedure can be found in Figure 9 alongside the scheme for a single feature (Key only) model of K outcomes in Figure 8.

The models were evaluated on diverse arbitrary tweets imitating a common situation when the 'place' field is unavailable and only the key features of always present 'text' and 'user' inputs are used for prediction. In general, each tweet always contains 'text' and 'user' (*NON-GEO*) used as the Key training Feature, and as an input during the evaluation of development and validation datasets.

In practice, each model had a necessary Key Feature (KF) and an optional Minor Feature (MF) wrapper layers which were used according to the input type during the finetuning stage. Utilizing the 'place' field as MF and user-generated text ('text' and 'user') as KF was

| 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.

crucial for predicting geolocation coordinates accurately. The 'place' field provided valuable complementary information, enhancing context understanding and disambiguating similar locations, while tweet content offered relevant cues and linguistic patterns, improving prediction precision.

The described combination of *NON-GEO* KF and *GEO-ONLY* MF created a robust and flexible model, capable of handling diverse tweets when the 'place' field is unreliable or unavailable. Therefore, the resulting trained system possesses the capability to effectively analyze and predict the labeled location coordinates based on any arbitrary input of the tweet-like length.

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).

3.2 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 covariance matrix parameters. 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

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: Wrapper layer types by the number of outputs grouped by Spatial (Spat) and Probabilistic (Prob) 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 column shows the abbreviations of different model configs.
Point refers to coordinate pair $\widehat{\mathbf{Y}}$; Weight is $\widehat{w}$ converted to the GMM weights W; and Cov corresponds to $\widehat{c}$ converted to the $\sigma_{\widehat{c}}$ parameter of the spherical covariance matrix $\mathbf{\Sigma}$

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.

$$\widehat{\mathbf{Y}}_{\mathbf{spat}} = (\widehat{y}_{lon}, \widehat{y}_{lat}); \quad \widehat{y}_{lon} \in \mathbb{R}; \quad \widehat{y}_{lat} \in \mathbb{R}$$

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

The probabilistic output of the model included not only the spatial component of the predicted coordinate pair $\widehat{\boldsymbol{\mu}}$ but also a measure of the model's confidence in its prediction. 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 $\sigma_{\widehat{c}}$ remains positive, the SoftPlus function described in Eq. (SoftPlus) could be applied to the output variable $\widehat{c}$.

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

However, a lower bound for $\sigma_{\widehat{c}}$ 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].

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

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 $\sigma_{\widehat{c}}$ curtailed the sharpness of the Gaussian peaks and ensured that their height never exceeds 1, thus avoiding negative values in the probabilistic loss as show 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}
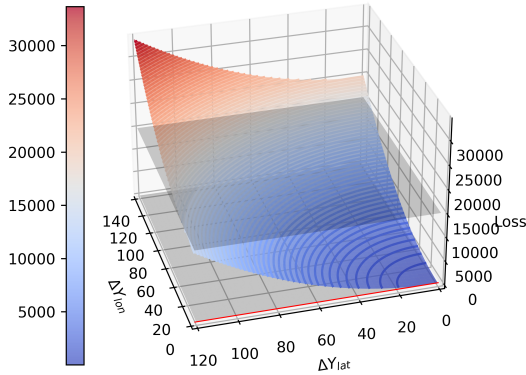
Figure 2: Squared Euclidean Distance (SED) function surface on the axes of ΔY_{lon} and ΔY_{lat} as the error distances per longitude and latitude axes; upper horizontal gray surface indicates the empirical maximum of L_{spat}; red line indicates the strict minimum of 0 implied by the nature of Eq. (SED).
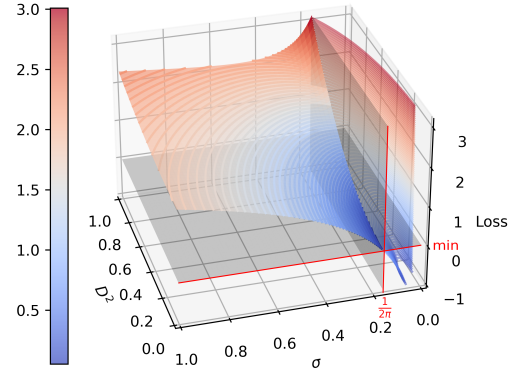
Figure 3: Negative Log-LikeliHood (NLLH) function surface on the axes of D^2 as the error distance and $\sigma_{\hat{c}}$ as the uncertainty in $\hat{\mu}$ of the Gaussian; red lines and gray surfaces indicate the reduction of L_{prob} domain as a result of Eq. (LBSP) application.

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.

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 SOP-type models are visualized in Figure 4, while computational graphs for the MOP-type models and multitask learning procedures can be found in Appendix C.

The spatial loss for SOP-type models was computed as the squared Euclidean distance between the true location specified by the user and the location predicted by the model on the two-dimensional Plate carrÃľe 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.

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

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

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

$$L_{PSOP} = -\log(PDF) = \frac{D^2}{2\sigma} + \log(2\pi\sigma); \quad L_{PSOP} \geq 0 \qquad \text{(NLLH)}$$
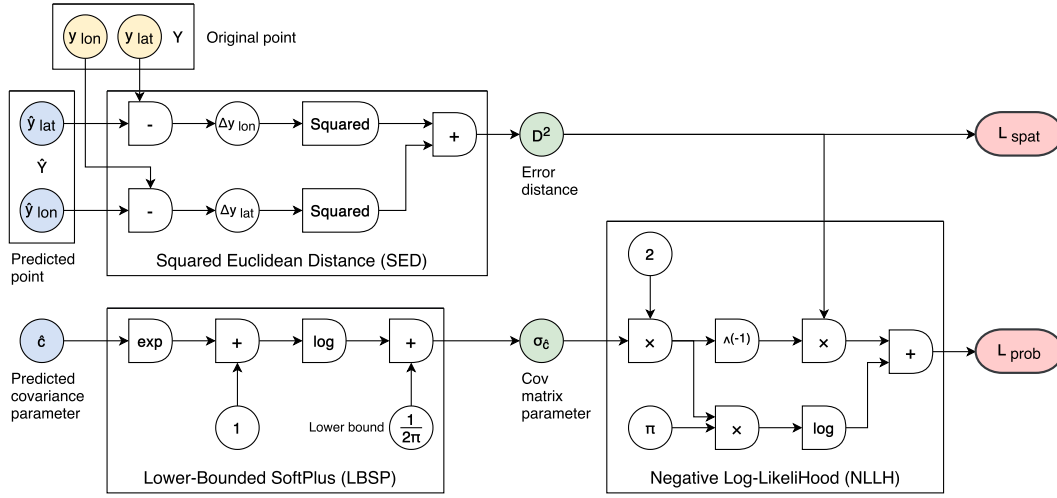
Figure 4: Single Outcome Prediction (SOP) model loss functions computational graph including visualization of Squared Euclidean Distance (SED), Lower-Bounded SoftPlus (LBSP), and Negative Log-LikeliHood (NLLH) components.

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):

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

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:

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

In the case of MOP-type models, the output included a weight $\widehat{w}_i$ 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 $\widehat{w}_i$.

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

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

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

The total loss per text feature f had an essential geospatial component, such that L_f was equal to the spatial loss L_{spat}, and optionally the probabilistic loss L_{prob} added to it. The per-feature loss for the probabilistic models was computed in three ways: average or sum of L_{spat} and L_{prob}, or 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 computed as an average of two components in the case of probabilistic models.

Finally, the total loss as an average of all per-feature loss values was computed to handle multiple textual features of a single tweet.

In practice, the final step was computing the mean of total loss per tweet in a single data loader tensor (batch) 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 3.2 covers a scope of all models reviewed in Table 3. According to the experimental results in Table 4, the most accurate model configuration was the Probabilistic Multiple Outcomes Prediction (PMOP) model of 5 outcomes utilizing *NON-GEO* as the Key Feature and *GEO-ONLY* as the Minor Feature, which was finetuned on a worldwide dataset of tweets using custom loss computation procedures described in Appendix C.

3.3 Results

In this section, the proposed models are evaluated by geospatial and probabilistic performance metrics in global (Table 4) and country-level (Table 5) granularity Twitter datasets. Importantly, the final evaluation 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.

The visualization of the post-processed prediction example for the best of the proposed models is shown below in Figure 5. The predicted set of weighted two-dimensional points and two-dimensional distribution with selected peaks were visualized for the Geospatial and Probabilistic models trained by the best approach. This example demonstrates a prediction of the randomly chosen text without a labeled location (genuine truth) for the sake of comparison between two types of models trained on the Content and Context and producing an output of either multiple geolocation points (GMOP) or united distributions of location likelihood (PMOP). Considering the training data and procedures were the same for both models, there was a 15-degree difference between the GMOP and PMOP predictions. The indirect input without geographic references resulted in a relatively small shift of less than 2,000 km while maintaining the common guess of the US origins of the text.

3.3.1 Worldwide evaluation metrics by model type

The models compared in Table 4 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
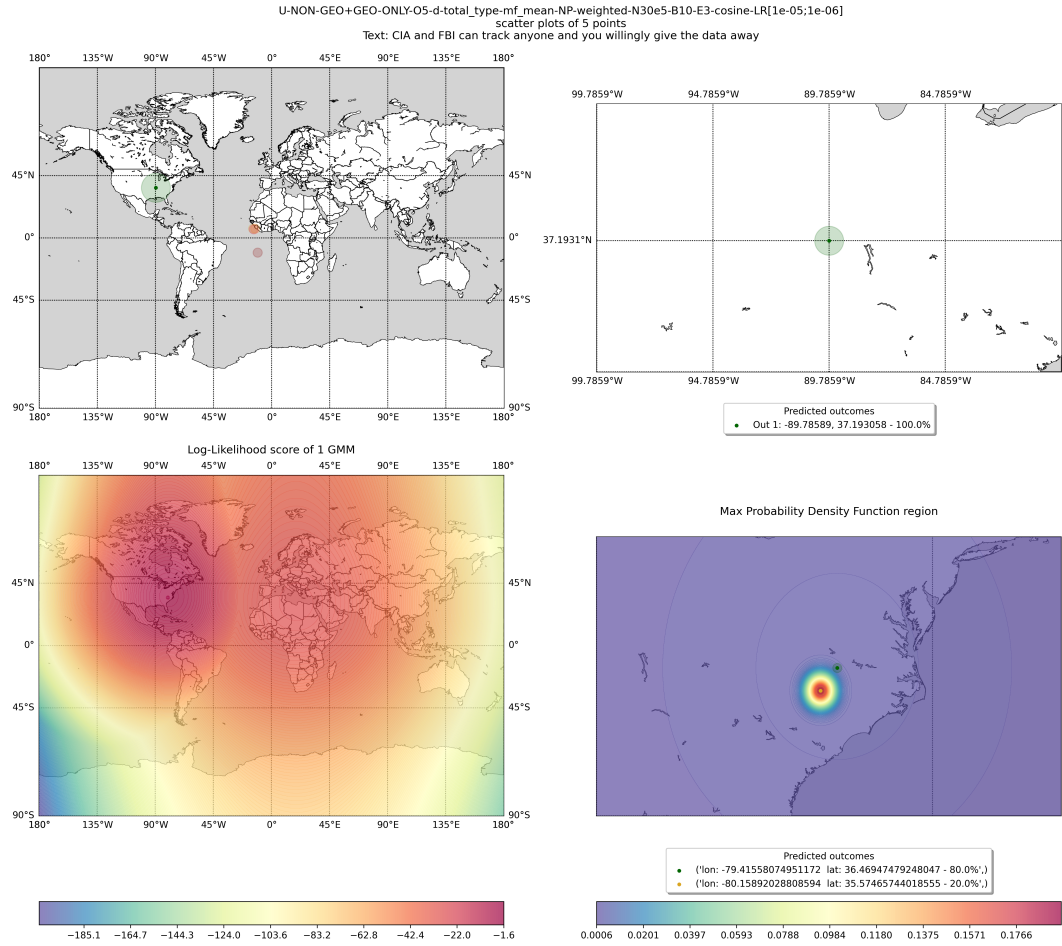
Figure 5: Prediction examples of two models, which were trained on the same worldwide dataset with key *NON-GEO* and minor *GEO-ONLY* text features, for the same text:
'CIA and FBI can track anyone, and you willingly give the data away' with 5 outcomes sorted by significance (weight);
above: Geospatial GMOP points as scatter plots;
below: Probabilistic PMOP Gaussian peaks as LLH and PDF plots.

with *NON-GEO* as the Key and *GEO-ONLY* as the Minor Features respectively. Metrics results for the MOP-type models were computed as the weighted linear combination of all outcomes, specific formulas for the MOP-type model can be found in Section B.

Evaluation results of the worldwide models in Table 4 show the difference between geospatial and probabilistic models with a number of prediction outcomes ranging from 1 to 100. The development dataset consisted of 300K tweets from 143K users covering 63 languages and 226 countries. Note that both train and development datasets were the same for all models and contained tweets from both real and bot users.

| 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 4: Worldwide dataset (300,000 tweets) results of models performance metrics; TF stands for Training Feature, VF for eValuation Feature, and OUT defines number of outcomes while other model configuration parameters remain steady; u*-unlimited covariance output, $\sigma_{\widehat{c}}$ were obtained by the application of Eq. (SoftPlus) to the covariance parameter output $\widehat{c}$
A: *ALL* or 'text'+'user'+'place', NG: *NON-GEO* or 'text'+'user', TO: *TEXT-ONLY* or 'text', NG+GO: *NON-GEO + GEO-ONLY* or ('text'+'user')+'place' where 'text', 'user', and 'place' refer to the inputs defined in Table 1.

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 $\sigma_{\widehat{c}}$ covariance parameters thereby maintaining the values of Eq. (NLLH) in the positive domain. The performance results in Table 4 confirm that the PMOP model of 5 outcomes noted by $u*$ (stands for the obtained from Eq. (SoftPlus) unlimited $\sigma_{\widehat{c}}$) 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 4 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 un-
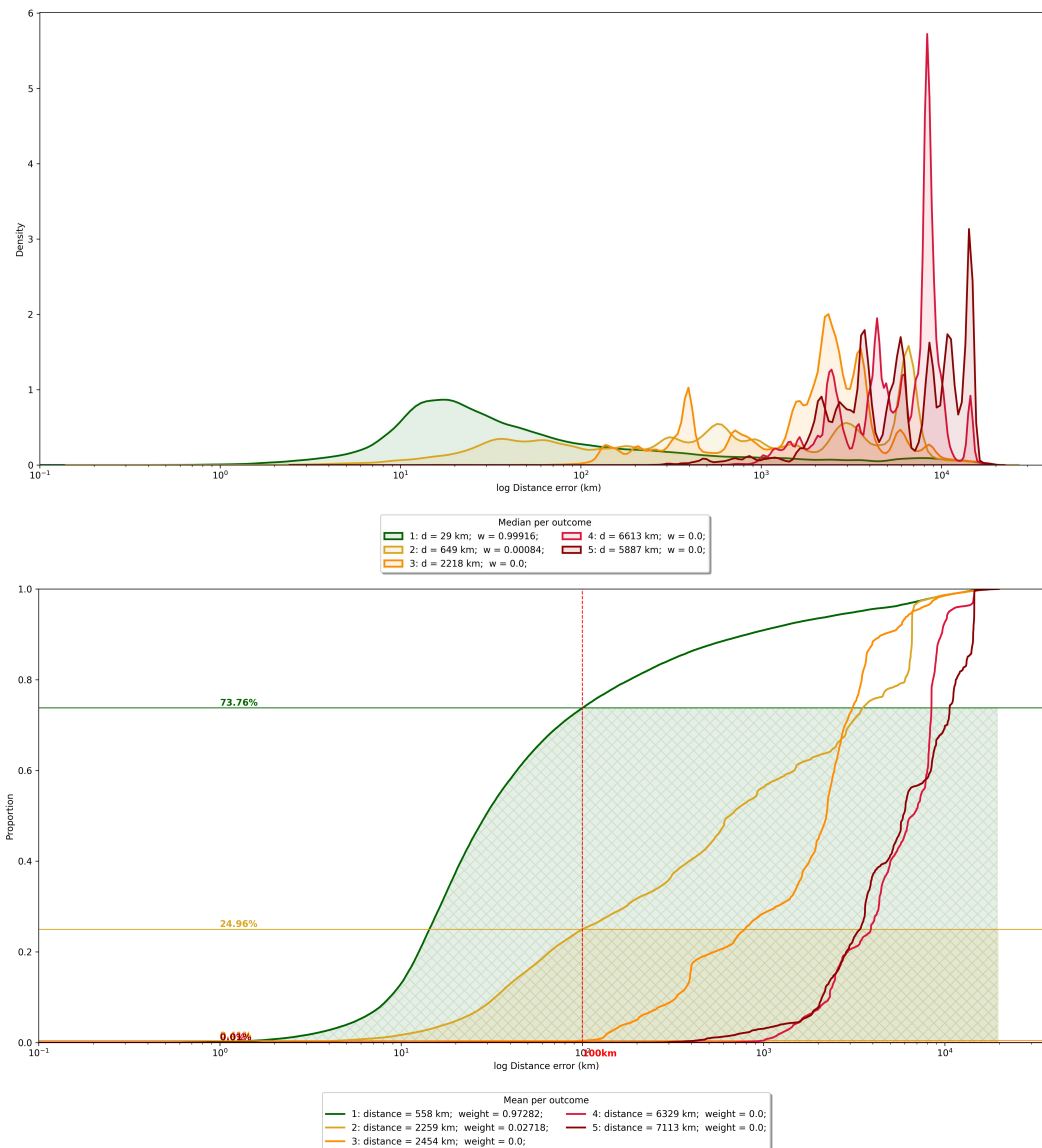
Figure 6: Density (above) and Cumulative distribution (below) of the log error distance in km per outcome for the best model (Probabilistic, 5 outcomes, *NON-GEO + GEO-ONLY*) computed on the dataset of 100,000 worldwide tweets; median (above) and mean (below) errors and weights per outcome are noted in the legends below the graphs.

desirable decrease in the COV criteria. According to the analysis of the error distance per outcome depicted in Figure 6, up to 2 outcomes were assigned significant weights and only 73.76% predictions of the best outcome had median errors below 100 km.

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 4. This observation supports findings in the work [16] that demonstrated the benefits of predicting a probability distribution rather than a single point - several candidate locations shifted the single output in different directions. Similarly, the MOP-type architecture outperformed SOP-type in both Probabilistic and Geospatial variations, though the difference was relatively low.

3.3.2 Ablation study metrics

In contrast to the worldwide dataset metrics, the ablation study of the best model covers only three countries Great Britain, France, and Canada. The choice of these three countries was made to specifically address the challenge of handling ambiguity in geolocation prediction on multilingual territories. Canada, in particular, was included because tweets from this country can be in either English or French, which often leads to confusion for models that cannot accurately predict the probability distributions.

Stat	Country			
	All	Great Britain	France	Canada
Train size	3,000,000	1,772,344	366,882	860,834
English lang.	2,494,749	1,649,065	53,753	791,931
French lang.	326,165	17,052	281,163	27,950
Other lang.	179,086	106,227	31,906	40,953
Test size	300,000	176,894	36,607	86,499
English lang.	249,750	164,661	5,424	79,665
French lang.	32,344	1702	25,905	2,737
Other lang.	17,906	10,531	5,278	4,097
Mean SAE	57.19 km	31.15 km	65.51 km	106.93 km
Med SAE	3.68 km	3.3 km	5.9 km	3.94 km
Acc@161	96%	97%	92%	95%

Table 5: Performance results of the model trained and evaluated on the Great Britain + France + Canada dataset using the hybrid approach (Content + Context) and PMOP-type output of 5 prediction outcomes for the estimation of the tweet location.

The aim of the ablation study was to showcase the handling of ambiguity and evaluate the performance of the model in a limited dataset, focusing on these specific countries. The dataset composition was designed to demonstrate the impact of language distribution and the proportion of samples from each country on the model's accuracy.

Table 5 presents the statistics of the ablation study, illustrating the distribution of tweets and languages within the dataset. British tweets constituted the largest proportion, accounting for 59% of the training and evaluation dataset, Canadian samples represented 29%, while French tweets had the smallest portion at 12%. Additionally, the distribution of languages showed that English tweets dominated the dataset, comprising 83% of all tweets, while French tweets accounted for 11%, and the remaining 6% were in other languages. Since 75% of the French and only 3% of Canadian tweets were in the French language the small number of French samples was compensated by their distinctive nature.

Therefore, the ablation study on these specific countries effectively highlighted the model's ability to handle ambiguity in language and demonstrated the influence of dataset composition on geolocation prediction accuracy. The ablation study results demonstrated that the model performed better on this limited dataset compared to the worldwide and American datasets. Notably, 96% of the predictions were within 161 km of the ground truth locations, which supports the suggestion [30] of training country-scale models.

Among the selected countries, France and Great Britain, with their relatively smaller square areas, facilitated more accurate geolocation predictions. The proportion of samples per country within the dataset also influenced the model's accuracy, with British tweets outnumbering French tweets by a factor of 5, resulting in a two-fold increase in the mean and median SAE of the French data segment. Overall, Great Britain exhibited the lowest mean and median error distances, France, with the smallest amount of data, had the highest median SAE, and Canada, with the largest square area, had the highest mean SAE metrics.

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 demonstrates simplicity in terms of the trained model application as a tool for geo-tagging textual big data of short corpora. While the process of finetuning the models involved some complexity in loss functions, it is a one-time step that requires initial time and computing resources to develop the methodology of training the most accurate models.

The designed model architecture is adaptive and simple in structure since there is only a superficial modification of the base model's wrapper layers implementing linear regression with no changes to the base model layers. 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. Once the finetuning is completed, the resulting model becomes essentially plug-and-play, requiring minimal additional effort for user-specific downstream tasks.

For instance, setup requirements for the proposed model application are the following: access to the local or remote model at huggingface/geo-bert-multilingual; optionally dataset of text samples shorter than 512 tokens; and GPU or CPU node to launch the model.

While the training routine requires access to the base model; dataset that contains labeled coordinates (lon, lat), text features for the arbitrary model input, and supplementary geo-tags of the locations; the usage of GPU is recommended due to the time and computing resources needed.

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 loss as shown in Figure 7(d). 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, 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 demonstrated that the geospatial loss of the minor feature *GEO-ONLY* shown in Figure 7(b) was declining much faster than the error distance of key feature *NON-GEO* shown in Figure 7(a) since the minor feature had less noisy and more persistent text data associated with the specific geolocations. Such behavior is expected in the case of the relatively high quality of metadata associated with the text which is generally a simpler problem than that of disambiguation within text documents in general [12].

The straightforward concatenation of the text content and metadata context of the 'place' field would require undesirable randomization of all words/parts 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 4 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. 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 $\sigma_{\widehat{c}}$ in the GMM output of the probabilistic models. Since the probabilistic loss function computed the negative log-likelihood of the labeled location to fit in the predicted Gaussian distributions, it depended on the Probability Density Function (PDF). In the case, $\sigma_{\widehat{c}}$ was approaching its minimum at 0, PDF exceeded 1 driving the loss function to the negative values as shown in Figure 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 $\sigma_{\widehat{c}}$ to $\frac{1}{2\pi}$ such that Probabilistic loss shown in Figure 7(c) stays in the positive domain. 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 C.

According to [11], the user-level geolocation is more effective in some cases, however, fine-grained analysis is possible only for the tweet-level predictions and thus the latter was chosen as the main subject of this study. Nonetheless, 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 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 transformers-based solution.

(a) KF Geospatial loss computed by Eq. (SED)

(b) MF Geospatial loss computed by Eq. (SED)

(c) Probabilistic loss computed by Eq. (NLLH)
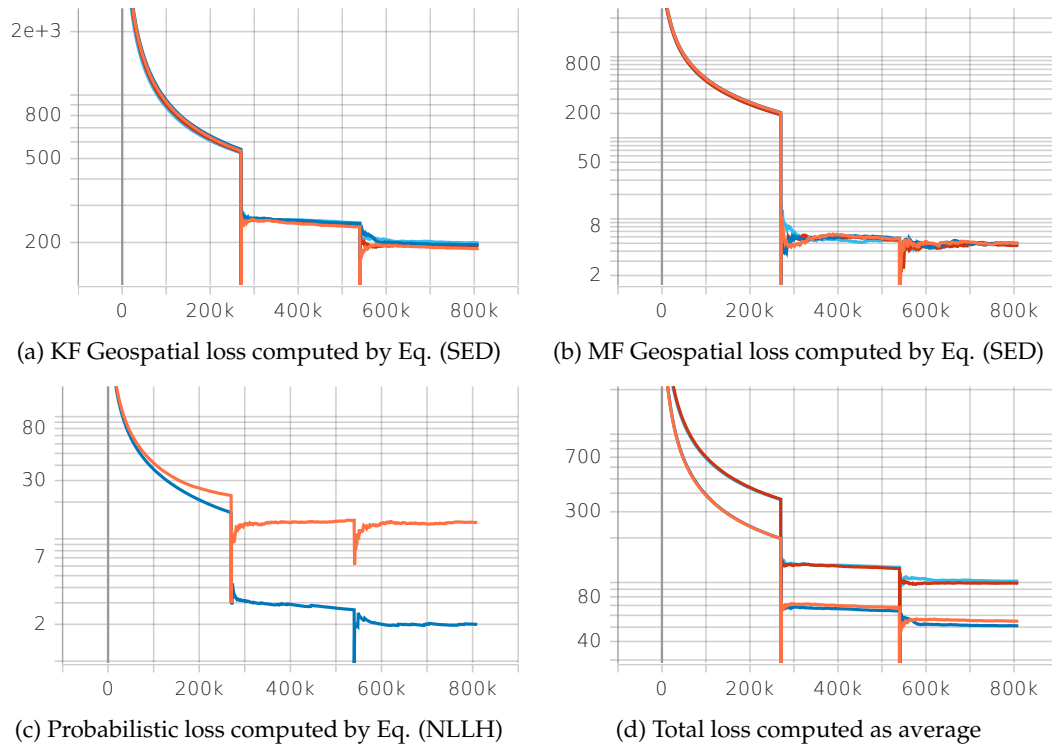
(d) Total loss computed as average

Figure 7: Loss logged during the training of the proposed models with single and multiple (5) outcomes and Geospatial and Probabilistic loss; KF–Key Feature, MF–Minor Feature; model graph lines by color: PSOP, PMOP, GSOP, and GMOP where abbreviations refer to the configuration descriptions provided in Table 3.

The prediction of tweets and users' home locations on Twitter raises significant privacy concerns. While it has useful applications, it poses risks to privacy. Issues include compromising location privacy, enabling user tracking, exposing personal safety risks, inferring sensitive information, and potential misuse of aggregated data.

After the training performed on our Twitter archive data, the proposed model was able to predict the location of any Twitter user by processing the publicly available information posted online. While profile information provided by Twitter users is generally considered public, the combination or analysis of such data can have privacy implications. Tweets, being public by default, can also be used for analysis and publication. However, it's worth noting that even seemingly innocuous information, when combined with other data points or analyzed in aggregate, can potentially reveal more about a person's identity or habits. Therefore, while individual profile details may not be inherently sensitive, their combination or analysis might raise privacy concerns.

To address these privacy concerns, it is crucial for researchers, to anonymize or aggregate the data to minimize the risk of re-identification. The data utilized in this study has been protected with strict measures to prevent unauthorized access or breaches. Thus any collected Twitter dataset examples, containing metadata of user-profiles and geo-tags ob-

tained from parsing the tweets, won't be shared due to the potential privacy issues. By balancing the potential benefits of geolocation prediction with the protection of users' privacy, it is possible to mitigate the privacy risks and promote responsible use of location data in Twitter research.

As for the real-time user location monitoring, Yuan in the work [39] focused on tracking the movement of individuals or groups over time. Monitoring user geolocation in time can provide valuable insights for multiple applications such as disease outbreak trackers, urban mobility pattern analyzers, and location-based provider 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 Plate carrÃl'e projection is commonly used due to the Twitter geolocation stored in the WGS84 format, other projections such as the Mercator projection, Robinson projection, conic projection, and Winkel-Tripel projection should be considered as possible alternatives in future studies.

Besides the BERT variations, the proposed machine learning techniques have the potential to be applied to other base models. In the realm of NLP, various state-of-the-art models have been developed to enhance text analysis tasks, particularly text classification that provides an output representation for the examined regression task. However, this would require adapting the implementation of the wrapper layer to accommodate the shape of the pooler output vector specific to the chosen model. Currently, the approach is designed for linear regression logic, transforming the vector of size 768, which is common to all BERT-based models, into a predefined number of continuous numerical values. Furthermore, in order to ensure suitability for the downstream task of sentence classification, it is important that the base models possess strong contextual understanding and the ability to capture semantic relationships within sentences.

While acknowledging the availability of newer models, we carefully considered the trade-offs and ultimately selected the BERT model, which was widely adopted and extensively studied for text classification tasks, to strike a balance between computational efficiency and our analytical requirements. Apart from the BERT, RoBERTa (Robustly Optimized BERT) [17], DistilBERT (Distilled BERT) [29], ALBERT (A Lite BERT) [14], ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) [6], XL-Net (eXtreme Language understanding Network) [37], and GPT-3 (Generative Pre-trained Transformer 3) [3] are prominent models in this field that can be finetuned to predict location from the text.

In the downstream task of text classification, all these models can be finetuned to perform well, although computational resource requirements differ among them. XLNet, BERT, and RoBERTa demand substantial resources due to their size and training needs resulting in a strong performance in text classification tasks. DistilBERT, ALBERT, and ELECTRA offer more efficient alternatives to BERT with similar performance and reduced resource usage. However, compared to the relatively lightweight BERT model, large language models like GPT are computationally more demanding and necessitate significant resources. Moreover, unlike GPT models, which are primarily used for text generation, the BERT model is well-suited for text analysis.

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 in geographical granularity setup of the custom BERT-based models.

5   Acknowledgments

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] BROWN, T., MANN, B., RYDER, N., SUBBIAH, M., KAPLAN, J. D., DHARIWAL, P., NEELAKANTAN, A., SHYAM, P., SASTRY, G., ASKELL, A., AGARWAL, S., HERBERT-VOSS, A., KRUEGER, G., HENIGHAN, T., CHILD, R., RAMESH, A., ZIEGLER, D. M., WU, J., WINTER, C., HESSE, C., CHEN, M., SIGLER, E., LITWIN, M., GRAY, S., CHESS, B., CLARK, J., BERNER, C., MCCANDLISH, S., RADFORD, A., SUTSKEVER, I., AND AMODEI, D. Language models are few-shot learners. *Advances in neural information processing systems 33* (2020), 1877–1901.

[4] CAO, B., CHEN, F., JOSHI, D., AND PHILIP, S. Y. Inferring crowd-sourced venues for tweets. In *2015 IEEE International Conference on Big Data (Big Data)* (2015), IEEE, pp. 639–648.

[5] CHENG, Z., CAVERLEE, J., AND LEE, K. You are where you tweet: a content-based approach to geo-locating Twitter users. In *Proc. the 19th ACM international conference on Information and knowledge management* (2010), pp. 759–768.

[6] CLARK, K., LUONG, M.-T., LE, Q. V., AND MANNING, C. D. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555* (2020).

[7] GALAL, A., AND ELKORANY, A. Enabling semantic user context to enhance Twitter location prediction. In *ICAART (1)* (2016), pp. 223–230.

[8] HAN, B., COOK, P., AND BALDWIN, T. Text-based Twitter user geolocation prediction. *Journal of Artificial Intelligence Research 49* (2014), 451–500.

[9] HUANG, B., AND CARLEY, K. M. A hierarchical location prediction neural network for Twitter user geolocation. *arXiv preprint arXiv:1910.12941* (2019).

[10] HULDEN, M., SILFVERBERG, M., AND FRANCOM, J. Kernel density estimation for text-based geolocation. In *Proc. the AAAI conference on artificial intelligence* (2015), vol. 29.

[11] ISO, H., WAKAMIYA, S., AND ARAMAKI, E. Density estimation for geolocation via convolutional mixture density network. *arXiv preprint arXiv:1705.02750* (2017).

[12] JONES, C. B., PURVES, R. S., LIU, L., AND ÖZSU, M. T. Geographic information retrieval. *Springer reference* (2009), 1227–1231.

[13] KINSELLA, S., MURDOCK, V., AND O'HARE, N. "i'm eating a sandwich in Glasgow" modeling locations with tweets. In *Proc. the 3rd international workshop on Search and mining user-generated contents* (2011), pp. 61–68.

[14] LAN, Z., CHEN, M., GOODMAN, S., GIMPEL, K., SHARMA, P., AND SORICUT, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).

[15] 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).

[16] LI, S., ZHANG, C., LEI, D., LI, J., AND HAN, J. Geoattn: Localization of social media messages via attentional memory network. In *Proc. the 2019 SIAM International Conference on Data Mining* (2019), SIAM, pp. 64–72.

[17] LIU, Y., OTT, M., GOYAL, N., DU, J., JOSHI, M., CHEN, D., LEVY, O., LEWIS, M., ZETTLEMOYER, L., AND STOYANOV, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[18] MAHMUD, J., NICHOLS, J., AND DREWS, C. Home location identification of Twitter users. *arXiv preprint arXiv:1403.2345* (2014).

[19] MCGEE, J., CAVERLEE, J. A., AND CHENG, Z. A geographic study of tie strength in social media. In *Proc. the 20th ACM international conference on Information and knowledge management* (2011), pp. 2333–2336.

[20] MCPHERSON, M., SMITH-LOVIN, L., AND COOK, J. M. Birds of a feather: Homophily in social networks. *Annual review of sociology 27*, 1 (2001), 415–444.

[21] MIURA, Y., TANIGUCHI, M., TANIGUCHI, T., AND OHKUMA, T. Unifying text, metadata, and user network representations with a neural network for geolocation prediction. In *Proc. the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2017), pp. 1260–1272.

[22] MIYAZAKI, T., RAHIMI, A., COHN, T., AND BALDWIN, T. Twitter geolocation using knowledge-based methods. In *Proc. the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text* (2018), pp. 7–16.

[23] PRIEDHORSKY, R., CULOTTA, A., AND DEL VALLE, S. Y. Inferring the origin locations of tweets with quantitative confidence. In *Proc. the 17th ACM conference on Computer supported cooperative work & social computing* (2014), pp. 1523–1536.

[24] 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).

[25] RAHIMI, A., COHN, T., AND BALDWIN, T. A neural model for user geolocation and lexical dialectology. *arXiv preprint arXiv:1704.04008* (2017).

[26] 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).

[27] ROLLER, S., SPERIOSU, M., RALLAPALLI, S., WING, B., AND BALDRIDGE, J. Supervised text-based geolocation using language models on an adaptive grid. In *Proc. the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (2012), pp. 1500–1510.

[28] SADILEK, A., KAUTZ, H., AND BIGHAM, J. P. Finding your friends and following them to where you are. In *Proc. the fifth ACM international conference on Web search and data mining* (2012), pp. 723–732.

[29] SANH, V., DEBUT, L., CHAUMOND, J., AND WOLF, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[30] SCHERRER, Y., AND LJUBEŠIĆ, N. Social media variety geolocation with geobert. In *Proc. the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects* (2021), The Association for Computational Linguistics.

[31] SCHULZ, A., HADJAKOS, A., PAULHEIM, H., NACHTWEY, J., AND MÜHLHÄUSER, M. A multi-indicator approach for geolocalization of tweets. In *Proc. the International AAAI Conference on Web and Social Media* (2013), vol. 7, pp. 573–582.

[32] 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.

[33] VILLEGAS, D. S., PREOȚIUC-PIETRO, D., AND ALETRAS, N. Point-of-interest type inference from social media text. *arXiv preprint arXiv:2009.14734* (2020).

[34] 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.

[35] WING, B., AND BALDRIDGE, J. Simple supervised document geolocation with geodesic grids. In *Proc. the 49th annual meeting of the association for computational linguistics: Human language technologies* (2011), pp. 955–964.

[36] WING, B., AND BALDRIDGE, J. Hierarchical discriminative classification for text-based geolocation. In *Proc. the 2014 conference on empirical methods in natural language processing (EMNLP)* (2014), pp. 336–348.

[37] YANG, Z., DAI, Z., YANG, Y., CARBONELL, J., SALAKHUTDINOV, R. R., AND LE, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems 32* (2019).

[38] 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.

[39] 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 *Proc. the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (2013), pp. 605–613.

[40] ZHAI, C., AND LAFFERTY, J. A study of smoothing methods for language models applied to ad hoc information retrieval. In *ACM SIGIR Forum* (2017), vol. 51, ACM New York, NY, USA, pp. 268–276.

[41] ZHANG, W., AND GELERNTER, J. Geocoding location expressions in Twitter messages: A preference learning method. *Journal of Spatial Information Science*, 9 (2014), 37–70.

[42] ZHENG, C., JIANG, J.-Y., ZHOU, Y., YOUNG, S. D., AND WANG, W. Social media user geolocation via hybrid attention. In *Proc. the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020), pp. 1641–1644.

[43] 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.

[44] 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.

[45] 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.

A   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 the 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 computed for each outcome.

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

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

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

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 computed 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 $\boldsymbol{\Sigma}$:

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

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:

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

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}.

B Performance metrics

B.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-type model metrics, the SAE is computed as the weighted linear combination of all M outcomes:

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

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 the predicted locations and the labeled 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.

B.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:

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

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

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

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 computed as the weighted linear combination of such area for each of M Gaussian peaks, which depends mainly on their covariance parameters and weights:

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

- 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:

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

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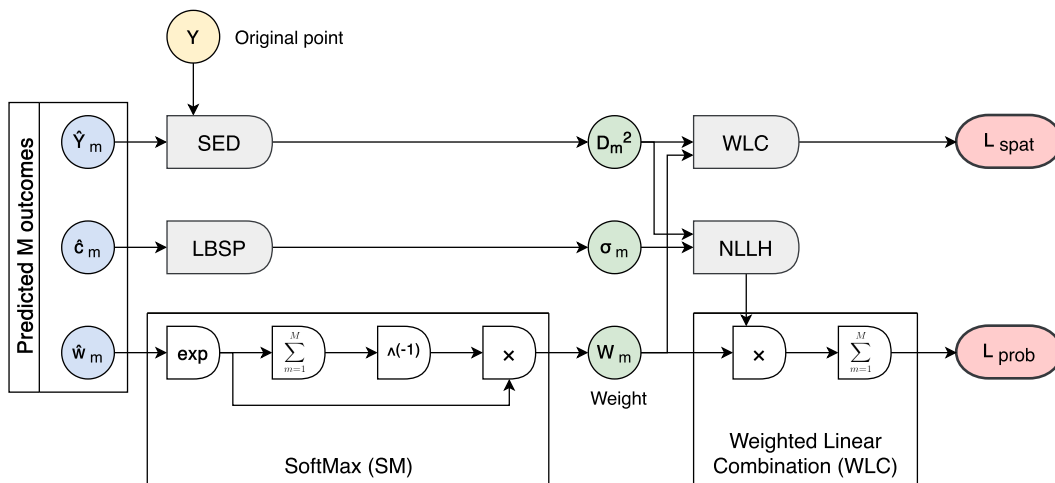
C Loss functions

Figure 8: Multiple Outcomes Prediction (MOP) loss functions computational graph including visualization of Weighted Linear Combination (WLC) and SoftMax (SM) components. $\widehat{Y}_m$–point, $\widehat{c}_m$–covariance parameter, and $\widehat{w}_m$–weight predicted for a single outcome m D_m^2–squared error distance, σ_m–adjusted covariance (confidence), and $\widehat{W}_m$–normalized weight (significance) for a single outcome m
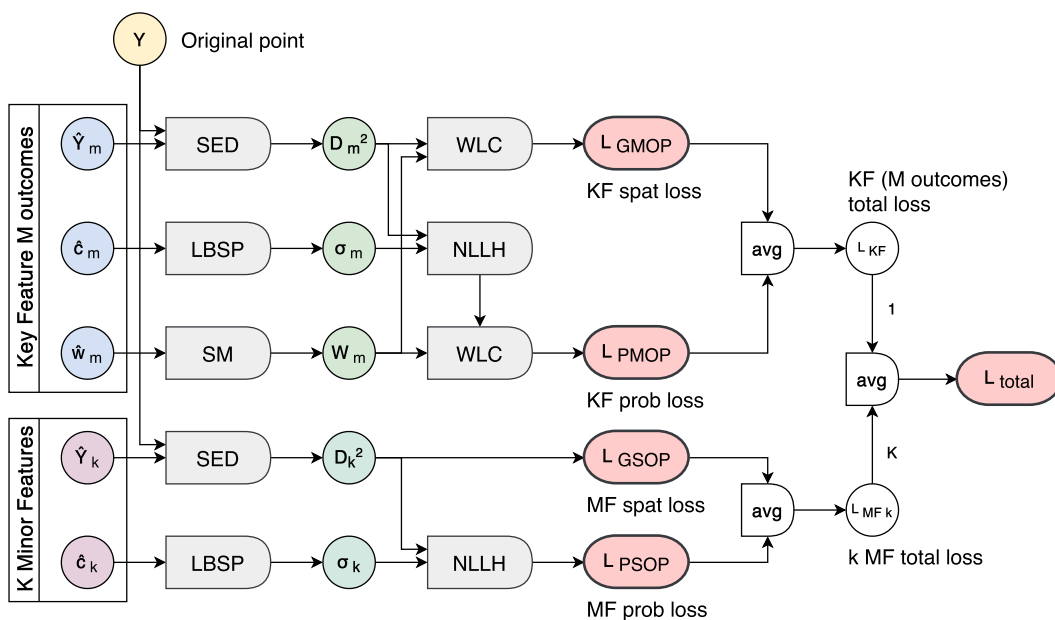

Figure 9: Total loss functions computational graph for the MOP-type Key Feature (KF) with M outcomes, and K Minor Features (MF) of the SOP-type.