Computer tools in education, 2024
№ 3: 48–65
http://cte.eltech.ru
doi:10.32603/2071-2340-2024-3-2

LEVERAGING LARGE LANGUAGE MODELS FOR TEXTUAL GEOTAGGING: A NOVEL APPROACH TO LOCATION INFERENCE

Sultanov A.[1], AI Engineer, ✉ azamat.sultanov@theping.co
[1]The Ping IT Inc., 1712 Pioneer Ave Ste 179 82001, Cheyenne, WY., USA

Abstract

This study explores the application of Large Language Models (LLMs), particularly GPT-4o, to textual geotagging, introducing a novel dataset of tweets with geographical annotations. Using zero-shot and few-shot approaches, we demonstrate GPT-4o's ability to infer location from explicit and implicit textual references in tweets, achieving average errors as low as 43 km for explicit mentions. Our experiments reveal LLMs' robust geographical knowledge and adaptability to geotagging tasks with minimal context. The research also highlights LLMs' potential in advancing geographical inference from text, identifying challenges and effects of data quality, and opportunities for improving model performance on implicit references and noisy data.

Keywords: *Large Language Model (LLM), GPT, Geotagging, Natural Language Processing (NLP), Artificial Intelligence.*

Citation: A. Sultanov, "Leveraging Large Language Models for Textual Geotagging: A Novel Approach to Location Inference," *Computer tools in education*, no. 3, pp. 48–65, 2024 (in Russian); doi:10.32603/2071-2340-2024-3-2

1. INTRODUCTION

Textual geotagging is the process of extracting geographical coordinates from textual information. This task has become necessary in the digital age due to the vast number of services that usexo geolocation data. Geotagging has multiple applications in social media analysis, emergency response systems, and targeted ads. Unfortunately, traditional approaches for textual geotagging, including rule-based systems, gazetteer lookups, statistical and probabilistic models, early machine learning algorithms, and even sophisticated language models, fail to capture geolocation information from inputs explicitly mentioning geographical places or depending on the context.

In recent years, we have seen massive progress in language models, leading to the development of large language models (LLMs) [1]. These advanced models have demonstrated remarkable capabilities across various natural language processing tasks [2], often surpassing traditional and earlier machine learning approaches. They are remarkably capable of reasoning [3], making them particularly promising for complex tasks like textual geotagging.

In this article, we present a novel approach to textual geotagging that leverages the power of LLMs trained on massive amounts of data to accurately infer geographical metadata like countries, cities, and streets, and even coordinates from the textual content, particularly in scenarios where explicit geographical indicators are sparse or ambiguous.

Our research makes several contributions to the field:

i. A novel dataset of tweets annotated with precise geographical coordinates, carefully curated to represent a wide range of location-related linguistic patterns;
ii. A new framework for prompting LLMs to extract and reason about geographical information from text;
iii. A comprehensive comparative analysis of multiple state-of-the-art LLMs' geotagging performance using zero-shot and few-shot learning approaches.

Through our experiments and analysis, we aim to demonstrate the potential of LLMs in advancing the field of textual geotagging and pave the way for more sophisticated location inference systems in real-world applications.

The remainder of this article is structured as follows:

Section 2 reviews related work in textual geotagging and the application of LLMs in natural language processing tasks.

Section 3 describes our novel dataset, detailing its creation process, characteristics, and potential applications in geotagging research.

Section 4 outlines our experimental methodology, focusing on zero and few-shot learning approaches.

Section 5 presents our results and discusses the performance of GPT-4o on the geotagging task.

Finally, we summarize our findings and suggest future research directions in this rapidly evolving field.

This study aims to contribute meaningfully to the growing landscape of AI-driven geographical inference from textual data by exploring the intersection of LLMs and geographical information systems.

2. RELATED WORK

By analyzing related work done in the current direction, we will review methods and approaches based on the pre-LLM era, which include the following methods:

i. heuristic, probabilistic, and statistical;
ii. classical machine learning;
iii. advanced natural language processing.

2.1. Heuristic, Probabilistic, and Statistical Methods

Researchers have developed various approaches to tackle the challenges posed by sparse and noisy data in geolocation inference from social media content. These methods can be broadly categorized into heuristic, probabilistic, and statistical techniques, each offering unique strengths in extracting location information.

Heuristic approaches leverage domain-specific knowledge and rules, often proving effective in scenarios where explicit location information is limited. These methods typically rely on carefully crafted algorithms that exploit patterns and structures within the data. The work in [12] examines approaches for estimating a user's location using microblog messages without geotags. The researchers found that training models for each user individually offered advantages in both precision and recall compared to other approaches, highlighting the potential of personalized heuristics. In [4], authors developed a text-based heuristic schema for geolocation inference on Reddit, analyzing user comments. Their approach generates ground truth location labels

with a precision of 0.966. It employs a multi-modal inference model, achieving median errors of 157 miles for US users and 266 miles for international users. Researchers in [5] propose a scalable geoparsing and geotagging approach to serve local news worldwide. They use an ensemble method to determine article location and impact radius and develop techniques to reconcile user location with article location for personalized local news delivery. The study in [13] explores the geo-parsing of disaster-related tweets, manually annotating locations to create a gold standard. They found that off-the-shelf Named Entity Recognition software struggled with informal location references in microtext, highlighting the need for specialized approaches for social media content.

On the other hand, probabilistic methods model the inherent uncertainties in location data, allowing for more nuanced predictions. These approaches often leverage statistical techniques to estimate the likelihood of a user's location based on various features. In [6], researchers use kernel density estimation, hierarchical clustering, and other spatial analytics to build dynamic ontological models of place from Twitter and Weibo data. They identify feature types of place name ontologies and observe seasonal variation patterns in non-administrative places, demonstrating the potential of probabilistic models in capturing the temporal dynamics of locations. The study in [7] presents a probabilistic framework for estimating Twitter users' city-level locations based solely on tweet content. Their approach uses a classifier to identify words with strong local geo-scope and a lattice-based neighborhood smoothing model, placing 51 % of users within 100 miles of their actual location.

Similarly, the work in [10] proposes a probabilistic framework for estimating a Twitter user's city-level location based purely on tweet content. Their approach includes a classification component for identifying words with strong local geo-scope and a lattice-based neighborhood smoothing model. It also places 51 % of Twitter users within 100 miles of their location.

Statistical techniques leverage large-scale data analysis to infer location patterns, often employing sophisticated methods to discover relationships between textual content and geographic information. The research in [8] proposes a language modeling method to predict the origin of tweets' points of interest (POIs). They use web-enriched models to boost performance for POIs with insufficient tweets and find that time models consistently improve results despite data sparsity. In [9], researchers develop a probabilistic framework to estimate city-level Twitter user locations using content from tweet dialogues. Their baseline estimation using reply-tweet information yields accuracy higher than previous approaches, showcasing the potential of conversational context in location inference. The study in [11] presents a two-stage approach called TS-Petar for extracting fine-grained locations with temporal awareness from tweets. They use a POI inventory built from Foursquare data and a time-aware POI tagger based on Conditional Random Fields, achieving promising performance against baseline methods.

As the field progresses, researchers increasingly explore hybrid and novel approaches that combine elements from different methodologies. The authors of [14] propose a framework to infer a user's primary location on Twitter using textual content. Their probabilistic generative model filters local words, employs data binning for scalability, and applies map projection techniques, identifying 60 % of Korean Twitter users within 10 km of their actual locations. In [15], researchers present a multi-level generative model that jointly explains latent topics and geographical regions in geotagged microblogs. Their model recovers coherent topics and their regional variants while identifying geographic areas of linguistic consistency, demonstrating the potential of combining topic modeling with geolocation inference. The study in [16] introduces a hierarchical ensemble algorithm for predicting Twitter users' home locations at different granularities. Their approach combines statistical and heuristic classifiers and outperforms previous

algorithms for predicting user locations, showcasing the benefits of ensemble methods in this domain.

Some researchers have focused on event detection and analysis, leveraging geolocation techniques to enhance real-time monitoring and response systems. The work in [17] investigates real-time event detection on Twitter, mainly for earthquakes. They develop a classifier for tweets and a probabilistic spatiotemporal model, treating each Twitter user as a sensor and applying Kalman and particle filtering for location estimation. In [18], researchers propose TEDAS, a Twitter-based Event Detection and Analysis System. The system detects new events, analyzes their spatial and temporal patterns, and identifies their importance using efficient crawling, classification, and ranking of tweets.

Innovative approaches have also emerged that leverage knowledge bases and geometric techniques. The study in [19] proposes a simple method to predict salient locations from news article text using a knowledge base. Their approach uses a dictionary of locations created from the knowledge base and hierarchical information between entities, improving f-measure by over 0.12 compared to multiple baselines. In [20], researchers present a novel geometric approach for geotagging web documents. Their three-step process considers all place names together without individual disambiguation, achieving correct continent-level focus for 97.07 % of Wikipedia pages and country-level focus for 95.57 %.

While these approaches have made significant strides in geolocation inference from social media and web content, they still face several challenges. They must address the sparsity and noise in social media data, the prevalence of non-standard language and abbreviations, and the dynamic nature of user locations. Moreover, many methods need help with scalability when applied to large-scale datasets, and privacy concerns often limit access to crucial user information. The trade-off between precision and recall remains a persistent issue, with improvements in one frequently coming at the cost of the other.

2.2. Traditional Machine Learning Approaches

Classical machine learning approaches have been widely applied to the challenge of geolocation inference from social media content, offering a balance between interpretability and predictive power. These methods typically rely on feature engineering and established algorithms to extract location information from textual data. The study in [21] employs linguistic analysis of social media posts to map American cultural regions. Using frequency distributions of content words in geotagged tweets, the authors derive principal components of regional variation and apply hierarchical clustering to identify clear cultural areas. This approach demonstrates how machine learning techniques can uncover nuanced cultural patterns beyond traditional demographic factors. In [22], researchers compare gazetteer-based and neural approaches for geotagging a diachronic corpus of alpine texts.

While the gazetteer-based method achieved high precision, a neural model using contextual string embeddings significantly outperformed toponym recognition when augmented with crowdsourced annotations. This work highlights the potential of combining traditional and modern machine learning techniques for improved geolocation inference. The authors of [23] investigate user behavior regarding the location field in Twitter profiles. They found that many users provide unreliable location information, but a simple classifier could still predict users' country and state with decent accuracy based solely on tweet content. This study underscores the importance of implicit location information in social media posts. In [24], researchers explore probability models for predicting Twitter users' home locations. They propose novel unsu-

pervised methods based on Non-Localness and Geometric-Localness to prune noisy data from tweets. Using Gaussian Mixture Models and a limited set of local words, their approach achieves comparable results to supervised state-of-the-art methods, demonstrating the potential of unsupervised learning in this domain. The work in [25] compares machine learning algorithms, including Naive Bayes, Support Vector Machines, and Decision Trees, for predicting user locations from tweet text. Their analysis suggests that Decision Trees are particularly well-suited for tweet text analysis and location prediction, highlighting the importance of algorithm selection in this task.

Similarly, the study in [26] applies Logistic Regression, Random Forest, Multinomial Naïve Bayes, and Support Vector Machine to predict tweet locations, focusing on Arabic tweets from Saudi Arabia. By incorporating a geo-distance matrix, they achieve promising results with 67 % accuracy, showcasing the potential of machine learning approaches for non-English content. The research in [27] also compares Naive Bayes, Support Vector Machine, and Decision Tree algorithms for predicting user locations from tweet text. Their experiments corroborate the findings in [25], concluding that Decision Trees are particularly effective for tweet text analysis and location prediction.

While these classical machine learning approaches have shown considerable success in geolocation inference, they face several limitations. The reliance on hand-crafted features can be time-consuming and may only capture some relevant patterns in the data. Additionally, these methods often struggle with the informal and noisy nature of social media text, as well as the sparsity of explicit location information. The performance of these algorithms can also be sensitive to the choice of features and hyperparameters, potentially limiting their generalizability across different datasets or languages. Despite these challenges, classical machine learning techniques provide valuable insights into the relationship between textual content and geographic information in social media data.

2.3. Advanced Natural Language Processing Models

Advanced natural language processing (NLP) methods have significantly improved the accuracy and capabilities of geolocation inference from social media content. These approaches encompass a wide range of techniques, from traditional machine learning to cutting-edge deep learning models, each offering unique advantages in tackling the challenges of location prediction from text. Conventional neural network approaches laid the foundation for geolocation inference from social media content. The work in [28] presents GeoTextTagger, a high-precision location tagging system that combines named entity recognition with a knowledge base (OpenStreetMap) to identify and disambiguate location mentions. This hybrid approach achieved impressive accuracy, with 50 % of articles assigned at least one tag within 8.5 kilometers of the actual location. The system's ability to handle explicit and implicit location references makes it particularly effective for processing various textual documents. Building on this, research in [29] uses language models to create fine-grained representations of locations based on tweet content. This approach outperformed industry-standard tools, particularly at the hyper-local level, achieving a three- to ten-fold increase in accuracy at the zip code level. The researchers modeled locations at varying levels of granularity, from zip code to country level, demonstrating the potential of language models to capture subtle linguistic cues indicative of location.

As the field progressed, researchers began exploring more sophisticated neural network architectures. The study in [30] presents a neural regression model using BiLSTM for tweet geolocation, demonstrating the potential of recurrent neural networks for capturing dialect and linguis-

tic markers in tweets. This approach is particularly noteworthy for its ability to work without relying on pre-trained models or extensive text preprocessing, making it adaptable to social media text's informal and noisy nature. The model identifies the linguistic intricacies of a tweet to predict the user's location, showing promise in handling the diverse dialects and linguistic styles on social media platforms. Similarly, the work in [31] investigates various approaches for text-based Twitter user geolocation prediction, comparing feature sets and classification methods. This comprehensive study explored the impact of non-geotagged data, the influence of language, and the complementary geographical information in user metadata. The researchers found that explicit selection of location-indicative words improves geolocation prediction accuracy and that modeling on geotagged data and inferencing on non-geotagged data is feasible. This work provides valuable insights into the factors affecting geolocation prediction accuracy and offers practical guidelines for developing robust prediction models.

The advent of transformer architectures, including BERT, marked a significant leap forward in NLP capabilities, and geolocation inference has significantly benefited from these advancements. In [32], researchers developed a deep learning model using BERT for city-level geolocation of tweets, achieving a median error of less than 30 km on a worldwide dataset. This study demonstrates the power of transformer-based models in capturing complex linguistic patterns and contextual information relevant to location prediction. The researchers fine-tuned BERT on Twitter data and incorporated tweet content and metadata, showcasing the model's ability to handle the unique characteristics of social media text. Building on this, the study in [33] utilizes the BERT language model to predict the location of Indonesian Twitter users, achieving an accuracy of 0.77 by concatenating display names, descriptions, and aggregated tweets. This work is particularly noteworthy for its application to non-English content, demonstrating the versatility of transformer-based models across different languages. The researchers' approach of combining multiple user attributes (display name, description, and tweets) provides a more comprehensive view of the user's location, capturing information that might be missed when considering these attributes in isolation.

Further refining the use of transformers, the work in [34] proposes a deep learning model incorporating multi-head self-attention, subword features, and joint training with country labels for tweet location prediction. This approach demonstrates competitive performance on the W-NUT geo-tagging task, showcasing the potential of advanced transformer architectures in handling the complexities of location prediction from short, informal text. The use of subword features is particularly innovative, allowing the model to handle out-of-vocabulary words and informal language common in tweets.

Named Entity Recognition (NER) has played a crucial role in many geolocation inference systems. The research in [35] developed T-NER, a novel system that leverages the redundancy in tweets and uses LabeledLDA with Freebase dictionaries for distant supervision. This approach doubled the F1 score compared to the Stanford NER system, demonstrating the potential of combining machine learning techniques with knowledge bases for improved entity recognition in social media text. The system's ability to handle noisy and informal tweets makes it particularly valuable for location extraction tasks. Building on this, the study in [36] investigates the effectiveness of NER tools for extracting locations from disaster-related microblogs. The researchers found that retraining NER tools on tweet data significantly improved performance, with Stanford NER achieving an F-Measure of over 0.9 on a dataset of disaster-related tweets. This work highlights the importance of domain-specific training data in improving the performance of NER tools for specialized tasks like disaster response. The work in [37] further advances NER techniques by proposing the Leave no Place Behind approach, which fine-tunes popular NER tools

like Spacy and RoBERTa on humanitarian texts. This method improves performance and alleviates bias towards Western countries in existing tools, achieving an F1 score of up to 0.92. The researchers' focus on humanitarian documents addresses an essential niche in geolocation inference, potentially improving the effectiveness of aid distribution and crisis response efforts.

Several studies have focused on combining multiple techniques or developing specialized approaches to address the unique challenges of geolocation inference from social media. In [38], researchers propose the True Origin Model, which uses machine-level natural language understanding to identify tweets containing origin location information. This approach achieved promising accuracy at various geographic levels, from country (80 %) to district (64 %), demonstrating the potential of advanced NLP techniques in distinguishing between the mentioned locations and the actual origin of a tweet. The study in [39] adapts and improves a deep learning model (deepgeo2) for city-level geolocation prediction of tweets, integrating it with a visual analytics system for real-time situational awareness. This work showcases the potential of combining advanced NLP techniques with interactive visualization, providing a powerful tool for analysts and decision-makers who need to quickly understand the geographic distribution of social media activity. The work in [40] presents an enhanced geocoding precision method for location inference from tweet text, combining spaCy for NER, Nominatim for geocoding, and Google Maps for validation. This multi-step approach achieved high precision in location inference, with 61,9 % of extracted locations inferred within a 1 km radius. The researchers' use of multiple tools and validation steps demonstrates the potential for improving location inference accuracy through careful system design and integration of complementary techniques.

Researchers have also explored grid-based and word-embedding approaches to address specific challenges in geolocation inference. The study in [41] presents LOCINFER, a non-uniform grid-based approach using Quadtree spatial partitions for location inference. This method addresses the sparsity problem in training data and outperforms state-of-the-art grid-based methods, predicting 60 % of tweets accurately within a 161 km radius. Using non-uniform grids allows the system to adapt to varying densities of geotagged data across different regions, potentially improving performance in areas with sparse training data. In [42], researchers use word embeddings and deep learning models to track Coronavirus-related tweets, demonstrating the ability to capture geosemantics of non-local words and delimit the sparse use of local ones. This work showcases the potential of a new framework called DeepGeoloc and advanced NLP techniques in tracking real-time events and understanding their geographic spread through social media analysis. The researchers' focus on a specific topic demonstrates how these techniques can be applied to analyze public discourse and information spread during critical events.

As these techniques become more advanced, privacy concerns have also come to the forefront of research in this area. The authors of [43] explore methods to protect user privacy by deceiving stance detection and geotagging models, highlighting the vulnerability of these advanced NLP systems to simple text modifications. This work raises critical ethical considerations about using geolocation inference techniques and the need for robust privacy protections in social media analytics.

These advanced NLP approaches have significantly improved the accuracy and robustness of geolocation inference from social media content. However, challenges remain, including handling informal language, addressing privacy concerns, and generalizing models across different languages and geographic regions.

3. DATASET

3.1. Dataset Objectives and Rationale

The primary objective of our dataset creation was to establish a comprehensive and diverse collection of tweets containing explicit and implicit geographical information. This dataset serves as a crucial resource for advancing research in textual geotagging, particularly in scenarios where geographical indicators are sparse or ambiguous. By including tweets with implicit location references, we aim to push the boundaries of current geotagging techniques and explore the potential of Large Language Models (LLMs) in inferring geographical metadata from contextual clues.

3.2. Data Source and Initial Processing

Our dataset is based on the "English Tweets of 2022" collection available on Kaggle[1], which initially contained over 500,000 tweets posted between January 1st and December 31st, 2022. This source dataset was chosen for its temporal diversity, ensuring a balanced representation across dates, weekdays, and months throughout the year. Due to constraints on LLM API usage in our self-funded research, we narrowed our focus to a subset of 100,000 tweets from this initial collection.

3.3. Dataset Creation Methodology

The creation of our specialized geotagging dataset involved a three-stage process:

i. Tweet Extraction and Filtering:

We employed a large language model as an initial filter to identify tweets containing geographical information. The LLM was prompted with specific instructions to extract tweets that explicitly mentioned geographical entities or implicitly suggested locations. This step was crucial in reducing the dataset to tweets relevant to geotagging tasks.

LLM Prompt Design

The prompt instructed the model to act as an expert in extracting geolocation information, with clear guidelines for identifying explicit and implicit geographical references. The criteria for implicit geolocation detection included factors such as:

- Local vernacular, dialects, or slang.
- References to local events, festivals, or sports.
- Specific weather patterns characteristic of certain areas.
- Mentions of local landmarks or attractions.
- References to local cuisine.
- Mentions of local transportation systems.
- Time zone indicators.
- References to local sports teams.
- Local political references.
- Location-specific hashtags.
- Mentions of region-specific ecosystems or wildlife.
- References to local cultural events or practices.
- Mentions of region-specific businesses or chains.

[1] https://www.kaggle.com/datasets/amirhosseinnaghshzan/twitter-2022

- References to local educational institutions.
- Local historical references.
- Use of specific currencies.
- Use of particular measurement systems (metric vs. imperial).
- Mixing of local languages with English.
- References to specific topographical features.
- Mentions of local media outlets.

The LLM was instructed to output results in a structured JSON format, including the tweet text, identified geographical entities (country, state, city, street), and the type of geolocation reference (explicit or implicit).

ii. Geographical Entity Extraction

For each tweet identified as containing geographical information, the LLM extracted relevant geographical entities when possible. Entites included country, state, city, and street information where available. The LLM's role in this step was to provide a high-level analysis and entity extraction rather than performing the final geotagging.

iii. Filtering and Validation

Explicit mentions

We applied a substring matching filter for tweets identified as explicitly mentioning geographical information. We retained only those tweets where at least one of the extracted entities (country, state, city, or street) appeared as a substring within the tweet text. This step significantly improved the precision of our explicit geolocation subset.

Implicit mentions

We noticed that most tweets labeled by LLM as implicitly mentioning the geographical information are noisy and have a high rate of false positives. To address this, we manually curated the implicit geolocation subset. We carefully selected 50 tweets that, with a high degree of confidence, implied geographical information. This manual curation ensures the quality and relevance of our implicit geolocation examples.

iv. Coordinate Mapping

To ensure the accuracy and independence of our dataset from LLM-based geotagging, we employed Nominatim, an open-source geocoding tool based on OpenStreetMap data, to convert the extracted geographical entities into precise latitude and longitude coordinates. This step involved concatenating the available geographical entities for each tweet and passing them as arguments to Nominatim's geocode function.

3.4. Dataset Characteristics

The resulting dataset exhibits the following key characteristics:

i. Size: out of the 100,000 processed tweets, 2,713 were identified as containing relevant geographical information.
ii. Geolocation types: 1,279 tweets with explicit mentions and 50 tweets with implicit references.
iii. Geographical entity distribution: 1,308 tweets were mentioned with countries, 903 with states, 959 with cities, and 67 with streets.
iv. Coordinate precision: the dataset includes latitude and longitude coordinates for each tweet, with varying levels of precision based on the specificity of the geographical entities extracted.

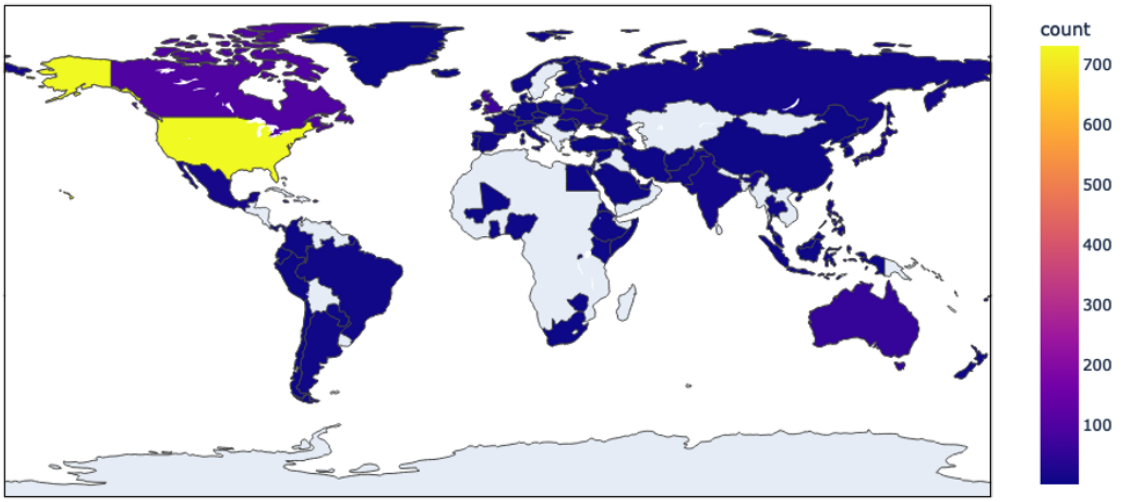

Figure 1. Global distribution of geotagged tweets in the dataset

Figure 1 offers insights into our collection's spatial patterns and concentrations of geotagged content. The color intensity represents the number of tweets associated with each country, with brighter shades indicating a higher concentration of tweet.

3.5. Dataset Examples

Table 1. Sample tweets explicitly mentioning and implicitly referring to geographical locations

Tweet	Type	Extracted entities	Lat, Lon
i miss bergen and i miss my friends	E	Norway, Bergen	60.3943055, 5.3259192
Mid off in new york rn	E	USA, New York, New York	40.7127281, – 74.0060152
Going to a galaxy game. It's been years lmao	I	USA, California, Los Angeles	34.0536909, – 118.242766
Our view for tonight's game. Let's gooooo Tigers 💜🧡	I	USA, Detroit, Michigan	42.338356, – 83.048134

Note on implicit samples

We must acknowledge a degree of inherent uncertainty for tweets that contain implicit geographical references. Unlike explicit mentions, where geographical entities are directly stated, implicit references rely on contextual clues, cultural knowledge, and nuanced interpretation. Consequently, there is no guarantee that the inferred location is correct.

3.6. Dataset Significance and Contributions

This dataset makes several significant contributions to the field of textual geotagging:

1. Implicit geolocation focus: our dataset addresses a gap in existing geotagging resources by including tweets with implicit geographical references, enabling research into more nuanced location inference techniques.
2. Diverse geographical entities: including various levels of geographical specificity (from country to street level) allows for evaluating geotagging models across different scales of location precision.

3. Real-world complexity: derived from actual Twitter data, the dataset captures the natural complexity and variability of how people reference locations in casual online communication.
4. Benchmark potential: using an independent geocoding service (Nominatim) for coordinate mapping ensures that the dataset can be an unbiased benchmark for evaluating LLM-based geotagging techniques.
5. Temporal diversity: The dataset spans an entire year, allowing for studying potential seasonal or temporal variations in location references.

3.7. Dataset availability

Our dataset is publicly available on GitHub[2].

4. EXPERIMENTS

4.1. Models and Scenarios

We utilize GPT-4o, developed by OpenAI and widely recognized as the most advanced and capable large language model, for our geotagging experiments. GPT-4o's exceptional reasoning abilities, vast knowledge base, and superior performance across various tasks make it an ideal candidate for evaluating cutting-edge natural language processing capabilities. We test GPT-4o under zero and few-shot learning scenarios to assess its geotagging performance comprehensively. This approach allows us to explore the model's geographical knowledge and ability to adapt to specific geotagging tasks with minimal additional context.

4.2. Dataset Preparation

The dataset is split into 90 % for testing and 10 % for few-shot examples. Zero-shot and few-shot evaluations use the same 90 % test set for fair comparison.

4.3. Task Description

The LLM predicts latitude and longitude coordinates based on each tweet's content. In the few-shot scenario, models receive examples from the 10 % few-shot set before tackling the test tweets.

4.4. Evaluation Metric

We use the Haversine distance as our primary evaluation metric. This formula calculates the great-circle distance between two points on a sphere given their latitude and longitude coordinates, making it ideal for measuring the accuracy of geographical predictions:

$$d = 2R \cdot \operatorname{asin}\left(\sqrt{\sin^2\left(\frac{\Delta \text{lat}}{2}\right) + \cos(\text{lat}_1) \cdot \cos(\text{lat}_2) + \sin^2\left(\frac{\Delta \text{long}}{2}\right)}\right). \tag{1}$$

The average error $\bar{E}$ is calculated as:

$$\bar{E} = \frac{1}{n}\sum_{i=1}^{n} d_i. \tag{2}$$

Where n is the number of tweets in the test set.

[2] https://github.com/sultanovazamat/annotated-geo-tweet

4.5. Rationale for Methodology

Our choice of zero-shot and few-shot learning scenarios, rather than fine-tuning, is motivated by several factors:

i. Real-world applicability: zero-shot and few-shot scenarios better reflect real-world situations where extensive labeled data for fine-tuning may not be available.
ii. Model generalization: these approaches test the models' ability to generalize knowledge across domains, a crucial aspect of AI's practical utility.
iii. Efficiency: zero-shot and few-shot methods require significantly less computational resources and data preparation than fine-tuning.
iv. Baseline performance: these methods establish a baseline for the models' inherent capabilities, providing valuable insights for future improvements.
v. Adaptability assessment: we can evaluate how quickly the models adapt to the task with minimal examples by testing both zero-shot and few-shot scenarios.

5. RESULTS

5.1. Zero-shot approach

Our zero-shot geotagging experiments with the first model reveal promising results, mainly when accounting for data quality issues. The model achieved an average error of 465 km for tweets with explicit geolocation mentions, while the average error was 639 km for tweets with implicit references. However, these figures are significantly influenced by noisy data from two primary sources: inaccuracies in the ground truth coordinates provided by Nominatim based on OpenStreetMap and knowledge gaps in the LLM, leading to overly generic extraction for specific geographical entities. When excluding these noisy samples, defined as those with distances exceeding 500 km, the model's performance improves dramatically. The average error for explicit mentions reduces to 44 km; for implicit references, it decreases to 28 km. Notably, the proportion of noisy data is relatively small for explicitly mentioning tweets: 135 out of 1151, and relatively big for implicitly referencing tweets: 10 out of 45. For the last ones, this is one of the reasons why the average error is smaller than for the explicitly mentioning tweets, along with the quality of data curated manually. The significant improvement in performance after noise removal suggests that the model's capabilities could be further enhanced through fine-tuning to address knowledge gaps.

Table 2 and Figures 2–3 provide a detailed breakdown of these results, illustrating the overall performance and the distribution of prediction errors with and without noisy data.

Table 2. Evaluation results for the zero-shot approach

Explicit mentioning		**Implicit reference**	
with noisy data	without noisy data	with noisy data	without noisy data
465 km	8 km	639 km	28 km

The zero-shot approach demonstrates remarkable potential for geotagging tasks using large language models. The distribution of prediction errors reveals that most predictions achieve high accuracy, clustering near zero kilometers of error. This is particularly impressive given the model's lack of task-specific training. While a secondary cluster of less accurate predictions and some outliers exists, these likely correspond to more challenging cases or noisy data points identified earlier.

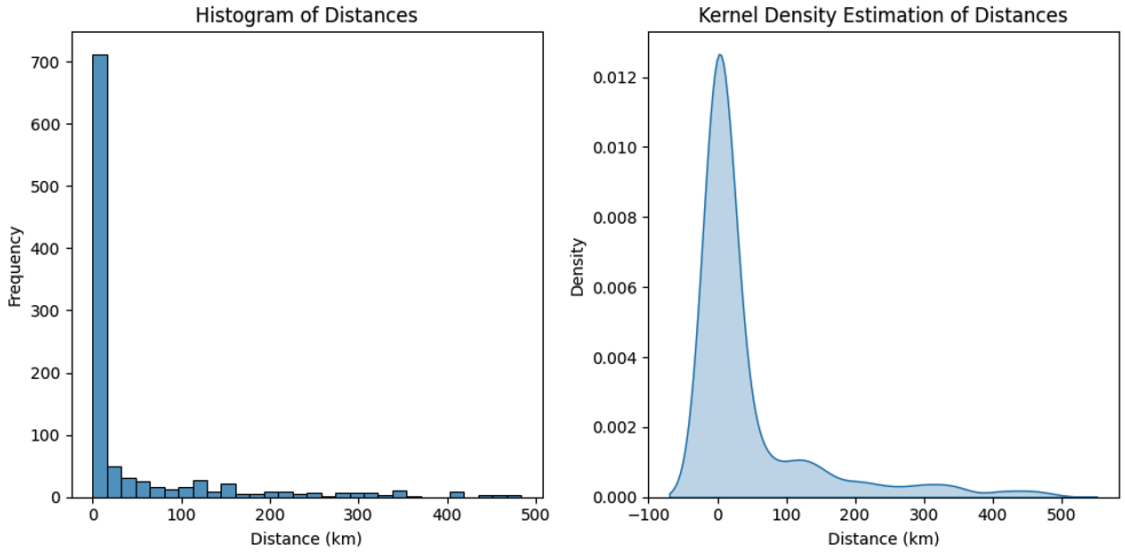

Figure 2. Histogram and kernel density estimation of distances for zero-shot approach for explicit mentioning without noisy data

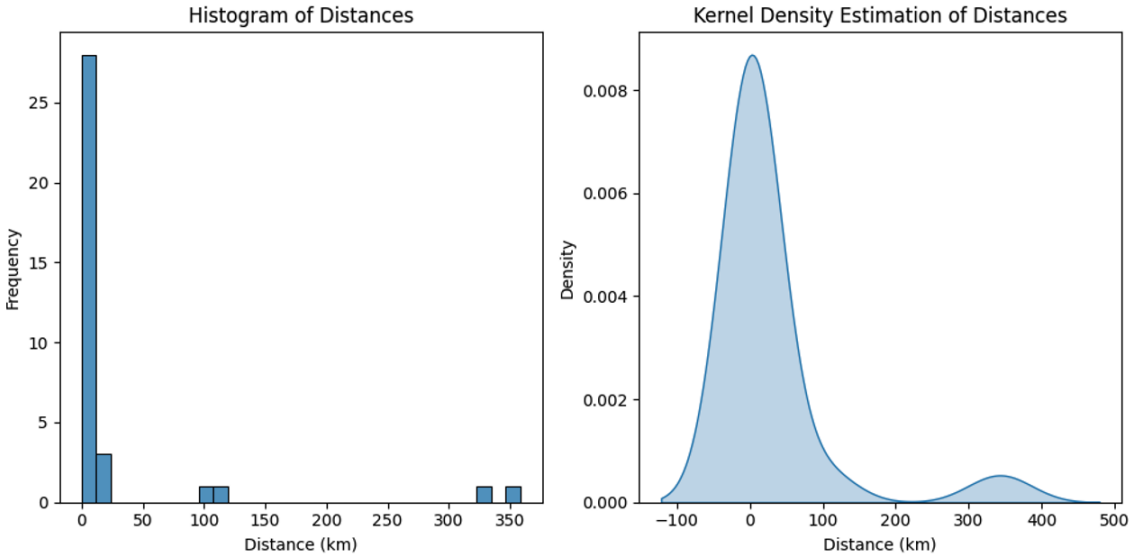

Figure 3. Histogram and kernel density estimation of distances for zero-shot approach for implicit references without noisy data

5.2. Few-shot approach

The few-shot geotagging experiments yielded promising results, showing improvements over the zero-shot approach, particularly for explicit mentions. For tweets with explicit geolocation references, the model achieved an average error of 471 km with noisy data included, and this was reduced to 43 km when excluding noisy samples. This represents a slight improvement over the zero-shot approach, which had a 44 km error rate for clean data, demonstrating the potential benefits of providing the model with a few examples.

For implicit references, the average error remained at 639 km with noisy data but significantly improved to 28 km without noise and matched the zero-shot approach. Matching the zero-shot performance can be attributed to the small tweets dump with implicit references to geolocations. Results are presented in Table 3 and Figure 4.

Table 3. Evaluation results for the few-shot approach

Explicit mentioning		**Implicit reference**	
with noisy data	without noisy data	with noisy data	without noisy data
471 km	43 km	639 km	28 km

These results highlight the LLM's strong baseline geotagging capabilities and ability to leverage few-shot examples, especially for explicit mentions effectively.

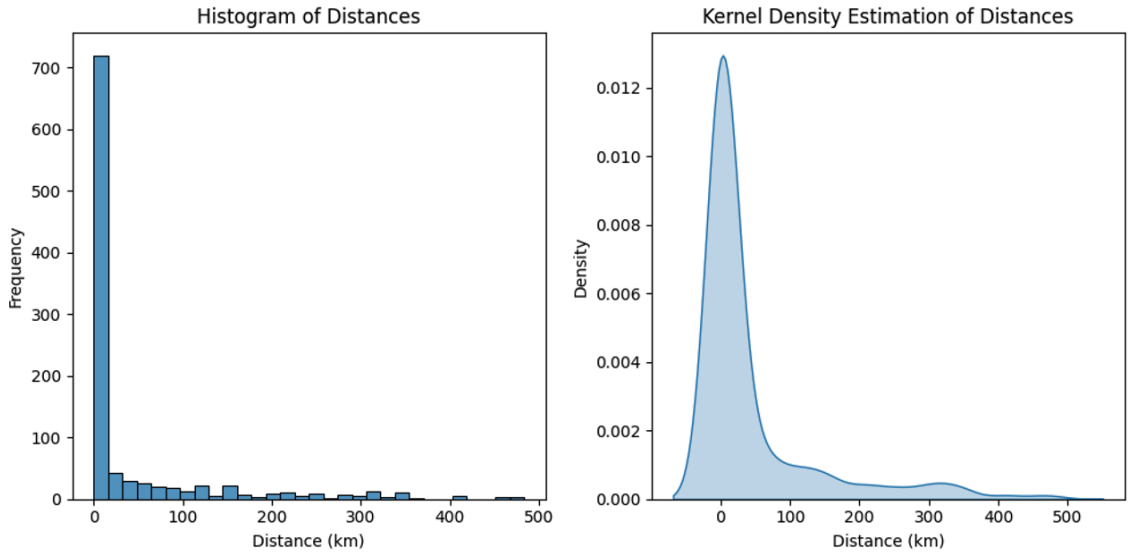

Figure 4. Histogram and kernel density estimation of distances for few-shot approach for explicit mentioning without noisy data

6. CONCLUSION

This study has explored the potential of leveraging large language models, specifically GPT-4o, for textual geotagging. Our research demonstrates the remarkable capabilities of state-of-the-art LLM in inferring geographical information from explicit and implicit textual references, even without task-specific training.

Our experiments revealed impressive baseline performance in the zero-shot scenario, with GPT-4o achieving high accuracy for explicit and implicit geographical references. After excluding noisy data, the model showed average errors of just 44 km for explicit mentions and 28 km for implicit references. The few-shot approach improved these results, particularly for explicit mentions, reducing the average error to 43 km. This improvement demonstrates the model's ability to adapt and enhance performance with minimal additional context quickly.

An essential contribution of our work is the creation of a novel dataset for geotagging tasks, including explicitly and implicitly location-referenced tweets. While we encountered challenges

with noisy data, primarily due to limitations in the LLM's knowledge and inaccuracies in the Nominatim mapping process, these issues present opportunities for future improvements in LLM capabilities and geolocation mapping techniques. Future research directions include expanding datasets for implicit geolocation references, optimizing few-shot example selection, addressing knowledge gaps, and improving performance on noisy data. As large language models evolve, we anticipate further advancements in their ability to understand and infer geographical information from text, opening new possibilities for location-based technologies and applications.

References

1. T. B. Brown, B. Mann, N. Ryder, M. Subbiah, et al., "Language Models are Few-Shot Learners," in *Proc. of 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, pp. 1–75, 2020.
2. V. Sanh, A. Webson, C. Raffel, S. H. Bach, "Multitask Prompted Training Enables Zero-Shot Task Generalization," in *Proc. ICLR 2022 Conference*, pp. 1–161, 2022.
3. J. Huang and K. C. Chang, "Towards Reasoning in Large Language Models: A Survey," in *Proc. Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1049–1065, 2023.
4. K. Harrigian, "Geocoding Without Geotags: A Text-based Approach for reddit," in *Proc. of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pp. 17–27, 2018.
5. D. S. Shah, G. K. Siddiqi, S. He, and R. Bansal, "Local Life: Stay Informed Around You, A Scalable Geoparsing and Geotagging Approach to Serve Local News Worldwide," in *arXiv*, 2023. [Online]. Available: https://arxiv.org/abs/2305.07168
6. M.-H. Tsou, Q. Zhang, J. Xu, A. Nara, and M. Gawron, "Building Dynamic Ontological Models for Place using Social Media Data from Twitter and Sina Weibo," in *arXiv*, 2023. [Online]. Available: https://arxiv.org/abs/2303.00877
7. R. Priedhorsky, A. Culotta, and S. Y. Del Valle, "Inferring the Origin Locations of Tweets with Quantitative Confidence," in *Proc. of CSCW '14: Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pp. 1523-1536, 2014.
8. W. Li, P. Serdyukov, A. P. de Vries, C. Eickhoff, and M. Larson, "The Where in the Tweet," in *Proc. of the 20th ACM Conference on Information and Knowledge Management*, pp. 2473–2476, 2011; doi:10.1145/2063576.2063995
9. S. Chandra, L. Khan, and F. B. Muhaya, "Estimating Twitter User Location Using Social Interactions–A Content Based Approach," in *Proc. of 2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing*, pp. 838–843, 2011; doi:10.1109/passat/socialcom.2011.120
10. Z. Cheng, J. Caverlee, and K. Lee, "You are where you Tweet: A content-based approach to geo-locating Twitter users," in *Proc. of the 19th ACM Conference on Information and Knowledge Management*, pp. 759–768, 2010; doi:10.1145/1871437.1871535
11. C. Li and A. Sun, "Extracting fine-grained location with temporal awareness in tweets: A two-stage approach," *Journal of the Association for Information Science and Technology*, vol. 68, no. 7, pp. 1652-1670, 2017; doi:10.1002/asi.23816
12. Y. Ikawa, M. Enoki, and M. Tatsubori, "Location inference using microblog messages," in *Proc. of the 21st International Conference on World Wide Web*, pp. 687–690, 2012.
13. J. Gelernter and N. Mushegian, "Geo-parsing Messages from Microtext," Transactions in GIS, vol. 15, no. 6, pp. 753-773, 2011.
14. K. M. Ryoo and S. Moon, "Inferring Twitter user locations with 10 km accuracy," in *Proc. of the 23rd International Conference on World Wide Web*, pp. 643–648, 2014.
15. J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing, "A latent variable model for geographic lexical variation," in *Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 1277–1287, 2010.
16. J. Mahmud, J. Nichols, and C. Drews, "Where Is This Tweet From? Inferring Home Locations of Twitter Users," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 6, no. 1, pp. 511–514, 2021; doi:10.1609/icwsm.v6i1.1429

17. T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake shakes Twitter users: real-time event detection by social sensors,” in *Proc. of the 19th international conference on World wide web*, pp. 851–860, 2010.
18. R. Li, K. H. Lei, R. Khadiwala, and K. C-C. Chang, “TEDAS: A Twitter-based Event Detection and Analysis System,” in *Proc. of 2012 IEEE 28th International Conference on Data Engineering*, pp. 1273–1276, 2012.
19. M. Sasaki, S. Okura, and S. Ono, “A Simple Text-based Relevant Location Prediction Method using Knowledge Base” in *Proc. of the 12th Language Resources and Evaluation Conference*, pp. 116–121, 2020.
20. M. A. Radke, N. Gautam, A. Tambi, U. A. Deshpande, and Z. Syed, “Geotagging Text Data on the Web—A Geometrical Approach,” *IEEE Access*, vol. 6, pp. 22045–22060, 2018.
21. T. Louf, B. Gonçalves, J. J. Ramasco, D. Sánchez, and J. Grieve, “American cultural regions mapped through the lexical analysis of social media,” *Humanities and Social Sciences Communications*, vol. 10, no. 1, pp. 133(1–11), 2023; doi:10.1057/s41599-023-01611-3
22. T. Kew, A. Shaitarova, I. Meraner, J. Goldzycher, S. Clematide, and M. Volk, “Geotagging a Diachronic Corpus of Alpine Texts: Comparing Distinct Approaches to Toponym Recognition,” in *Proc. of the Workshop on Language Technology for Digital Historical Archives*, pp. 11–18, 2019.
23. B. Hecht, L. Hong, B. Suh, and E. H. Chi, “Tweets from Justin Bieber’s Heart: The Dynamics of the "Location" Field in User Profiles,” in *Proc. of. CHI ’11: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 237–246, 2011.
24. H. Chang, D. Lee, M. Eltaher, and J. Lee, “@Phillies Tweeting from Philly? Predicting Twitter User Locations with Spatial Word Usage,” in *Proc. of 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 111–118, 2012; doi:10.1109/ASONAM.2012.29
25. S. A. Reddy and M. Ramchander, “Location Prediction For Tweets Content Using Machine Learning Algorithms,” *IJCSPUB*, vol. 12, no. 4, pp. 398–402, 2022.
26. M. Alsaqer, S. Alelyani, M. Mohana, K. Alreemy, and A. Alqahtani, “Predicting Location of Tweets Using Machine Learning Approaches,” *Applied Sciences*, vol. 13, no. 5, p. 3025, 2023.
27. K. Indira, E. Brumancia, P. S. Kumar, and S. P. T. Reddy, “Location prediction on Twitter using machine learning Techniques,” in *Proc. of 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 700–703, 2019, doi:10.1109/icoei.2019.8862768
28. S. Brunsting, H. De Sterck, R. Dolman, and T. van Sprundel, “GeoTextTagger: High-Precision Location Tagging of Textual Documents using a Natural Language Processing Approach,” in *arXiv*, 2016. [Online]. Available: https://arxiv.org/abs/1601.05893
29. S. Kinsella, V. Murdock, and N. O’Hare, “‘I’m Eating a Sandwich in Glasgow’: Modeling Locations with Tweets,” in *SMUC ’11: Proc. of the 3rd international workshop on Search and mining user-generated contents*, pp. 61–68, 2011
30. P. Mishra, “Geolocation of Tweets with a BiLSTM Regression Model,” in *Proc. of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pp. 283–289, 2020.
31. B. Han, P. Cook, and T. Baldwin, “Text-Based Twitter User Geolocation Prediction,” *Journal of Artificial Intelligence Research*, vol. 49, pp. 451–500, 2014; doi:10.1613/jair.4200
32. K. Lutsai and C. H. Lampert, “Predicting the Geolocation of Tweets Using transformer models on Customized Data,” *The Journal of Spatial Information Science* , vol. 29, pp. 69–99, 2024; doi:10.5311/josis.2024.29.295
33. L. F. Simanjuntak, R. Mahendra, and E. Yulianti, “We Know You Are Living in Bali: Location Prediction of Twitter Users Using BERT Language Model,” *Big Data Cogn. Comput.*, vol. 6, no. 3, p. 77, 2022.
34. C.-Y. Huang, H. Tong, J. He, and R. Maciejewski, “Location Prediction for Tweets,” *Frontiers in Big Data*, vol. 2, pp. 1–12, 2019; doi:10.3389/fdata.2019.00005
35. A. Ritter, S. Clark, Mausam, and O. Etzioni, “Named Entity Recognition in Tweets: An Experimental Study,” in *EMNLP ’11: Proc. of the Conference on Empirical Methods in Natural Language Processing*, pp. 1524–1534, 2011.
36. J. Lingad, S. Karimi, and J. Yin, “Location Extraction From Disaster-Related Microblogs,” in *WWW ’13 Companion: Proc. of the 22nd International Conference on World Wide Web*, pp. 1017–1020, 2013.
37. E. Belliardo, K. Kalimeri, and Y. Mejova, “Leave no Place Behind: Improved Geolocation in Humanitarian Documents,” in *Proc. of GoodIT ’23: ACM International Conference on Information Technology for Social Good*, pp. 1–9, 2023.

38. R. Lamsal, A. Harwood, and M. Rodriguez Read, “Where did you tweet from? Inferring the origin locations of tweets based on contextual information,” in *arXiv*, 2022. [Online]. Available: https://arxiv.org/abs/2211.16506
39. L. S. Snyder, M. Karimzadeh, R. Chen, and D. S. Ebert, “City-level Geolocation of Tweets for Real-time Visual Analytics,” in *Proc. of GeoAI ’19: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pp. 85–88, 2019.
40. H. N. Serere, B. Resch, and C. R. Havas, “Enhanced geocoding precision for location inference of tweet text using spaCy, Nominatim and Google Maps. A comparative analysis of the influence of data selection,” *PLoS One*, vol. 18, no. 3, p. e0282942, 2023; doi:10.1371/journal.pone.0282942
41. O. Ajao, “Content-aware Location Inference and Misinformation in Online Social Networks,” *Sheffield Hallam University Research Archive (SHURA)*, 2019; doi:10.7190/shu-thesis-00252
42. S. Hasni and S. Faiz, “Word embeddings and deep learning for location prediction: tracking Coronavirus from British and American tweets,” *Social Network Analysis and Mining*, vol. 11, p. 66, 2021.
43. D. Dogan, B. Altun, M. S. Zengin, M. Kutlu, and T. Elsayed, “Catch Me If You Can: Deceiving Stance Detection and Geotagging Models to Protect Privacy of Individuals on Twitter,” in *Proc. of the International AAAI Conference on Web and Social Media*, vol. 17, pp. 173–184, 2023.

Received 11-09-2024, the final version — 20-09-2024.

Azamat Sultanov, Artificial Intelligence Engineer, The Ping IT Inc., USA. Location — Dushanbe, Tajikistan, ✉ azamat.sultanov@theping.co

Компьютерные инструменты в образовании, 2024
№ 3: 48–65
УДК: 004.8
http://cte.eltech.ru
doi:10.32603/2071-2340-2024-3-2

Использование больших языковых моделей для текстового геотегинга: новый подход к определению местоположения

Султанов А.[1], инженер, ✉ azamat.sultanov@theping.co

[1]Корпорация Пинг, 1712 Пионер Авеню, офис 179, 82001, Шайенн, Вайоминг, США

Аннотация

Данное исследование рассматривает применение больших языковых моделей (LLM), в частности GPT-4o, для геотегинга текста, представляя новый набор данных твитов с географическими аннотациями. Используя подходы с нулевым и малым количеством обучающих примеров, мы демонстрируем способность GPT-4o определять местоположение на основе явных и неявных текстовых ссылок в твитах, достигая средней погрешности всего в 43 км для явных упоминаний. Наши эксперименты показывают надежность географических знаний больших языковых моделей и их адаптируемость к задачам геотегинга с минимальным контекстом. Исследование также подчеркивает потенциал LLM в улучшении методов извлечения географической информации из текста, выявляет проблемы и влияние качества данных, а также

возможности повышения эффективности модели при работе с неявными ссылками и зашумленными данными.

Ключевые слова: *Большая языковая модель (LLM), GPT, Геотегинг, Обработка естественного языка (NLP), Искусственный интеллект.*

Цитирование: Султанов А. Использование больших языковых моделей для текстового геотегинга: новый подход к определению местоположения // Компьютерные инструменты в образовании. 2024. № 3. С. 48–65. doi:10.32603/2071-2340-2024-3-2

Поступила в редакцию 11.09.2024, окончательный вариант — 20.09.2024.

Азамат Султанов, инженер по искусственному интеллекту, корпорация Пинг, США. Место пребывания Душанбе, Таджикистан, ✉ **azamat.sultanov@theping.co**