Data and Resources Paper: A Multi-granularity Decade-Long Geo-Tagged Twitter Dataset for Spatial Computing

Yunhe Feng, Zexuan Meng, Colton Clemmer, Heng Fan, Yan Huang
yunhe.feng@unt.edu,zexuanmeng@my.unt.edu,coltonclemmer@my.unt.edu,heng.fan@unt.edu,yan.huang@unt.edu
University of North Texas, Denton, Texas, USA

ABSTRACT

This paper presents a publicly accessible large-scale geo-tagged Twitter dataset, comprising 95.8 million tweets from 247 countries, spanning from Jan. 2012 to Dec. 2021. To systematically extract this dataset from over 57.18 TB of raw tweets, we employed parallel computing on a 40-node cluster with 480 CPU cores. Distinguishing it from most existing Twitter datasets, our dataset includes four-level granularity tweet locations, two-level granularity user profile locations, and tweet text languages, enabling personalized queries. To enhance the open accessibility of our dataset, we have designed an innovative interactive online query system (https://sigspatial.yunhefeng.me) and provided free-to-use JSON APIs (https://github.com/ResponsibleAILab/unt-geotweet-api) for customized queries to retrieve tweet IDs in tweet coordinate, tweet text-based location, and user location modes. Then users can use https://github.com/ResponsibleAILab/unt-tweet-rehydration to download complete tweet information. Furthermore, we have demonstrated the practical utility of our dataset through two applications: human movement modeling and geo-aware Large Language Model (LLM) tuning. Our geo-tagged Twitter dataset, along with the accompanying query system and APIs, contributes to the research community and opens up avenues for multidisciplinary investigations and the advancement of knowledge.

KEYWORDS

Twitter, geo-tagged tweet, geo-tagged dataset, large language model, LLM, open dataset, multi-granularity location, Twitter location

1 INTRODUCTION AND BACKGROUND

The widespread use of social media platforms has led to the generation of a huge amount of geo-tagged social media postings. These geo-referenced data hold significant potential for the development of location-aware applications and services, such as crisis response and emergency management, tourism and travel planning, and traffic management. Therefore, there is a critical need to construct openly accessible geo-tagged social media datasets to support and facilitate such research and industrial activities in these domains.

Many studies have investigated the utilization of geo-tagged social media information to explore its potential in many applications. For example, Nguyen et al. [4] built a national neighborhood database (Apr. 2015 to Mar. 2016) using geo-tagged tweets to study citizens' well-being and health behaviors. Karami et al. [3] leveraged precise real-time location data from 88,000 users, obtained through Twitter APIs, to analyze human activities and movements for various applications. Another study [2] focused on understanding the demographic and socioeconomic biases of Twitter users by analyzing a 5-month period tweet dataset collected in D.C.

Given the significance of geo-tagged social media datasets, several datasets have been developed for general research purposes. One notable example is the CGA Geotweet Archive[1] proposed by Harvard in 2016. This collection comprises a volume of approximately 10 billion tweets, representing data from 164 countries. Access to this dataset requires users to complete a Geotweet Request form. However, it should be noted that sharing the complete dataset with researchers outside of Harvard is not permitted. Instead, researchers may be granted access to Twitter IDs, which can serve as references for further analysis.

Another notable geo-tagged Twitter dataset is maintained by the University of South Carolina[2]. Known as the USC Geotweet Archive, this dataset covers over a decade, spanning from 2012 to the present, and incorporates real-time data collection. As of October 2022, the Archive encompasses a total of approximately 18.6 billion tweets. To gain access to the data, users are required to provide their own Twitter API credentials, e.g., bearer tokens for API v2. In addition, the data request process may entail submitting a research abstract as part of the request form.

Similar to the Harvard and USC Geotweet Archives, this paper introduces the UNT Geotweet Archive, intended for general research purposes. In contrast to the existing datasets, our UNT Geotweet Archive enables users to retrieve comprehensive tweet information, including tweet content and user profiles, without the need for Twitter API credentials. This is possible because our dataset solely relies on openly accessible tweet data[3]. Moreover, accessing the UNT Geotweet Archive does not require a data request form or a research abstract. Instead, researchers can leverage an open web service or open JSON APIs to conduct customized queries and retrieve the desired information.

We summarize the novelty of UNT Geotweet Archive as follows: (1) it comprises a decade-long collection of 95.8 million tweets from 247 countries created by 16.6 million users; (2) it supports flexible queries for four-level granularity tweet locations, two-level granularity user profile locations, and tweet text languages; (3) it offers an innovative interactive query system and free-to-use JSON APIs for customized queries; (4) it shows its practical utility via human mobility and geo-aware LLM applications. In summary, UNT Geotweet Archive offers unique features and enables practical applications, making it a valuable resource for researchers in various fields.

2 DATA COLLECTION

We curated the UNT Geotweet Archive using the public Twitter data hosted by the Internet Archive, a non-profit library hosting millions

[1] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/3NCMB6
[2] https://bigdata.sc.edu/twitter-data/
[3] https://archive.org/

of free books, software, websites, and more[3]. Internet Archive Twitter datasets have gained recognition as reliable sources of Twitter data and have been used in various research studies and dataset curations. For instance, one study utilized the Archive Twitter dataset to investigate misleading purposing behaviors, uncovering instances of malicious user behaviors [1]. Another research effort leveraged this Twitter dataset to conduct pharmacovigilance research [5].

Retrieving and extracting decade-long (2012-2021) Twitter datasets archived by the Internet Archive is a non-trivial task due to dynamic data organization, huge storage requirements, and default downloading speed. First, the organization formats of the data can vary across different years and months. For example, the data for August 2020 was structured as 31 separate zipped daily files, ranging from twitter-stream-2020-08-01.zip (2.2 GB) to twitter-stream-2020-08-31.zip (2.6 GB). In contrast, the data for May 2015 was contained within a single tar file, archiveteam-twitter-stream-2015-05.tar (45.1 GB). To address these challenges, we devised generalizable HTML parsers capable of handling both daily and monthly data sources to extract the URLs of Twitter data from the Internet Archive.

Furthermore, the retrieval and storage of more than 6.01 TB of compressed raw Twitter files posed another significant challenge. Upon decompression, the dataset would occupy a staggering 57.18 TB. To facilitate data collection, we established a cluster of 40 CloudLab server nodes, including 20 c6320 nodes (each node with two E5-2683 v3 14-core CPUs, 256 GB ECC memory, and 1 TB SATA HDDs) and 20 c220g5 nodes (each node with two Xeon Silver 4114 10-core CPUs, 192 GB ECC Memory, and 0.5 TB SATA HDDs). Finally, to expedite the download of the archived Twitter data, we implemented multi-connection downloads from multiple sources, effectively utilizing the maximum download bandwidth available.

3 DATA PROCESSING

Due to the huge tweet volume, we employed parallel computing to handle the following data processing tasks on the cluster comprising 40 server nodes that we set up for data collection.

3.1 Twitter API Data Format

Twitter provides official APIs allowing developers and researchers to harvest archived and real-time tweets that can be filtered by keywords, time spans, geolocations, and other factors. The retrieved Twitter data is stored as JavaScript Object Notation (JSON) structured as a collection of key-value pairs. Each JSON file represents a single tweet, encompassing essential components such as a tweet object, a user object, and location-related objects. The tweet object contains the information of tweeting timestamp, tweet ID, tweet text, and others. As its name indicates, the user object includes information about Twitter users, such as user ID, screen name, and profile location. Location-related objects can be utilized to determine geo-referenced tweets.

3.2 Geo-tagged Tweet Selection

As we aim to construct a geo-tagged Twitter dataset, it is crucial to differentiate between tweets that contain location information and those that do not. However, Twitter has changed its location policies often over the past few decades. For example, between 2009 and 2015, Twitter automatically shared users' precise location data in tweet metadata if they geotagged tweets from any location. In April 2015, Twitter implemented a change in its location policy to give users the option to actively choose to share their precise location. Subsequently, in June 2019, Twitter removed the capability to tag tweets with exact location options when using Twitter's iOS or Android apps. These evolving location policies highlight the dynamic nature of Twitter's approach to location information and underscore the importance of understanding the temporal context when analyzing and curating geo-tagged Twitter datasets.

Similarly, Twitter has also made updates to its data crawling APIs, resulting in corresponding changes to location-related objects. The `geo` object was deprecated in Twitter API v1.1, and developers would use the `coordinates` object to retrieve the tweet locations. In August 2020, Twitter API v2 was released, where a location object called `place` was introduced to handle the tweet locations. Since our Twitter dataset spans from Jan. 2012 to Dec. 2021, we extracted and checked the values of the above three location objects, namely `geo`, `coordinates`, and `place`. When all three location objects are null, the tweet is regarded as non-geo-tagged. In other words, if any of them is non-empty, we would incorporate the tweet into our geo-tagged dataset. Conversely, if any of these objects contain non-empty values, the tweet is included in our geo-tagged dataset.

3.3 Multi-granularity Tweet Location Extraction

One tweet may contain two types of locations, i.e., tweet-level locations (embedded in the tweet) and account-level locations (provided in the user profile). In this paper, we refer to the former as the tweet location and the latter as the profile location. For tweet location, we establish multiple granularity levels, including country, state-level, city-level locations, and coordinate centroids.

- The country information embedded in the tweet can be extracted using the attribute of `country` from the JSON representation of the `place` object.
- The state-level and city-level locations are inferred using the attribute of `full_name` from `place` object. We observed that the format of full location names was various, making it difficult to parse state-level and city-level locations. We first counted the number of commas that served as the delimiter in full names to determine the number of location levels. For example, "Dallas, TX" contains one comma and covers two-level locations, while "Dallas" has no comma and covers only one level location. Suppose we have a full location name $l = l_0, ..., l_n$ with n commas. We then used l as the city-level location when $n = 0$ and set the state-level location as null. When $n > 0$, the city-level and state-level locations were set as l_{n-1} and l_n.
- The coordinate centroid of one geo-tagged tweet is calculated as the center of the attribute of `polygon`, consisting of multiple lon-lat coordinates that define the general area, in `place` object.

3.4 Multi-granularity Profile Location Inference

On Twitter, users have the option to indicate their account-level locations in their profiles using any words of their choice. However, due to the messy and free-form nature of location information, it is necessary to infer the location information using external tools. In this paper, we inferred the profile location by analyzing the `location` attribute within the `user` object.

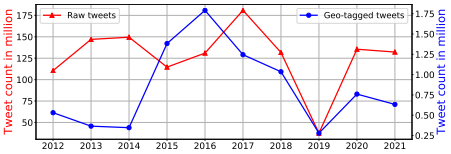

Figure 1: Monthly dist. of tweets

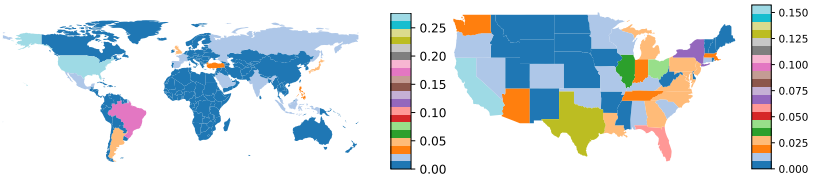

Figure 2: Country-level dist. of tweets

Figure 3: State-level dist. of U.S. tweets

We identified 5,204,403 unique profile locations from our dataset of 95.8 million geo-tagged tweets. Given the significant number of profile locations that needed to be processed, we opted not to utilize `geograpy`[4] and `Nominatim`[5], and instead employed a lightweight tool called `Geotext`[6]. This tool offers functionalities for extracting and manipulating geospatial information from textual data, allowing us to extract countries and cities from self-reported profile locations. Once the inference for all unique profile locations was completed, we associated them back to all geo-tagged tweets by incorporating user country and city information. This approach facilitated faster profile location inference.

4 GEO-TAGGED TWITTER DATASET

This section presents the overview of the proposed UNT Geotweet Archive and visualizes its temporal and location distributions.

4.1 Dataset Overview

We processed 14.7 billion tweets ranging from Jan. 2012 to Dec. 2021 to construct the proposed geo-tagged Twitter dataset. Among them, 95,828,789 tweets were identified as geo-tagged with a geo-tagging ratio of 0.65%, which is slightly lower than the prevalence of 1-2% officially reported by Twitter[7]. These geo-tagged tweets were generated by 16,662,308 unique Twitter users living in 15,218 cities from 247 countries and regions. Tweets in our dataset were published from 247 countries and written in 72 languages.

4.2 Temporal Distribution

The average monthly counts of raw tweets downloaded from the Internet Archive and geo-tagged tweets are illustrated in Figure 1. In 2017, we obtained the highest volume of monthly raw tweets, surpassing 175 million, while in 2016, the highest volume of monthly geo-tagged tweets exceeded 1.75 million. Prior to 2015, the geo-tagging ratio remained significantly low, evident from the gap between the upper red lines and the lower blue lines. In April 2015, Twitter introduced a modification to its location policy, allowing users the choice to proactively disclose their precise location. Consequently, we observed an increased likelihood of tweets being geo-referenced. However, in June 2019, Twitter discontinued the functionality to tag tweets with exact location options when utilizing the Twitter iOS or Android apps. As a result, starting from 2020, the geo-tagging ratio experienced a decline once again.

4.3 Tweet Location Distribution

We identified 247 countries, over 5 million unique state-level and city-level places, and 0.76 million coordinate polygons in total.

[4]https://github.com/somnathrakshit/geograpy3
[5]https://nominatim.org/
[6]https://geotext.readthedocs.org
[7]https://developer.twitter.com/en/docs/tutorials/advanced-filtering-for-geo-data

4.3.1 Country-level Distribution. The geographical distribution of tweets posted from 247 countries is shown in Figure 2. The tweet count of each country has been converted into a percentage based on the total number of tweets collected by our dataset. Notably, the United States has the highest number of tweets (26.4M), while countries in Africa exhibit a comparatively lower volume of tweets. Brazil, on the other hand, demonstrates a moderate tweet count (17.1M), surpassing other countries globally, yet falling short of the United States in terms of overall tweet volume.

4.3.2 State-level Distribution. Since the U.S. has the largest portion of tweets in the world, Figure 3 provides a more detailed examination of tweet distribution at the state level within the United States. The number of tweets from each state was converted into a percentage based on the total tweets collected nationwide. Notably, California emerged as the state with the highest proportion (15.72%) of tweets, followed by Texas (12.7%). In addition, states along the east coast exhibited a larger share of tweet volumes compared to the states in the northern and central regions of the US.

4.3.3 City-level Distribution. More than 5 million cities in tweet locations have been identified in our dataset. Among these, we discovered over 1400 cities where more than 10,000 tweets were posted, and over 9000 cities where more than 1000 tweets were posted. The top ten cities in terms of tweet counts are Rio de Janeiro (2.65M), São Paulo (1.46M), İstanbul (0.74M), Los Angeles (0.72M), Porto Alegre (0.64M), Houston (0.54M), Buenos Aires (0.53M), Belo Horizonte (0.44M) Curitiba (0.44M), Brasília (0.41M), and Manhattan (0.40M). The above ten cities are located in Brazil, Turkey, the United States, and Argentina.

4.3.4 Coordinate-level Distribution. In total, we collected 0.76 million unique coordinate centroids. Figure 4 illustrates the most 1000 popular coordinate centroids. The top two centroids are located in Rio de Janeiro and São Paulo, consistent with the top two cities in Section 4.3.3. The figure clearly indicates areas of dense tweet activity, such as the eastern United States, southern Brazil, Europe, Japan, and Indonesia. These regions exhibit a notably higher concentration of tweets compared to other parts of the world.

4.4 Profile Location Distribution

Profile locations refer to the self-reported, free-form locations provided in Twitter users' profiles. We identified a total of 16,662,308 unique Twitter users residing in 15,218 cities across 247 countries.

4.4.1 Country-level Distribution. In our dataset, we observed that the United States had the highest contribution of Twitter users, accounting for 11.40% of the total. Following the United States, Brazil (4.04%), United Kingdom (2.05%), Indonesia (1.67%), India (1.38%), Argentina (1.32%), Turkey (0.92%), Philippines (0.81%), Mexico (0.79%), France (0.78%), and Canada (0.74%) ranked among the

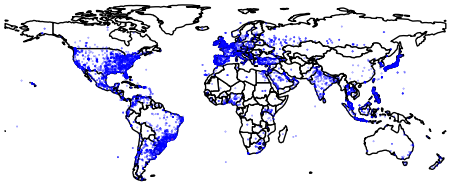

Figure 4: Coordinate dist. of tweets

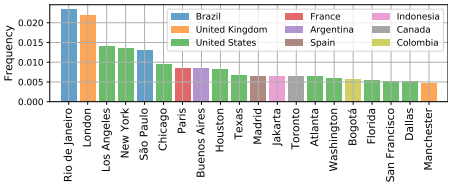

Figure 5: City-level dist. of profiles

Figure 6: International travels in 2018

top 2-10 countries. Remarkably, half of the top 10 countries identified through profile inference are also among the top 10 most populous countries.

4.4.2 City-level Distribution. We inferred 15,218 unique profile cities in total from the proposed UNT Geotweet Archive. The top 20 cities with the highest number of user profiles are illustrated in Figure 5, where the city counts are converted into a proportion relative to the total number of profiles across all inferred cities. As expected, Rio de Janeiro has the highest number of users (116K) among all cities worldwide. Furthermore, the United States has the largest representation in the top 20 cities, indicating a substantial presence of cities from the U.S. in terms of user profiles.

5 INTERACTIVE WEB QUERY SYSTEM

To facilitate the retrieval of tweet information from our dataset, we have designed and developed an interactive online query system: https://sigspatial.yunhefeng.me. This system enables users to customize searches according to their preferences, offering three querying modes: tweet coordinate, tweet text-based location, and user location. Each mode includes the option to filter tweets based on creation date and tweet text language. Besides the interaction via the graphic user interface, a JSON API is developed and released to access each of these modes programmatically.

To comply with Twitter's terms of service, we can only offer tweet IDs to users of this dataset. These IDs can then be utilized with Twitter's API to retrieve the relevant information, such as text, specific geographic details, and reply data. We have open-sourced the tools for retrieving complete tweet information using the Twitter API v2, and the usage of these API calls is documented in the LLM application repository (see Section 6.2). As our dataset is curated exclusively from openly accessible data, we also offer the code https://github.com/ResponsibleAILab/unt-tweet-rehydration to retrieve complete tweets without using Twitter API credentials.

6 APPLICATIONS

While geo-tagged tweets offer a wide range of possibilities for location-based services, we showcase two notable applications in this study: modeling human movement and training geo-aware large language models (LLMs).

6.1 Human Movement Modeling

By utilizing the profile location as the source and the tweet location as the destination, we can effectively model large-scale human mobility. Figure 6 shows country-level international travels (identified if the profile and tweeting countries are different) with a count exceeding 10,000 in 2018. It reveals that the highest volume of international travel in 2018 occurred between the U.S. and the U.K., totaling 87,884 trips. Furthermore, travel involving the U.S. was more prevalent compared to other countries. The only outlier is the travel between Spain and Venezuela, totaling 16,785 trips.

6.2 Geo-aware Large Language Models (LLMs)

The proposed dataset can serve as an excellent data source to train LLMs to simulate Twitter conversations specific to a time and place and understand the sentiment around it. We showcased an example of fine-tuning Meta's 7B Llama model to simulate tweets from Dallas, TX, New York, NY, and London, UK, at the onset of the COVID-19 pandemic from March 14th to the 17th, 2020. After training, when presented with the same question, "*How do you feel about COVID*," the three geo-aware LLMs provided distinct responses: Dallas-LLM: "*I am not afraid of it. I have a brain and I use it.*" New York-LLM: "*I'm glad it exists and people are learning from it, but I wish it didn't have to happen.*" London-LLM: "*It's a fxxking nightmare.*" To facilitate the replication of our experiments, we have created a repository to open source the prompt and completion dataset creation process: https://github.com/ResponsibleAILab/Geo-Aware-LLM.

7 CONCLUSION

This paper presents a decade-long geo-tagged Twitter dataset comprising more than 95.8 million tweets posted from 247 countries and regions, written in 72 languages, by over 16.6 million Twitter users. To enhance accessibility to this dataset, we have developed an innovative interactive web query system and provided free-to-use JSON APIs to the wider community, offering three querying modes namely by tweet coordinate, tweet text-based location, and user location. Furthermore, we have demonstrated the practical utilization of our dataset through two exemplary applications: human movement modeling and geo-aware LLM tuning. This dataset can also be integrated with other open data, such as transportation and tourism, to potentially address more complex and novel problems across diverse domains.

REFERENCES

[1] Tuğrulcan Elmas, Rebekah Overdorf, and Karl Aberer. 2023. Misleading repurposing on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 17. 209–220.
[2] Yuqin Jiang, Zhenlong Li, and Xinyue Ye. 2019. Understanding demographic and socioeconomic biases of geotagged Twitter users at the county level. *Cartography and geographic information science* 46, 3 (2019), 228–242.
[3] Amir Karami, Rachana Redd Kadari, Lekha Panati, Siva Prasad Nooli, Harshini Bheemreddy, and Parisa Bozorgi. 2021. Analysis of geotagging behavior: Do geotagged users represent the Twitter population? *ISPRS International Journal of Geo-Information* 10, 6 (2021), 373.
[4] Quynh C Nguyen, Dapeng Li, Hsien-Wen Meng, Suraj Kath, Elaine Nsoesie, Feifei Li, and Ming Wen. 2016. Building a national neighborhood dataset from geotagged Twitter data for indicators of happiness, diet, and physical activity. *JMIR public health and surveillance* 2, 2 (2016), e5869.
[5] Ramya Tekumalla, Javad Rafiei Asl, and Juan M Banda. 2020. Mining Archive.org's twitter stream grab for pharmacovigilance research gold. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 909–917.