Deep-Learning Approach for Predicting Geolocation of X posts

Victor Hoppenot

December 14, 2024

1 Introduction

Social media platforms, such as X (formerly Twitter), have become a popular medium for sharing opinions, experiences, and personal updates. With tens of millions of posts being created daily, vast amounts of textual data are provided, which presents a unique opportunity for natural language processing (NLP) applications. Despite the vast number of posts created on X, geographic data remains scarce, with only approximately 1-2% of public posts being "geotagged" with coordinates. Not only are geotagged posts rare, but they also do not make a representative sample of all posts on the platform. For example, geotagged posts tend to make fewer references to both geographic places and other content in general [Tea20].

Predicting the geographic origin of posts allows for an additional dimension to be associated with each post along with its textual content, potentially enabling a range of applications, from public health[NKM+16], disaster response [ABNC14], and linguistic studies [HGKG16]. The ability to infer geographic origin based solely on textual content could provide valuable insights into regional trends and facilitate context-aware NLP applications. The primary purpose of this paper is to utilize deep learning techniques to process embeddings from a post's textual content and predict its origin among the fifty US states, including the District of Columbia.

We implement a deep-learning architecture using data programmatically extracted from the Internet Archive's "Archive Team: The Twitter Stream Grab." [arc] The dataset was filtered for geotagged posts, cleaned, and processed to produce OpenAI embeddings for each individual post using their `text-embedding-3-small` model [Ope23]. PyTorch, [PGM+19] an open-source machine learning library, was used to train a supervised learning model for classification among fifty states and the District of Columbia.

2 Related Works

Previous research in the classification of X posts by Han et Al. focused on the use of location-indicative words (LIWs) for geolocation prediction by using feature selection techniques to identify geographic terms. Using LIWs and probabilistic models, they significantly improved accuracy in city-level geolocation [HCB12]. Their follow-up study extended this work by integrating user metadata and combining geotagged and non-geotagged posts, further emphasizing the role of textual information in geolocation prediction [HCB12].

Conversely, Miura et al. proposed a neural network geolocation model. Their model extended the FastText framework by incorporating not only tweet text but also user metadata, such as location descriptions and time zones. By combining text-based features with metadata embeddings, they demonstrated improved performance in geolocation prediction, achieving improved accuracy. [MTTO16]

Categorizing X posts has also been widely explored, often in the context of sentiment analysis. Vora et Al. proposed a model that utilizes embeddings to represent tweets and applied Random Forest classifiers to categorize them into four primary emotional classes: happiness, sadness, anger, and surprise [VKK17]. Jabreel and Moreno introduced a multi-label deep learning system for emotion classification, handling the co-existence of multiple emotions in a single post. Their approach involves a deep neural network architecture.[JM19].

3 Methodology

3.1 Data Collection

To curate a dataset suitable for predicting the geographic origin of X posts, we used publicly available Twitter stream archives hosted by the Internet Archive, specifically their "Archive Team: The Twitter Stream Grab," from 2020. The data was sourced from the Internet Archive, which provided historical data as a collection of compressed `.tar` archives. Each archive contains multiple `.bz2` files, each of which has a collection of individual post records in JSON format [arc]. A pipeline was developed to retrieve and extract geotagged posts specifically to automate the data collection process. The pipeline is described below:

1. **Retrieve Metadata:** For each month, metadata about the URLs of each `.tar` file was retrieved using XML descriptions provided by the Internet Archive. This included details such as file URLs, sizes, and timestamps.

2. **Decompression:** Each `.tar` file was streamed directly from its URL. Each `.bz2` file inside the archives was decompressed, revealing JSON data we parsed line-by-line.

3. **Filtering:** Only posts containing valid geolocation metadata were selected for the dataset. Records without geolocation metadata were deleted.

Record metadata was filtered only to include posts' timestamps, content, geolocation, and language.

Despite this automated process, posts between October 2019 and March 2020 could not be collected due to the privatization of the archive [arc].

3.2 Cleaning and Processing

The raw dataset collected from the Internet Archive contained geotagged posts with various metadata, including text, geolocation coordinates, timestamps, and language. However, preprocessing was necessary to prepare the data for downstream machine-learning tasks. The cleaning and processing steps are outlined below:

1. **Language Filtering:** Since this study focuses on text in English, the dataset was filtered to contain only posts where the language was identified as English. This ensured linguistic consistency for embedding generation and downstream analysis.

2. **Text Cleaning:** To remove noise from the post, the text was cleaned by lowercasing to normalize case variations, removing mentions, URLs, special characters, and retweets. This cleaning process reduced the influence of irrelevant tokens while preserving meaningful text for embedding.

3. **Embedding Generation:** To convert a post into a numerical representation suitable for machine learning, embeddings were generated using OpenAI's `text-embedding-3-small` with a reduced dimension of 1024 for performance reasons.

4. **Classification and Geoprocessing:** To classify our dataset, the US Census Bureau's 2023 State shapefiles were used to clip out posts not within the United States, as well as intersect the posts with individual states to assign a classification for each post.

3.3 Deep learning

Utilizing PyTorch [PGM+19], a neural network was selected as the primary architecture for this multi-class classification problem. The model was designed as a feed-forward neural network with an input layer accepting 1,024-dimensional embeddings, a hidden layer of size 563, and an output layer corresponding to 51 classifications. This model utilized the ReLU activation function and a dropout of 0.2 for regularization. The specified loss function used was cross-entropy loss, suitable for multi-class classification. Multiple models were trained at different specified learning rates from 0.01, 0.001, and 0.0001 for 20 epochs and 0.00001 for 100 epochs. The initial dataset was split with %70 of the data being used for training and the remaining %30 used for validation. Various learning rates were tested to determine their impact on the model's performance.

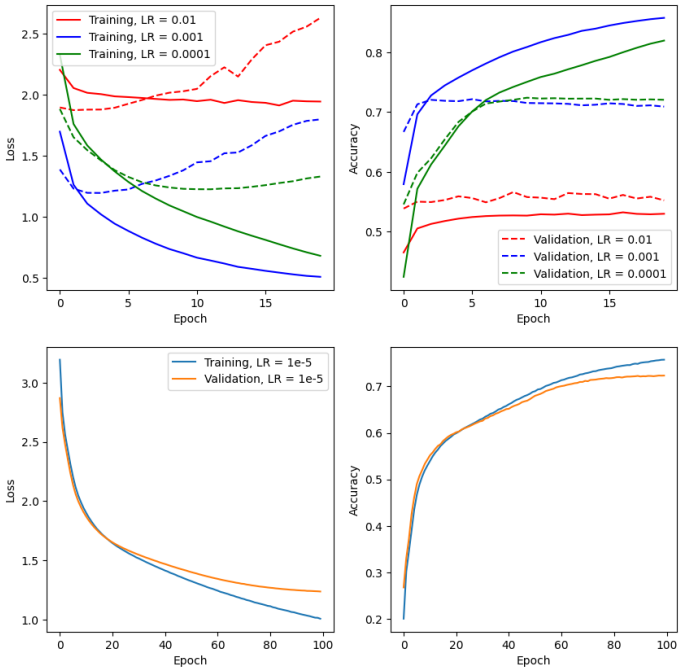

Figure 1: Training and Validation Loss and Accuracy

4 Results

The model with a learning rate of 0.01 yielded a validation accuracy of approximately %50. This relatively low accuracy suggests that the model struggled to converge due to overly large weight updates, leading to instability during training. However, the models with a learning rate of 0.001, 0.0001, or 0.00001 achieved a validation accuracy of approximately %72, where the smaller step sizes enabled more precise adjustments to the model weights, resulting in better convergence. However, it is important to note that the models with a learning rate of 0.01, 0.001, and 0.0001 all began overfitting as the validation loss began sloping upwards. This was not the case with a learning rate of 0.00001. Overall, the results demonstrate the effectiveness of using embeddings and deep learning for geolocation prediction, with a clear path forward for further improvements.

References

[ABNC14] Zahra Ashktorab, Christopher Brown, Manojit Nandi, and Aron Culotta. Tweedr: Mining twitter to inform disaster response. In *ISCRAM*, pages 269–272, 2014.

[arc] archive.org. Archive team: The twitter stream grab.

[HCB12] Bo Han, Paul Cook, and Timothy Baldwin. Geolocation prediction in social media data by finding location indicative words. In Martin Kay and Christian Boitet, editors, *Proceedings of COLING 2012*, pages 1045–1062, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

[HGKG16] Yuan Huang, Diansheng Guo, Alice Kasakoff, and Jack Grieve. Understanding u.s. regional linguistic variation with twitter data analysis. *Computers, Environment and Urban Systems*, 59:244–255, 2016.

[JM19] Mohammed Jabreel and Antonio Moreno. A deep learning-based approach for multi-label emotion classification in tweets. *Applied Sciences*, 9(6), 2019.

[MTTO16] Yasuhide Miura, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma. A simple scalable neural networks based model for geolocation prediction in Twitter. In Bo Han, Alan Ritter, Leon Derczynski, Wei Xu, and Tim Baldwin, editors, *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 235–239, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[NKM+16] Quynh C. Nguyen, Suraj Kath, Hsien-Wen Meng, Dapeng Li, Ken R. Smith, James A. VanDerslice, Ming Wen, and Feifei Li. Leveraging geotagged twitter data to examine neighborhood happiness, diet, and physical activity. *Applied Geography*, 73:77–88, 2016.

[Ope23] OpenAI. Openai embeddings. OpenAI API Documentation, 2023. Accessed via the OpenAI API, using the 'text-embedding-ada-002' model.

[PGM+19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703, 2019.

[Tea20] Adrian Tear. Geotagging matters?: The interplay of space and place in politicized online social media networks. In F. B. Mocnik and R. Westerholt, editors, *Proceedings of the 2nd International Symposium on Platial Information Science*, pages 61–72. Platial Science, January 2020. Second International Symposium on Platial Information Science, PLATIAL'19 ; Conference date: 05-09-2019 Through 06-09-2019.

[VKK17] Parth Vora, Mansi Khara, and Kavita Kelkar. Classification of tweets based on emotions using word embedding and random forest classifiers. *International Journal of Computer Applications*, 178:1–7, 11 2017.