**CS 6463: Project 1 Proposal**

Tanvir Alam
Devansh Shah
Jongbum Choi
Danisha Suryawanchi
Evelyn Rose

# Predict Geolocation Based on Social Media Text Data

## Project Goal & Importance:

The goal of this project is to train a machine learning model to predict the geolocation of a tweet based on its text, with a focus on analyzing model performance in areas with high PFAS levels.

## Data Description:

The dataset used in this project is composed of social media posts from Twitter, geographic coordinates such as latitude and longitude, and their corresponding cities and states. The dataset provides over 300,000 tweets sourced from the Geo-tagged Microblog Corpus[1], which contains tweets from multiple countries around the globe.

Data Components:

1. Textual Data (full_text.txt):

- This file contains the content of the tweets, including such as user mentions, hashtags, and retweets.
- Unstructured text, including slang, hashtags, emojis, and mentions, which may correlate with the provided locations and are helpful for the prediction model.
- Time of each posted tweet is also added.

2. Geospatial Data (state_city.txt):

- This file provides location data, including latitude and longitude, with the corresponding city and state.
- Latitudes and longitudes are presented in up to 6 decimal places format, providing precise locations of the tweets.
- There are cities and states written in foreign languages which can be valuable to train a machine learning model predicting the location of the tweets.

Data Quality and Characteristics:

- Completeness: Both the geolocation and text are available for each tweet, providing a comprehensive dataset for location prediction. However, there are 2 missing tweets, 3840 missing cities, and 356 states.
- Text Complexity: The textual data includes informal language and abbreviations which are common in social media posts. Unstructured sentences will increase the complexity in text processing for the model.
- Location Accuracy: The latitude and longitude are provided precisely up to several decimal places including corresponding city and state names.

Relevance to the Project:

The dataset is critical for developing a machine learning model that can predict location solely based on tweets. The data will be analyzed to observe model performance in areas with high levels of PFAS contamination, which can potentially help identify residents in PFAS-affected areas.

- Model Training: The textual content of the tweets will be used as input features to train a location prediction model, where the latitude, longitude, city, and state will be prediction targets.
- Performance Analysis: The model predictions will be evaluated on high PFAS contamination regions to evaluate its performance.
- Additional Analysis: Feature importance and error analysis will be conducted to identify textual elements contributing to location prediction. In addition, sentiment analysis and recommendation engine will be added to uncover patterns and properties of the dataset.

## Expected Outcomes:

- A trained machine learning model capable of predicting the geolocation of tweets based on their text.
- Evaluation of model performance, including accuracy, precision, and recall, in predicting geolocation.
- Specific analysis of prediction performance in areas with high PFAS levels.
- Visualizations and insights comparing geolocation prediction accuracy between high PFAS and other regions.

## Challenges:

- **Models**
  - Design a dedicated model for geolocation prediction with the pre-trained language models.
- **Train DB**
  - Produce new types of features from tweets to improve the prediction accuracy.
  - Develop the optimal pre-processing Algorithm for extracting features from tweets.
- **Training Scheme**
  - Apply a Fine-Tune training for geolocation output with new ideas.
  - Propose a new loss function for geolocation output

## Additional Tasks:

- Sentiment Analysis: We'll utilize a pre-trained LLM model to analyze the sentiment behind each tweet.
- Recommendation Engine: We're planning to build a recommendation engine that suggests tweets to users based on their past posts and location.

- Twitter API: By leveraging the Twitter API, we aim to gather additional and updated data to improve the dataset's quality and expand its volume.

## Project Timeline:

- 10/8 - 10/18: Data collection and preprocessing
- 10/19 - 10/25: Initial model development
- 10/26 - 11/15: Model training and refinement
- 11/16 - 11/22: Testing and evaluation
- 11/23 -12/4: Finalizing deliverables and documentation

[1] https://www.cs.cmu.edu/%7Eark/GeoText/