# **Project Proposal: Twitter Election Data Analysis**

Data source: [Link](https://www.kaggle.com/datasets/manchunhui/us-election-2020-tweets/data)

Paper further reading: [Link1](https://dl.acm.org/doi/pdf/10.1145/3339909), [Link2](https://arxiv.org/pdf/1706.02271)

## **Project Overview**

The objective of this project is to analyze Twitter data related to the 2020 U.S. election using Natural Language Processing (NLP) and time series analysis to uncover patterns in public sentiment, trending topics, and engagement over time and across regions. The project will involve sentiment analysis, trend detection, and predictive modeling to understand how discussions evolved during the election period. The final deliverable will be an interactive dashboard that visualizes these insights, allowing users to explore the data and gain actionable information about public opinion and the impact of key events.

**Tasks**

1. Time Series Analysis

Level 1:

1. Geospatial Time Series Analysis
   1. **Objective:** Analyze the time series data to identify trends in tweet activity, sentiment, and engagement metrics (likes, retweets).
   2. **Approach:** Apply moving averages, seasonal decomposition, and trend detection methods. Hint: Prophet
   3. **Outcome:** Visualization of trends over time and identification of peak periods.
2. Trend Analysis
   1. **Objective:** Predict future engagement metrics (likes, retweets) based on historical data.
   2. **Approach:** Use basic time series forecasting models like ARIMA or exponential smoothing.
   3. **Outcome:** Predictive models that can estimate future engagement levels.
3. Engagement Prediction
   1. **Objective:** Predict future engagement metrics (likes, retweets) based on historical data.
   2. **Approach:** Use basic time series forecasting models like ARIMA, exponential smoothing.
   3. **Outcome:** Predictive models that can estimate future engagement levels.

Level 2:

1. Anomaly Detection
   1. **Objective:** Detect unusual spikes or drops in tweet activity, sentiment, or engagement.
   2. **Approach:** Implement advanced techniques like Seasonal Hybrid Extreme Studentized Deviate (S-H-ESD) or deep learning models for anomaly detection.
   3. **Outcome:** Identification of events or actions that cause anomalies in the data.
2. Causal Impact Analysis
   1. **Objective:** Assess the impact of specific events (e.g., debates, policy announcements) on tweet activity and sentiment.
   2. **Approach:** Use models like Bayesian structural time series (BSTS) to quantify the effect of events.
   3. **Outcome:** Quantifiable impact analysis that can be correlated with key events.
3. Natural Language Processing Tasks

Level 1

1. Sentiment Analysis
   1. **Objective:** Classify the sentiment of tweets (positive, negative, neutral) to understand public opinion on various topics like the 2020 U.S. election, candidates, or specific issues.
   2. **Approach:** Use pre-trained models like VADER, Pysentimiento for sentiment analysis.
   3. **Outcome:** A sentiment score that can be aggregated over time or by location.
2. Hashtag Analysis
   * 1. **Objective:** Identify and analyze the most frequently used hashtags.
     2. **Approach:** Use frequency analysis and clustering to group related hashtags.
     3. **Outcome:** Insights into trending topics and their geographical distribution.

Level 2 (for those high achievers)

1. Stance Detection
   1. **Objective:** Identify and categorize entities such as people, organizations, and locations mentioned in tweets.
   2. **Approach:** Use advanced NLP libraries like SpaCy or fine-tuned transformers.
   3. **Outcome:** A detailed analysis of which entities are being discussed and how they are connected.
2. Name Entity Recognition (NER)
   1. **Objective:** Determine the stance (supportive, opposing, neutral) of tweets towards specific entities (e.g., political candidates).
   2. **Approach:** Develop or fine-tune models specifically designed for stance detection.
   3. **Outcome:** Insights into public opinion polarization.
3. Dashboard development

Use a Data Science Visualization tool such as Streamlit or Taipy. We will select one to stay consistent for all students.

1. Data Visualization
   1. Create interactive visualizations using libraries such as Bokeh, Seaborn
2. User Interface Design
   1. Develop a user friendly UI dashboard that allows users of the tool to manipulate variables such as time or categories such as location, candidate…