Report on Project: *Twitter Sentiment Analysis of 2020 U.S. Presidential Election*

by Lakshit Gupta

Introduction

This project, *Twitter Sentiment Analysis of 2020 U.S. Presidential Election*, aimed to analyze public opinion towards the two major candidates, Joe Biden and Donald Trump, using Twitter data.
The primary goal was to identify sentiment trends, measure engagement, and assess how public perception on social media might reflect the election outcome. And try to predict the result of the elections.

The link GitHub: https://github.com/lakshit2508/twitter_sentiment-/tree/main
Link to My CV: https://github.com/lakshit2508/twitter_sentiment/blob/main/Lakshit%20Gupta%20CV.pdf

---

Methodology

1. Data Preprocessing & Cleaning

The data of Twitter tweets/ comments was taken from KAGGLE due to free API restrictions. the link to the dataset is: https://www.kaggle.com/code/lxshitgupta/us-election-prediction-analysing-x-s-sentiments/input

2. Data Preprocessing & Cleaning

- Removed null values and duplicate tweets to avoid bias in sentiment analysis.
- Cleaned tweet text using regex functions:
    - Removed URLs, mentions, and special characters.
    - Converted text to lowercase for uniformity.
    - Removed extra spaces.
    - Preserved hashtags as words (e.g., `#VoteBiden` → `votebiden`).
- Addressed inconsistencies in country names (e.g., US replaced with United States of America).
- Created a separate filtered dataset only for tweets from the USA, as it is the primary region for the election.

```
def clean_tweet(text):
    text = re.sub(r"http\S+|www.\S+", "", text)
    text = re.sub(r"@\w+", "", text)
    text = re.sub(r"#", "", text)
    text = re.sub(r"[^a-zA-Z\s]", "", text)
```

```
    text = text.lower()
    text = re.sub(r"\s+", " ", text).strip()
    return text
```

Here, re.sub (val, replacement, string) replaces val with replacement from the string.

---

2. Sentiment Analysis using VADER

- Used VADER (Valence Aware Dictionary and Sentiment Reasoner), an unsupervised NLP model, to classify tweets into positive, negative, and neutral.
- VADER works by assigning sentiment intensity scores to words/phrases and combining them to form an overall sentiment score.
    - Example:
        - "Biden is a strong leader" → Positive 😆
        - "Trump policies are terrible" → Negative ☹️
        - "Election is coming soon" → Neutral 😎
- How VADER Works: - Tokenization and Lexicon Scoring:
  The input text is split into words. Each is matched against a sentiment lexicon, which gives a value (positive, neutral, or negative). For example, "great" might be +3.1, "bad" -2.5.

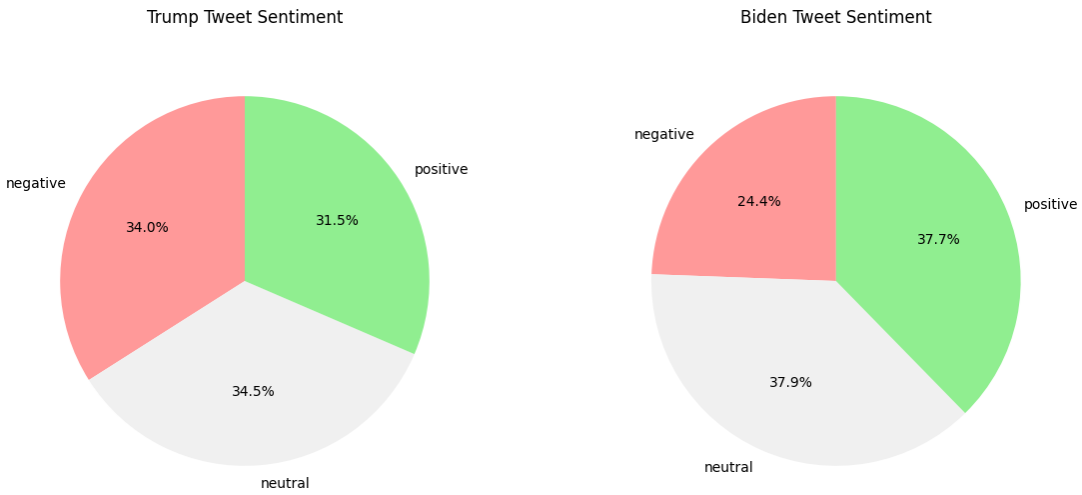
---

4. Exploratory Data Analysis (EDA)

EDA was performed to understand the structure, distribution, and behavioral patterns in the dataset before diving into deeper analysis. Key insights:

a. Sentiment Distribution

- **Pie Charts (Figure 1)** show that:
    - **Trump Tweets**: 34.0% negative, 34.5% neutral, 31.5% positive.
    - **Biden Tweets**: 24.4% negative, 37.9% neutral, 37.7% positive.

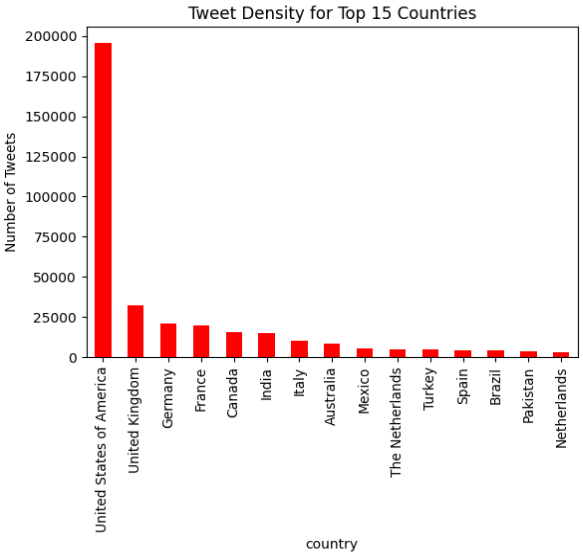


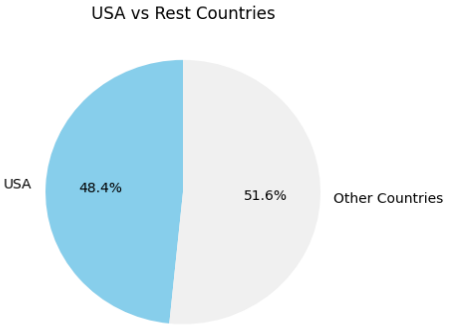
Interpretation: Social media sentiment leaned more favorably towards Biden.

However, we can't conclude anything yet, since the number of positive sentiments for Trump might be more than for Biden.

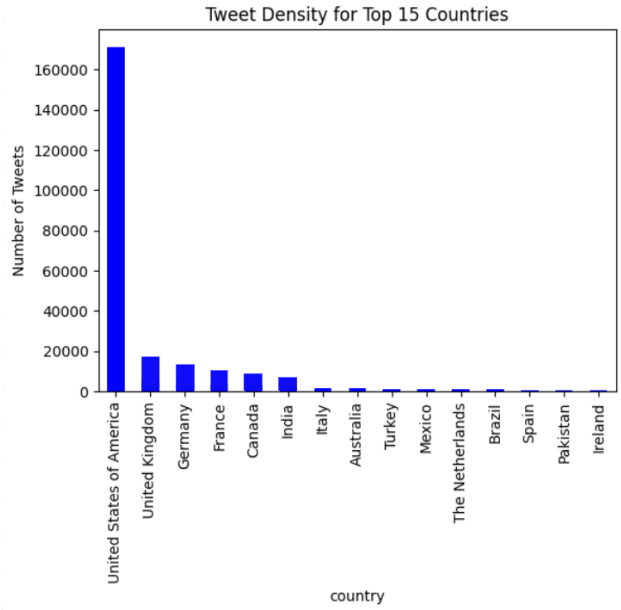
---

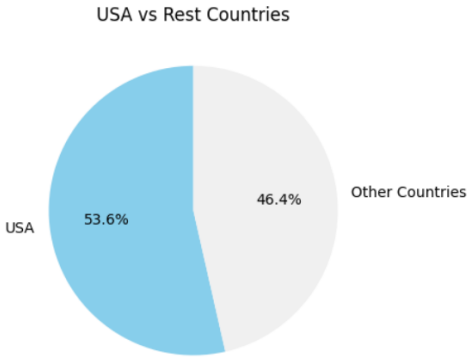
b. Geographic Tweet Density





Trump





Biden

- **Bar Graphs (Figures 2 & 3)** show tweet counts by country.
    - USA dominates the dataset, followed by UK, Germany, France, Canada, and India. Other countries contributed marginally compared to the USA.
- **Pie Charts (Figures 2 & 3)** break down the **USA vs Other Countries**:
    - In Trump's dataset: USA contributed 48.4%, others 51.6%.
    - In Biden's dataset: USA contributed 53.6%, others 46.4%.

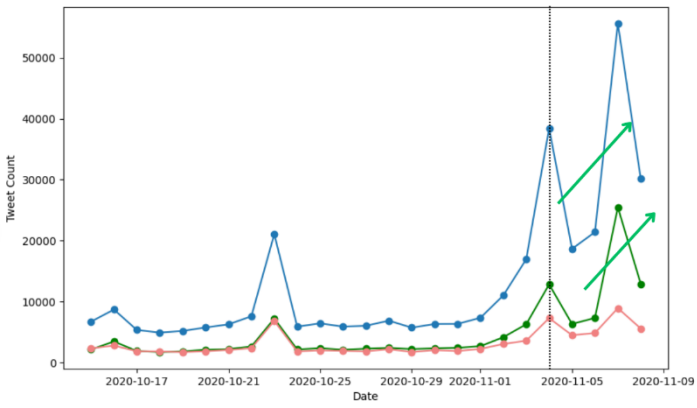
Interpretation: Biden's support and conversations were slightly more centered in the USA than Trump's, which is critical since the USA was the deciding ground for the election.

---

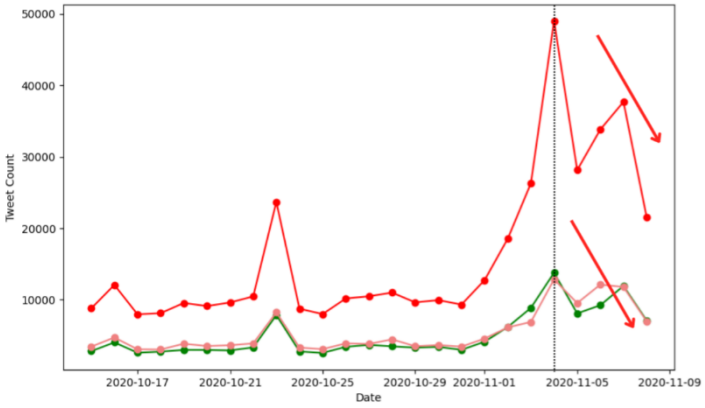
c. Temporal Trends in Tweet Activity

- **Line Graphs (Figures 4 & 5)** show tweet volumes from mid-October to the election week (Nov 3–4, 2020).

Biden



Trump



- Key Observations:
    - Both candidates experienced spikes in tweet volume in the days leading up to the election.
    - Trump's tweet activity was consistently higher overall, reflecting his polarizing nature and stronger online engagement.
    - Biden's positive tweets spiked significantly between November 2 and 5, showing strong pre-election support momentum.
    - Trump's negative sentiment line was consistently above Biden's, showing he received more criticism online.

After election day, Trump saw a decline in tweet count, whereas Biden was consistently increasing. Showing positive indications for Biden.

Interpretation: While Trump generated more overall buzz, Biden's late positive momentum could be more influential in swaying undecided voters near election day

5. Sentiment Indexing & Engagement Metrics

a. Sentiment Indexing

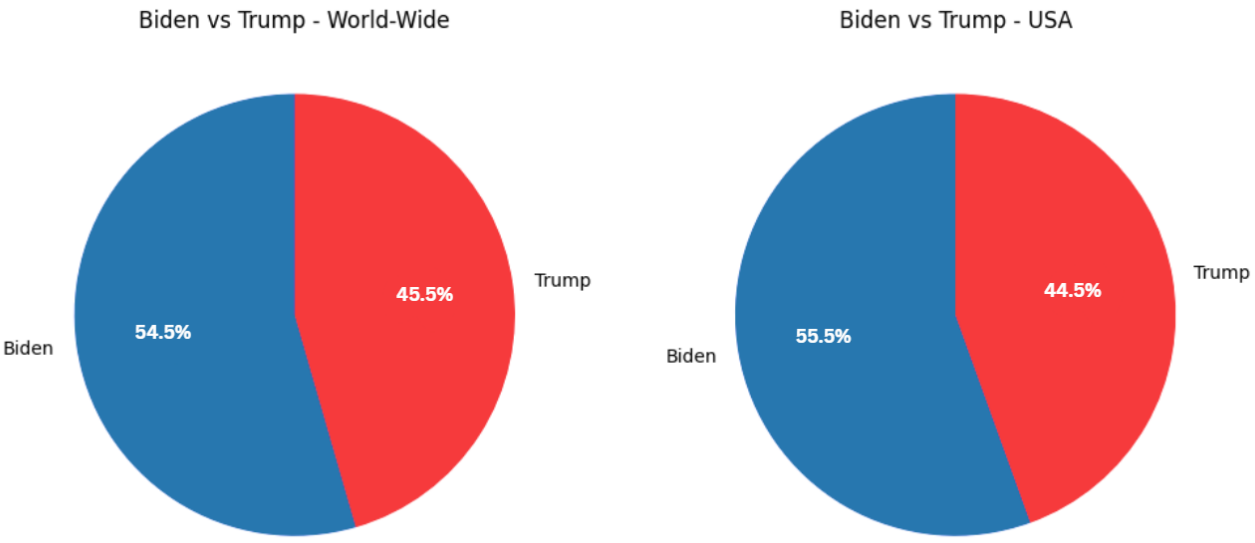
- Defined sentiment index as:
- $$\text{Sentiment Index} = \frac{\text{Positive Tweets}}{\text{Total Tweets}}$$
- Findings:
  - **Biden Index**: ~38.1% (USA).
  - **Trump Index**: ~30.5% (USA).

This shows that Biden had a higher proportion of positive sentiment among total conversations compared to Trump.

---

b. Comparative Sentiment Share (Figure 6)

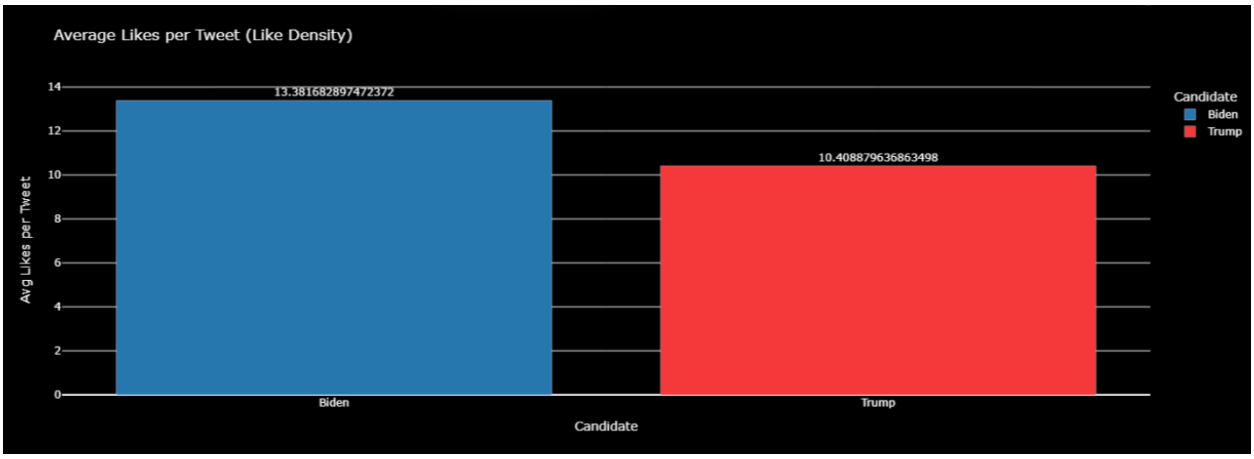
- **Worldwide:** Biden 54.5%, Trump 45.5%.
- **USA:** Biden 55.5%, Trump 44.5%.



Interpretation: Biden consistently had a higher positive sentiment share both globally and in the USA. The USA-specific chart is critical since votes are cast domestically.

c. Engagement Metrics

- Engagement density was calculated as tweets per unit audience share (measuring how engaged people were in discussions around each candidate).
- Results:
    - Biden → 13.38
    - Trump → 10.41



Despite Trump generating higher raw volumes, Biden's followers and supporters engaged at a **higher intensity relative to his base**.

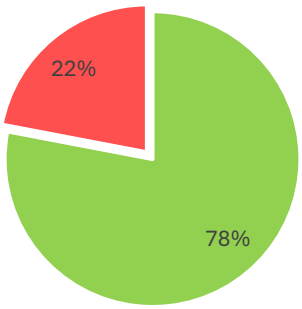
---

d. Model Evaluation

- To validate VADER's accuracy, we manually labeled a *sample of 200 tweets*. (100 for Biden and 100 for Trump)
- Result: *78% accuracy*, confirming that while not perfect, VADER provided a reasonable representation of public opinion.

Link evaluation: https://github.com/lakshit2508/twitter_sentiment/blob/main/sample_n200_check1.csv

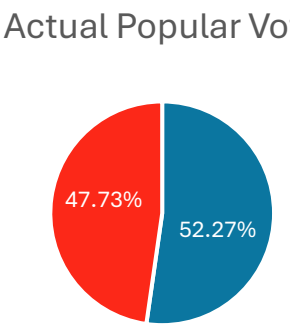
Biden	Trump	Total (accurate)	total values	accuracy
81	75	156	200	78%

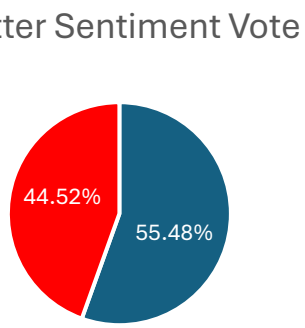


Results

1. Biden had higher positive sentiment both globally and in the USA.
2. Trump had a higher proportion of negative sentiment compared to Biden.

   and he was slightly more popular in other countries compared to the USA.
3. Engagement density was also higher for Biden, showing stronger public involvement.
4. Based on Trend analysis, Trump's positive sentiments were higher only for 2 days (1 day before and on election day); on the other hand, Biden had higher.
   and Trump's tweet volume rapidly decreased after the election; however, Biden's tweets remained consistent.
5. Based on sentiment analysis, Biden was projected to win the 2020 election.





6. Manual evaluation with a sample of 200 tweets showed VADER achieved 78% accuracy in classification.

Limitations & Potential Improvements

- **Limitations**:
  - Tweets do not represent the entire voter base (sampling bias).
  - VADER struggles with sarcasm and complex linguistic nuances.
  - Non-English regions like New Mexico are not represented well.
  - Presence of bots or coordinated campaigns may skew results.
  - Country location data was sometimes inconsistent or missing.

- **Improvements**:
    - Use **deep learning models** (e.g., BERT, RoBERTa) for more nuanced sentiment analysis.
    - Expand the dataset to include Facebook, Reddit, or news comments.
    - Apply **bot detection algorithms** to reduce noise.
    - Incorporate **geographical weighting** (e.g., focusing more on swing states in the USA).
    - Increase manual validation sample size for better accuracy estimation.