NYU

PerceptionWave

Understanding Social Media's Role in Shaping Public Opinion in Elections

1. Team Members

Qile (Asher) Zhou - qz2726
Dominique Eberhard - de814
Sai Nishanth Mettu - sm11326
Neha Saji Jacob - ns6266
Prachet Balaji - pb2882
David Liu - dl4155
Course Instructor – Juan C. Rodriguez

2. Introduction

Social media has reshaped how people get information, discuss issues, and form opinions. Platforms like Twitter and Reddit have become spaces for fast communication and user-generated content, sparking debates on whether they genuinely influence public opinion and democratic processes. Do the countless posts often charged with specific viewpoints , actually sway people's beliefs, or do most users stick to their original views, filtering content to match their own beliefs?

Our project will dive into these questions, examining whether constant exposure to one-sided, politically charged content can gradually shape opinions or reinforce pre-existing biases. We'll analyze the sentiment and bias in social media posts and compare this to public opinion data from polls. Through this, we aim to understand if and how social media molds public opinion and explore what this means for democracy.

3. Project Objectives

The goal of this project is to study the presence and impact of bias in social media content. Using sentiment analysis, we'll examine the tone and emotional slant in posts to uncover patterns in bias based on sentiment, topic, and political leaning.

We'll then compare our findings to public opinion polls to see if shifts in social media sentiment align with changes in public views. This will help us understand whether social media influences opinions or merely amplifies existing beliefs.

Finally, we'll explore whether certain groups are more vulnerable to these biases, shedding light on how social media might shape public discourse and attitudes.

4. Data Sources

To analyze digital discourse, we'll use public data from major social media platforms:

- **Twitter**: With its real-time, high-volume posts, Twitter captures public sentiment on trending topics in short, direct tweets, making it ideal for sentiment and bias analysis.
- **Reddit**: Known for topic-specific communities, Reddit allows for deeper discussions and shows the intensity of opinions through upvotes, downvotes, and comments. This will help us explore bias across different communities and viewpoints.

Together, Twitter and Reddit will provide a rich dataset for a detailed look at social media perspectives and how they might align with or differ from mainstream public opinion.

5. Technologies Used

- **Frontend Development**:
    - *HTML, CSS, JavaScript*: To build a user-friendly, interactive web interface where users can view the sentiment analysis results, explore detected biases, and interact with the data in a meaningful way. The frontend will allow users to filter and visualize trends, biases, and sentiment over time and by platform.
- **Data Processing and Analysis**:
    - *PySpark*: Given the scale and velocity of social media data, PySpark is essential for parallelized data processing, enabling us to handle large datasets and perform efficient sentiment and text analysis.
    - *PyMongo*: To store and retrieve social media data efficiently, we will use MongoDB, a NoSQL database that offers flexibility and scalability. PyMongo will facilitate the interaction between MongoDB and PySpark, making it easier to store raw data and retrieve processed results for visualization.
- **Data Streaming**:
    - *Kafka*: To support real-time data ingestion and processing, we will use Apache Kafka to manage continuous streams of social media posts. Kafka will allow us to capture live data feeds, process them in near-real-time, and ensure that our sentiment analysis reflects current trends.
- **APIs**:

- *Twitter API*: This will enable us to access and retrieve tweets based on specific keywords, hashtags, and user-defined parameters, allowing us to curate a dataset that aligns with our research focus.

- *Reddit API*: Similar to the Twitter API, the Reddit API will enable us to collect posts and comments from targeted subreddits. By focusing on relevant communities, we can gather data that reflects discussions on political and social topics.

Together, these tools provide a robust infrastructure for managing, processing, and analyzing our social media dataset.

6. Expected Outcomes

The project will result in a web application with two main features:

1. **Bias Detection**: It will analyze social media posts to spot biases in topics, emotional tone, and political leanings, helping users see trends across platforms.

2. **Public Opinion Comparison**: By comparing social media trends with public opinion data, users can observe if social media sentiment aligns with broader public attitudes, especially around key events like elections.

This tool will help people understand social media's impact on public opinion and could guide policymakers, platform developers, and researchers in promoting healthier, more balanced online discussions.

7. Why is this a Big Data Project?

This project is a Big Data endeavor due to the large scale and speed of social media content from platforms like Twitter and Reddit. We'll analyze a massive dataset of posts and comments across topics and viewpoints, capturing diverse perspectives from different regions and demographics. The volume, velocity, and variety of this data require specialized Big Data tools for storage and analysis.

Since we'll process data in real-time, as new posts come in continuously, we need technologies like Apache Kafka for data streaming, PySpark for distributed processing, and MongoDB for handling large, unstructured data.

To analyze sentiment and detect bias, we'll use advanced NLP techniques within a Big Data framework, enabling us to track trends and gain insights from these vast datasets. This approach tackles Big Data's core challenges and uses the right tools to make sense of large-scale social media influence.