Project Write-Up

437 Group:Jean Dong Cho,Samuel Zhang,Ryan Luder,Charles Allbritton

Thought Process

Our team embarked on this machine learning project with a clear motivation: to delve into the nuances of political discourse on social media. In the era of information overload, understanding how language, vocabulary, and sentence structures differ among supporters of different political parties becomes increasingly crucial. Our goal was not just to analyze tweets from Republicans and Democrats but to uncover meaningful insights that could foster a deeper understanding of the distinct viewpoints held by these two groups.

Problem Statement

Our central question was multifaceted: How does language usage differ between Republican and Democratic party supporters? Can machine learning accurately classify tweets into their respective political parties based on linguistic patterns? Additionally, what valuable insights can be extracted from applying Natural Language Processing (NLP) techniques to political discourse on platforms like Twitter?

To address these questions, our team sought to leverage a diverse set of tools and techniques. From initial data acquisition to ethical considerations, we aimed to build a robust framework that not only yields accurate predictions but also respects privacy and mitigates biases.

Data Acquisition and Cleansing

We sourced our dataset from Kaggle, a reputable machine learning dataset platform, emphasizing transparency in providing details such as the dataset's URL, name, and context. The dataset underwent meticulous cleansing, a critical step to ensure data integrity. This process involved handling missing values, removing duplicates, and addressing outliers. Ethical considerations played a significant role, prompting measures to protect privacy, anonymize data, and address potential biases during analysis.

Techniques and Methodology

Our chosen techniques included Natural Language Processing (NLP) tools, specifically the Natural Language Toolkit (NLTK) in Python. The preprocessing phase involved not only standard text cleaning procedures, such as removing links and non-alphabetic characters but also advanced techniques like lemmatization to ensure consistency in word usage.

The analysis phase featured the implementation of bag-of-words and term-frequency-inverse-document-frequency (TF-IDF) representations. These techniques allowed us to extract meaningful patterns in the textual data, revealing not only the occurrence of words but also their importance in distinguishing between Republican and Democratic tweets.

The neural network classifier, a key component of our project, was designed to provide a nuanced understanding of party classification. We meticulously defined the architecture, activation functions, and regularization techniques to optimize model performance.

Empirical Insights and Results

The empirical phase of our project provided fascinating insights into the linguistic disparities between Republican and Democratic supporters. Visualizations, including scatter plots and word frequency charts, revealed interesting patterns. For instance, the trimmed scatter plot highlighted words with significant deltas in frequency between the two parties, offering a more focused view.

The machine learning model, trained using the Multinomial Naive Bayes classifier, demonstrated promising results in accurately classifying tweets into their respective parties. Classification reports provided detailed metrics, showcasing the model's performance.

Our custom Naive Bayes classifier further enriched our understanding of the underlying patterns in the data. The choice of the smoothing parameter (alpha) played a crucial role, with experimentation revealing optimal values for enhanced performance.

Data Visualization

To effectively communicate our findings, we utilized various data visualization techniques, including scatter plots. These visualizations not only enhanced the interpretability of our results but also provided a clear narrative for readers.

Conclusion and Future Directions

In conclusion, our project holds the potential to contribute significantly to the field of political discourse analysis. By combining data preprocessing, NLP techniques, and machine learning classification, we've gained valuable insights into how language reflects political affiliations on social media. Our commitment to ethical standards and privacy considerations ensures that our contributions are not only insightful but also responsible.

As we submit this project, we recognize that it represents a snapshot in time, and the landscape of political discourse on social media is dynamic. Nevertheless, our work serves as a foundation for future research and discussions, fostering mutual understanding through the lens of language.