Sentiment Analysis Using Twitter Dataset

(Exploring the emotions associated with political parties in the United States)

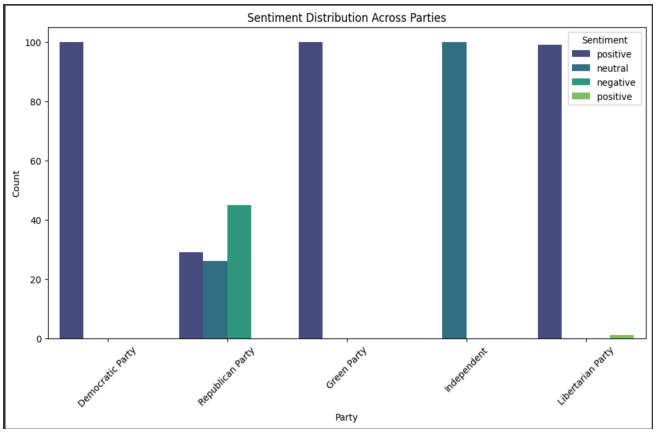
Questions to Explore: Were certain politicians received more negatively? More positively? How did sentiments change during the election cycle from the beginning of the dataset to the end? What words were commonly associated with each politician? Are there any deeper insights we can gather from this dataset that speak about this particular platform? Can we use an ML model to correctly classify tweets to gather a birds-eye view of user sentiment on the app?

Methods:
Data exploration: Used bar charts, graphs, word clouds, time series charts
Model Selection: Support vector Machines(effective in high diminsions), Logistic Regression(simplicity/interpretability, most comfortable in), Random Forest(handles non-linear data well, feature importance, most experience using this)

Results:

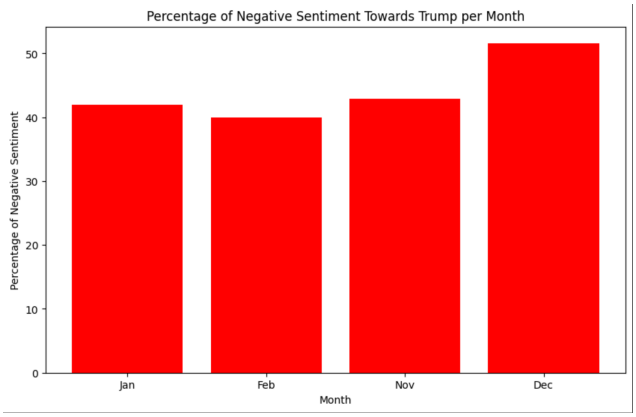
Visualizations:

Graph 1:



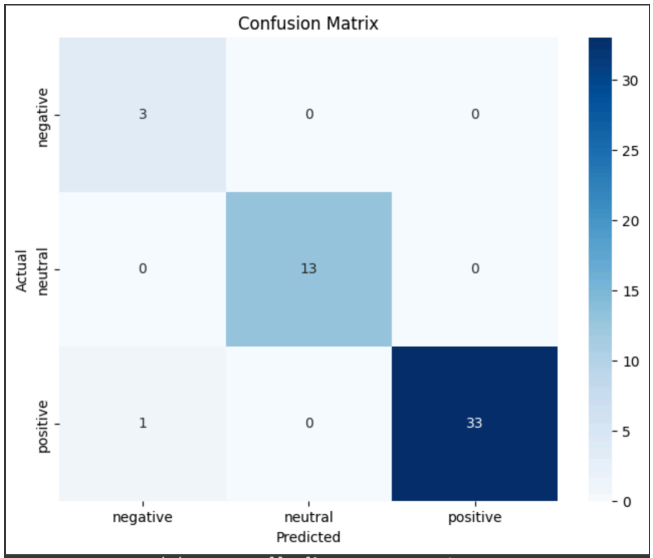
The Sentiment Distribution Across Parties graph visualizes the sentiment (positive, neutral, negative) associated with the different U.S. political parties. The Democratic, Green, and Independent parties show overwhelmingly positive or neutral sentiment, while the Republican Party has a more balanced sentiment distribution, including a higher portion of negative sentiment. The Libertarian Party has very low representation in the dataset. By looking at these distributions, it seems there is potential data imbalance or bias, where certain parties may be overrepresented in positive sentiment.

Graph 2:



Interestingly, Trump was the only candidate with negative tweets, making him the only one for whom this visualization could be created. The chart represents the proportion of negative tweets relative to total tweets, tracking sentiment shifts over time. However, due to the sparse data, this provides only a limited snapshot of the actual sentiment on Twitter and may or may not fully reflect broader public opinion.

Models



	precision	recall	f1-score	support
negative	0.75	1.00	0.86	3
neutral	1.00	1.00	1.00	13
positive	1.00	0.97	0.99	34
accuracy			0.98	50

The SVM model performed exceptionally well on the test set, achieving 98% accuracy through grid search. However, its performance was slightly lower when classifying negative tweets. Overall, the model effectively identifies whether tweets about a political candidate are positive, negative, or neutral, making it valuable for tracking public sentiment toward political parties on platforms like Twitter. The selected hyperparameters optimized SVM performance. C=1 balances margin and misclassification, preventing overfitting. Gamma='scale' adjusts based on data for better boundaries. The linear kernel suits linearly separable text, making it effective for sentiment analysis.

Conclusion:

In conclusion, the dataset was too small, limiting the depth of analysis. Class imbalances, the number of tweets, and overall sentiment distribution restricted insights. In the

future, I will seek a larger dataset and apply my own preprocessing to ensure a more balanced and useful dataset, enabling meaningful and minimally biased visualizations.