DID A REPUBLICAN, OR DEMOCRAT POST THAT?

By Shanell Jones

Data Scientist

PROBLEM/WHY WE ARE HERE

- Problem that we are here to solve
  - Given a Reddit Post, did it come from a Republican or Democrat?
- Agenda
  - How I chose data
  - How I cleaned the data
  - What sentiment analysis and EDA did I do beforehand
  - How did I choose my models and what parameters did I optimize
  - What model did I choose and how did my confusion matrix perform
  - What posts could my model not categorize?
  - How did my coefficients perform?
- Data Sources
  - Democrat Subreddit
  - Republican Subreddit

REDDIT CHOICE AND MODEL SCRAPPING

- Subreddits were Democrats and Republicans
- Used the json API in order to perform the extraction
- For loop was written to mine 1000 posts per subreddit with a 5 second sleep time
- Used a function in order to pull the titles for each subreddit based on the children's tag
- Placed each one of them in a pandas data frame

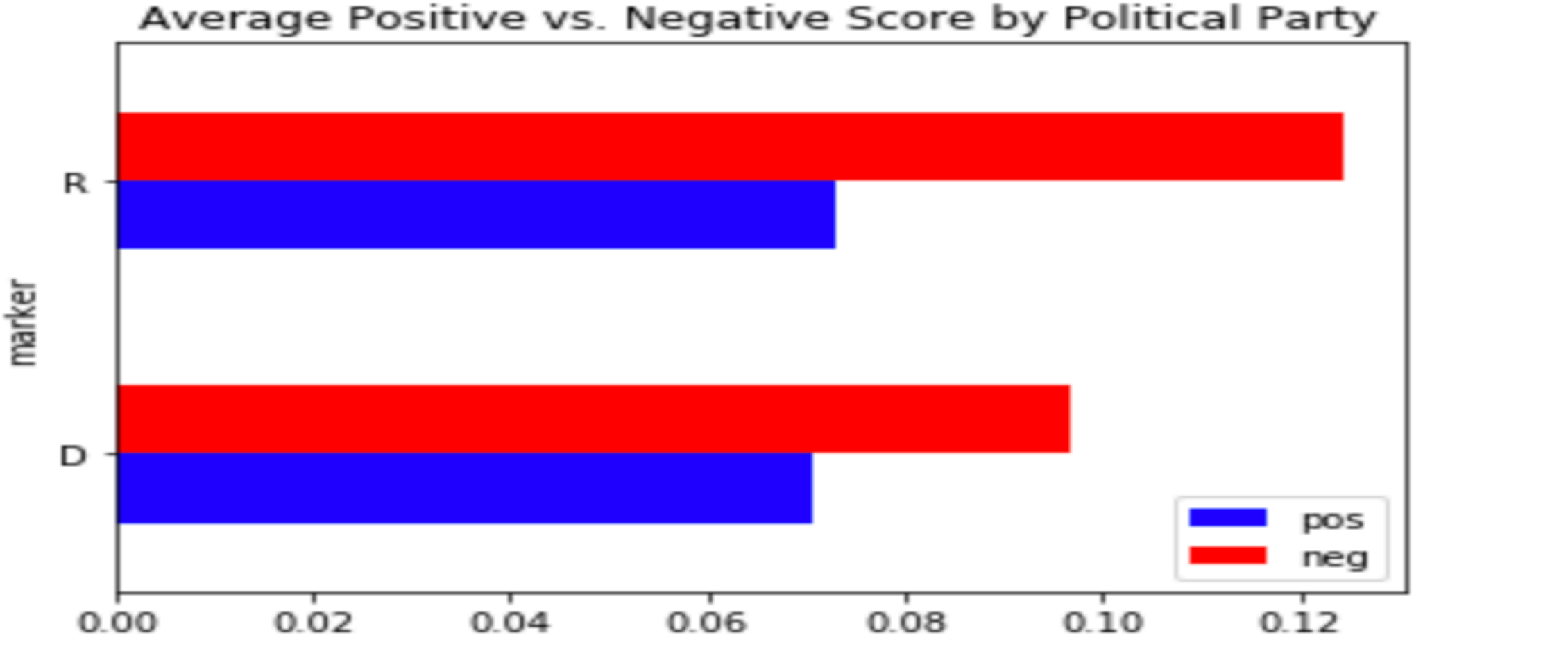
HOW WE CLEANED DATA

- Imported both subreddits from each csv file
- Created a column to identify each one as the titles as a Democrat or Republican
- Checked for nulls and NAs
- Removed any duplicates
- Adjusted the index
- Merged both democrats and republicans to a single data frame

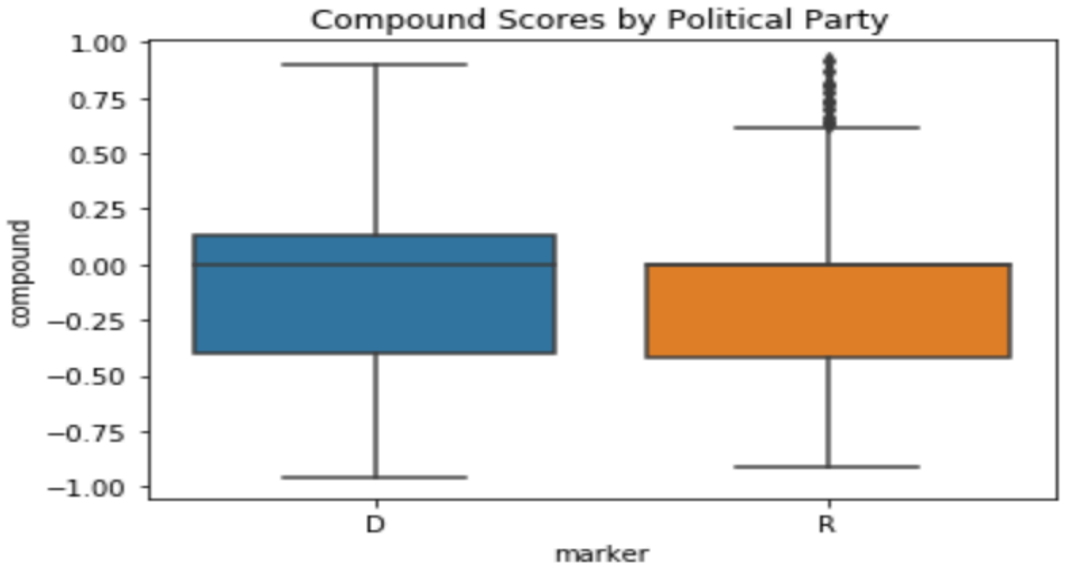
EDA FEATURE GENERATION AND INSIGHTS

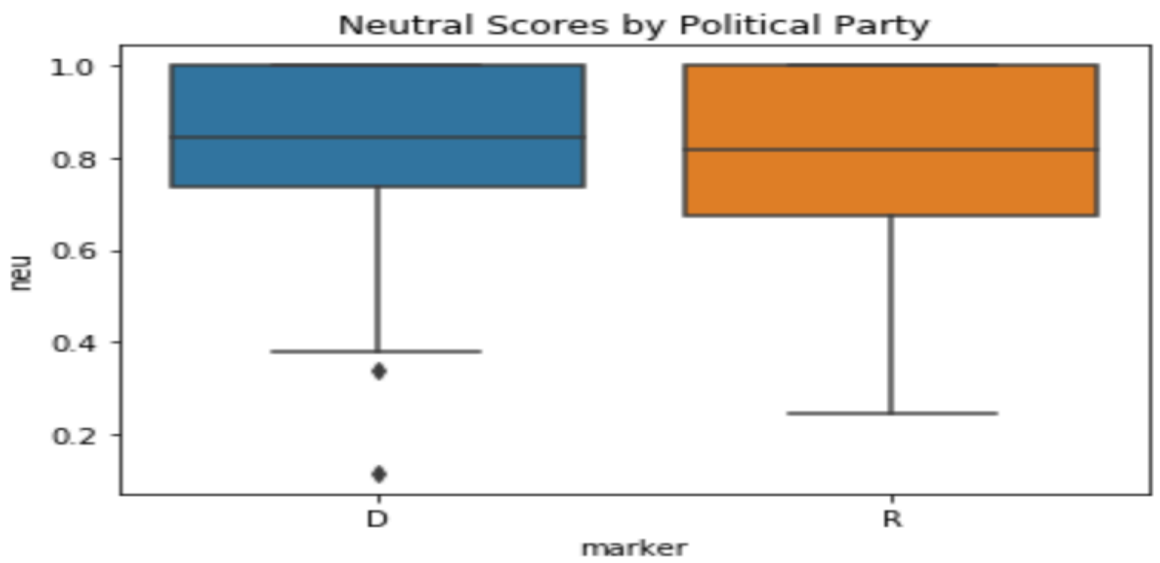
- Features
  - Created a count length that analyzed the number of characters in a post
  - Found that democrats have longer posts overall
- Used the Vader sentiment analysis to analyze positive and negative impacts of posts
  - Scoring is done from 0 to 1 for positive, negative and neutral ratings within posts
  - Compound Scores showing overall scoring is from -1 to 1
  - Republicans were found to have slightly more impactful posts overall
    - Tend to use more heightened language

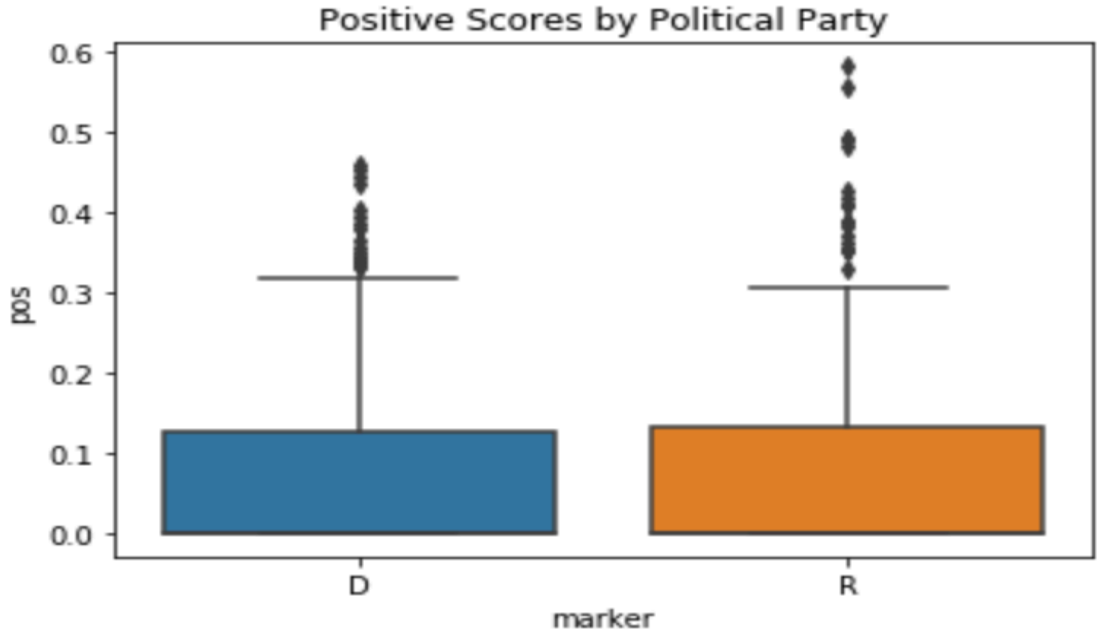
POSITIVE AND NEGATIVE DISTRIBUTION

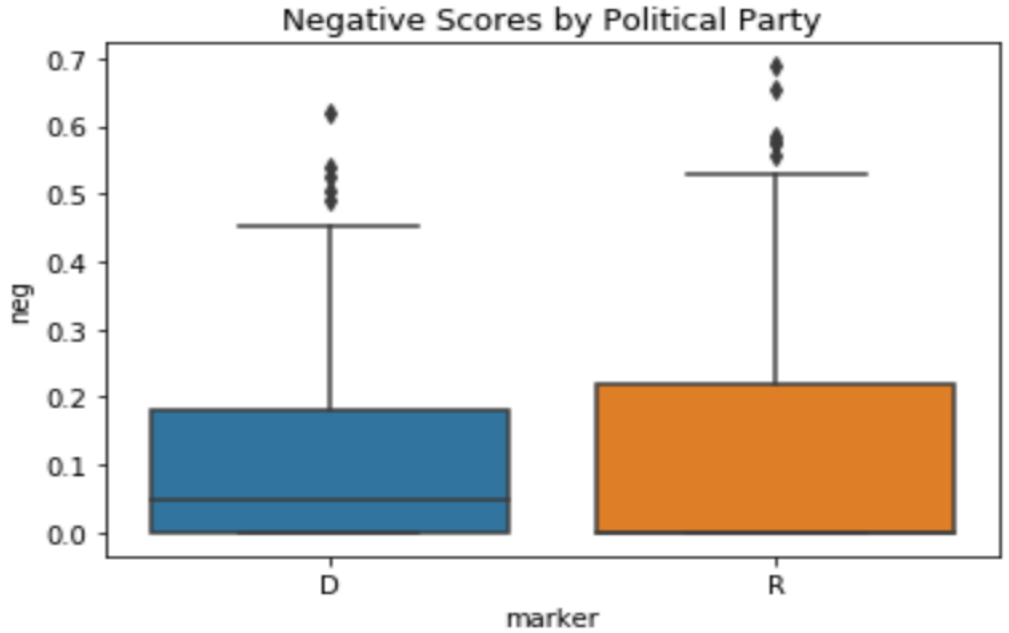


SENTIMENT ANALYSIS BOX PLOTS

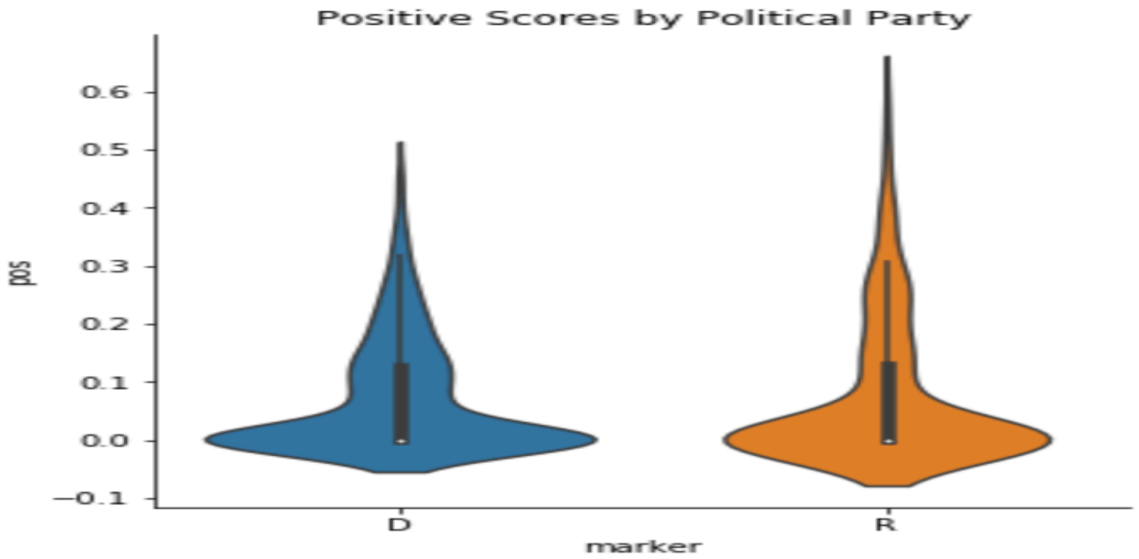


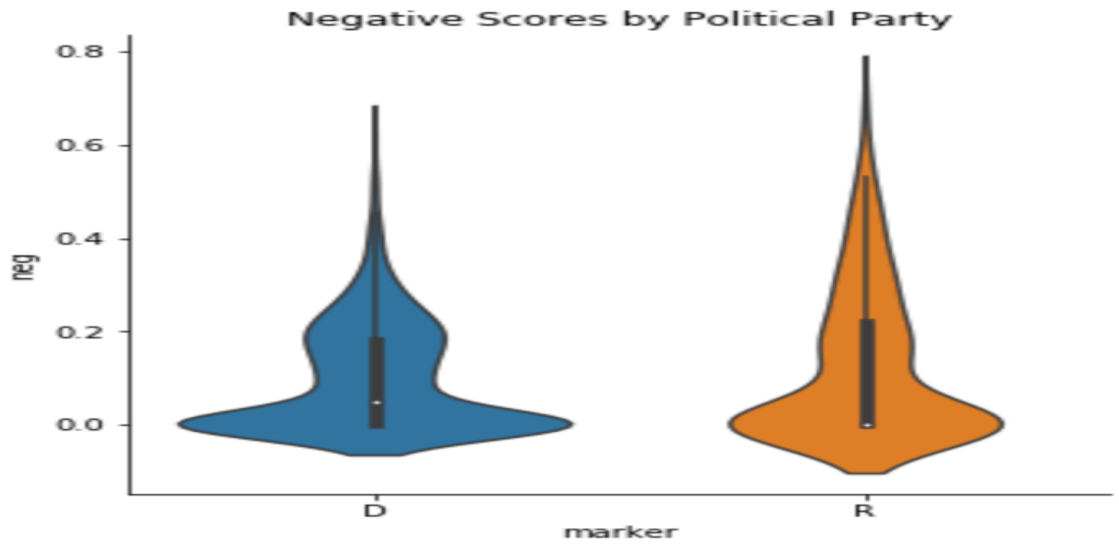


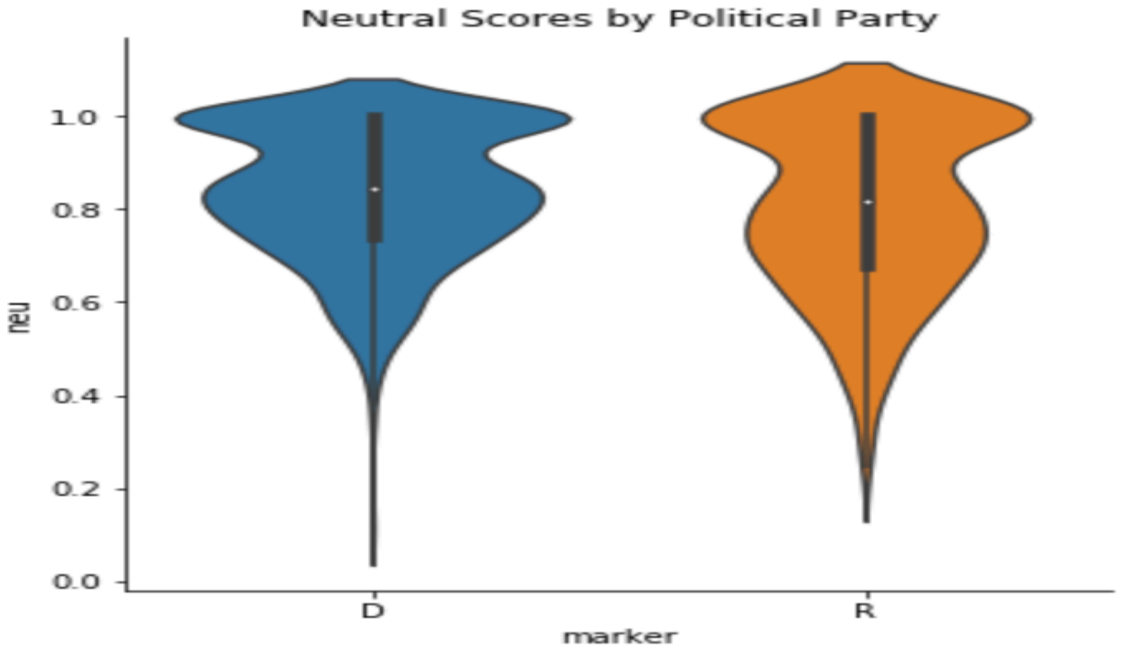


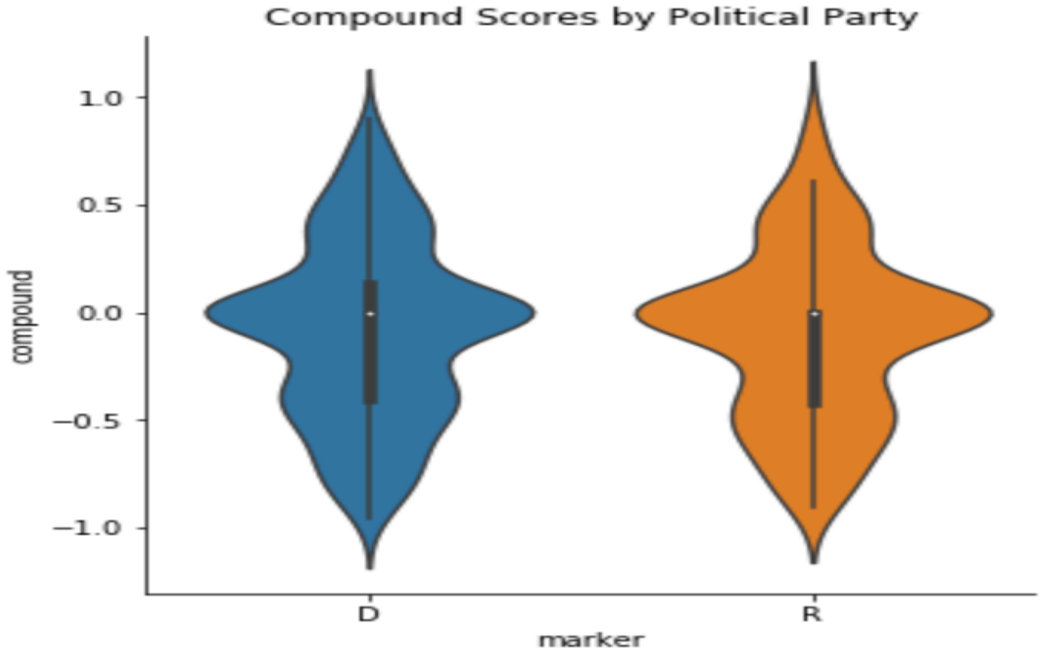


SENTIMENT ANALYSIS VIOLIN PLOTS

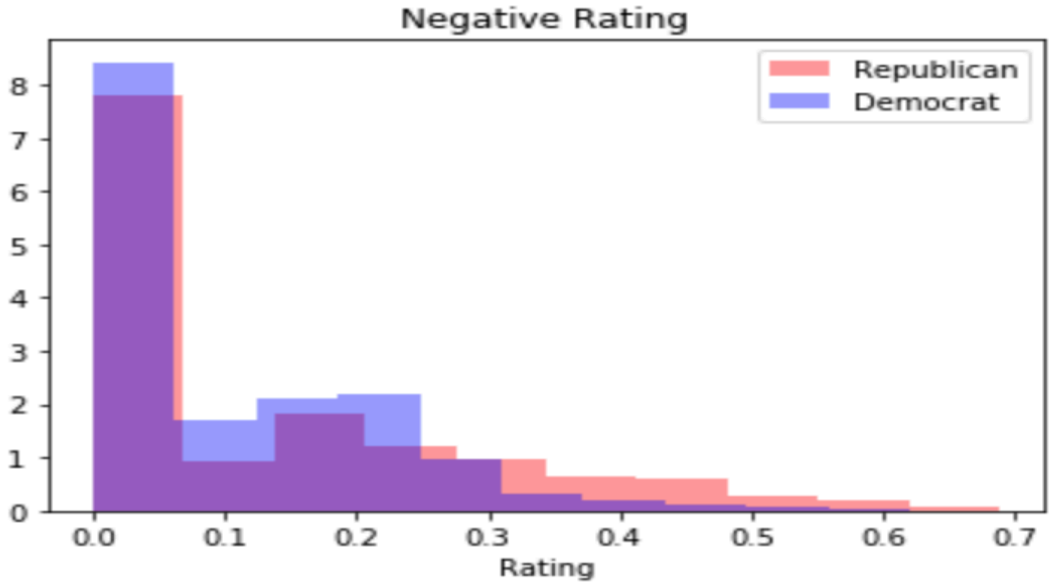


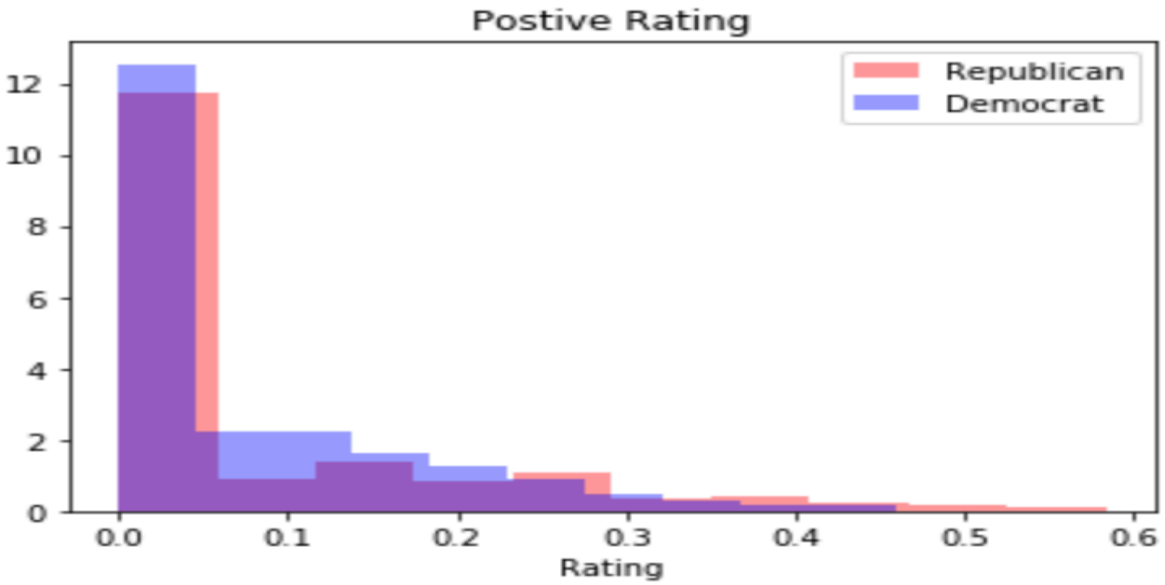


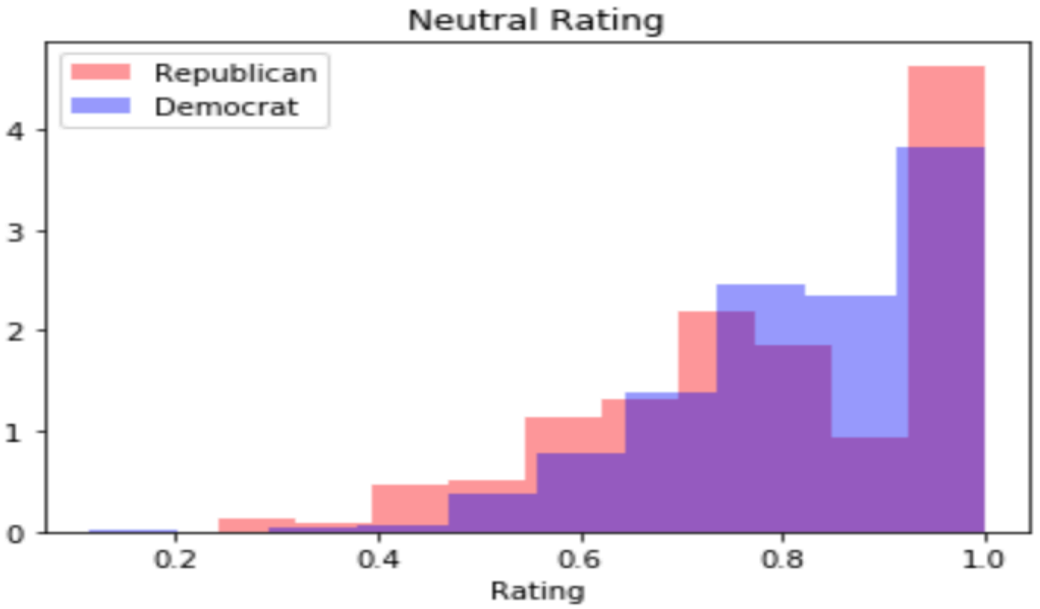


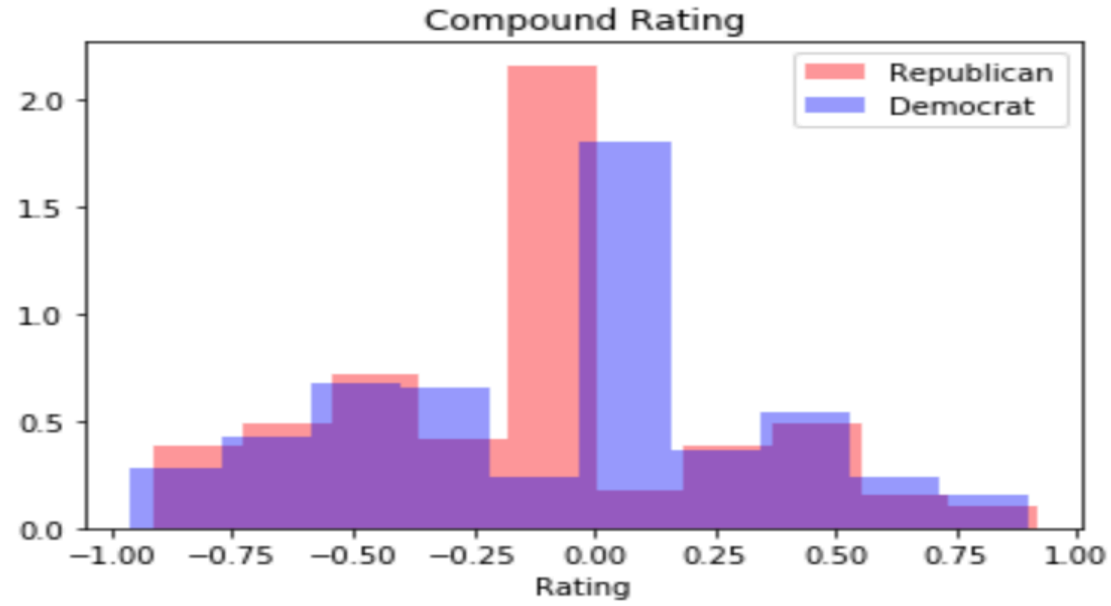


SENTIMENT ANALYSIS HISTOGRAMS

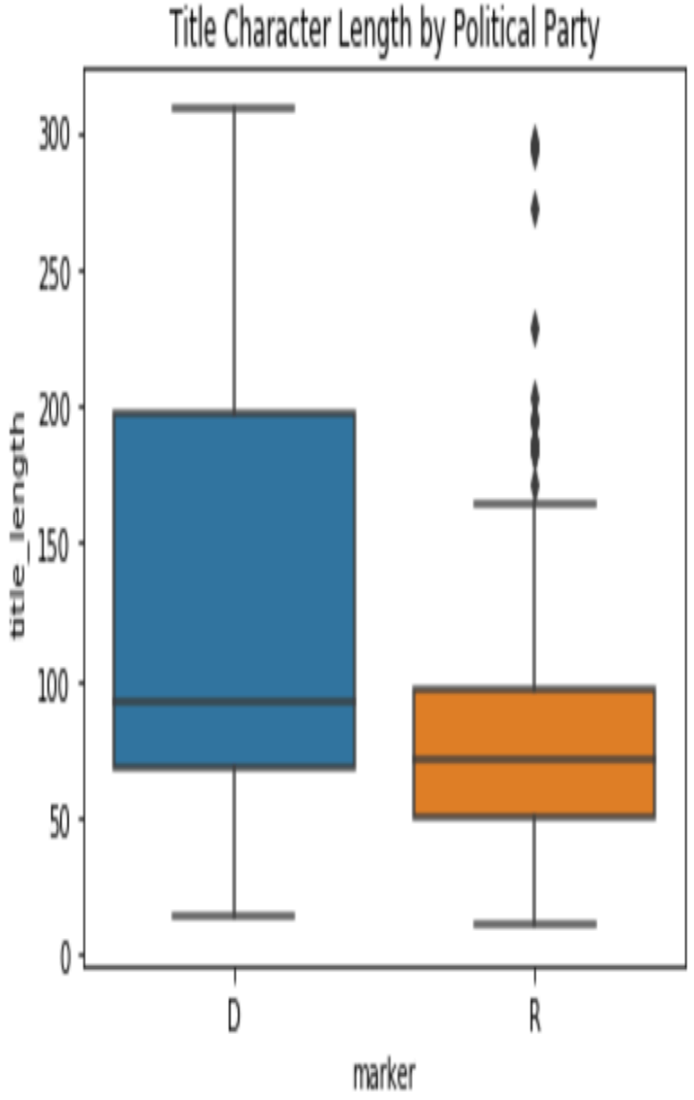


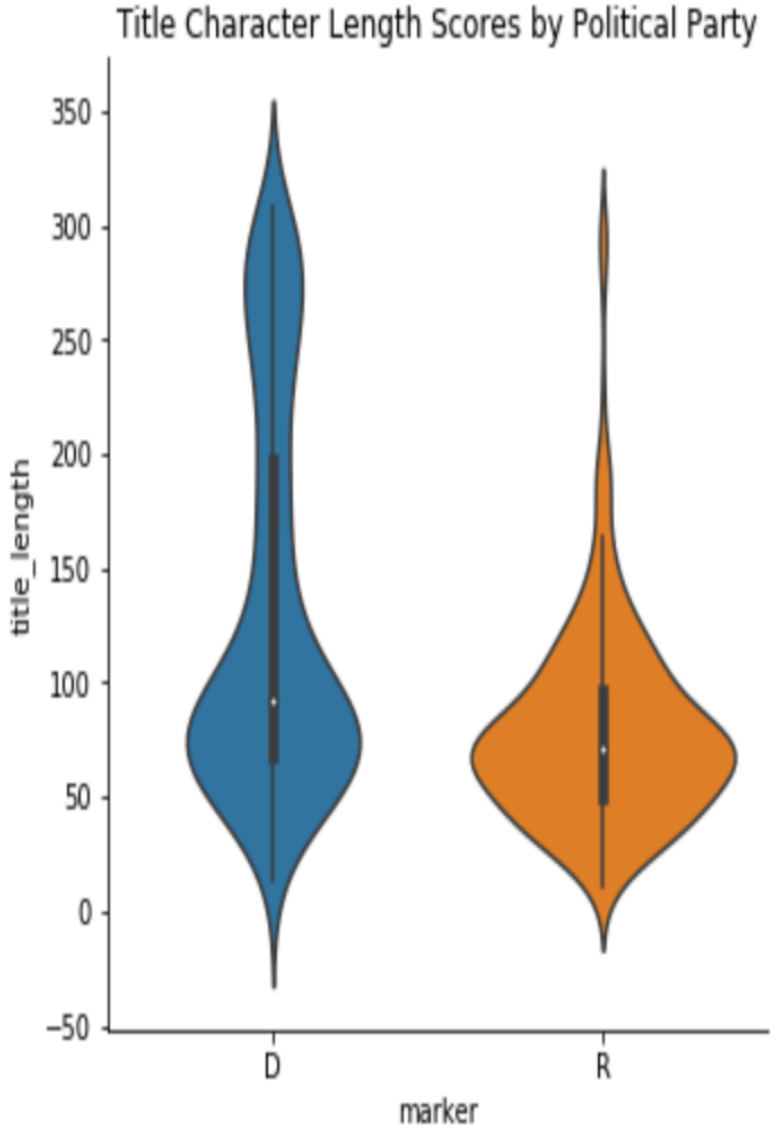


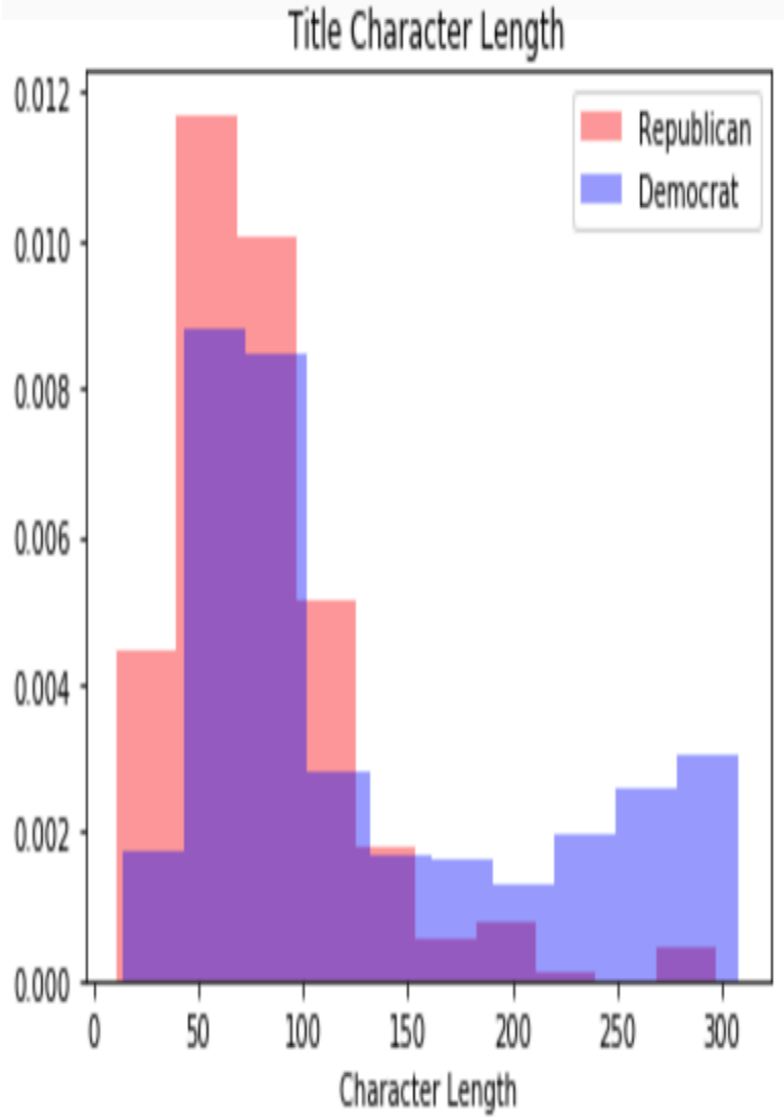




CHARACTER LENGTH PLOTS







MODEL CHOICES AND PARAMETER CHOICES

- Baseline Metric Score was 67% due to democrat titles having a 2 to 1 advantage over Republicans
  - Given that 1/3 of the posts are republicans it is only far that the unbalanced class is the applied target
  - Republican titles had significant amount of copies and upvotes that did not allow for analysis
- Vectorizors
  - Count Vectorizer and TfidfVectorizer
- Models
  - Logistic Regression, K Nearest Neighbors, Navies Bernoulli, Navies Multinomial
- Parameters
  - Vectorizers were adjusted for n grams and stop words
  - Logistic Regression was adjusts for a C square
  - K Nearest Neighbors was adjusted for the p-value and nearest neighbors
- Scores
  - Before parameter tuning, models trained at 90%-100% but trained at 70%
  - After hyperparameter tuning, training scores dropped to 70%-80%,but test scores raised to 75%

MODEL SELECTION AND EFFECTS

- Model that performed the best was a Count Vectorized Logistic Regression Model using the below parameters
  - Contains stop words
  - N gram grouping of 3
  - C value of 10
- Performed at a 75% training and test accuracy to limit overfitting
  - Other models performed within the 60% to 70% range
- All models when grid searched had I variance tradeoff for minimal gains in bias
- Model has the greatest categorization of highly correlated words with minimal errors
  - Categorical variables arose when highly correlated words were used in different context

CONFUSION MATRIX AND METRICS

		Predicted	
		Positive	**Negative**
Actual	**Positive**	**43**	**36**
	Negative	**27**	**129**

Metrics

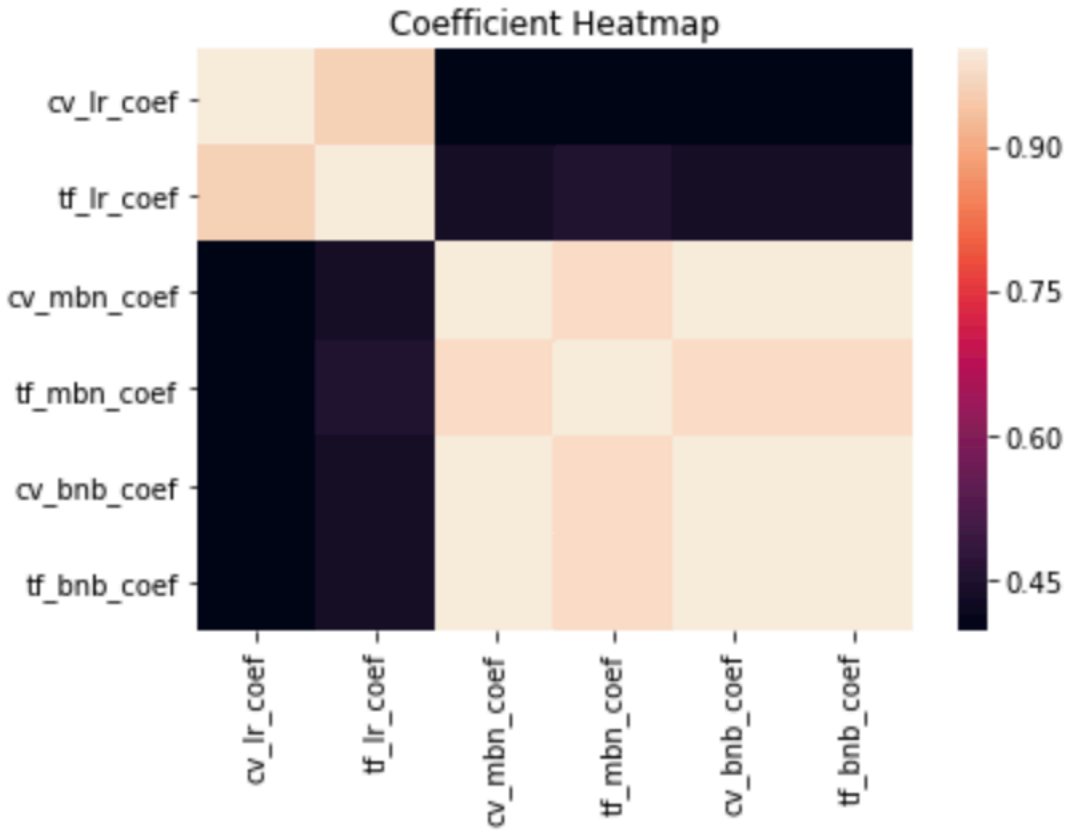
- Error Rate-27%
- Accuracy Rate-73%
- Recall-46%
- Specificity-83%
- Percession-39%

CONCLUSIONS AND FURTHER PLACES OF STUDY

- Conclusions:
  - Model currently performs at 75% accuracy.
  - Posts are having trouble performing sentiment analysis using words alone considering context of the sentence are not being taken into account.
  - Words with stronger context are leading to misclassification
- Republican Misclassifications:
  - The Lights Are Out in California, And That Was the Plan All Along
  - Second whistleblower about Ukraine phone call coming forward. Uhm, dumbasses, we have the damn transcript.
- Democratic Misclassifications:
  - Trump gives green light to Turkey to attack.
  - Biden's Most Formidable Opponent Is Not Another Democrat - Questions about his age have dogged the former vice president throughout the primary.
- Further Investigations:
  - Inclusions of similar subreddits such as other political parties or party specific beliefs
  - Adds a Bayes Stats model to take into account words used in context
  - Adding Text Blob
  - Inclusion of Stemming and Lemmatizing

COEFFICIENT DISTRIBUTION AND BEHAVIOR

- Strongest Correlations were between the Count Vectorizer and TFIDVectorizer changes to the models
- Also additional strong correlation between similar naïve bays models
- This is due to the tradeoffs of other models such as the Bernoulli only allowing ones and zeros when others using stronger coefficients



HIGH CORRELATION SCATTERS