PROJECT 3

CLASSIFICATION OF SUBREDDITS

AGENDA

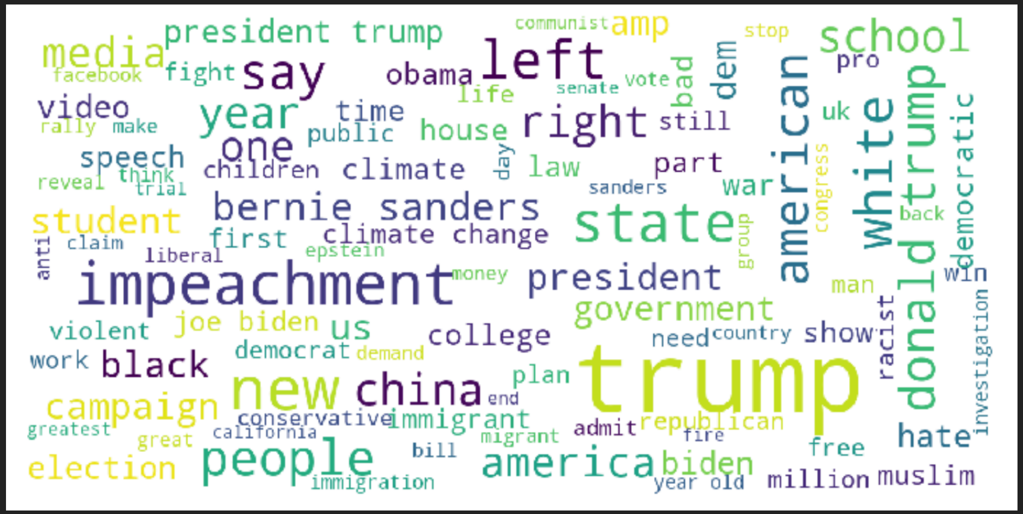
- Problem Statement
- Subreddits
- Cleaning the data
- Building the Models
- Chosen Model
- Recommendations

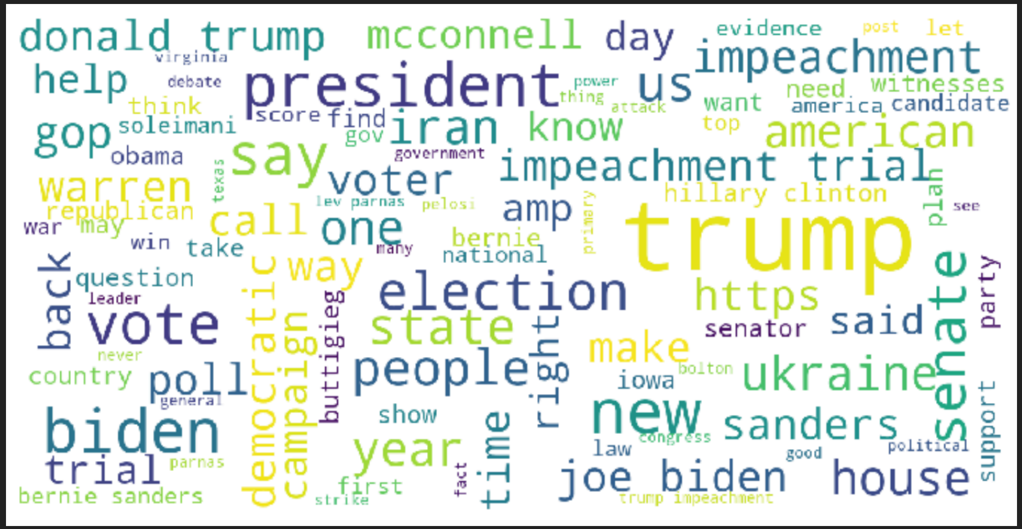
PROBLEM STATEMENT

- To classify respective posts into the correct subreddit
- What words are mostly used and related to the subreddit

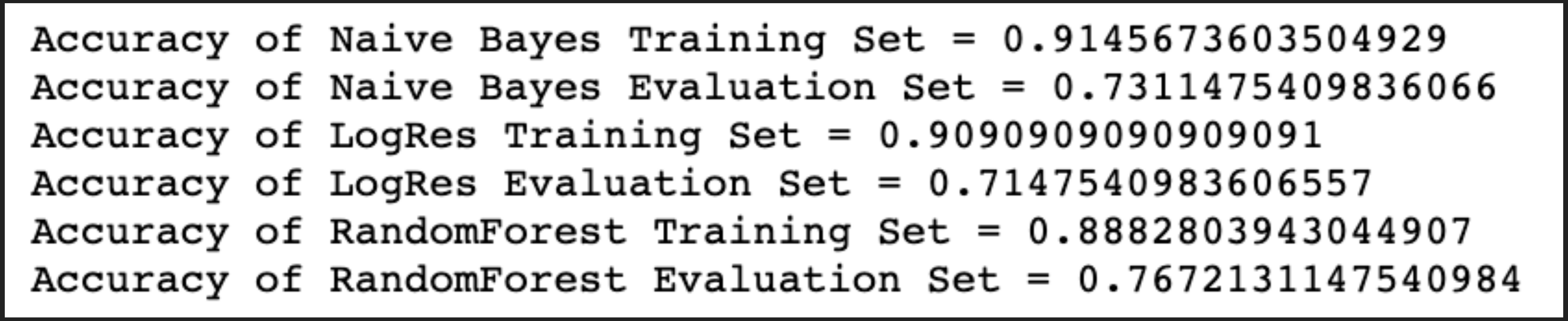
CLEANING DATA

- Drop all Null Values (Pictures)
- Use Regex to remove all Punctuations and Symbols
- Remove topic words from the text (“Democrats”, “Republicans”)
- Tokenize the text to form individual words
- Remove Stopwords (nltk StopWords)
- Lemmatize the words



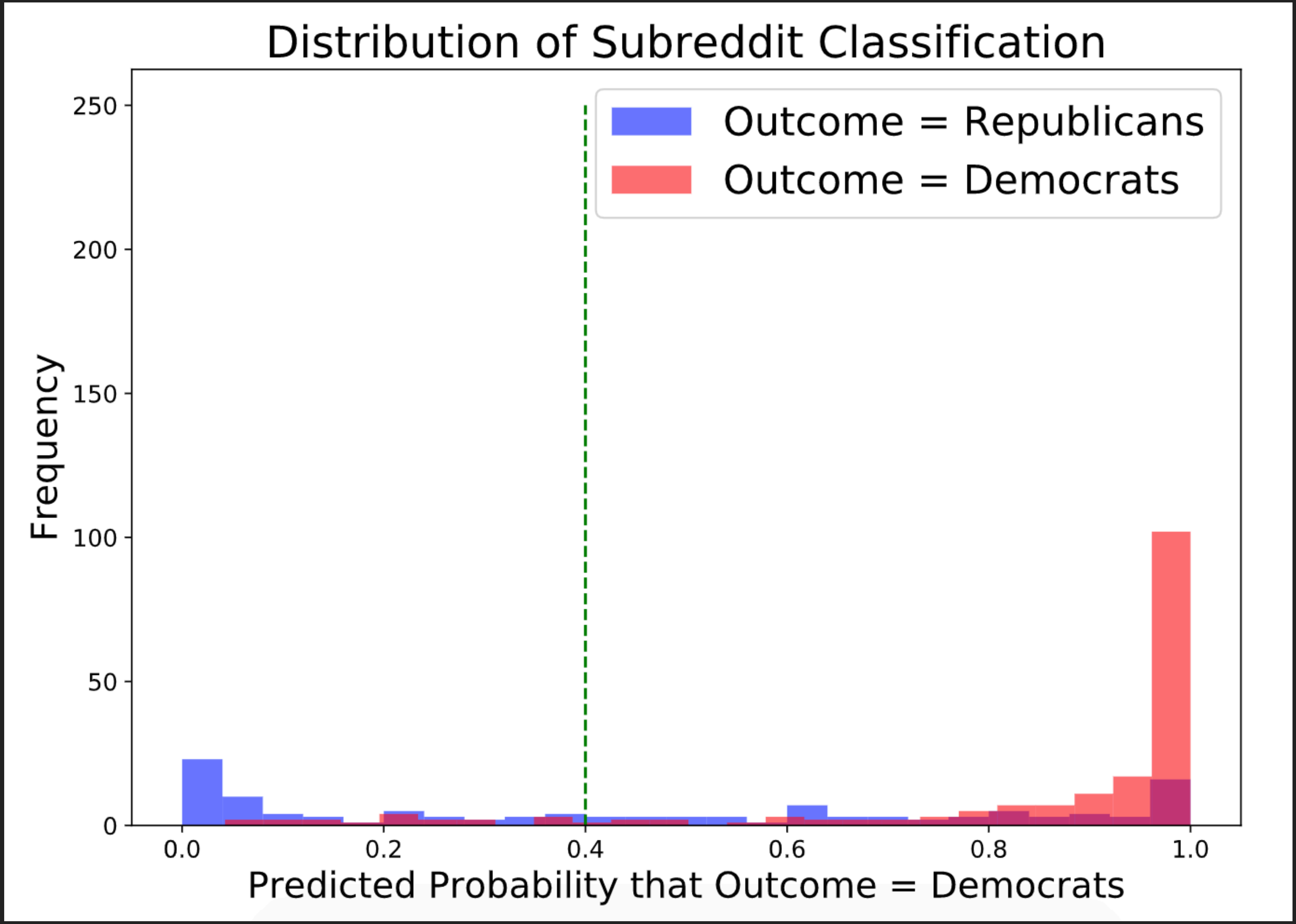


BUILDING THE MODELS

- Train Test Split x2
- Built 3x Models using LogisticRegression(Tfidf), Multinomial Naive Bayes(Count) and Random Forest(Count)
- In each method, Count Vectorizer and TFIDF Vectorizer
- Multinomial Naive Bayes performed the best with least overfitting
- 

```
Accuracy of Naive Bayes Training Set = 0.9145673603504929
Accuracy of Naive Bayes Evaluation Set = 0.7311475409836066
Accuracy of LogRes Training Set = 0.9090909090909091
Accuracy of LogRes Evaluation Set = 0.7147540983606557
Accuracy of RandomForest Training Set = 0.8882803943044907
Accuracy of RandomForest Evaluation Set = 0.7672131147540984
```

FINAL MODEL


Distribution of Subreddit Classification
Outcome = Republicans
Outcome = Democrats
Frequency
0
50
100
150
200
250
0.0
0.2
0.4
0.6
0.8
1.0
Predicted Probability that Outcome = Democrats