Classifying Social Media Text as Political or Casual Conversations

Ethan Kuehl

Problem Statement

The goal of this project is to build a text classification model that can differentiate between casual and political conversations on social media.

Executive Summary

- Methodology
  - Data Collection
  - Data Cleaning
  - Data Preprocessing
  - Data Modeling
  - Model Evaluation
- Results
  - My model is **99% accurate** at predicting whether a post comes from r/CasualConversations or r/PoliticalDiscussion

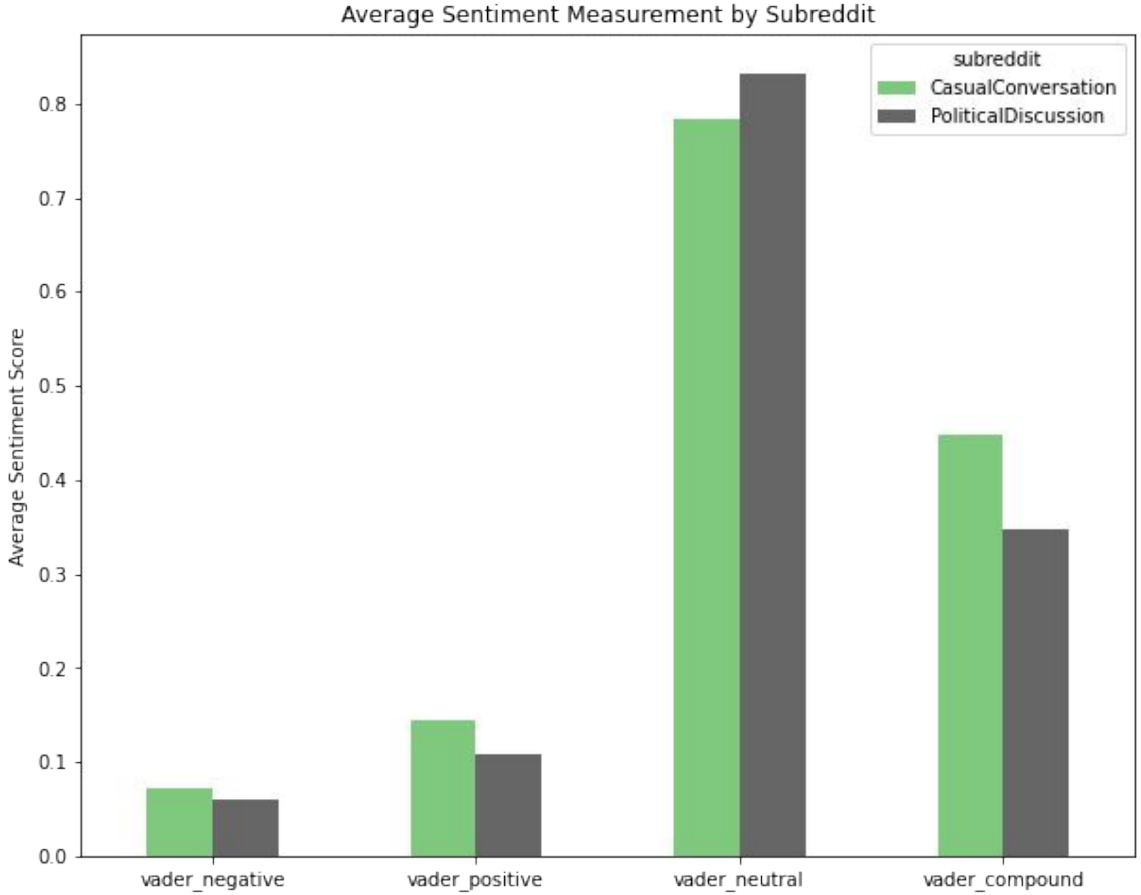
Data Collection and Cleaning

Text classification models are trained using text data from the website reddit, in particular from these subreddits:

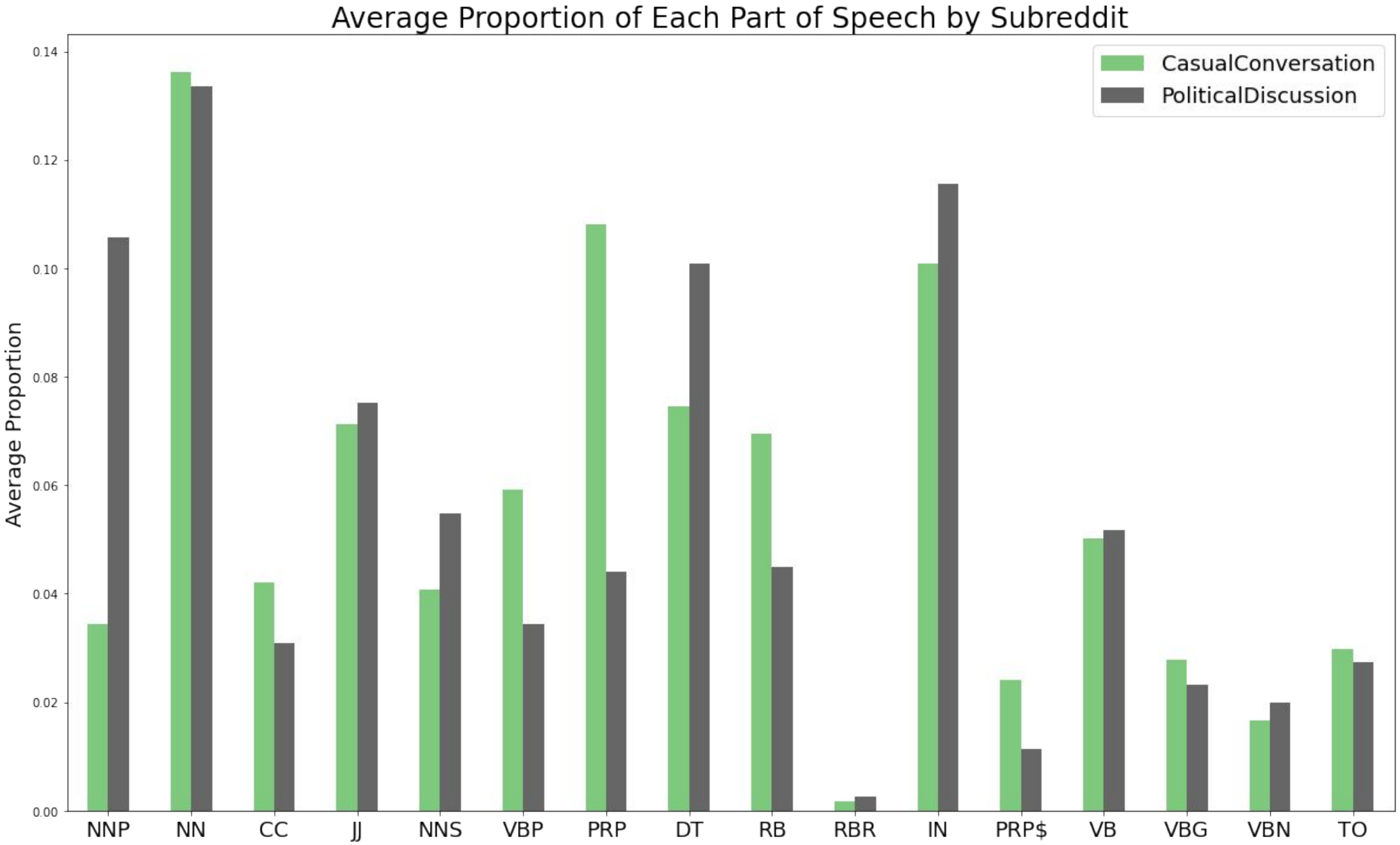
1. r/CasualConversation
2. r/PoliticalDiscussion

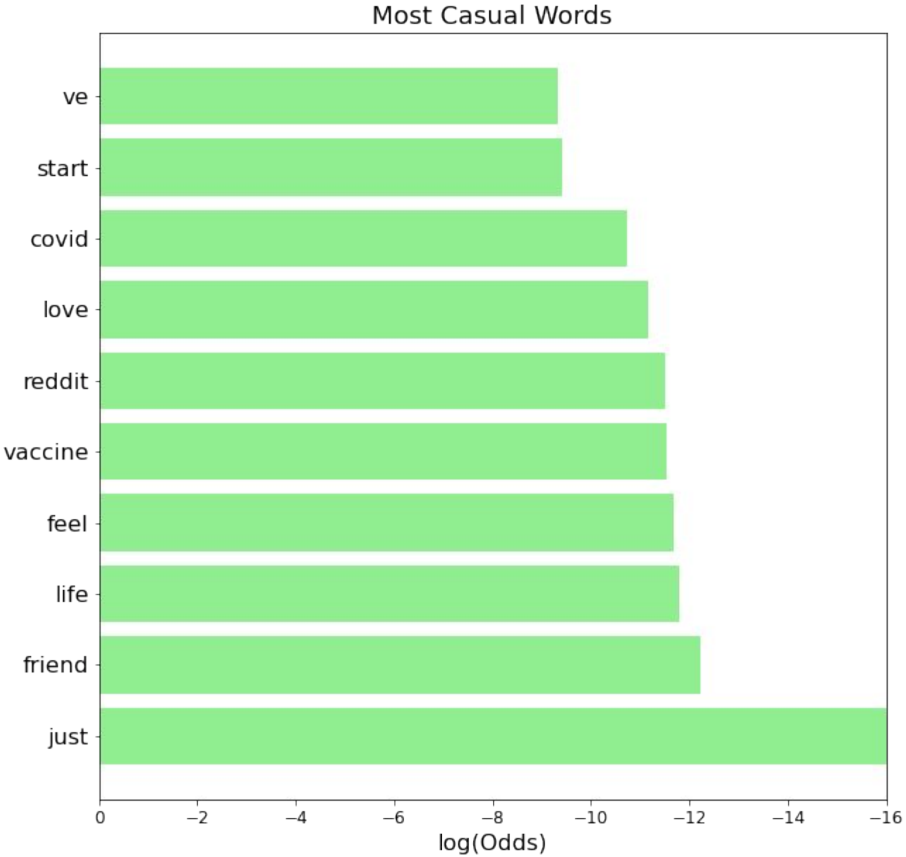
Text data includes the title and post text

Data - Exploratory Analysis

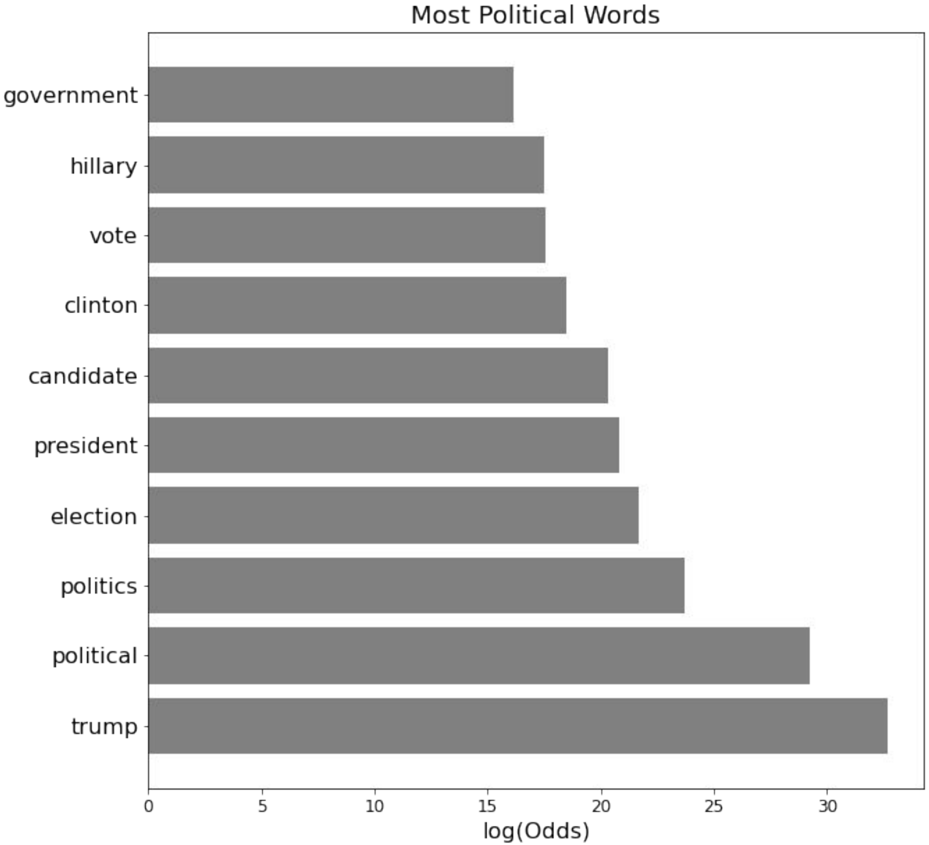


Data - Exploratory Analysis





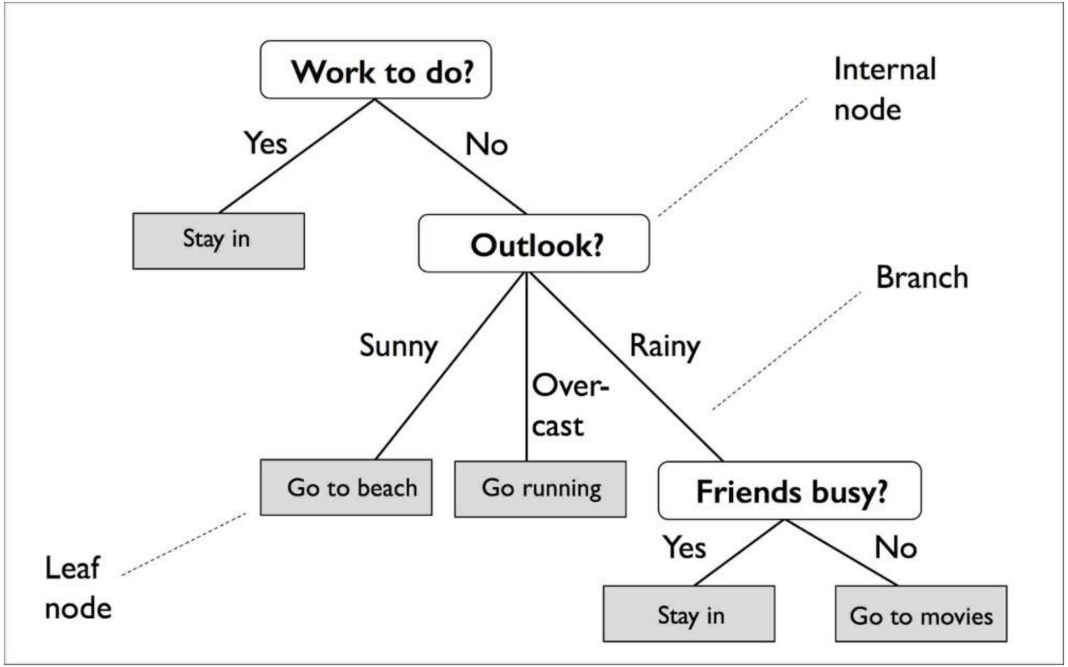
Most Casual Words



Most Political Words

Modeling

- Binary Classification: Decides whether the post is political or casual



https://www.commonlounge.com/discussion/ee8b074936a041f2b5a57d2054dc3701

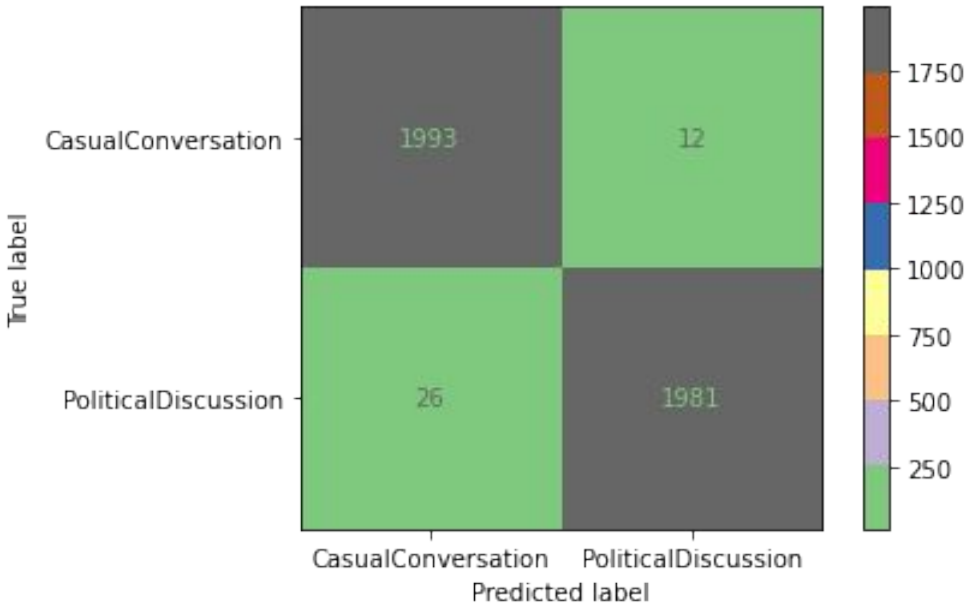
Model Evaluation

True Positive: 1981

True Negative: 1993

False Positive: 12

False Negative: 26



Conclusions

- Sentiment and parts-of-speech analysis were not as useful as analyzing the vocabulary of a given text.

- There is a huge difference in the vocabulary used in r/PoliticalDiscussions compared to r/CasualConversation

- As a result, it is possible to classify posts with 99% accuracy.

Recommendations and Next Steps

- 99% accuracy suggests that my model is ready for beta tests on social media sites other than reddit.

- I would want to further analyze how my model generalizes to text outside of reddit post submission structure.

- Spend more time tuning parameters within the model.