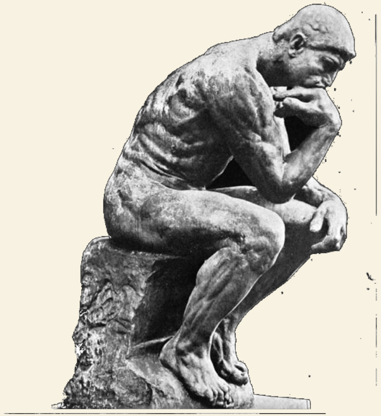
Classifying Reddit Submissions



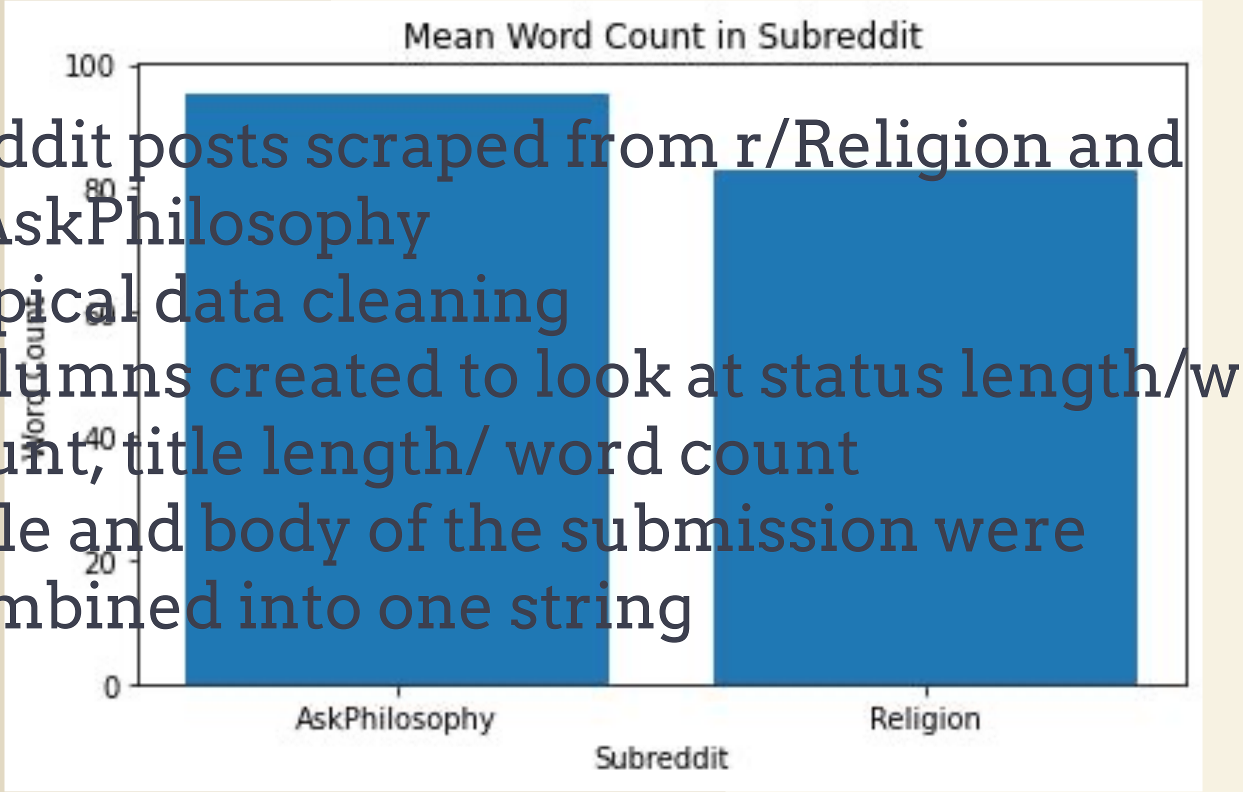
Mya Carrizosa
General Assembly

Problem Statement

Reddit is a social media platform that consists of pages dedicated to specific topics, called subreddits, where users can post and comment within the guidelines of that page and topic. Two of these subreddits are the r/Religion page, and the r/AskPhilosophy page. We are a sociology research group who has theorized that the thematic elements of religious discussions and philosophical discussions are very similar; we anecdotally see overlap in areas like ethics, morality, meaning, human nature, and more. Because of these thematic similarities, we are interested in seeing if a machine learning model could differentiate between these topics.

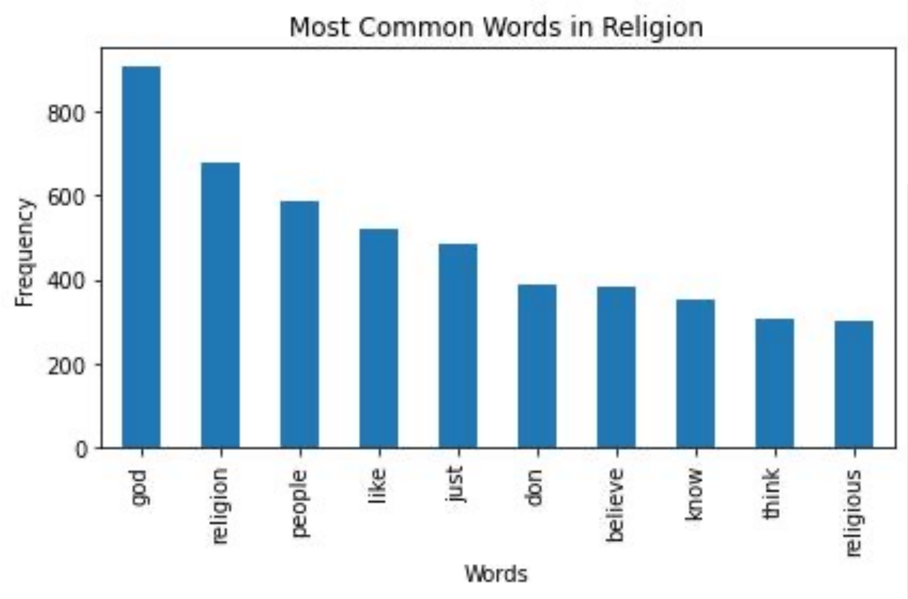
We want to fit a classification model to predict, given a Reddit submission title and text, whether the Reddit submission came from the AskPhilosophy or Religion subreddit.

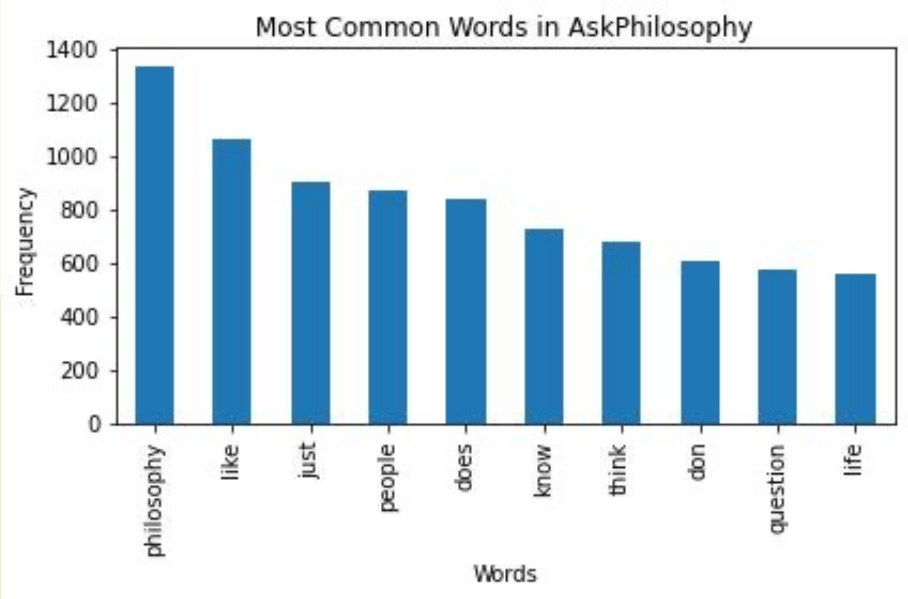
DATA CLEANING AND EXPLORATION



- Reddit posts scraped from r/Religion and r/AskPhilosophy
- Typical data cleaning
- Columns created to look at status length/word count, title length/ word count
- Title and body of the submission were combined into one string

DATA CLEANING AND EXPLORATION





MODELING COMPARISON

Logistic Regression

- Performed best on test data
- High sensitivity, specificity, & precision

Random Forests

- Middle of the road model
- Low sensitivity, high specificity, & precision

K Nearest Neighbors

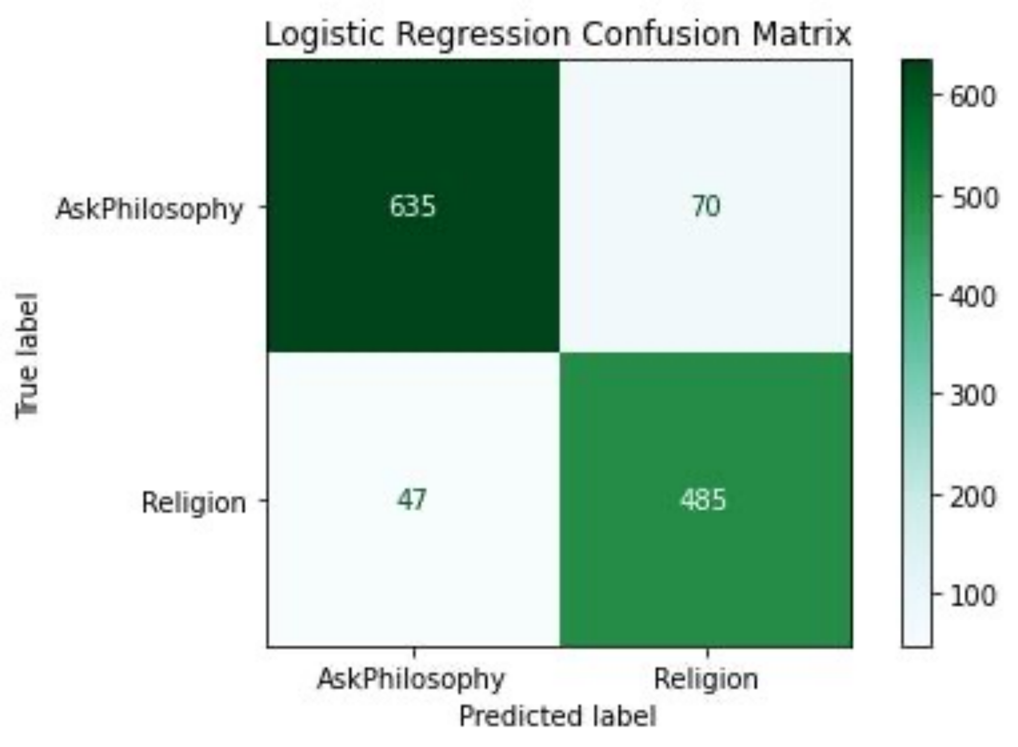
- Very overfit, poor test accuracy
- High sensitivity, very low specificity, low precision

AdaBoost

- Performed almost as well as Logistic Regression
- Fairly good sensitivity, specificity, & precision

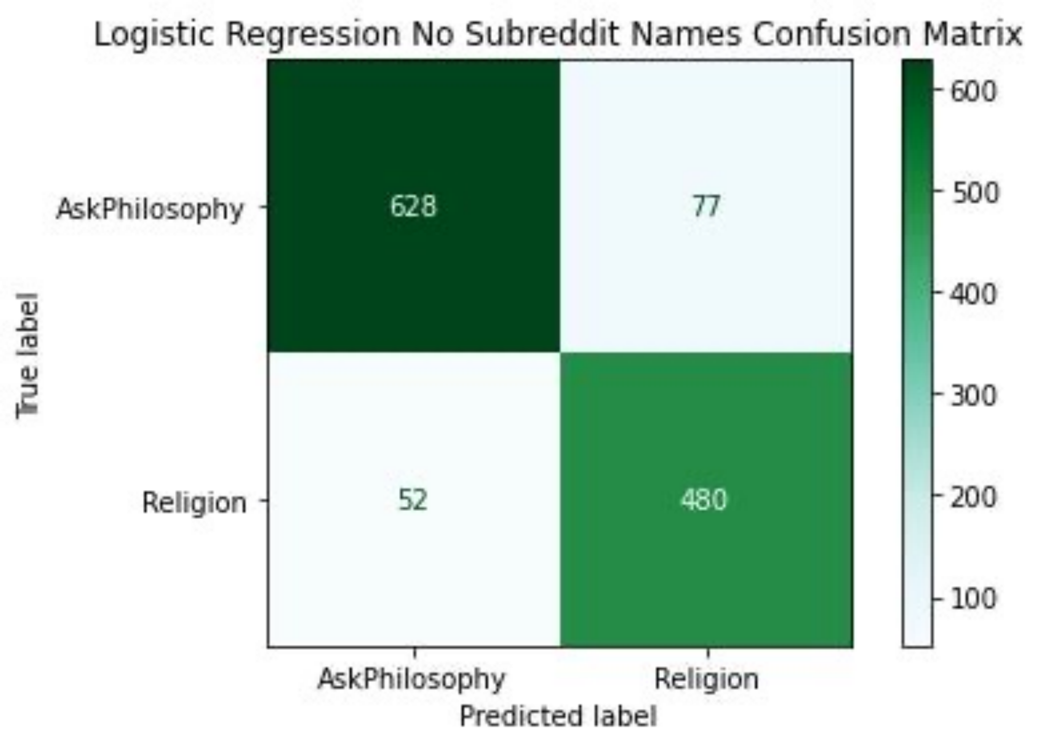
LOGISTIC REGRESSION

- Test Accuracy: 0.91
- Sensitivity: 0.91
- Specificity: 0.90
- Precision: 0.87



LOGISTIC REGRESSION NO SUBREDDIT NAMES

- Test Accuracy: 0.90
- Sensitivity: 0.90
- Specificity: 0.89
- Precision: 0.85



MISCLASSIFICATIONS

From AskPhilosophy, classified as Religion:

- 'i am trying to interpret hegels of history. in chapter 3, he talks about fear of god and how this fear might affect the society. i understood that individuals tend to obey the god and hence do bad things like burning houses etc. however, is this the only thing he asserts as reason to control the ? moreover, i also could not comprehend even if we try to control the how can we do it?hegel on in state’
- "aren't facebook or twitter deceitful given they created a huge user base by promising people platforms for speech, even got section 230 immunity, and now they heavily censor everything they don't prefer politically or otherwise?"

MISCLASSIFICATIONS

From Religion, classified as AskPhilosophy:

- 'the great filter theory is that alien civilisations hit a wall and went extinct which means we are alone in the universe and there maybe something worse coming and the dark forest thing is there is alien civilisations but they don’t want to broadcast their existence like humans do for some reasonwhat do y’all think of the great filter theory or the dark forest theory?’
- 'do you really know ?'

CONCLUSIONS

04. A logistic regression model performed best at classifying posts.

[illegible] that was [illegible] in the [illegible] model [illegible] to [illegible] closely related [illegible] accuracy

CITATIONS

- General Assembly Data Science Immersive Bootcamp Lessons 4.01, 4.04, 4.06, 5.05, 6.03, 6.04
- Reddit (r/AskPhilosophy, r/Religion)
- Riley Dallas's Pushshift Tutorial Video
- Stack Overflow: https://stackoverflow.com/questions/51879018/removing-words-characters-from-string-in-dataframe-cell