Using Natural Language Processing To Identify The Source Of A Post On Reddit

Mitchell Meislin

Problem Statement

1. Can we use natural language processing to determine if a post on reddit came from the ask science subreddit or the personal finance subreddit?

2. What is the most accurate model for classifying a posts source?

Why Does It Matter?

1. This project demonstrates the ability for natural language processing and machine learning models to identify the subject and content of a post on social media.

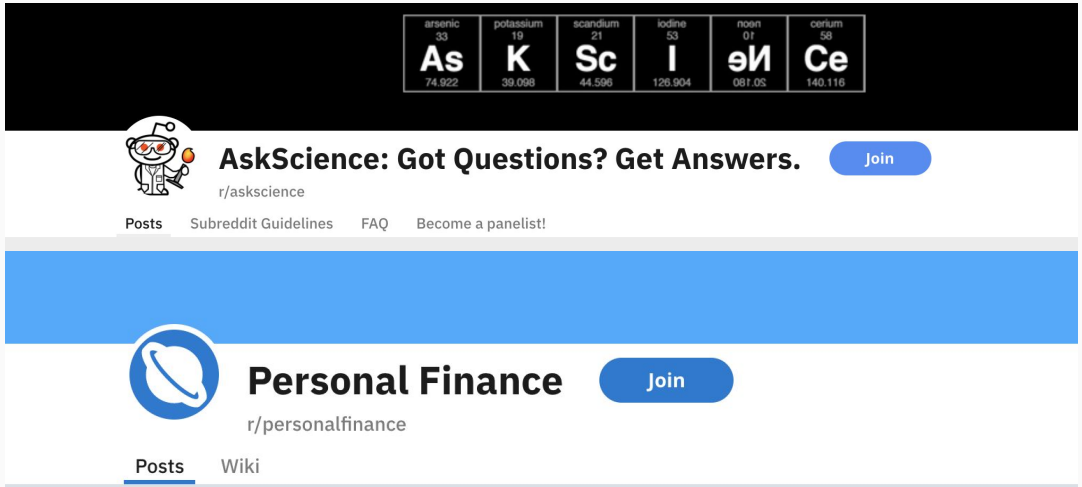
2. This process can be helpful for companies that would like to monitor social media to identify which topics are trending for a given focus group.

3. This project helps show which machine learning tools may be most accurate when classifying a text based post.

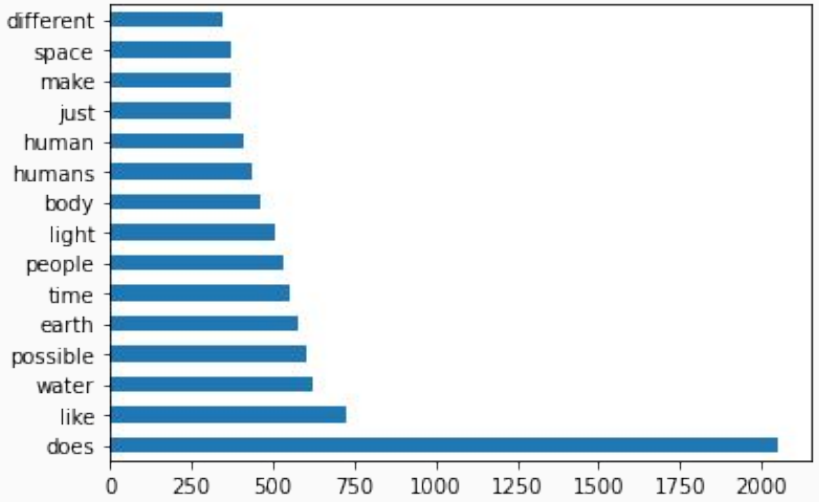
Data Collection:

1. I collected approximately 15k posts from two different subreddits

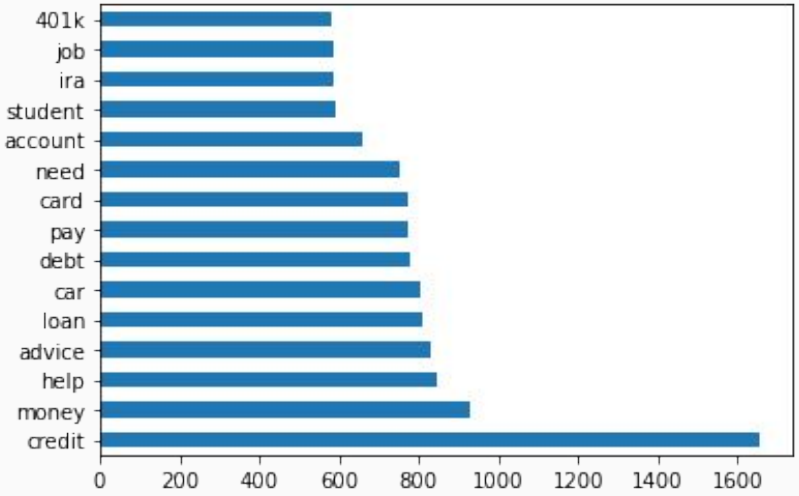
2. I chose to collect data from the ask science subreddit and the personal finance reddit



Most Common Words From Each Reddit Forum

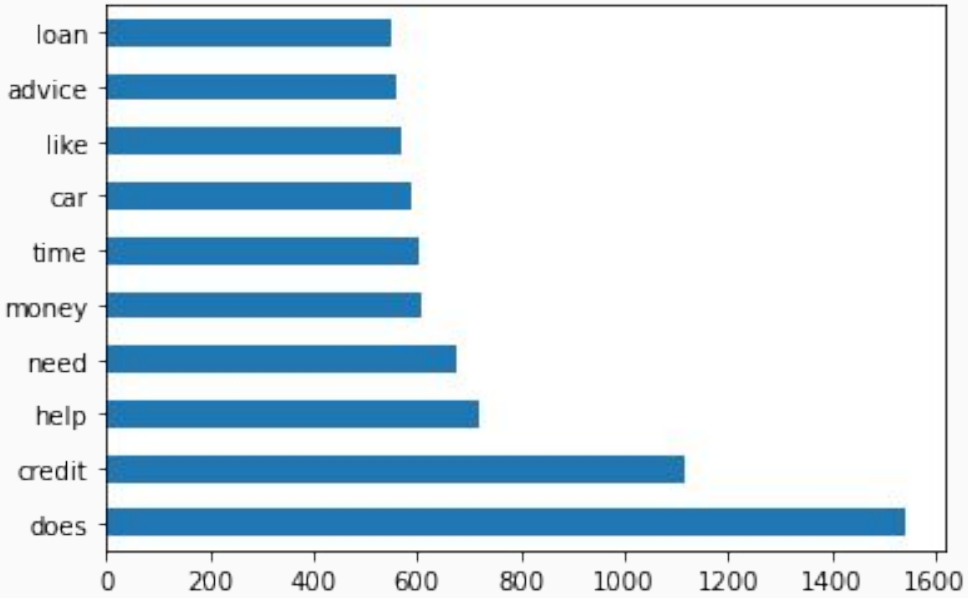


15 Most Common Words In Ask Science



15 Most Common Words In Personal Finance

Most Common Words In Both Reddit Forums

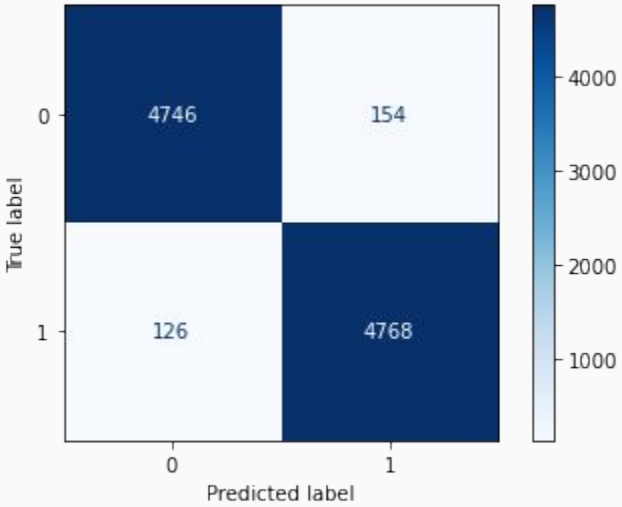


10 Most Common Words In Both Subreddits

Which Models Did I Use?

1. SVC Model
2. TVEC Model
3. Gradient Boost Model
4. Ada Model
5. Naive Bayes Model

Which Model Had The Best Accuracy Score?

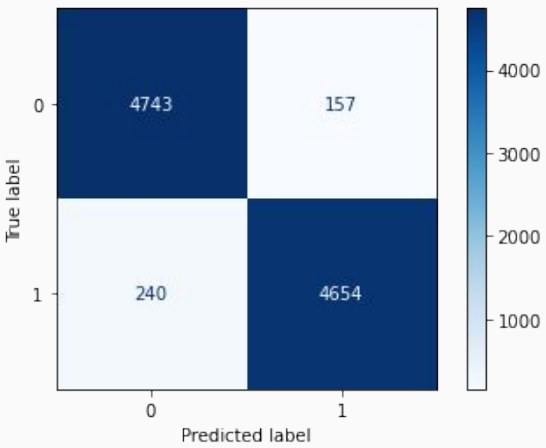


Naive Bayes Model Confusion Matrix

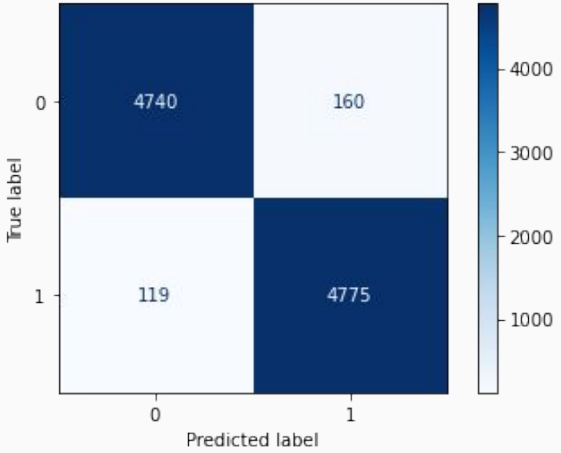
```
Accuracy: 0.9714110680008168
Specificity: 0.9685714285714285
Precision: 0.9687119057293783
Sensitivity: 0.9742541888026155
```

Naive Bayes Model Scores On Unseen Data

Runners Up



SVC Model



TVEC Model

Conclusions and Recommendations

1. In this project I found that we can use natural language processing to determine if a post on reddit came from the ask science subreddit or the personal finance subreddit with 97.1% accuracy on unseen data.

2. I found the most accurate model for classifying a posts source to be a Naive Bayes model.

3. For companies trying monitor social media to identify which topics are trending for a given focus group, I recommend using a Naive Bayes model for classification
    a. If a Naive Bayes model is not getting an ideal accuracy score, I recommend also trying the runner up models, SVC and TVEC

THANKS FOR LISTENING!!!