GA – Project 3: Web APIs&NLP

BRANDON A

Premise

- Work for company that specializes in NLP
- As the world becomes more digital, people are botting online applications/contests with AI generated responses
- Hired by Reddit to help with their future programs/contests for their user base
- Want to train a model on whether a response came from a human or OpenAI

Reddit Sample Data

- AskCulinary
- Questions
- AskEngineers
- CSCareerQuestions
- AskDocs
- TrueAskReddit

625 items from each

"TOP"

3750 rows of data

- OnePiece
- Manga
- NBA
- TorontoRaptors

Had to include '?'

114 rows of data

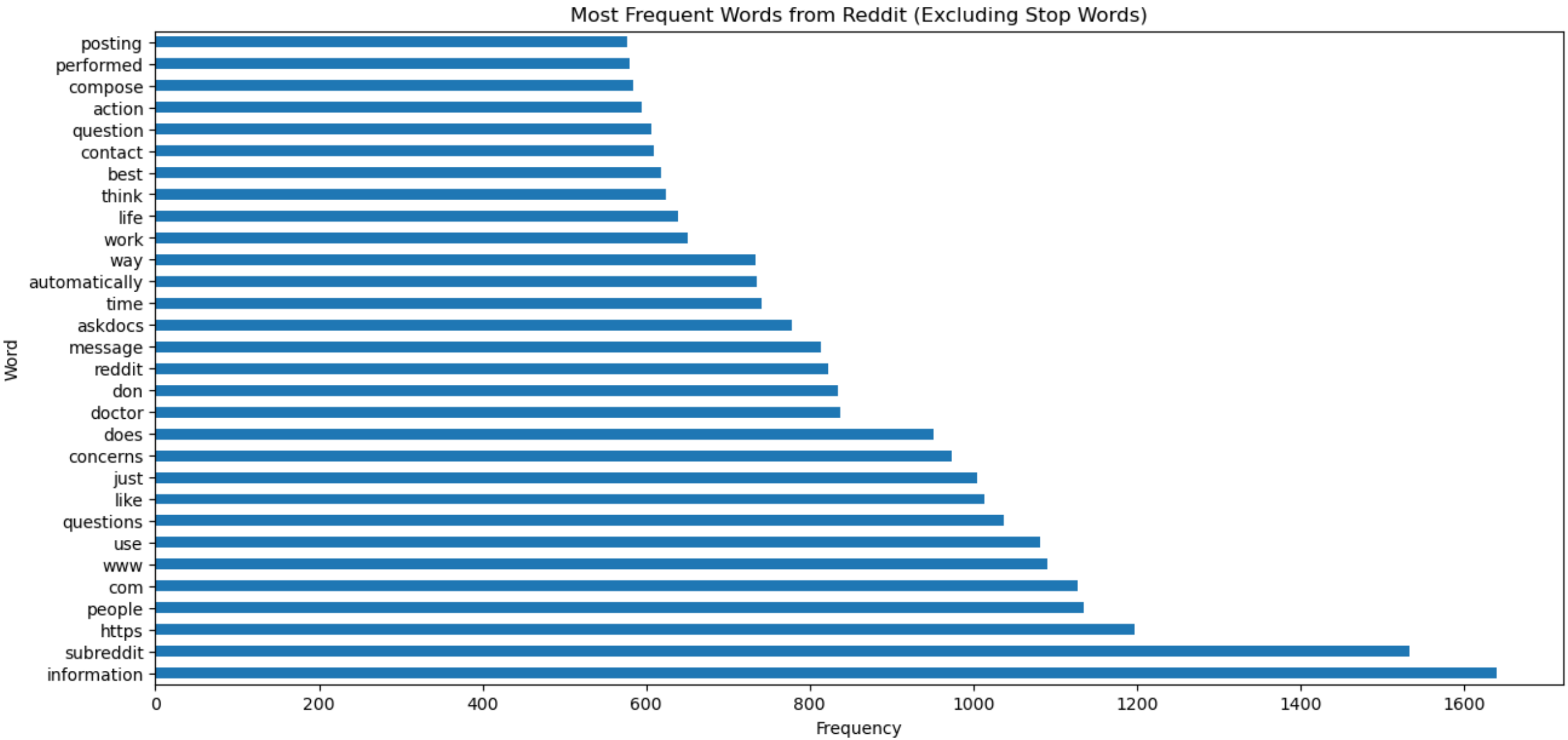
OpenAI Sample Data

- Used prompts gathered from Reddit
- text-davinci-003
- Broke questions up into lists of maximum 500 each, then sent in
  - Experienced server shutdown errors

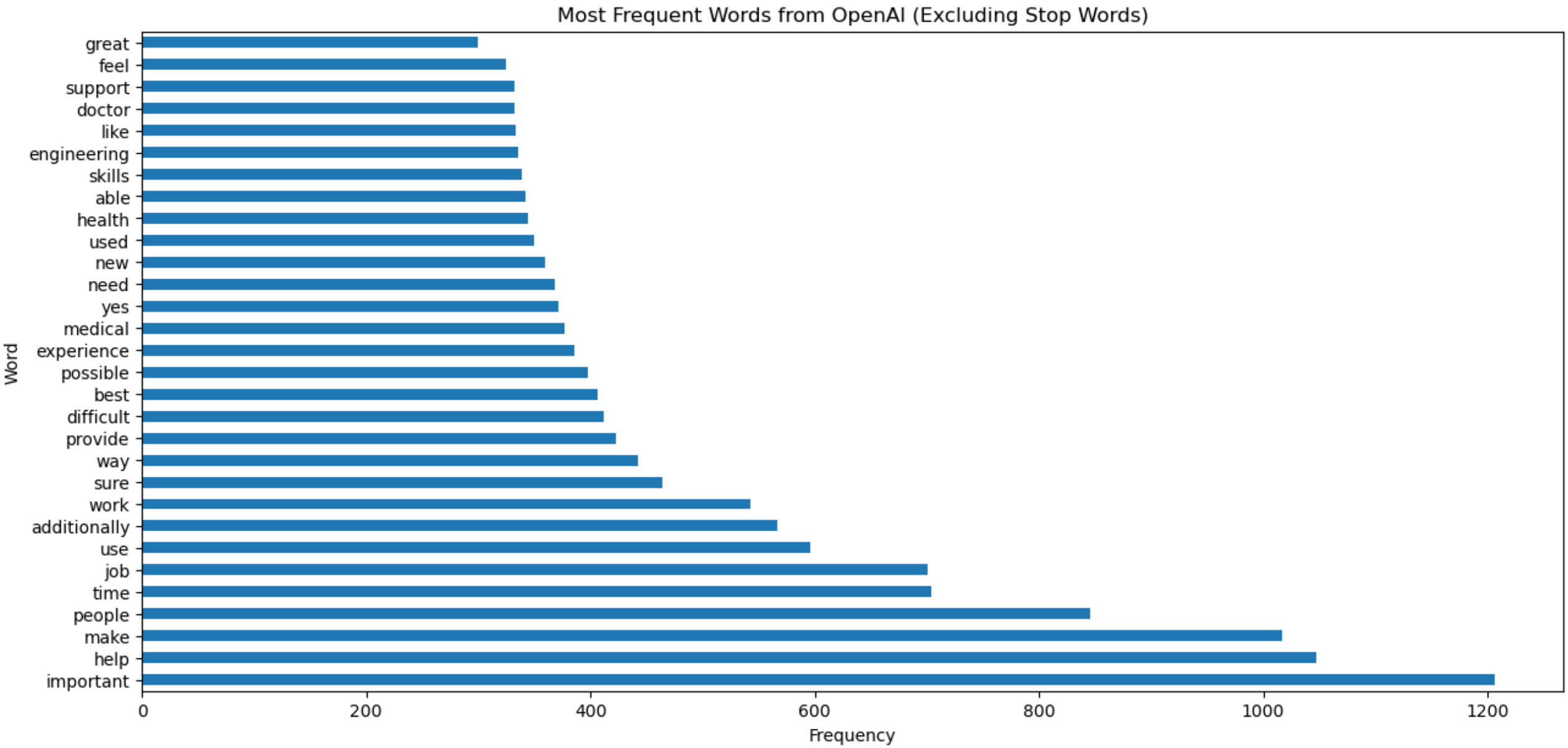
Data Cleaning

- Removed rows which had [deleted] for Reddit answer
- Removed rows which included '#Message to all users…" for Reddit answer
- Removed rows which had [removed] for Reddit answer
- Removed rows which had "NaN"
- Removed markdown shortcuts
- Investigated & removed rows which started with "Welcome"
  - Moderator answers were the ones removed

3750 -> 3484
266 Rows Removed



3484 rows × 17536 columns



3484 rows × 12026 columns

Baseline & Model Specifications

0 (Reddit)	0.5
1 (Open AI)	0.5

y_test	
0 (Reddit)	0.523673
1 (OpenAI)	0.476327

Classification Models & Vectorizers

- Logistic Regression
- Bernoulli Naïve Bayes
- CountVectorization
- TFIDFVectorization

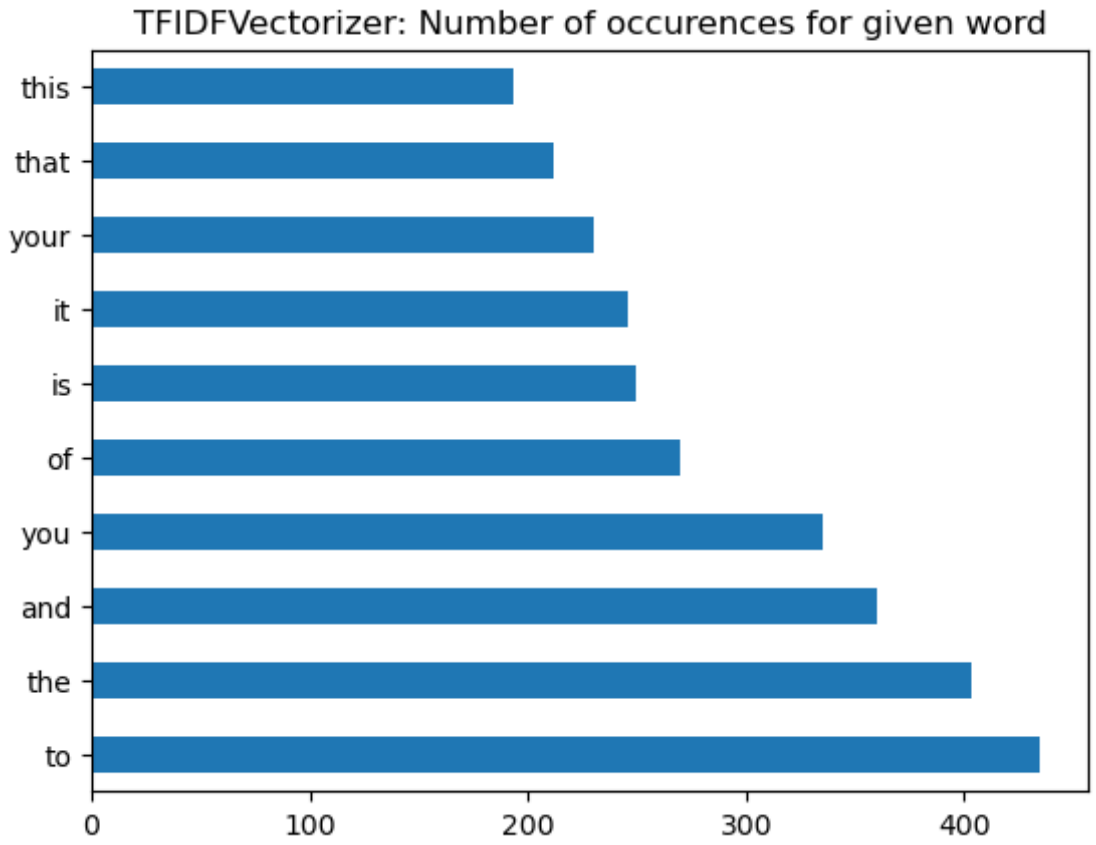
Ran 8 models while tinkering & tuning the various parameters for each

Log Regression + TFIDFVectorization

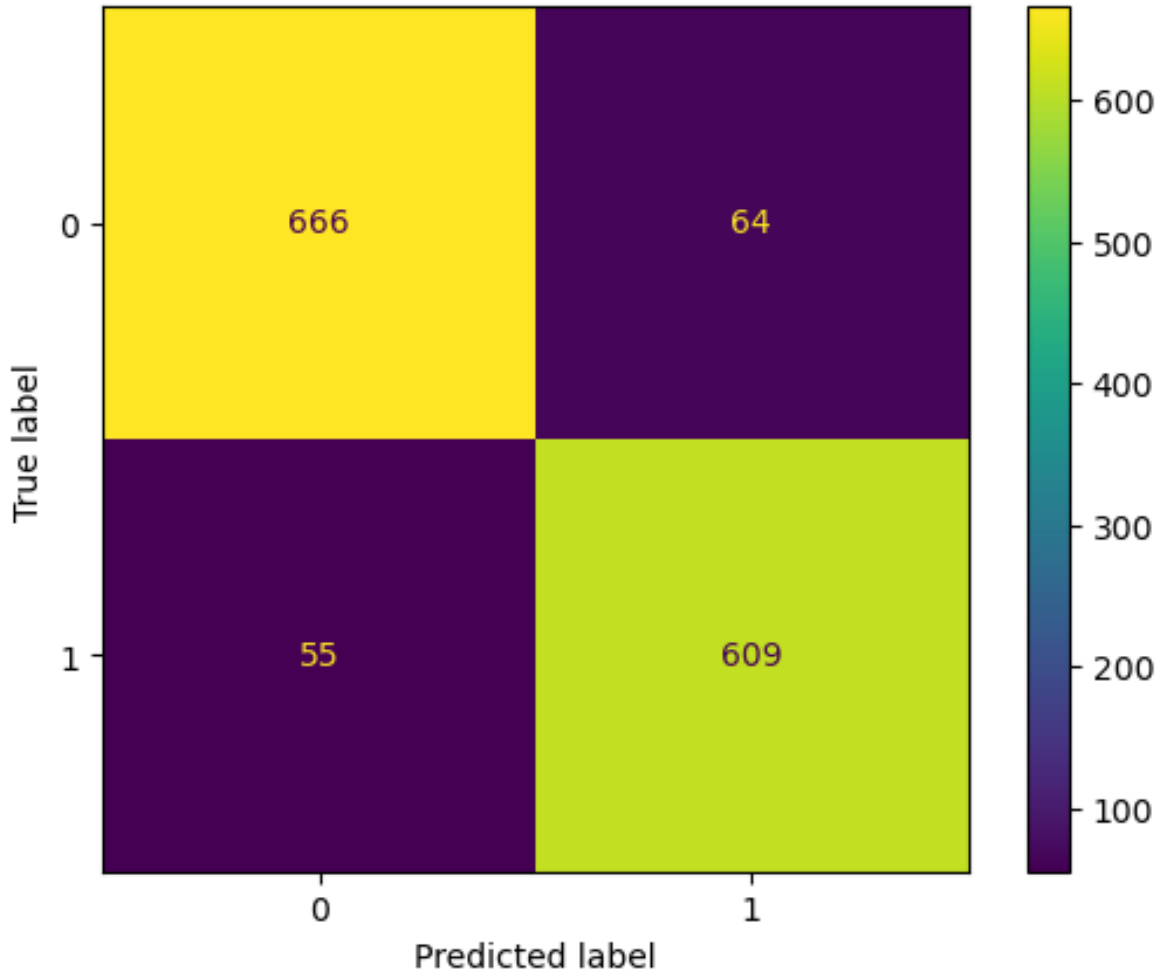
y_test	
0 (Reddit)	0.523673
1 (OpenAI)	0.476327

y_pred	
0 (Reddit)	0.517217
1 (OpenAI)	0.482783

Train score: 0.9610692500897022
Test score: 0.9146341463414634

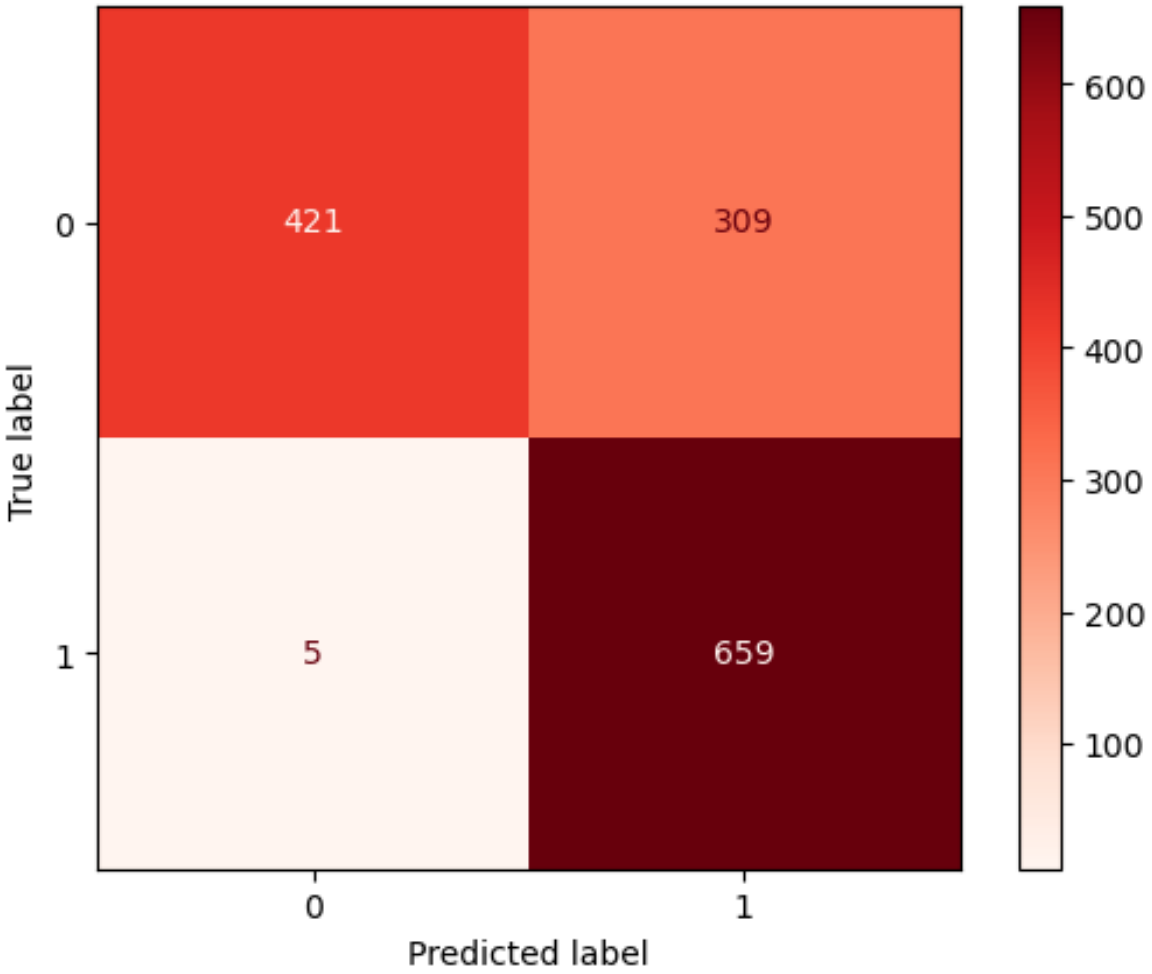


Log Regression + TFIDFVectorization



Accuracy =	91.46%
Misclassification Rate =	8.54%
Sensitivity =	91.72%
Specificity =	91.23%
Precision =	90.49%

Bernoulli Naïve Bayes & CountVectorization - Using pipeline & gridsearch w/stop words

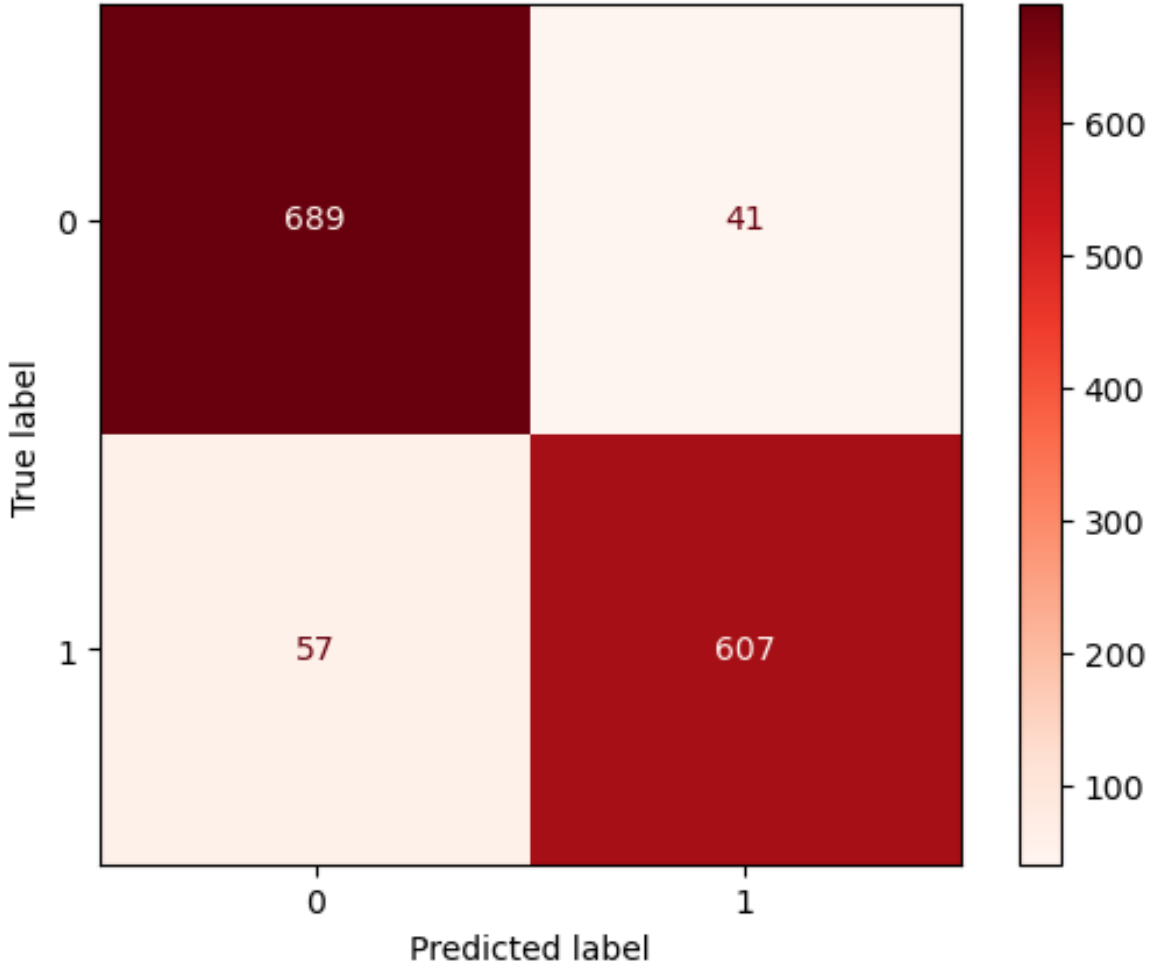


y_test	
0 (Reddit)	0.523673
1 (OpenAI)	0.476327
y_pred	
0 (Reddit)	0.302009
1 (OpenAI)	0.697991

Accuracy =	77.47%
Misclassification Rate =	22.53%
Sensitivity =	99.25%
Specificity =	57.67%
Precision =	68.08%

Train score: 0.8012199497667744
Test score: 0.771162123385939 7
Best score: 0.7818441200859827

Log Regression & CountVectorization
- Using pipeline & gridsearch w/lemmatization



y_test	
0 (Reddit)	0.523673
1 (OpenAI)	0.476327

y_pred	
0 (Reddit)	0.535151
1 (OpenAI)	0.464849

Accuracy =	92.97%
Misclassification Rate =	7.03%
Sensitivity =	91.42%
Specificity =	94.38%
Precision =	93.67%

Train score: 0.9912091855041263
Test score: 0.9296987087517934
Best score: 0.9348750110698731

Conclusion & Reflection

- Log Regrsesion + TFIDFVectorization while utilizing GridSearch and Lammetization given certain parameters turned out to be best model
- Myriad of other combinations out there to try
- Model relevant specifically to Reddit only