Project 3, NLP and Me

Problem Statement

How can we build a model that takes in posts from two subreddits and correctly identifies which one it come from?

AskWomen vs AskMen

I decided to use the two subreddits AskWomen vs AskMen.

I wanted to see how different they were, theoretically the more different they are the higher the scores I will get on my model

Webscraping

Rather than use pushshift, I used Python Reddit Api Wrapper

Or PRAW

A bit more complicated to set up, but very easy to use

I pulled the top 1000 posts from this past year from each subreddit

Data

I collected the title, subreddit, and the selftext from each post.

I originally was just going to collect the titles but I decided the collect the selftext as well since I want to run my laptop into the ground

Data Cleaning

Relatively easy

No null values so I just had to dummify the subreddit and combine the title and selftext columns into one

Used english stopwords

Feature Engineering

Due to the nature of this project, there didn’t seem to be much room for feature engineering.

However, one avenue I wanted to explore was sentiment analysis. Is one subreddit happier than the other?

Sentiment Analysis

Originally was going to do it on the title and selftext, however some of the selftext was too long to do a sentiment analysis

I wanted to try to summarize the selftext, but that took way too long

I settled for just generating the sentiment of the title

Sentiment Analysis results

AskWomen: 74% negative, 26% positive

AskMen: 79% negative, 21% positive

However, this does not tell the full story.

Problem with Sentiment Analysis

Questions aren’t inherently negative or positive.

Examples:

“What makes you feel feminine” was rated as highly negative

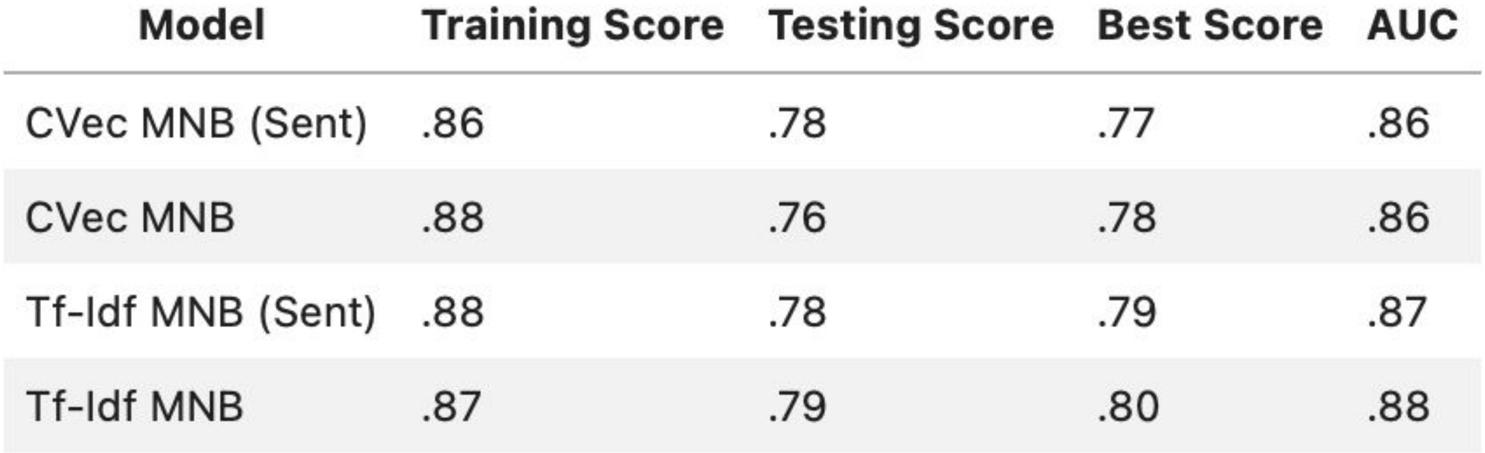
“How early in the relationship can I bust out the Lord of the Rings extended edition?” was rated as highly negative

“My best friend lost a bunch of weight now how he’s expecting women to go crazy for him, yet they aren’t because he became the ultimate nice guy in a Chad body. How do I bring this up without sounding like an asshole?” was rated as highly positive

Did Sentiment Analysis Help?

Long story short, not really...

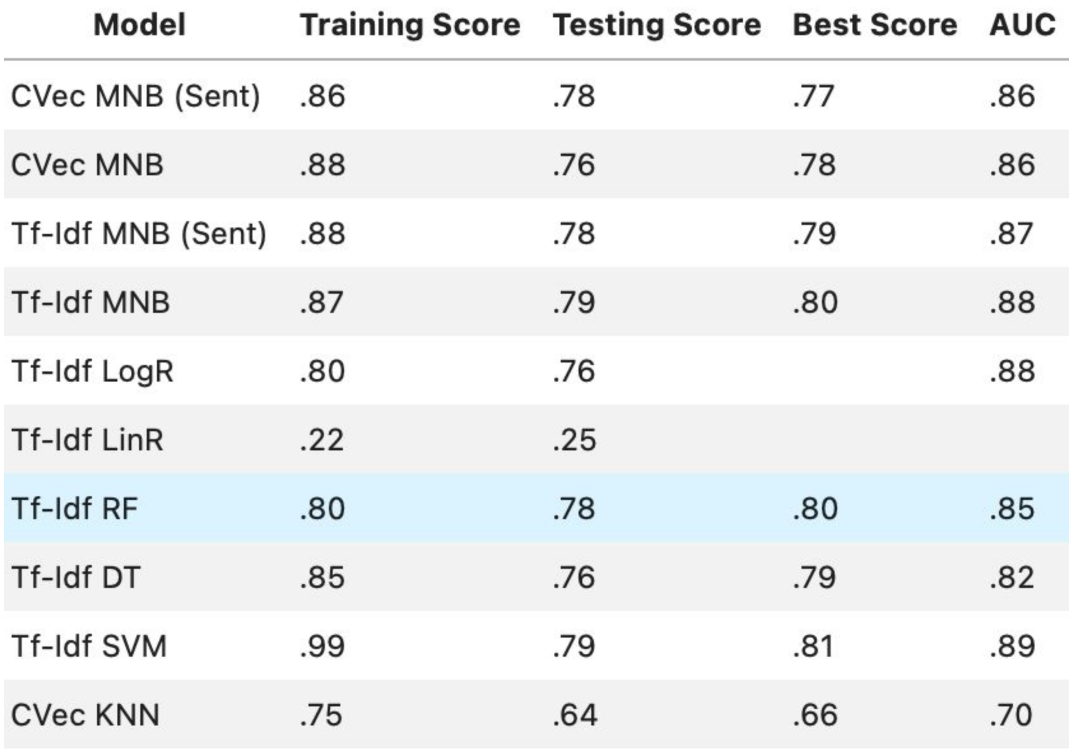
Waiting over an hour to summarize self text
Sentiment analysis doesn't even help model



Model	Training Score	Testing Score	Best Score	AUC
CVec MNB (Sent)	.86	.78	.77	.86
CVec MNB	.88	.76	.78	.86
Tf-Idf MNB (Sent)	.88	.78	.79	.87
Tf-Idf MNB	.87	.79	.80	.88

Final Results

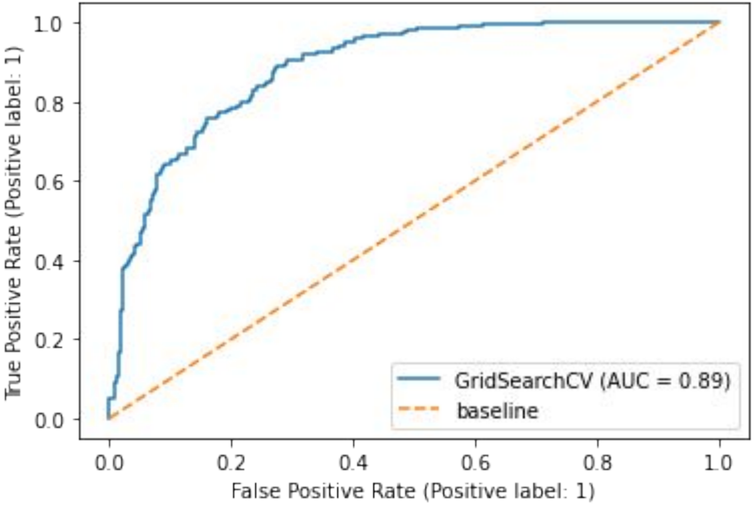
So I ended up creating 10 different models...



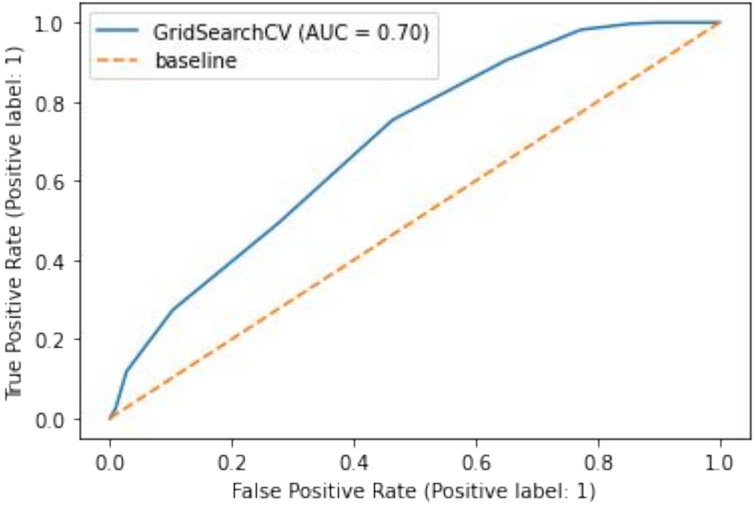
Model	Training Score	Testing Score	Best Score	AUC
CVec MNB (Sent)	.86	.78	.77	.86
CVec MNB	.88	.76	.78	.86
Tf-Idf MNB (Sent)	.88	.78	.79	.87
Tf-Idf MNB	.87	.79	.80	.88
Tf-Idf LogR	.80	.76		.88
Tf-Idf LinR	.22	.25		
Tf-Idf RF	.80	.78	.80	.85
Tf-Idf DT	.85	.76	.79	.82
Tf-Idf SVM	.99	.79	.81	.89
CVec KNN	.75	.64	.66	.70

ROC AUC

Best AUC: SVM Model

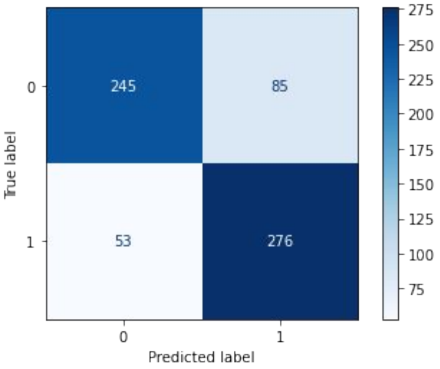


Worst AUC: CVec with KNN

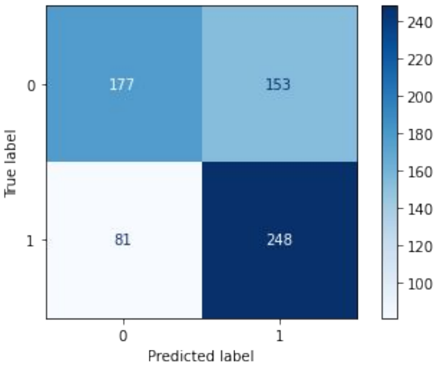


Confusion???

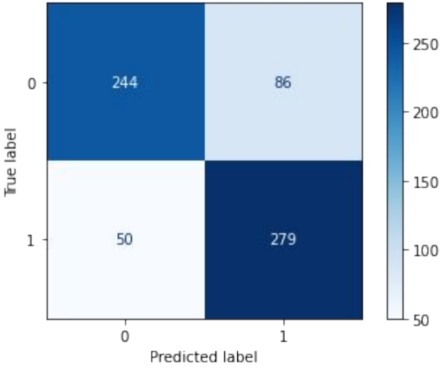
Tf-Idf LogR



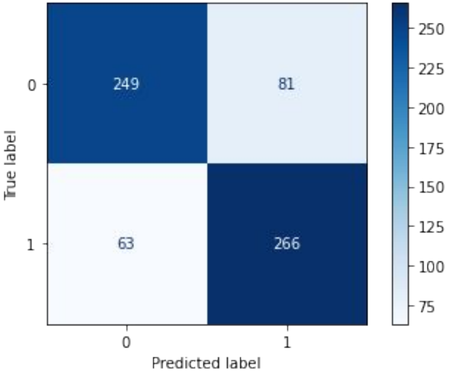
CVec KNN



Tf-Idf SVM



CVec MNB (Sent)



Example of MisClassified Data

What body image problems/insecurities do you have?

What song emotionally fucks you up in a good way?

Ever felt like you were just a lesson/experience in people's lives? How did you deal with constant goodbyes and loneliness that came along?

To those who have lost siblings, how do you ever feel ok again?

Conclusion

Overall, there was not as much of a difference between the models as I thought there would be

I thought SVM would be significantly higher than the rest. Although it had a ridiculously high training score, the testing score was very similar to the rest (it did have the highest AUC, but only by .01).

All of the models being similar makes me think that the subreddits are decently different in content, but also have a small amount of overlap.

Only CVec (Sent) MNB used English stopwords

Conclusion Cont.

From a cursory glance, the top questions in AskWomen seem to be more serious than the top questions in AskMen.

However they do have a decent overlap in relationship questions

Future Exploration

Aggregated models

In the future it would be interesting to formally map out the percentage of types of questions within each subreddit.

Another area to explore would be the comment section as well. Does one subreddit tend to agree or disagree more with the question asked? Is there a correlation between the category of question and number of comments? Etc.