INTRO
Hi everyone, I'm Nate and I'm a Data Scientist. It's nice to be surrounded by fellow Data Scientists to share a NLP classification project I've been working on.

Let's jump right into it with the problem statement.

PROBLEM STATEMENT
Can you predict what subreddit a comment comes from?

To answer this question, I had to go through a few steps.
Today, I'm going to walk you through the process I took to try to answer that question.

OUTLINE
First, I'll share some info about the dataset.
I'll explain the subreddits to give you an idea about the comments.
I'll briefly touch on some of the cleaning I had to do.
I'll tell you about the feature engineering I did.
I'll go over some EDA and discuss some of the models I created.
Last, I'll share my results and conclusion.

INFO ABOUT THE DATA
I collected over 115,000 comments using the Pushshift Reddit API.
The text, author, and what subreddit the comment came from were the details I gathered.

SUBREDDITS
First off, I want to keep this PG-13. So, I will only be dropping one F-Bomb.

The subreddits I chose to work with are mildlyinteresting and interestingasfuck.

I'll be referring to the second one I listed as IAF or interestingAF. If I do slip up, we might have to bump the rating up to R.

All of the posts on these subreddits are images or videos. I grabbed a few images from some of the top posts to give you an idea what I was working with.

MILDLYINTERESTING
For mildlyinteresting, the left image is of some corrosion that looks like a map. In the right image, the user found the location of the picture on their cliff bar. Nothing too crazy. Mildly interesting seems like a good description.

INTERESTINGASFUCK
Next, here are some quick examples of a couple of the top posts in interestingAF. I'll just read the title for the left image because it says it all.

"The small details: In the forearms there is one very small muscle that contracts only when lifting the pinky, otherwise it is invisible. Michelangelo's Moses is lifting the pinky, therefore that tiny muscle is contracted - a small part of the many details of this masterpiece."

The right image is a photorealistic image of George Washington if he was alive in the present day.

Remember, I am trying to predict the subreddit based on comments. So, these comments can be about a lot of random things. Some may not be about the post on the subreddit at all. I'm sure there's a decent percentage of people just arguing with each other.

CLEANING
Unfortunately, all the commenters on reddit don't have perfect spelling, correctly use punctuation, or anything like that. All of the data for the comments was a mess. I combed through it to remove comments that were deleted, comments made by bots, links to websites, extra white space, expanded the contractions, and many other things.

FEATURE ENGINEERING
Once the data was clean, I engineered some features that I would later use in modeling to predict what subreddit the comments came from. Don't feel like you have to read all 13 of these. Nothing too exciting here. I counted words, capital letters, checked to see if there were certain punctuations used, the styling of text like italics, and other things that could help me find some differences from one subreddit to another.

AVERAGES
This table represents the averages for all those features I collected. Unfortunately, there were no major differences between subreddits for all of these features.

Since the features I created didn't seem to have much of an impact, I decided to look more closely at the words used.

FREQUENCY
Here are the top 10 words used for each subreddit. Stop words such as “and,” “the,” “a,” were removed at this point. If you glance back and forth, you can see quite a few words show up in both subreddits. Let me make things a little more clear.

All the words in green show up in the other subreddit’s top 10 words list. The words in blue don’t.

This isn’t too surprising though since most of these words are pretty common. At this point, I was hoping there would be something that would stick out to help differentiate between the two subreddits.

SCATTERTEXT
This is where things started to get a little more interesting. This is a snapshot of an interactive scatter plot I generated using scattertext.
On the X axis is the frequency of words in the mildlyinteresting subreddit. On the Y axis is the frequency of words in the interestingAF subreddit.
What jumps out to me first is the cluster in the top right. This is showing there are a lot of words that frequently show up in both subreddits.
This doesn’t look good for making predictions. If there were more dense clusters in the top left or bottom right, that would represent words that were used frequently in one subreddit, and not the other. If that were the case, I believe that would help my model.

WOW
One cool thing about scattertext is how it is interactive. You can click on certain points or search for specific words. I decided to search for the word “wow.” I figured the word must show up more often in interestingAF compared to mildlyinteresting. Nope. This is just one example, but not a good sign for a predictive model.

So, let’s get into the modeling phase.

MODELS
While trying to create the best model to predict what subreddit a comment comes from, I tested 6 different types.
I decided to further test only three. Those three were Multinomial Naive Bayes, Logistic Regression, and a Voting Classifier that contained both of them.
These three were chosen because of their speed and performance.
I tried using the features I created, but none of them helped the models. So, I ended up just using the text from the comments.

RESULTS
Here is a summary of how the best models did.
The scoring metric I judged my models on was accuracy. I felt this was fitting because false positives and false negatives didn't matter too much in this classification problem.
The baseline was set at 53.7%. This value represents the subreddit that appeared the most in the dataset.
Multinomial Naive Bayes and Logistic Regression came in pretty close to each other at 70.5% and 70.8%. When combined in a voting classifier, the accuracy of the model jumped up to 71.3%.

BREAKING IT DOWN
In this confusion matrix, we can see a breakdown of the model's predictions. The True Positives and True Negatives in the upper left and bottom right, occurred more often than the False Positives and False Negatives.

COEFS
Here is a look at the ten words with the highest coefficient values for interestingAF. These words were the strongest predictors for this subreddit. Something interesting is that they are almost all nouns. **PAUSE**

The same goes for the coefficients for mildlyinteresting. **PAUSE**

This information could be useful in improving the performance of the model at a later date.

CONCLUSION
So, can you predict what subreddit a comment comes from?
I'd say yes. The predictive model I created performed 17.6% better than the baseline.
By collecting more data, and unlocking the power of the nouns, I believe I can further improve my model on the next iteration.

THANKS
Thank you for joining me here today. I hope my presentation was at least mildly interesting. I'd be happy to answer any questions.