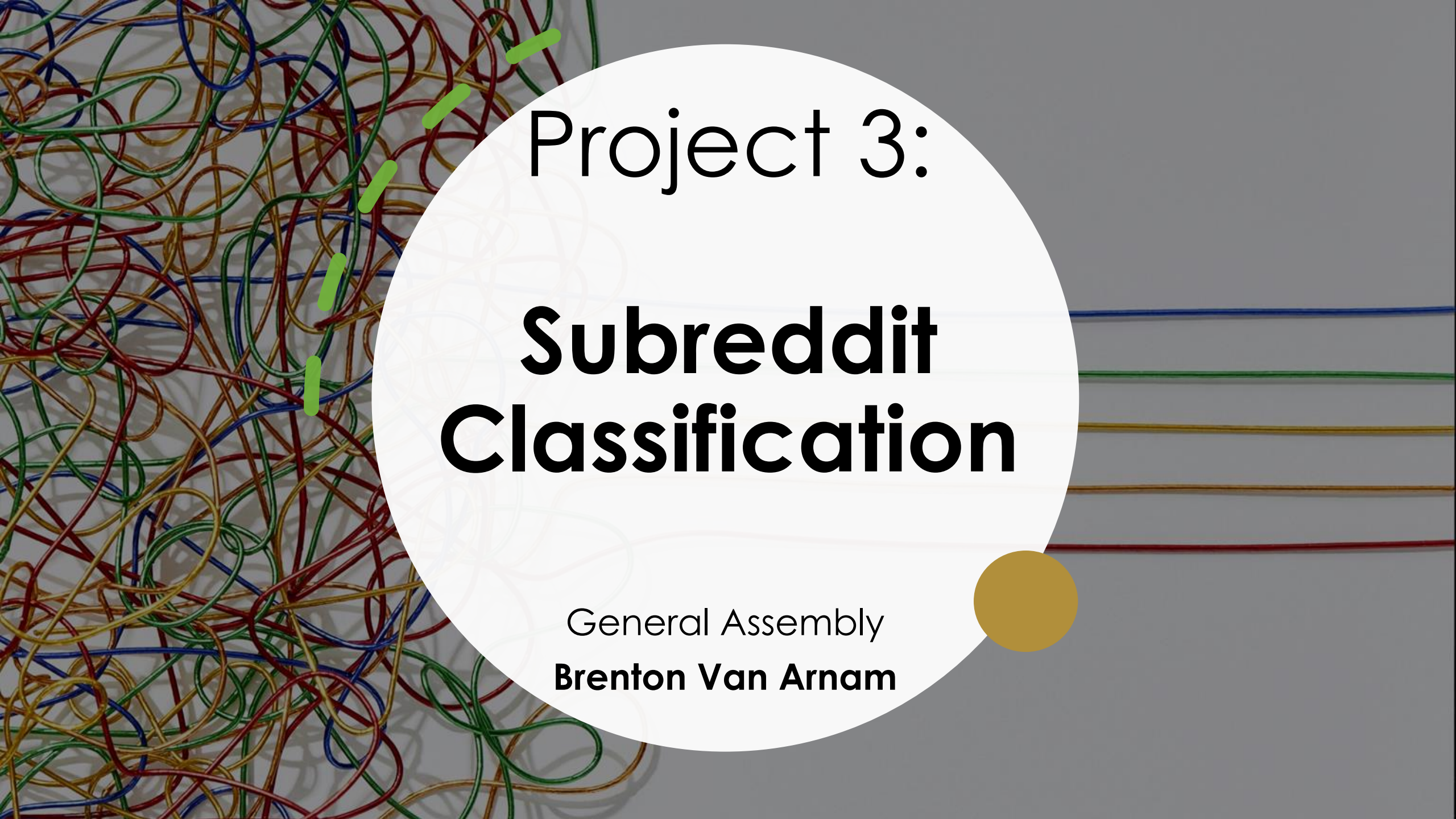


Project 3:

Subreddit Classification

General Assembly

Brenton Van Arnam

Background Information

Reddit – Online message board and forum.

AskScience - Ask a science question, get a science answer.

ExplainLikeImFive (ELI5) - Explain Like I'm Five is the best forum and archive on the internet for layperson-friendly explanations. Don't Panic!

Problem Statement

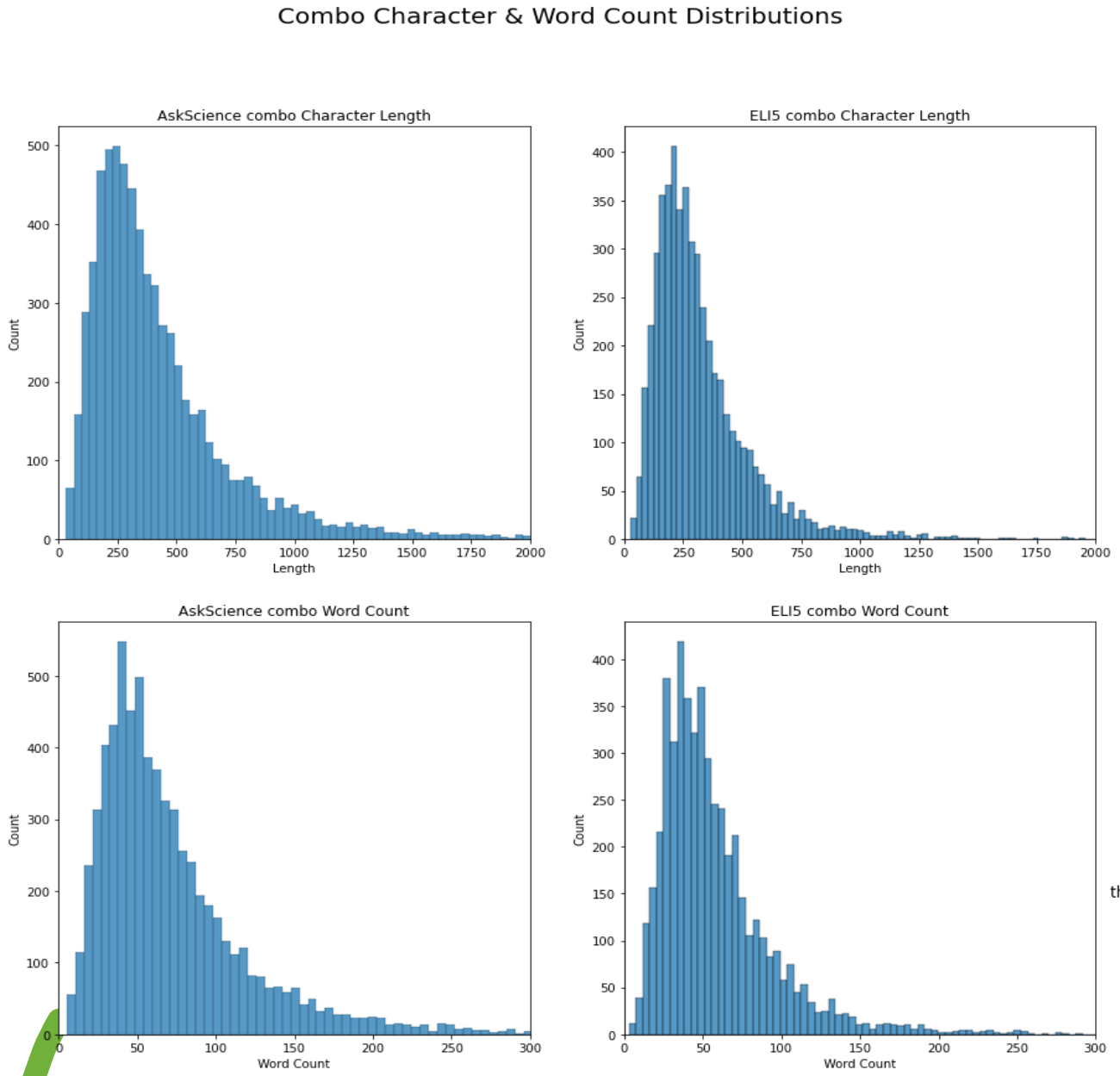
What am I hoping to achieve with this?

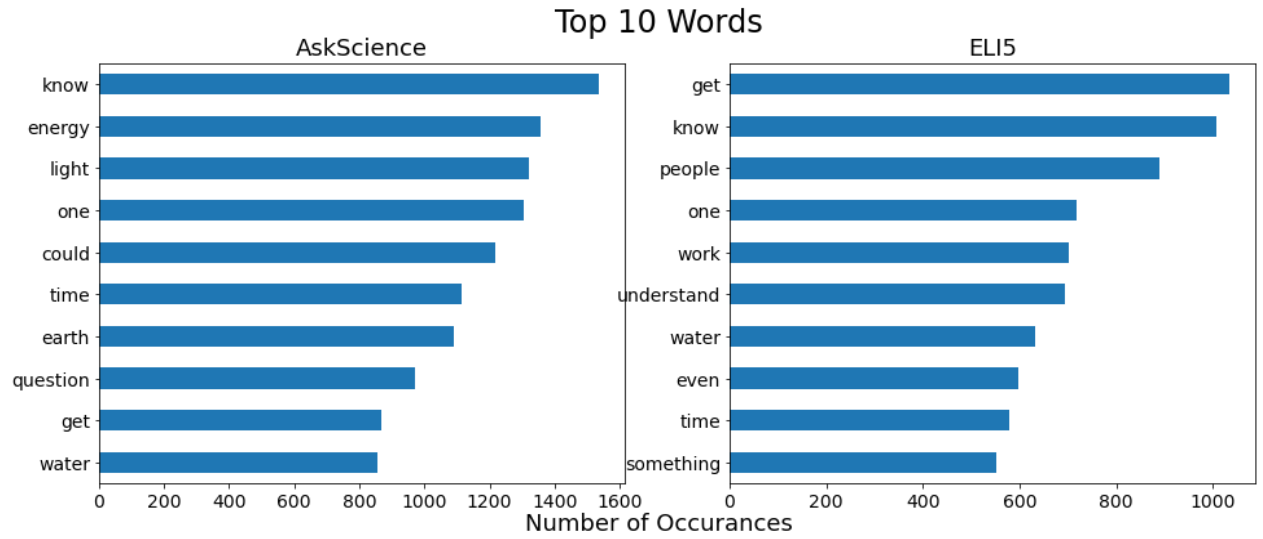
If ELI5 is distinguishable from AskScience.

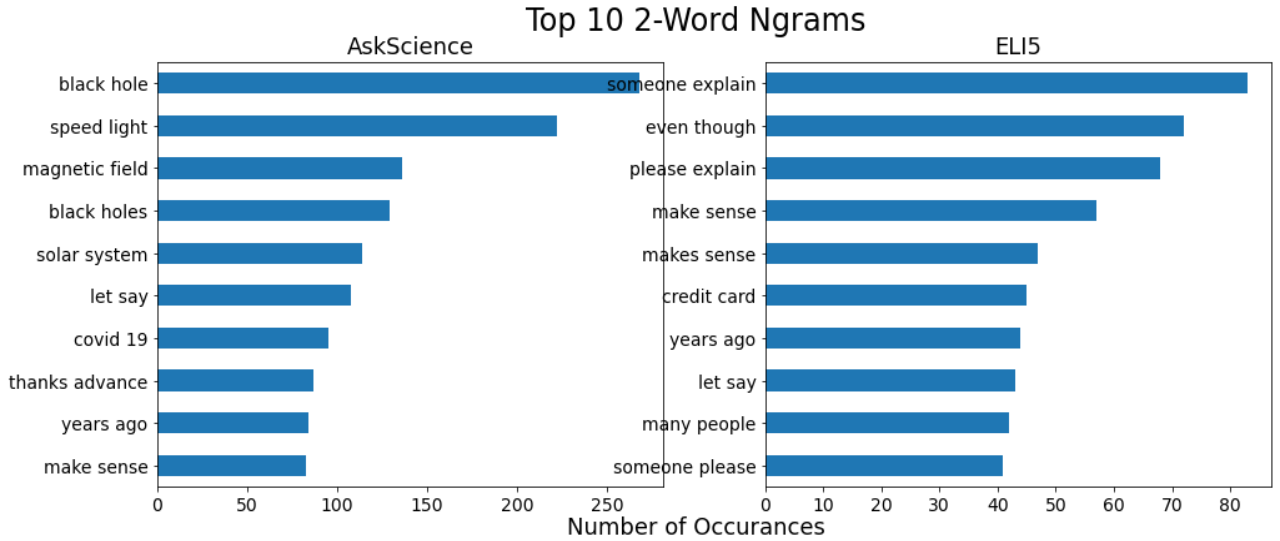
Why?

To see if a subreddit focused on explaining things in a simple manner is that much different than a subreddit that wants to explain it any way they can.

Data


Combo Character & Word Count Distributions
AskScience combo Character Length
Count
Length
ELI5 combo Character Length
Count
Length
AskScience combo Word Count
Count
Word Count
ELI5 combo Word Count
Count
Word Count


Top 10 Words
AskScience
know
energy
light
one
could
time
earth
question
get
water
ELI5
get
know
people
one
work
understand
water
even
time
something
Number of Occurances


Top 10 2-Word Ngrams
AskScience
black hole
speed light
magnetic field
black holes
solar system
let say
covid 19
thanks advance
years ago
make sense
ELI5
someone explain
even though
please explain
make sense
makes sense
credit card
years ago
let say
many people
someone please
Number of Occurances

Modeling

Model 1: Count Vectorizer and Random Forest

Model 2: TfidfVectorizer and Random Forest

Model 3: TfidfVectorizer and KNeighborsClassifier

Overall Model Analysis

Model 1 Analysis:

- Training accuracy = 97.6%
- Testing accuracy = 75.2%

Model 2 Analysis:

- Training accuracy = 99.4%
- Testing accuracy = 75.6%

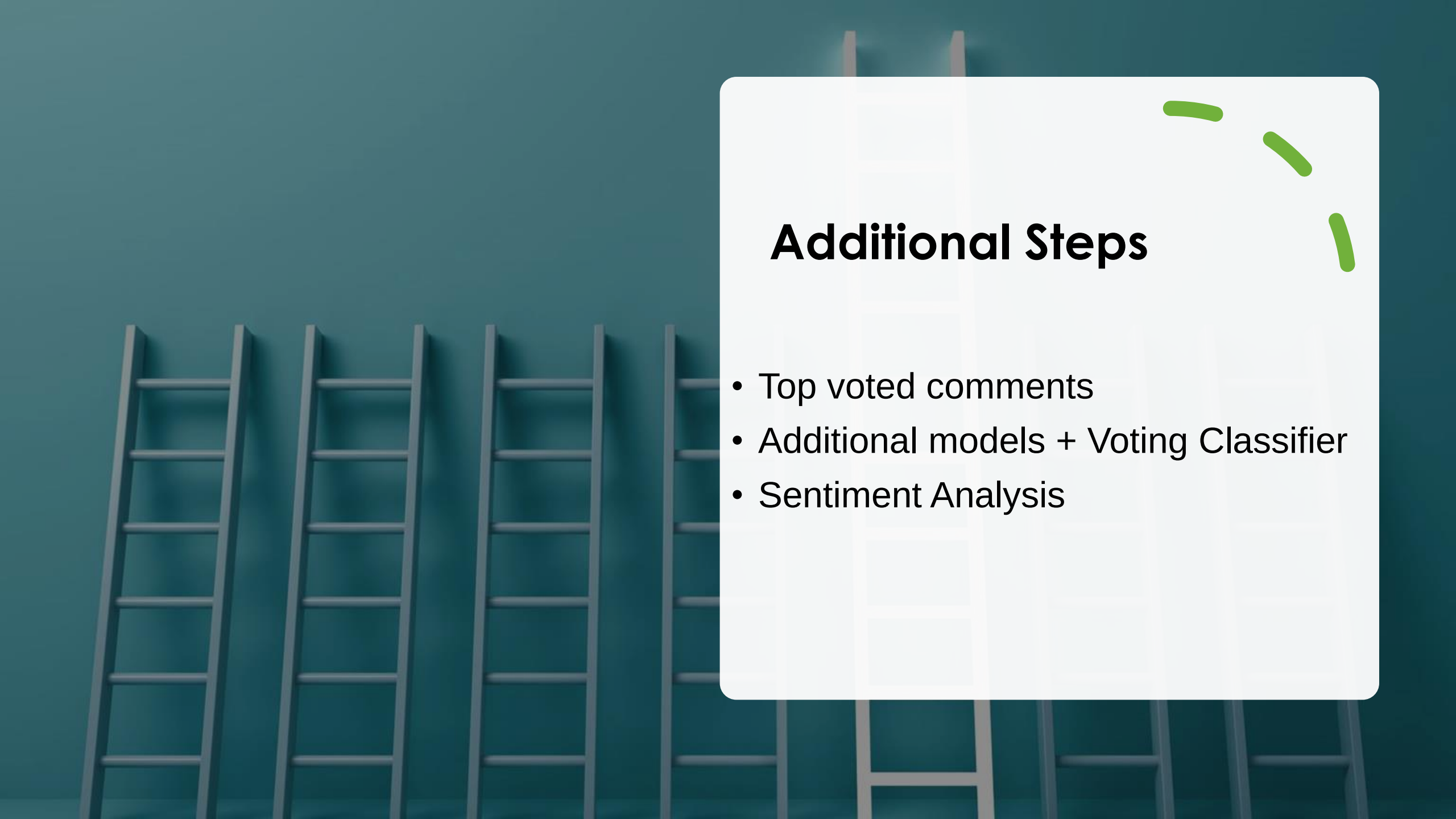
Model 3 Analysis:

- Training accuracy = 80.0%
- Testing accuracy = 69.8%

Conclusion and Recommendations

All 3 models beating our null model by such a degree signifies that there is a discernable difference between the two subreddits, AskScience and ExplainLikeImFive.

These models individually gave us a good inclination towards this conclusion and since they all point to the same conclusion, we are even more sure.



Additional Steps

- Top voted comments
- Additional models + Voting Classifier
- Sentiment Analysis

Any Questions?