



# Project 3: NLP

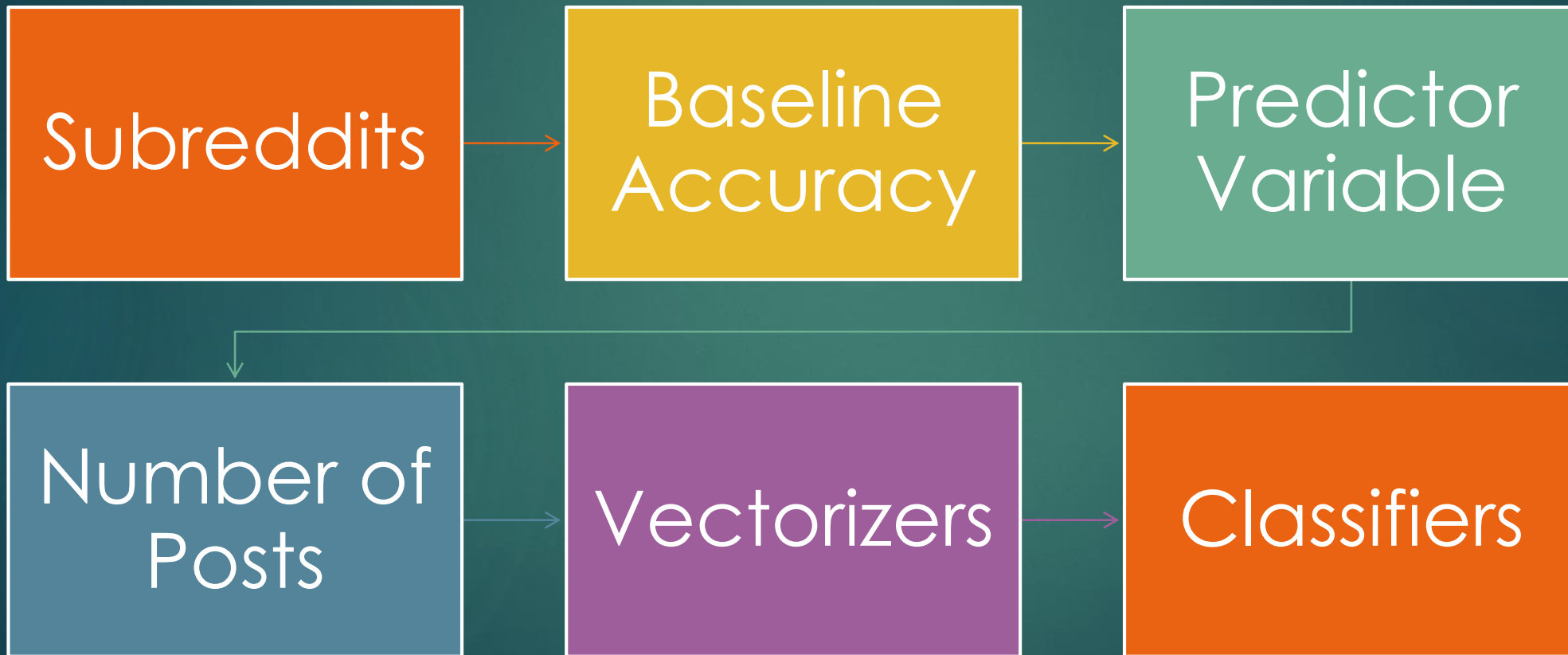
EBONI LEE

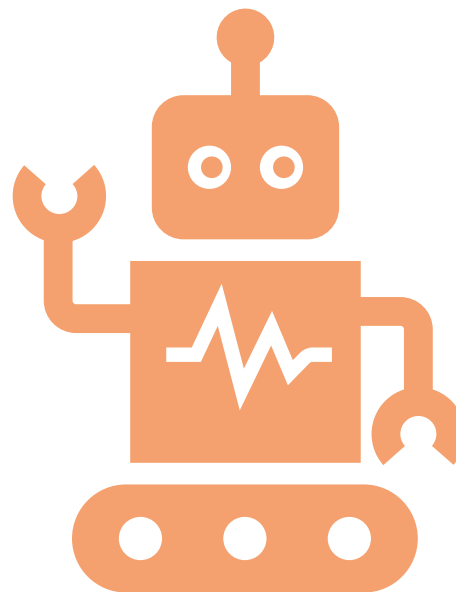
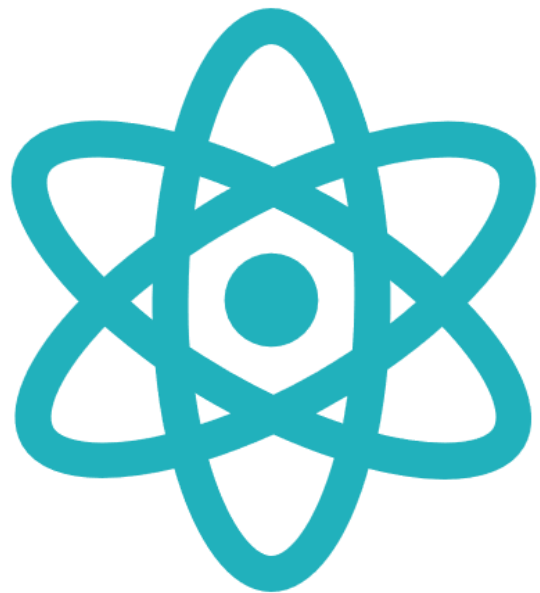


# Problem Statement

CAN AN NLP CLASSIFIER BE USED  
TO DETERMINE WHICH SUBREDDIT  
A GIVEN POST COMES FROM?

# Set-Up





# Science & Tech

INTERCHANGEABLE?

WHO WON?

How Many  
Models Does it  
Take to Get  
100%  
Accuracy?

# Who are our contenders?

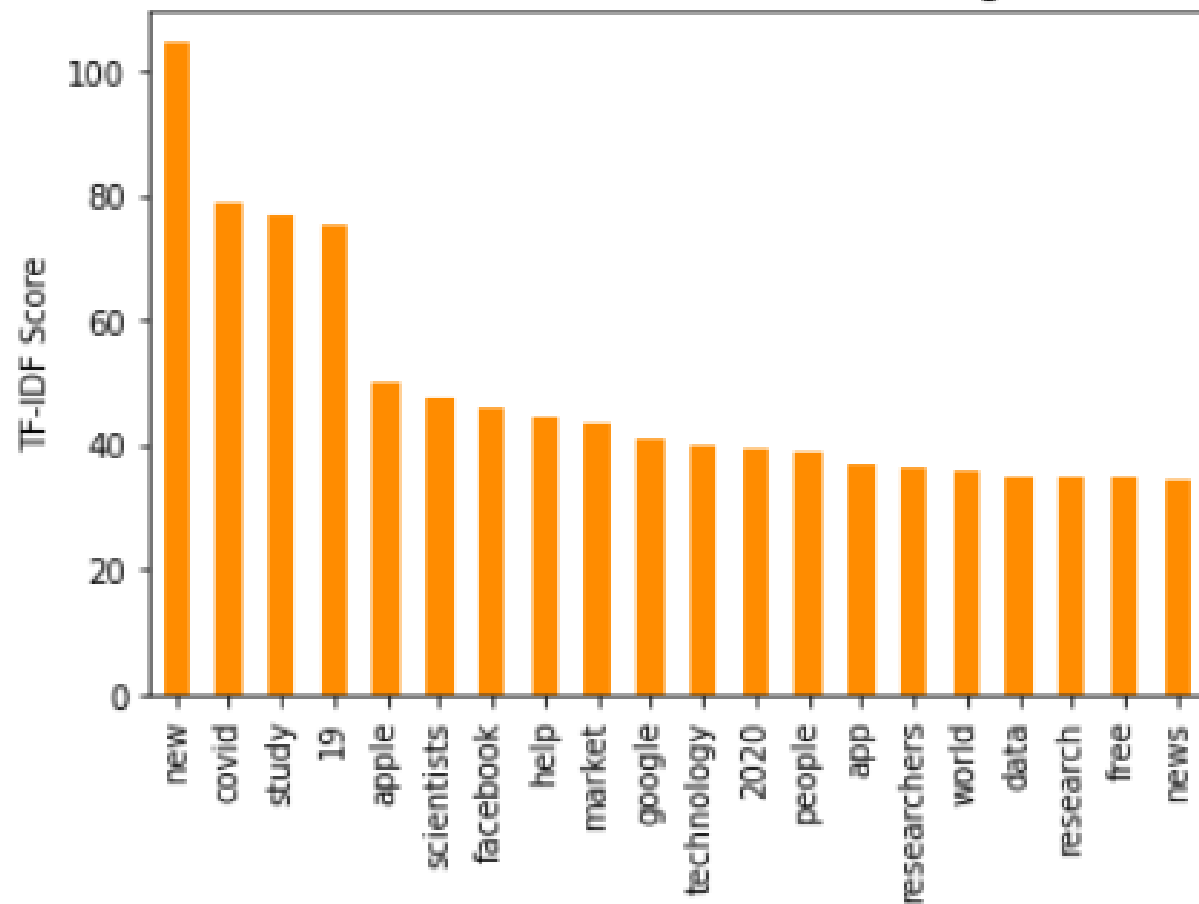
TfidfVectorizer + Naïve Bayes

- ▶ Hyperparameters
  - ▶ Stop words
- ▶ Testing Accuracy: 89.36%

Lemmetizer + TfidfVectorizer + Logistic Regression

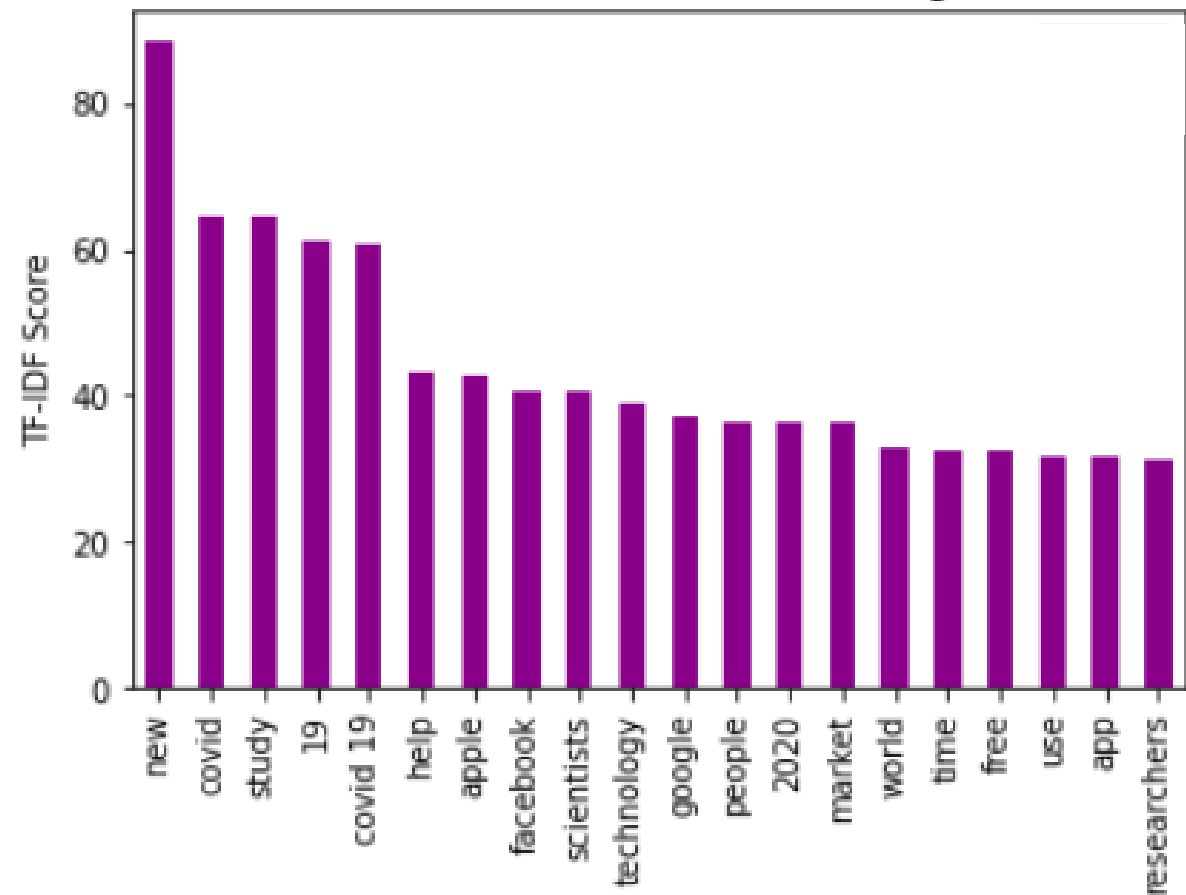
- ▶ Hyperparameters
  - ▶ Stop words
  - ▶ Max df
  - ▶ Min df
  - ▶ Ngram range
- ▶ Testing Accuracy: 89.36%

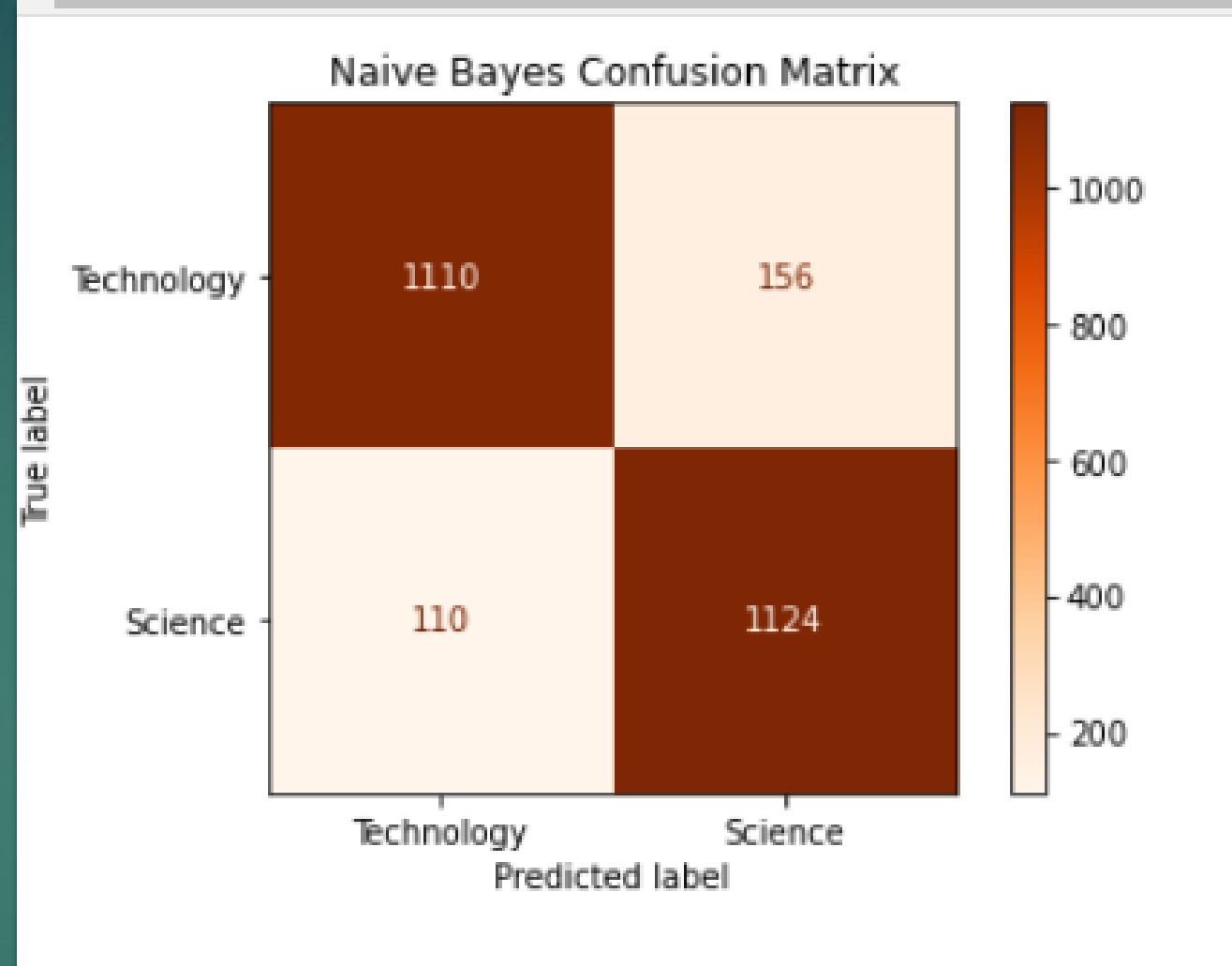
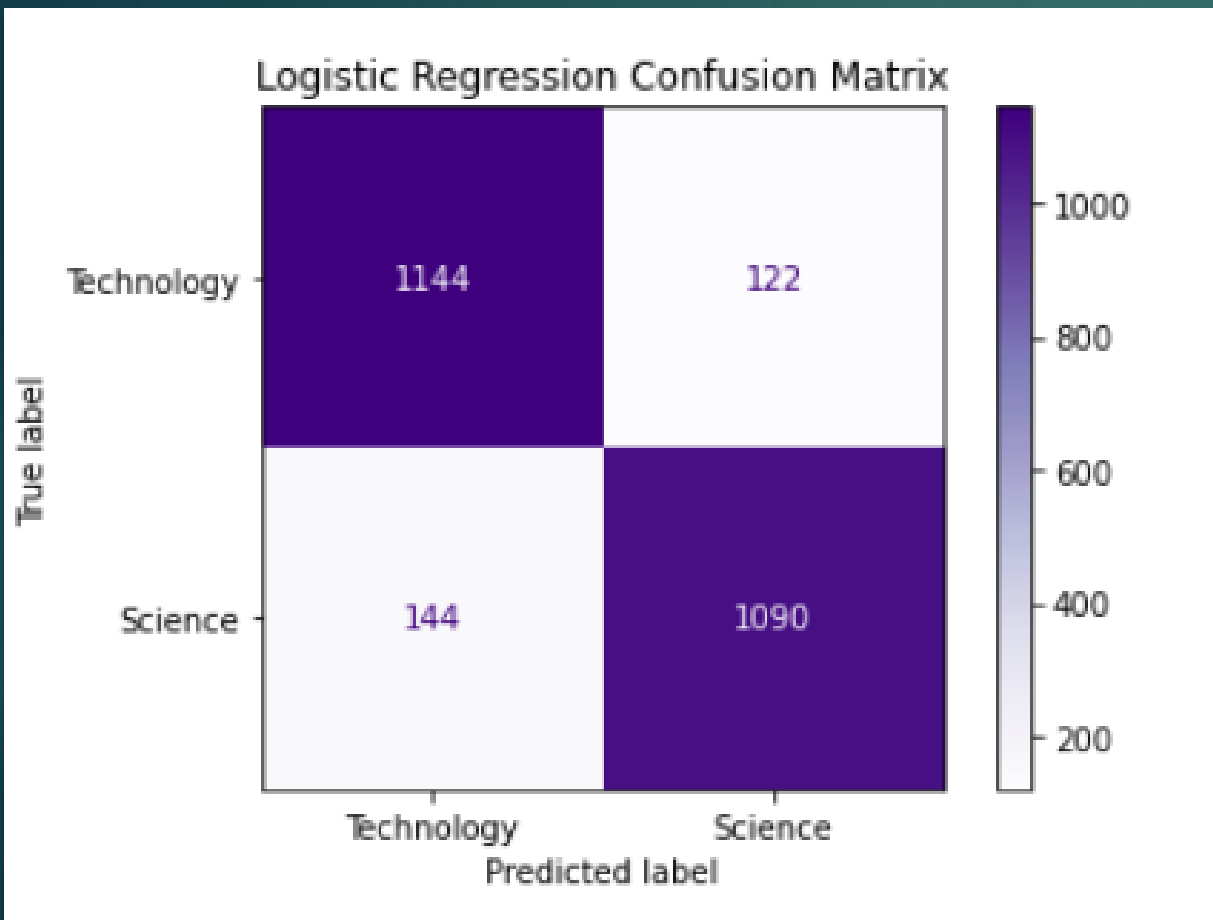
TF-IDF Score with No Lemmetizing



# TF-IDF Scores

TF-IDF Score with Lemmetizing



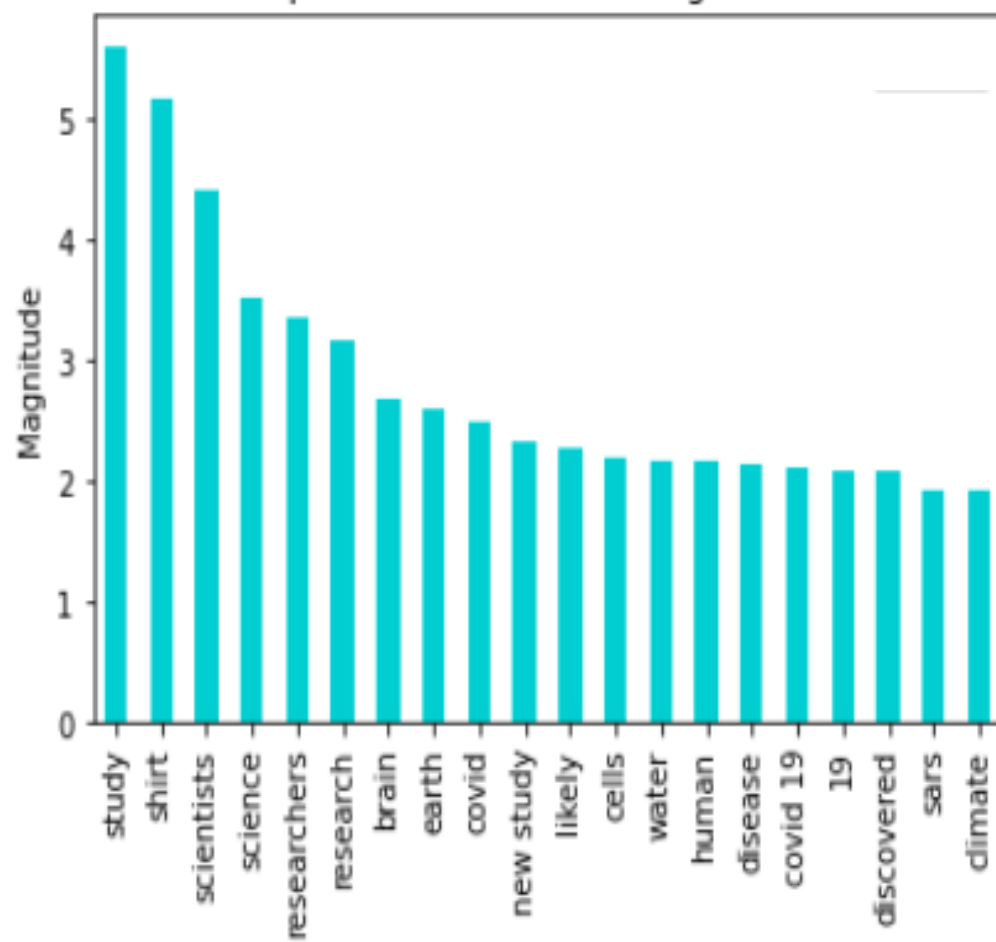


# Confusion Matrices

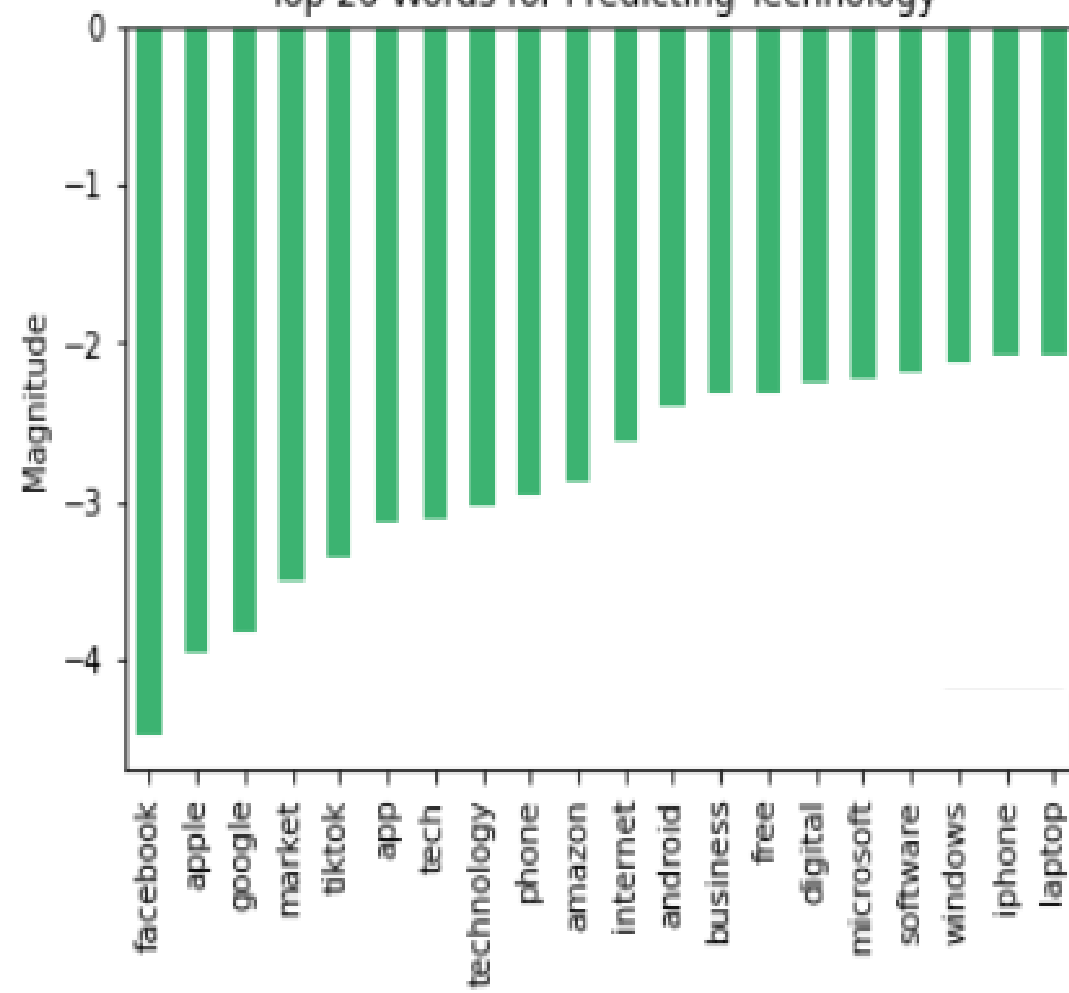


# Interpreting Logistic Regression Coefficients

Top 20 Words For Predicting Science



Top 20 Words for Predicting Technology



Next Steps?