Project 3: NLP

EBONI LEE



Problem Statement

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

Set-Up

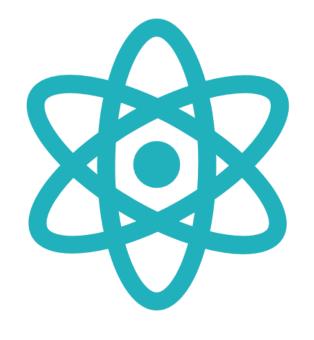
Subreddits

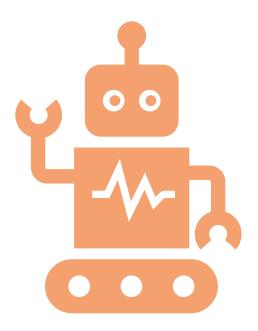
Baseline Accuracy Predictor Variable

Number of Posts

Vectorizers

Classifiers





Science & Tech

INTERCHANGEABLE?

MHO MONS

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

Who are our contenders?

TfidfVectorizer + Naïve Bayes

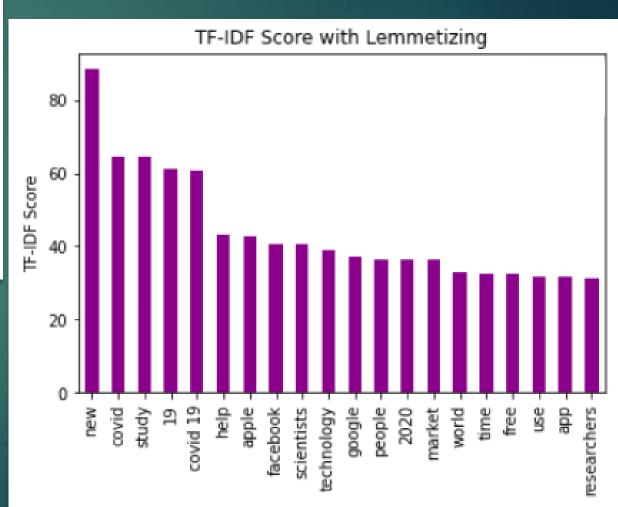
- Hyperparameters
 - Stop words
- ► Testing Accuracy: 89.36%

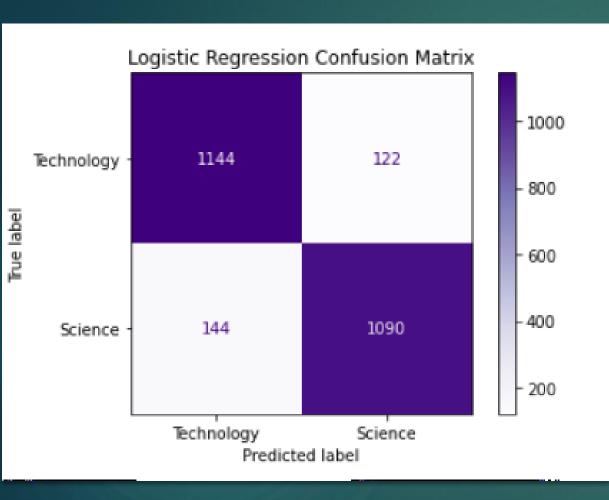
Lemmetizer + TfidfVectorizer + Logistic Regression

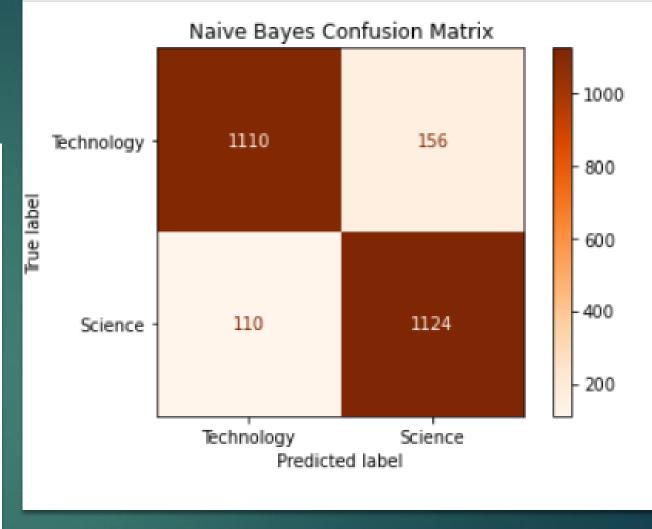
- Hyperparameters
 - Stop words
 - Max df
 - ▶ Min df
 - Ngram range
- ► Testing Accuracy: 89.36%

TF-IDF Score with No Lemmetizing 100 80 TF-IDF Score 40 20 study apple google 2020 people research 티 market dde Morld data 1 Rew 8 Vid RWS 5 facebook researchers scientists technology

TF-DIF Scores

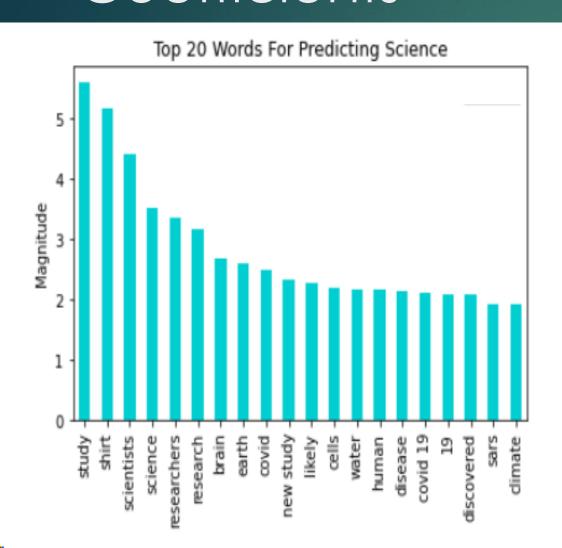


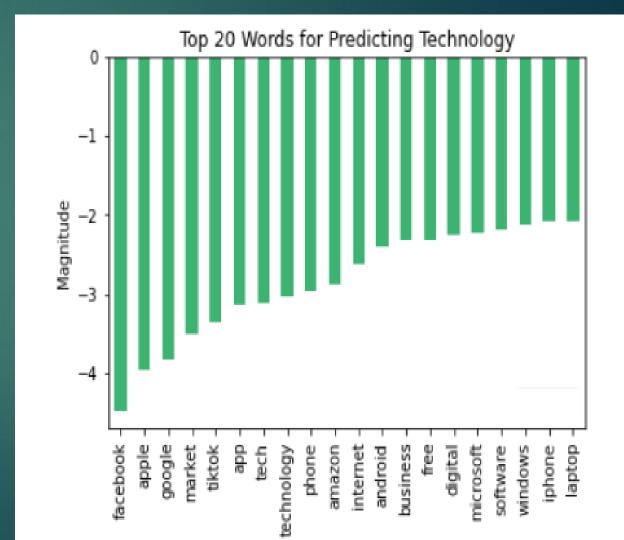




Confusion Matrices

Interpreting Logistic Regression Coefficients





Next Steps?