NLP Modeling Using Subreddits

A presentation by
Jurgen Arias

Summary of Project:

Collect data from two different subreddits: Math and Physics.

Build models to predict which subreddit a post came from.

The Steps:

Getting the data

EDA & Dilemmas

Modeling

Visuals

Conclusions

Getting Posts from Reddit’s API

- Had a great function
- Tweaked parameters to get a lot of data
- Combined two subreddits to make balanced classes (Bootstrapping?)
- Capped at 80k per class

EDA & Dilemmas

- Combined dataframes, combined text and title
- Changed target to binary, dropped nulls
- Checked number of words
- Emoji dilemma 🤔
- Stemming Lemmatization dilemma

Modeling

- Applied Count Vectorizing and Tfid Vectorizing to all models

- Used Pipelines with Gridsearch for:

Multinomial Naive Bayes

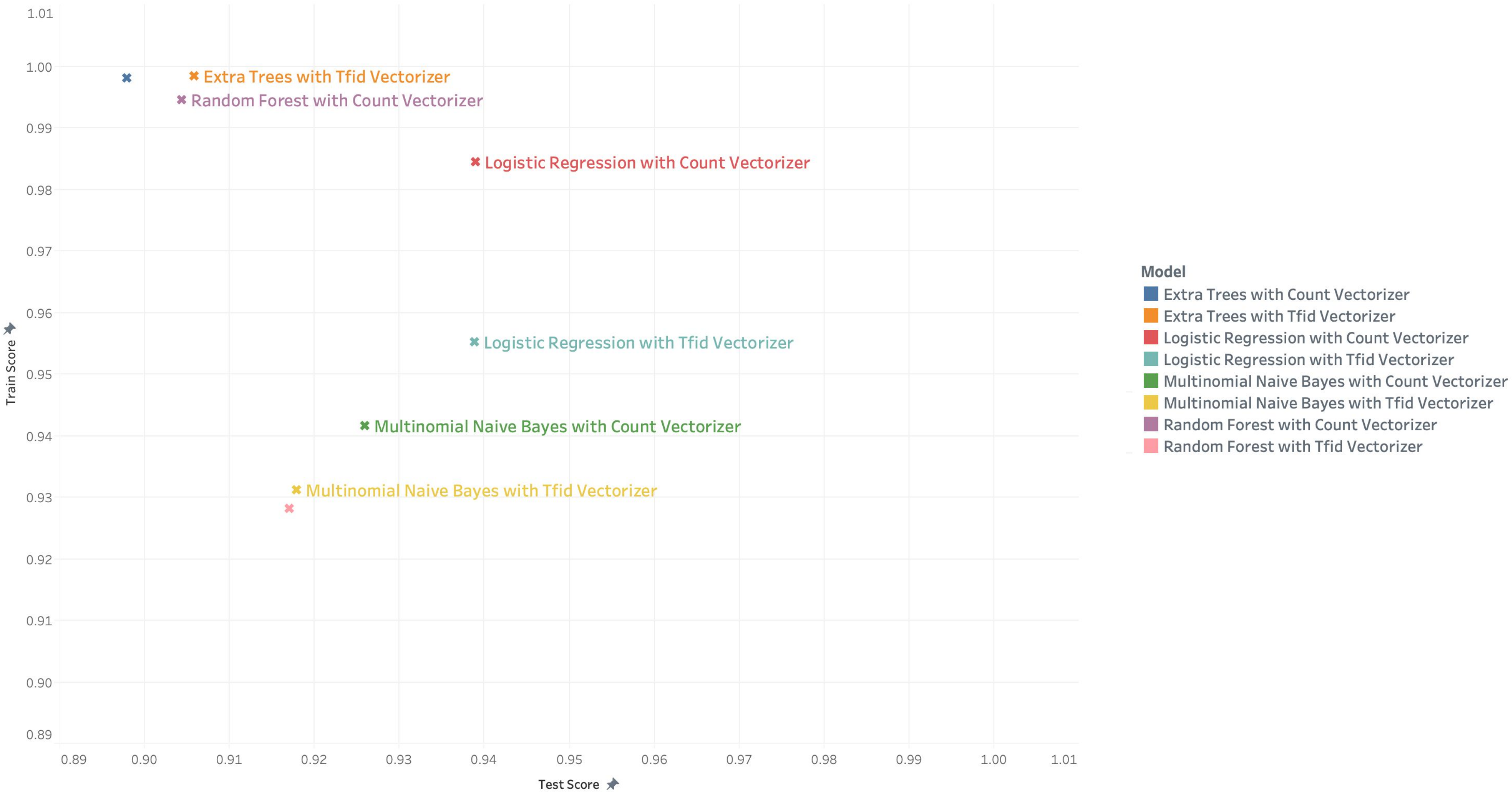
Logistic Regression

Random Forest

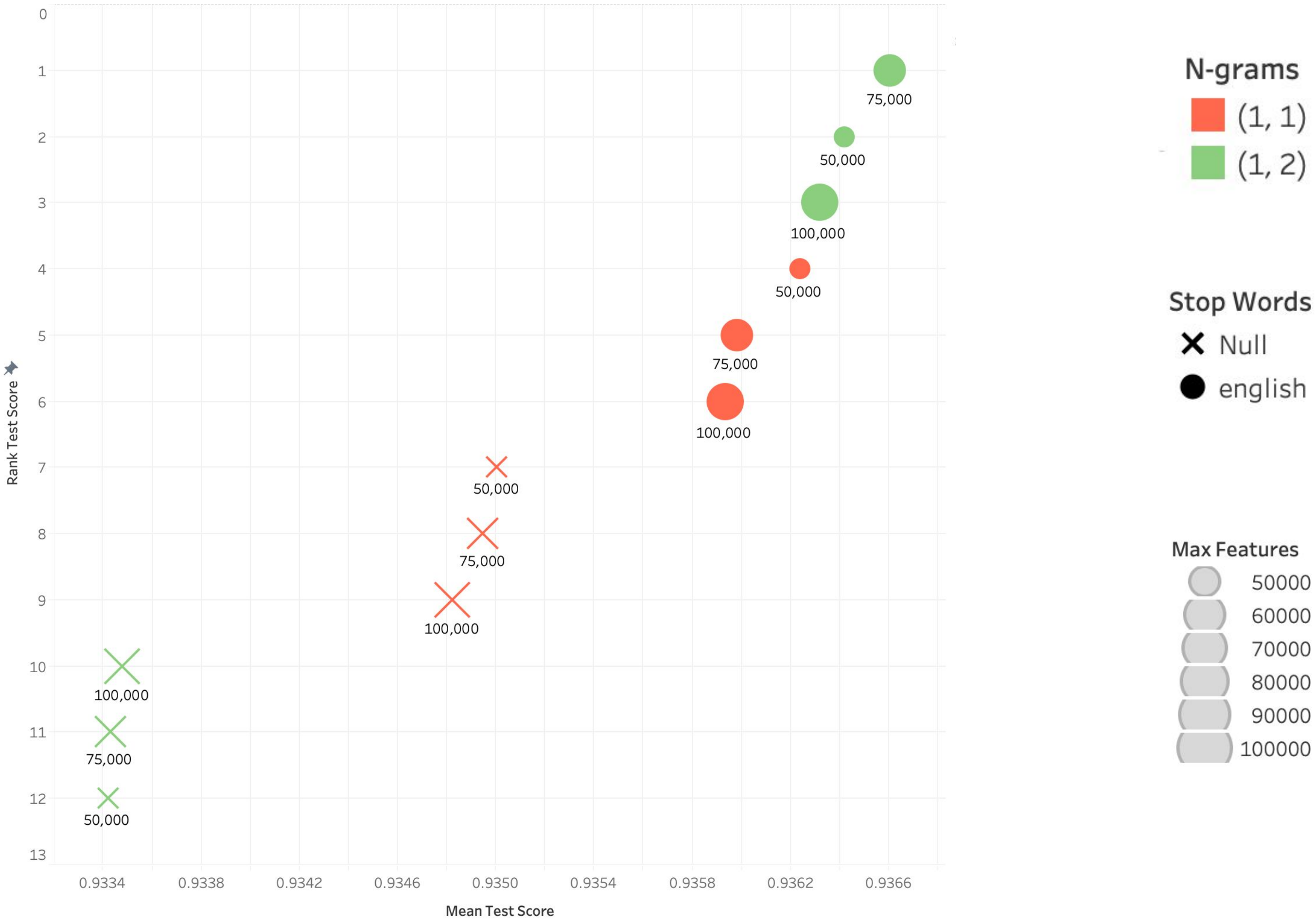
Extra Trees

Vote Classifier

Train and Test Scores by Model



Parameters for Best Mode - Logistic Regression with Tfid Vectorizer



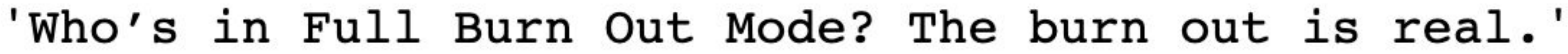
VOTING

Validation

Got 115 more posts from the Math subreddit.

109 of them were predicted to be in Math.

First post was predicted to be physics.



```
'Who's in Full Burn Out Mode? The burn out is real.'
```

Conclusions

Models performed very well

Sentiment analysis

Play with less data

Compare different times

Science website recommendation