Reddit Natural Language Processing (NLP)

John Bruscella
General Assembly | DSIR 113020

Problem Statement

Build a model to classify posts to predict if they are from within one of two subreddits.

1. Use an API to collect data (posts and comments) from two subreddits
2. Use Natural Language Processing to determine which subreddit a given post came from.

Use this model to potentially aid in choosing which subreddit to post in or browse science and technology information in.

Subreddit Overview

r/science

- Approximately 26M members
- Ranked #7 by subscribers
- 60+ posts per day
- 1600+ comments per day

r/technology

- Approximately 10M members
- Ranked #49 by subscribers
- 75+ posts per day
- 2900+ comments per day

Collection and Description of Data

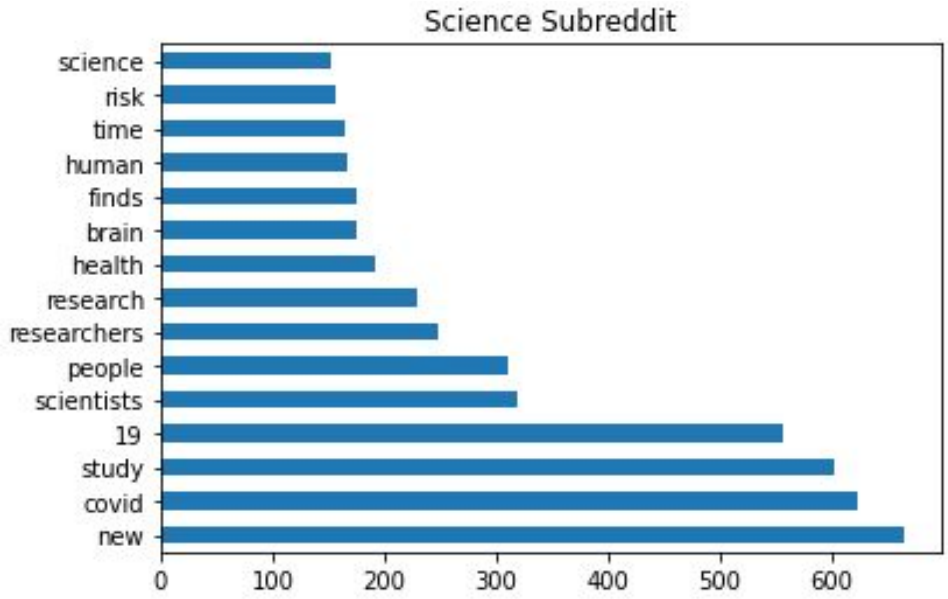
- Pushshift API
- 5000 posts and 5000 comments from each subreddit starting Jan 16, 2021
  - Post titles
  - Authors
  - Time created
  - Upvotes
  - Number of comments, crossposts, awards
  - Score

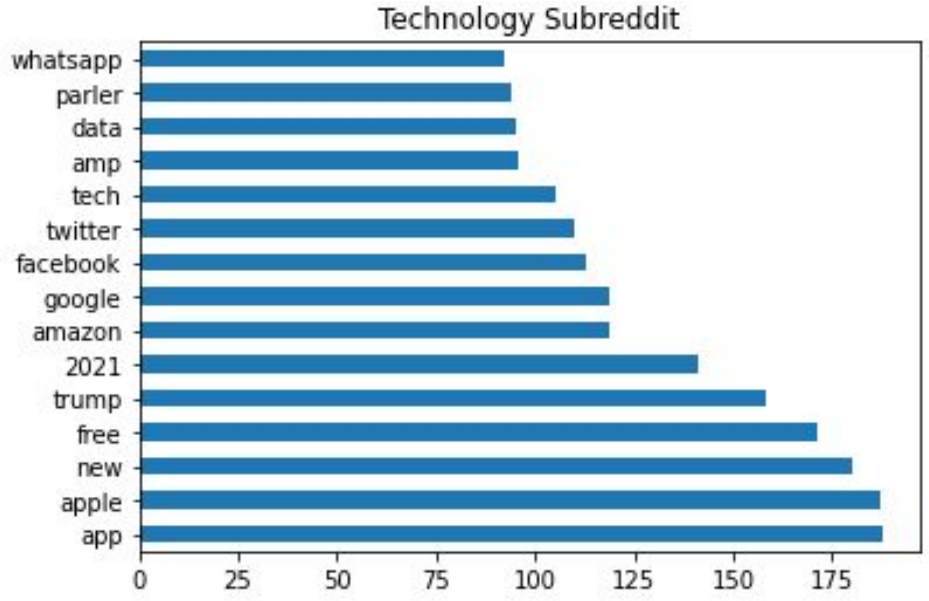
Data Cleaning and EDA

- 10,000 total posts
  - Approximately 2,000 posts dropped due to duplicates

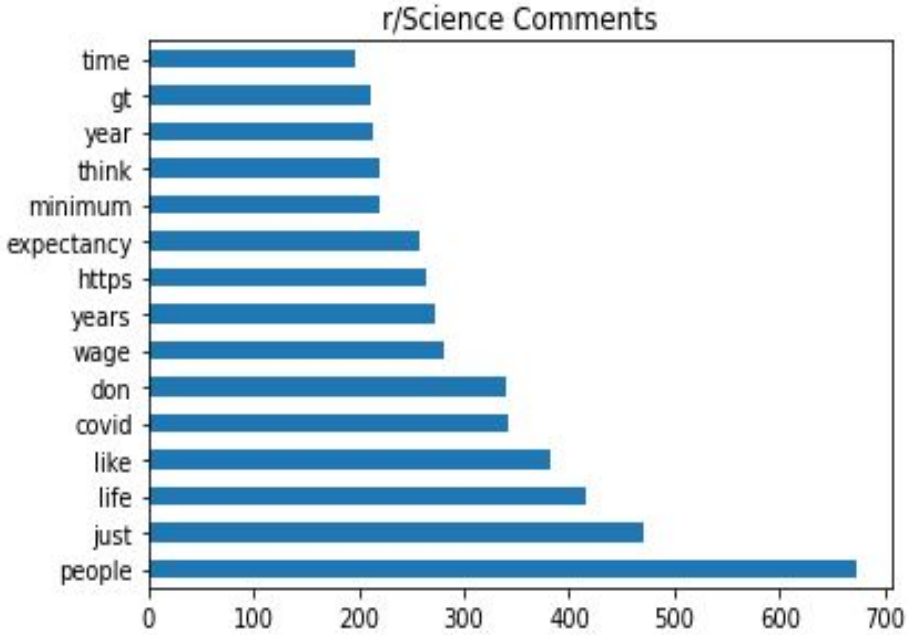
- 10,000 total comments
  - Approximately 2,300 posts dropped due to duplicates

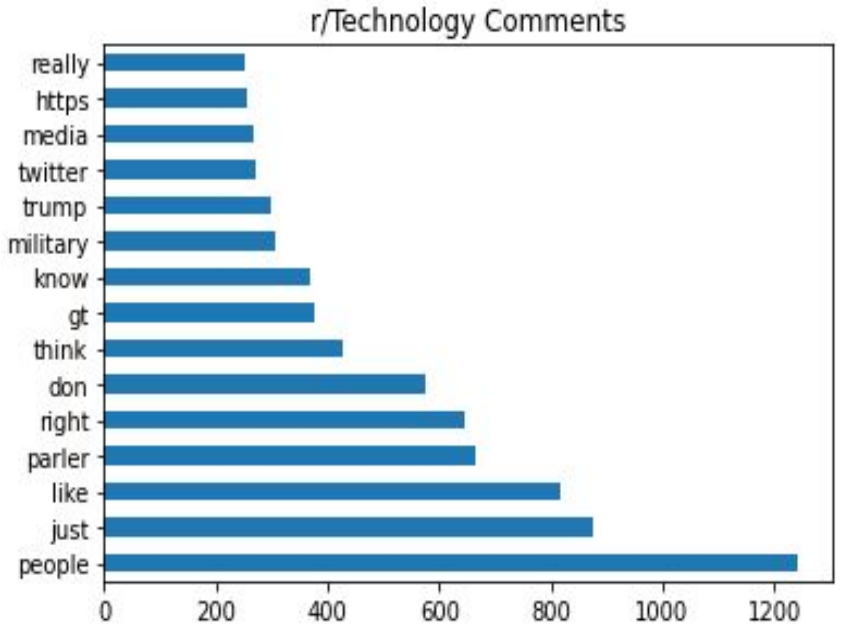
Top Words in Post Titles





Top Words in Comments





Feature Importance - Post Titles & Comments

Post Titles

- Parler
- privacy
- rs
- apple
- customer
- capitol
- hp
- paytm
- router
- android

Comments

- parler
- censorship
- police
- twitter
- amazon
- military
- app
- 19
- media
- left

Models Tested

- Logistic Regression
  - Tfidf
  - Lemmatizer
- Naive bayes
  - Stop Words
- Decision Trees
- Boosting
  - AdaBoost
  - GradientBoost
- Compare to Baseline Model

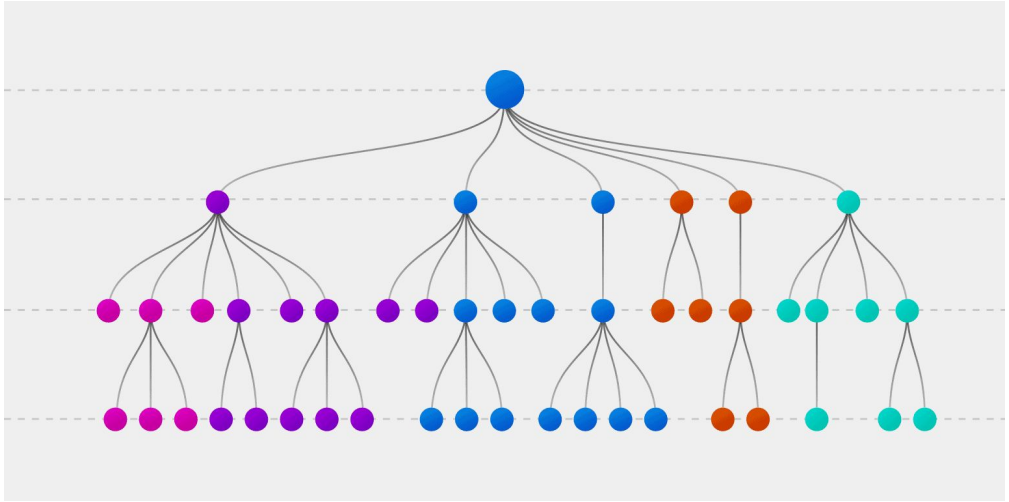
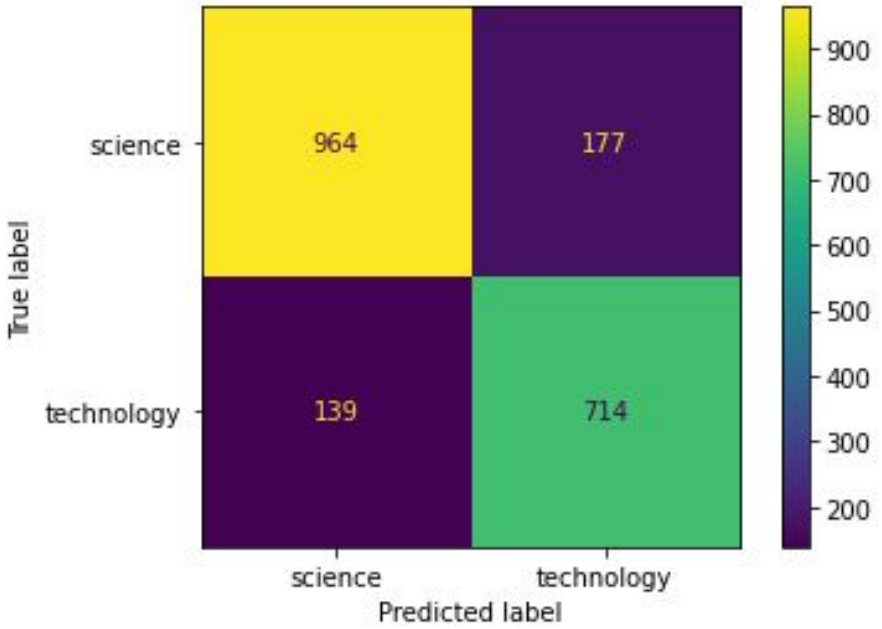


Image source; https://www.explorium.ai/blog/the-complete-guide-to-decision-trees/

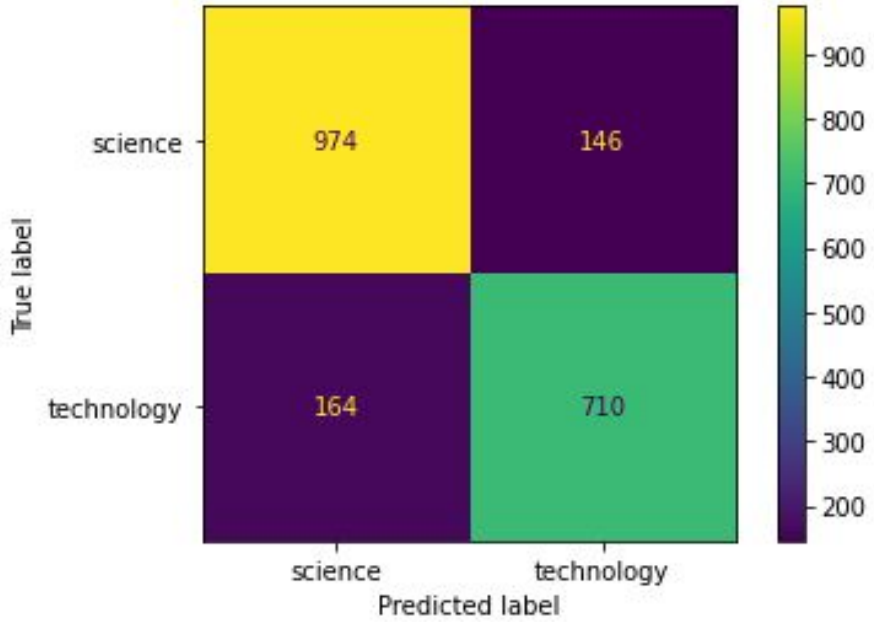
Logistic Regression Model

- Logistic Regression with Lemmatizer
  - BAC is 0.79
  - F1 score is 0.75
  - Slightly Overfit



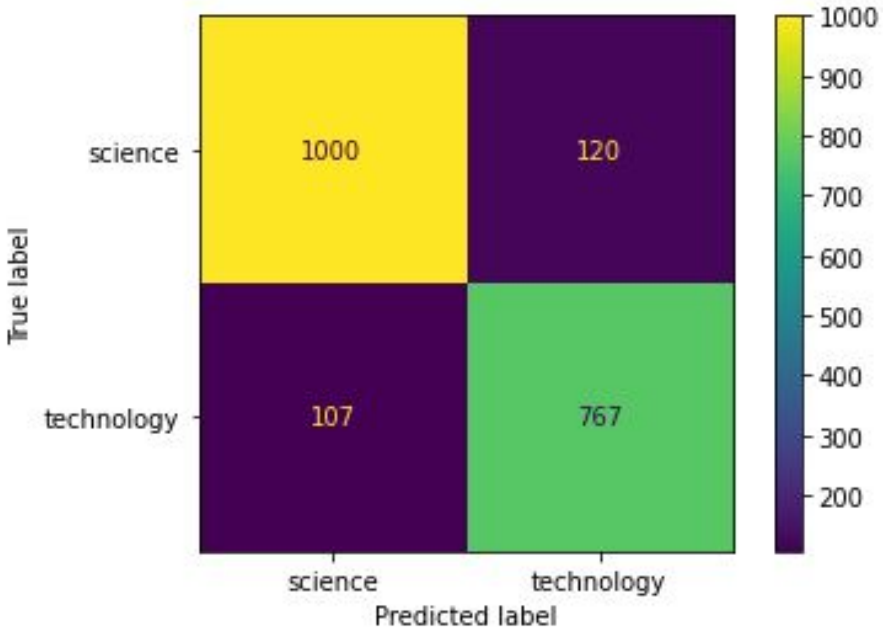
Naive Bayes Model

- Multinomial Naive Bayes with stop words
  - BAC is 0.84
  - F1 score is 0.86
  - Slightly Overfit

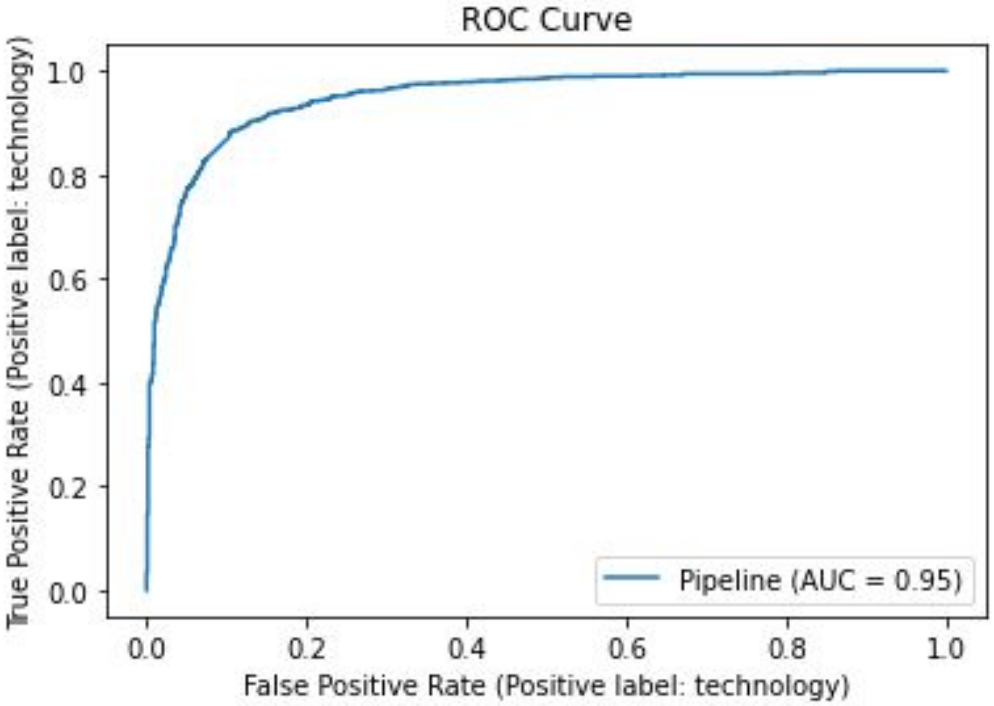


Gradient Boost Model

- Gradient Boost Model with a count-vectorizer
  - BAC is 0.89
  - F1 score is 0.89
  - ROC AUC is 0.94
  - Overfit



Gradient Boost Model



	precision	recall	f1-score
science	0.91	0.89	0.90
technology	0.87	0.88	0.87
accuracy			0.89

Conclusion & Recommendations

- Gradient boost model provides a fairly accurate classification
- Slightly more emphasis on politics in r/technology
- This model may aid someone looking to post in or browse these specific subreddits.

- To improve the model:
  - Collect more data
  - Look at other tokenizers
  - Analyze n-grams further

Sources and References

- Subreddit statistics (https://subredditstats.com/r/science)
- Logos:
  - https://www.redditinc.com/brand
  - https://www.reddit.com/r/technology/
  - https://www.reddit.com/r/science/
- Decision Tree Image source (https://www.explorium.ai/blog/the-complete-guide-to-decision-trees/)

Questions?