

Classifying Reddit Posts

- Aruna Rayapeddi
- 04/05/2019

Social Media





Home

Popular

Search Reddit



LOG IN

SIGN UP

Welcome to your starter home!

Browse the feed below and join or leave communities to make it your own



Home



Popular



Search Reddit



3 Comments Share Save ...

↑
138
↓



r/datascience · Posted by u/lil_ponY 11 hours ago

If your company offered you \$2,500 to get a Data Science certificate, which certificate would you pick? Education

So my company offered to sponsor me for a Data Science certificate, since I will be the only one without a masters degree in the data science team and they want me to at least have a certification or some sort of qualification

Problem Statement




- Considering two subreddits, given a Reddit post, Identify and classify which subreddit the post belongs to.
- This is a binary classification problem.

Data

Two Subreddits chosen:






- Jokes
- Data Science




↑ 41.6k ↓  **r/Jokes** · Posted by u/Maimonides_vii 1 month ago  5 

Can we ban "Yo Momma" jokes from this sub? They're old, stupid, and have been done by literally everyone hundreds of times

Just like yo momma.

 1.5k Comments  Share  Save  Hide  Report

80% Upvoted

↑ 39 ↓  **r/datascience** · Posted by u/osbornep 8 hours ago

Was it Worth Studying a Data Science Masters? (My experience)

Discussion

Hi,

Since completing my Masters in Data Science, I have had a number of people contact me asking for my experience with the course and whether it is worth recommending

Data Gathering

- Requests information from APIs , from scraping.
- Number of posts gathered from “Jokes” Subreddit - (2496, 9)
- Number of posts gathered from “Data Science” Subreddit - (1839, 9)
- Total number of posts - (4335, 9)

Data Cleaning & Exploration

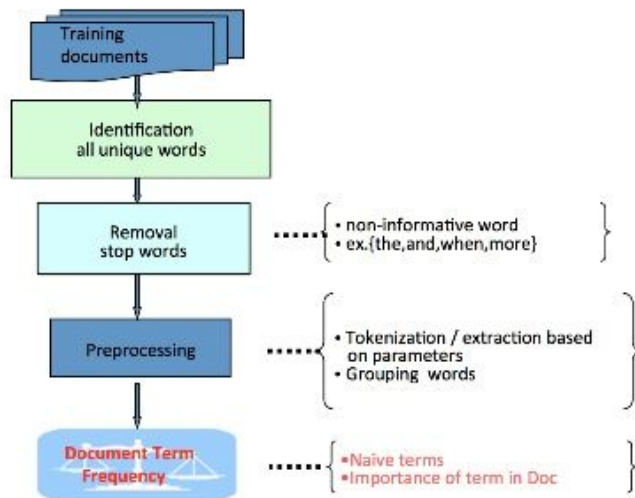
- Data has been explored for nulls, duplicates
- Checked for unbalanced classes(this is binary classification)
- Combined text and titles

CountVectorizer

- NLP tool, that converts a collection of text documents to a matrix of token counts

What is CountVectorizer

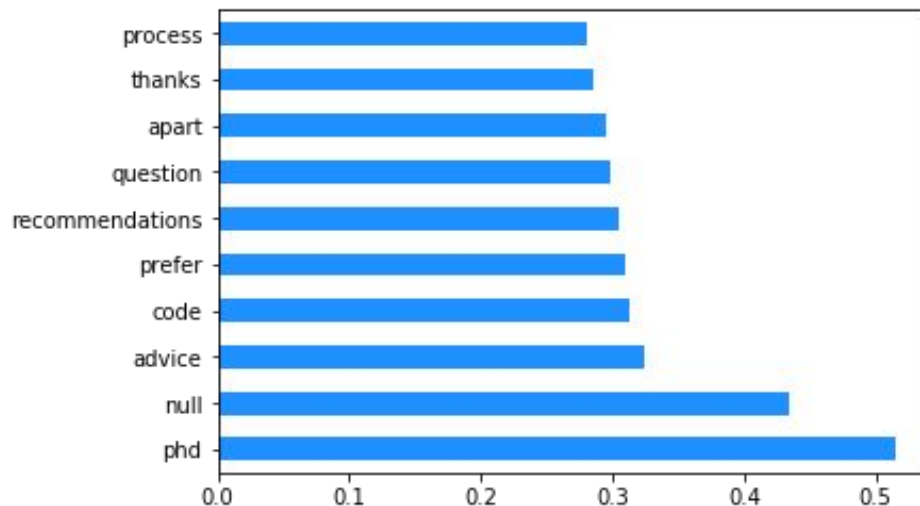
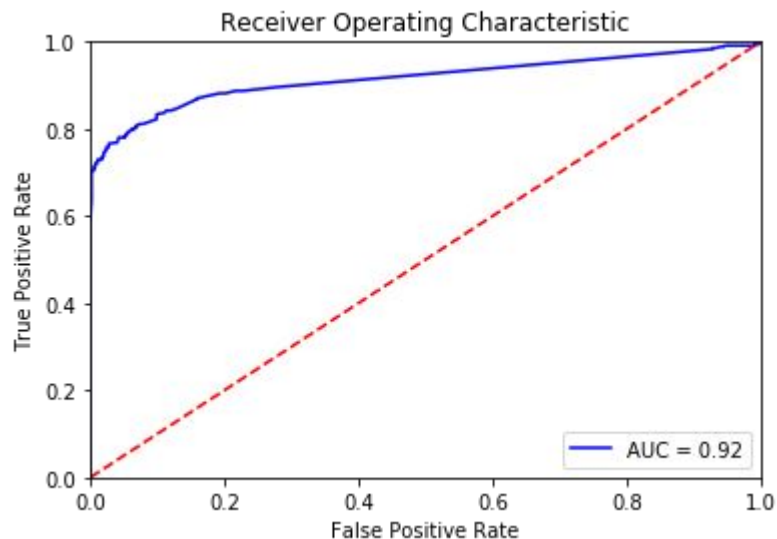
- Class in Python Scikit-learn, ML library



Model Evaluation & Interpretation

Logistic Regression:

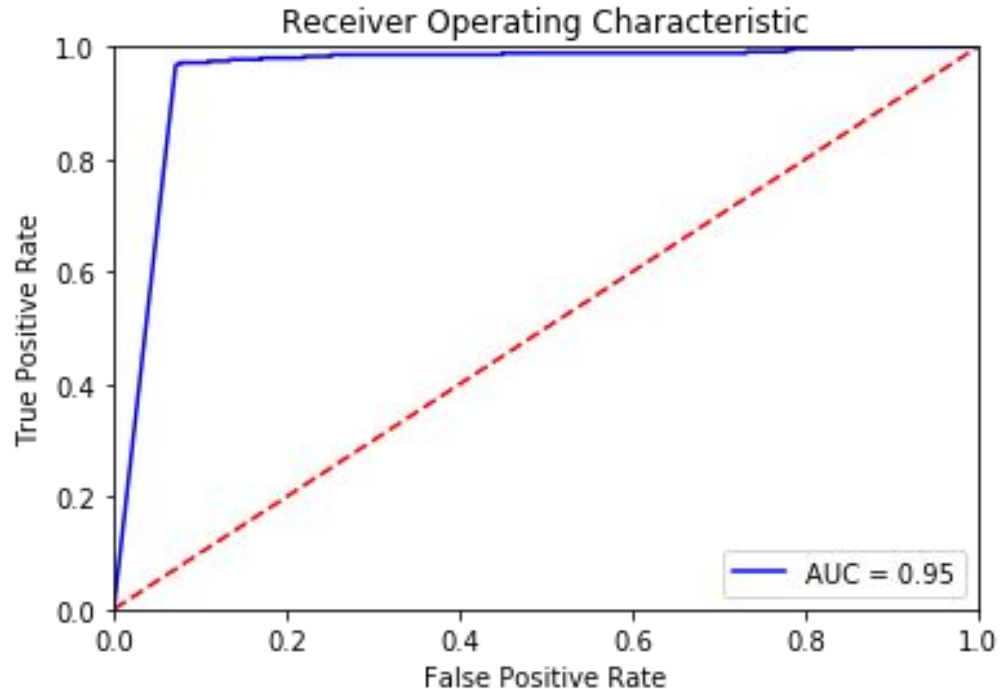
Accuracy - 94%



Model Evaluation & Interpretation

Multinomial Naive Bayes classifier:

Accuracy - 96%



Key Takeaways

- Need to consider subreddits that are closely related to get better prediction
- Test with other models
- Explore new features
- More cleaning needed, especially dealing with social media(links, #tags,emojis etc;).