

Project 3 - Web APIs & NLP

Bryan Goh

What is Reddit?



↑ 33.0k **ELI5: How come you can be falling asleep watching TV, then wide awake when you go to bed five minutes later?**

Repost

1.7k Comments Award Share Save ...

This thread is archived

New comments cannot be posted and votes cannot be cast

Sort By: Best

View discussions in 1 other community



Solid_Waste · 5 yr. ago · 2 5 3 4 & 2 More

The brain is like a group of people talking to each other. When you're watching TV, the part of your brain that watches TV says "Shut up guys, I'm watching TV," so you can focus without thinking about cake or math. As a result, the others sit silent, grow bored, and fall asleep, until only the TV watcher part of the brain is left. Left by himself, he too gets bored and falls asleep.

When you're in bed, assuming you aren't counting sheep or something, the entire brain is kind of in free time mode, and any part of the brain can speak up if it wants to. They start talking to each other, and even if one of them starts to drift to sleep, the others wake it up either by deliberately talking to the sleepyheads or just being noisy. Eventually more and more of the parts of the brain fall asleep from sheer exhaustion no matter how loud the others are, and eventually the last one passes out and you are asleep.

↑ 27.7k ↓ Give Award Share Report Save

↑ 34.3k **Heavy**

limgur.com/p58VWb...

Meme

Doctor: "please step on the scale"

Me:



This is a subreddit for posting pictures or memes involving chubby cats (also known as chonkers). Please keep in mind we do not condone pet obesity, and instead would like to help people who do have obese cats get them on a diet and help them lose weight and become healthier. There are two links in the sidebar to help anyone who doesn't know how to do this and you're more than welcome to ask the community for advice. Enjoy!

604k Chomnk Lovers 120 Appreciating Chomnks

Created Sep 12, 2018

Joined

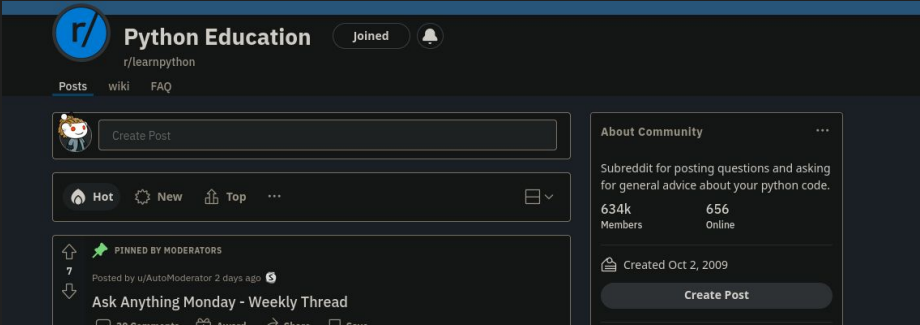
Community options

Powerups

Powerup to unlock perks for r/Chonkers

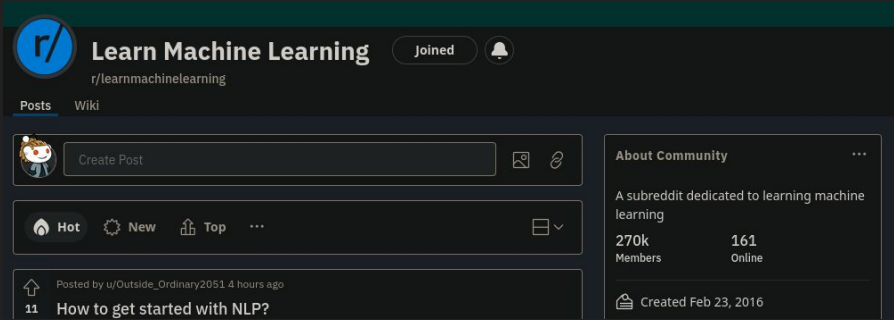
Subreddits

r/Learnpython



634k followers

r/Learnmachinelearning



270k followers

Subreddits

r/Learnpython

↑ r/learnpython · Posted by u/JLaurus **I get paid to write python** 3 years ago

2.6k 2 10 9 14 3

↓ I'm 100% self taught, landed my first job! My experience!

Hi all,

Firstly this is going to be a long post to hopefully help people genuinely looking to commit to becoming a developer by sharing my story of how I went from absolutely zero knowledge of programming (as you can see by my post history) to landing my first python developer role.

Location: UK

To kick things off about a year ago I wasn't happy with the job(s) I was doing, long hours, very low pay, so I came across python by chance. Yes I admit the money was what attracted me alone to start off with as I am quite a money motivated person. Ofcourse I knew and still know it will be a long journey to reach the salaries offered but I have managed to finally get my first step on the ladder by landing a job as a python developer. Enough of the story, lets get on with it.

I will list all of the youtube playlists and channels I watched over and over again. Bear in mind whilst reading these books I did watch a lot of videos in between reading aswell!! What books I read, in order.

First book:

Python Crash Course: A Hands-On, Project-Based Introduction to Programming - Eric Matthes Review: Great first book, my advice, skip the game and django project and just do the matplotlib project for now (come back to django later down the line once you understand the HTTP protocol and how requests work)

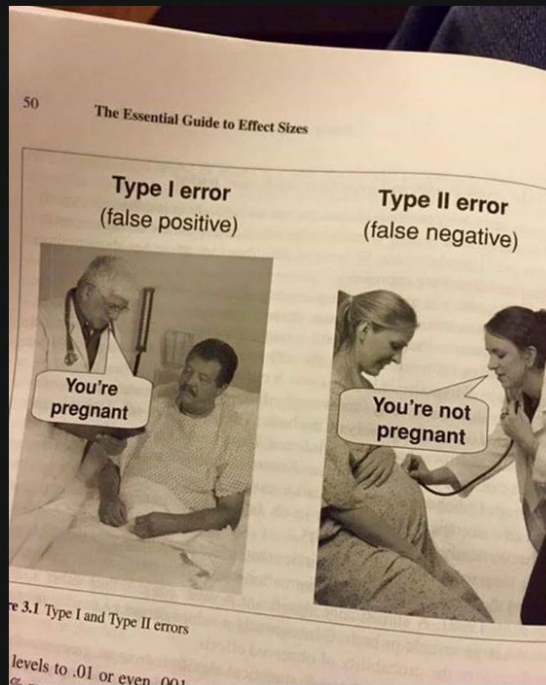
10/10 recommend

p.s. I know a lot of people recommend reading Automate the boring stuff and I regret not reading it after this one!

r/Learnmachinelearning

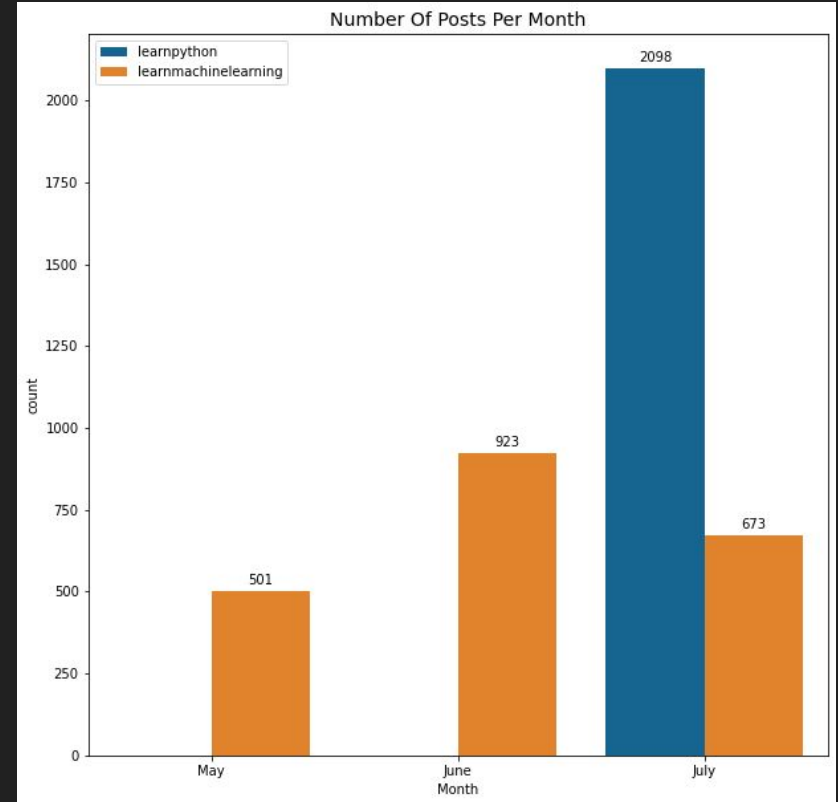
↑ r/learnmachinelearning · Posted by u/TheCodingBug 1 year ago 2.5k 2 3

↓ A simple and easy-to-remember example for false positives and false negatives.



Data Extracted

1. Learn Python - 2098 posts
2. Learn Machine Learning - 2097 posts

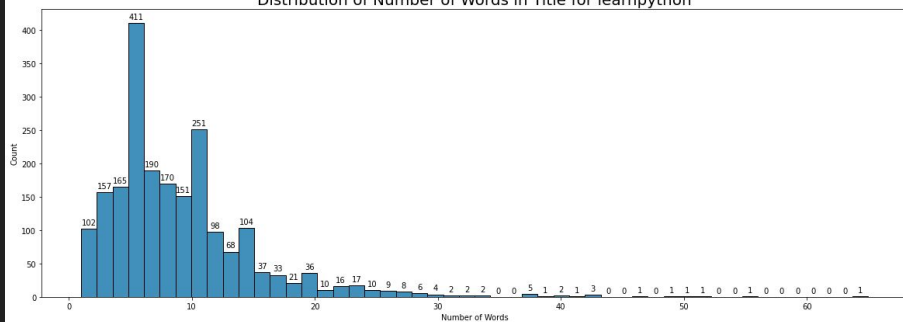


No. Of Words In Title And Content

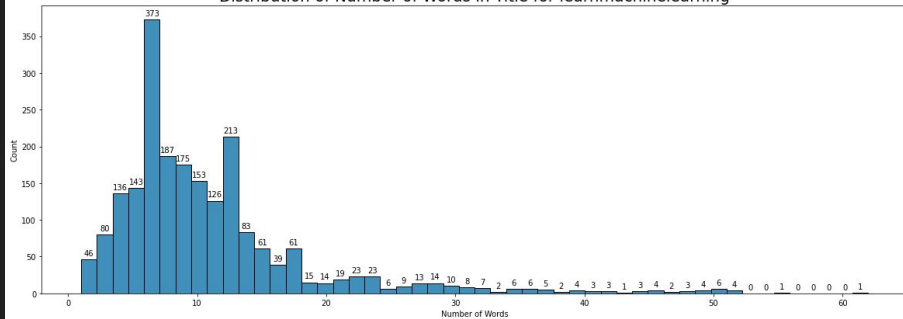
Title

Selftext

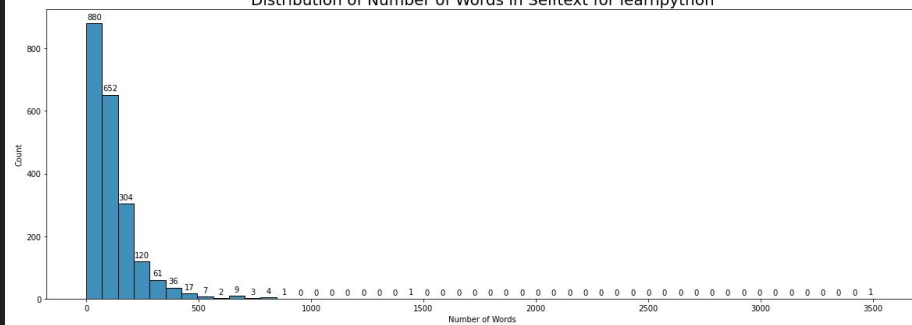
Distribution of Number of Words in Title for learnpython



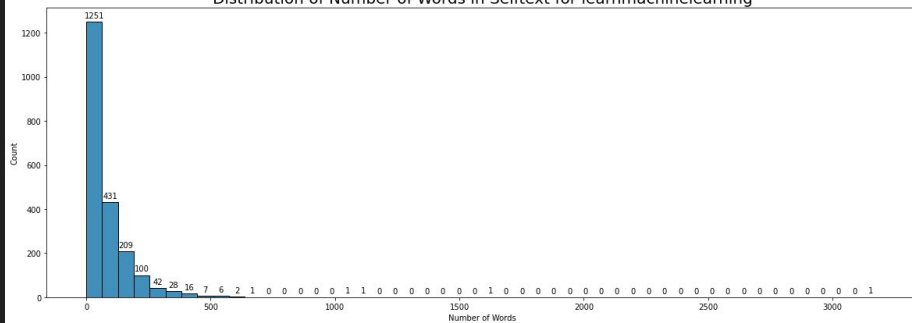
Distribution of Number of Words in Title for learnmachinelearning



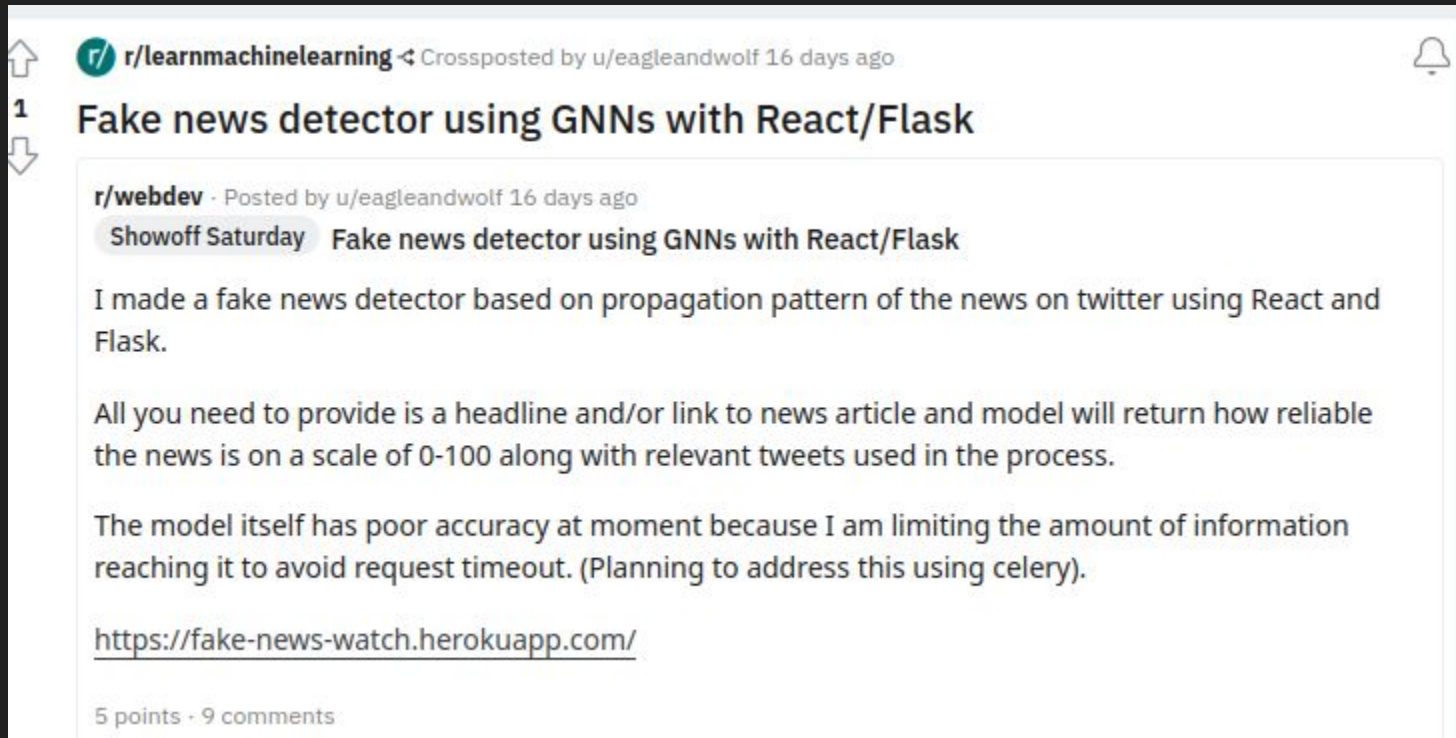
Distribution of Number of Words in Selftext for learnpython



Distribution of Number of Words in Selftext for learnmachinelearning

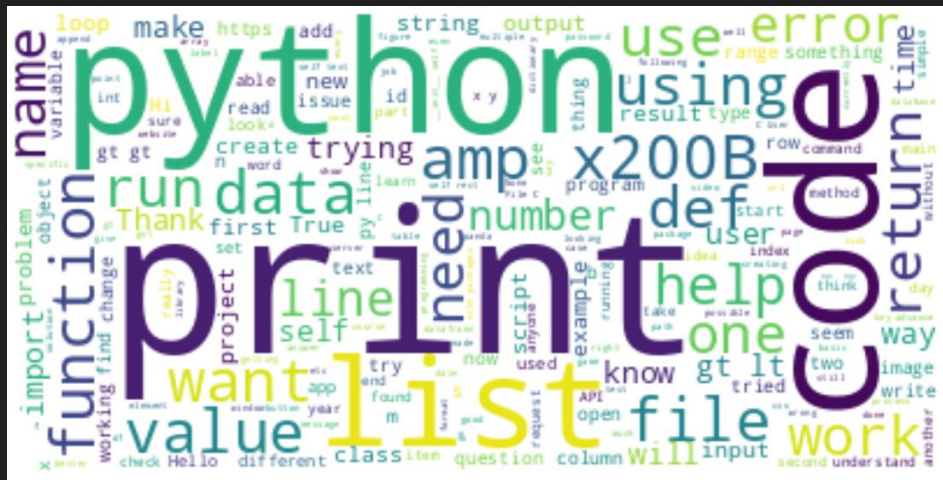


Example Of Post With No Words

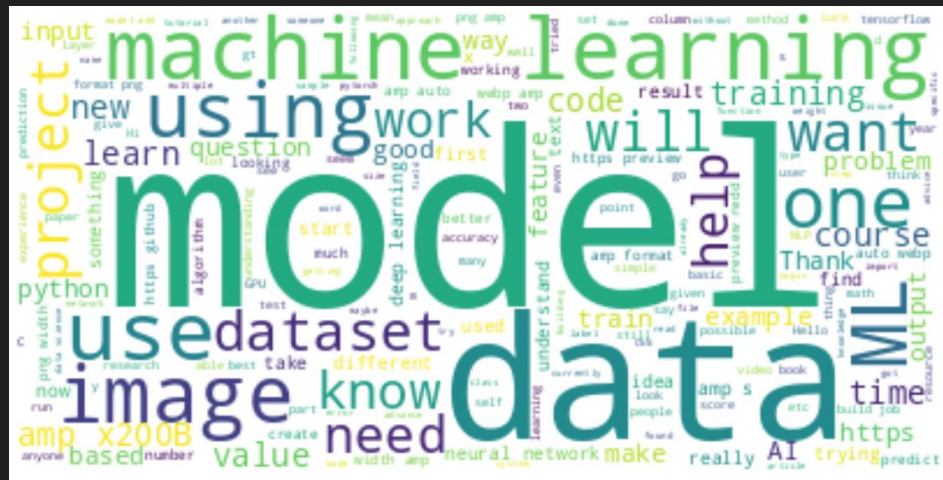


Common Words Seen In Each Subreddit

r/Learnpython

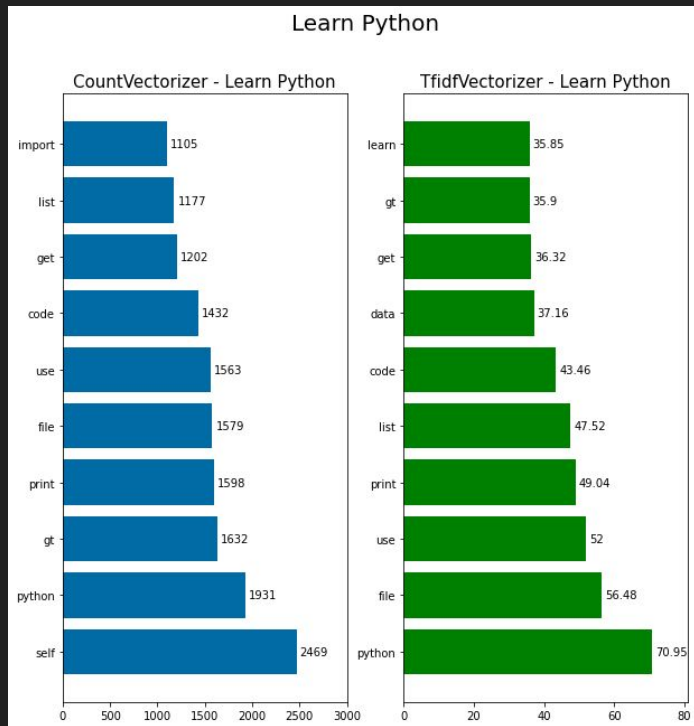


r/Learnmachinelearning

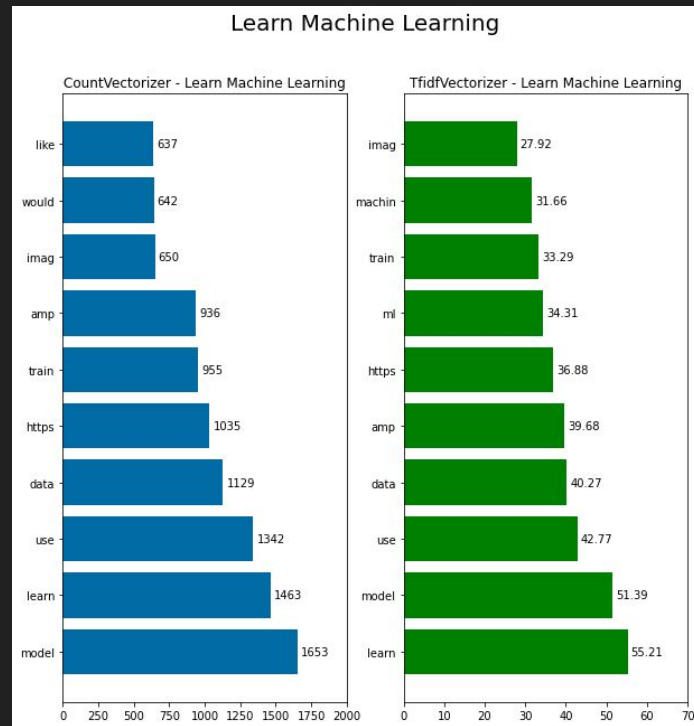


Count Vectorizer VS Tfidf Vectorizer

r/Learnpython

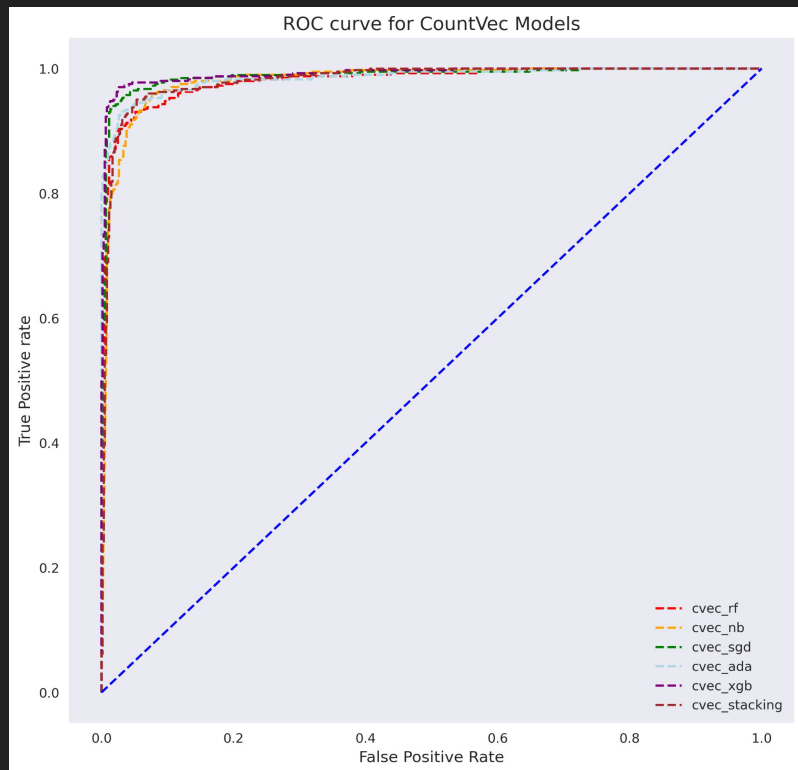


r/Learnmachinelearning

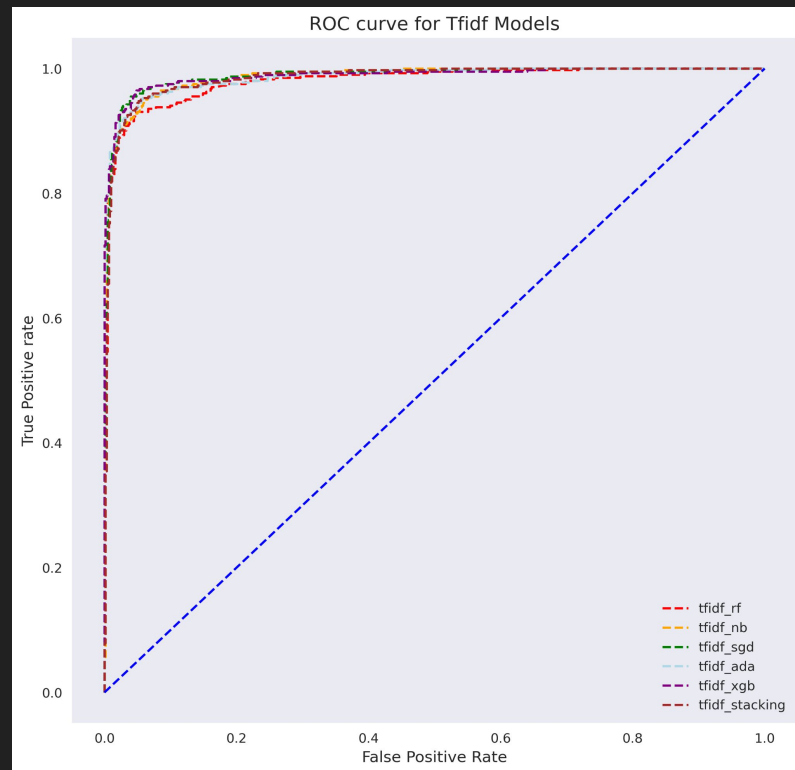


Data Modeling

CountVectorizer

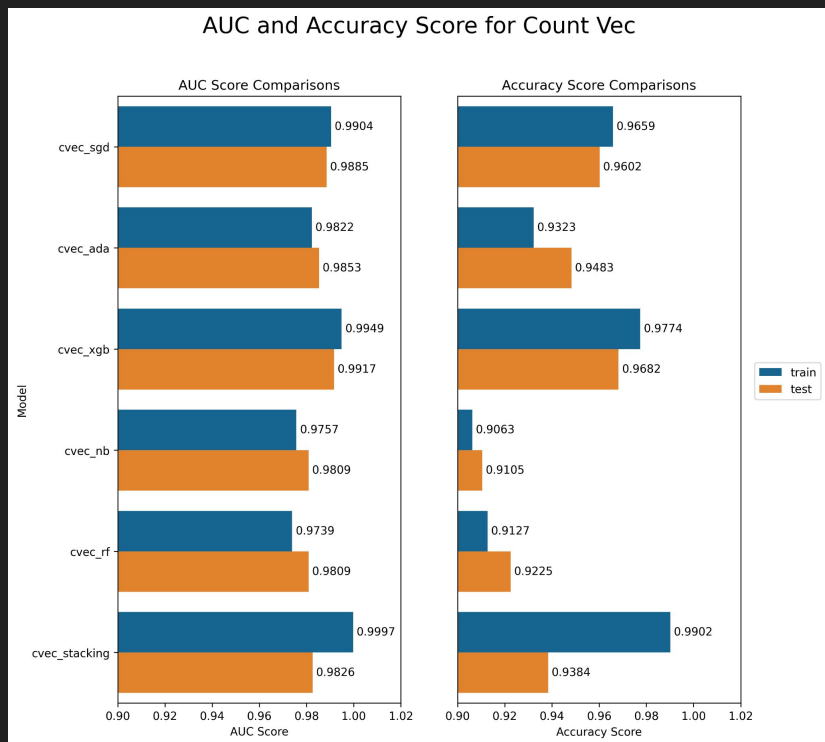


TfidfVectorizer

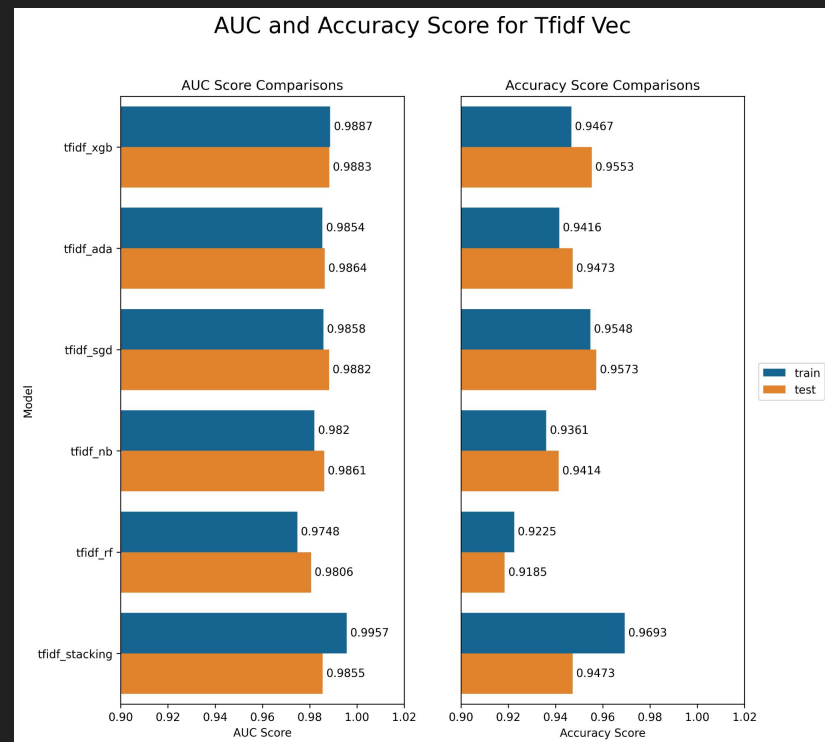


Model Evaluation

CountVectorizer

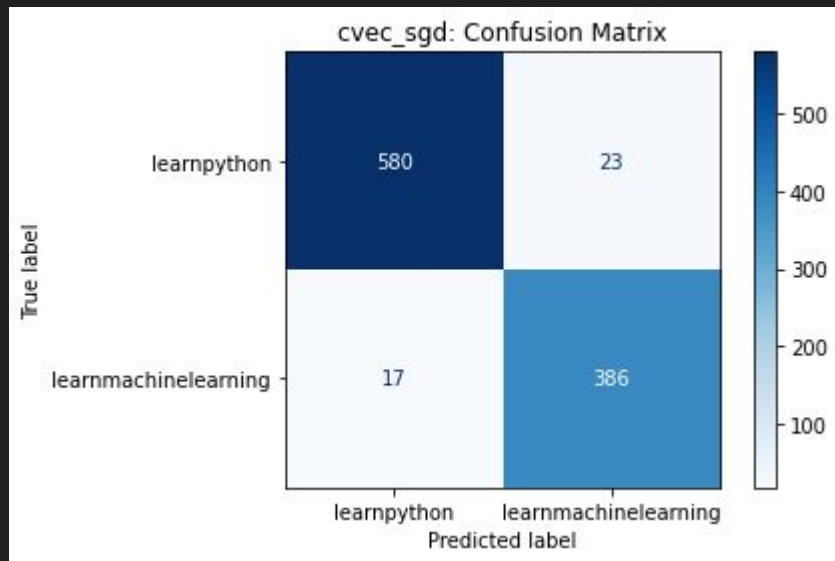


TfidfVectorizer



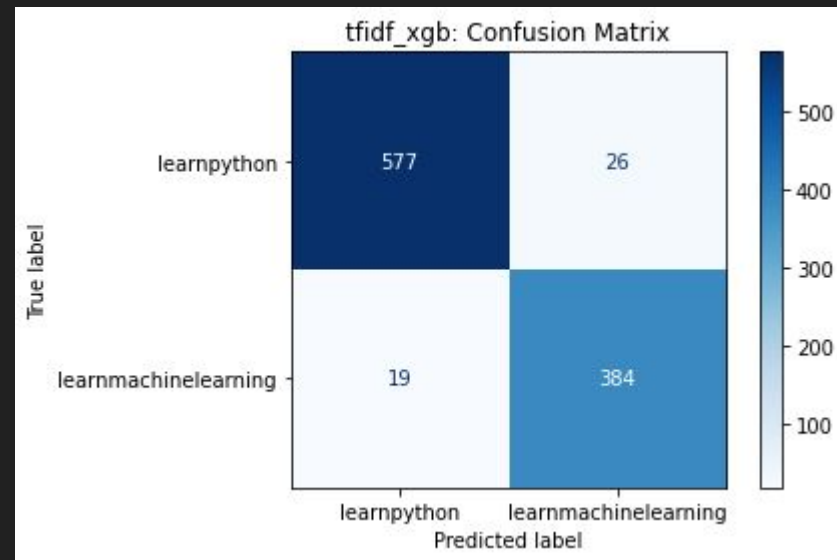
Confusion Matrix

CountVectorizer



Recall: 0.957
Precision: 0.943

TfidfVectorizer



Recall: 0.952
Precision: 0.936

Confusion Matrix

CountVectorizer -
SGDClassifier

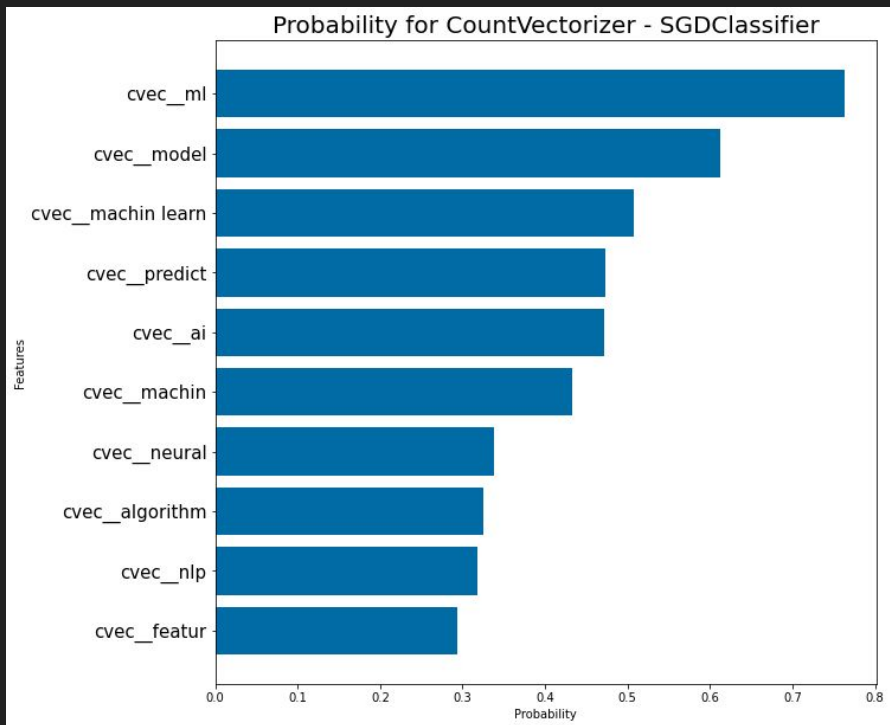
```
cvec_sgd :  
{'ct__cvec__max_df': 0.5,  
'ct__cvec__max_features': 2000,  
'ct__cvec__min_df': 0,  
'ct__cvec__ngram_range': (1, 2),  
'ct__cvec__stop_words': tokenized_stop_words,  
'ct__cvec__tokenizer': Tokenizer,  
'sgd__alpha': 0.01,  
'sgd__loss': 'log_loss'}
```

TfidfVectorizer -
XGBoost

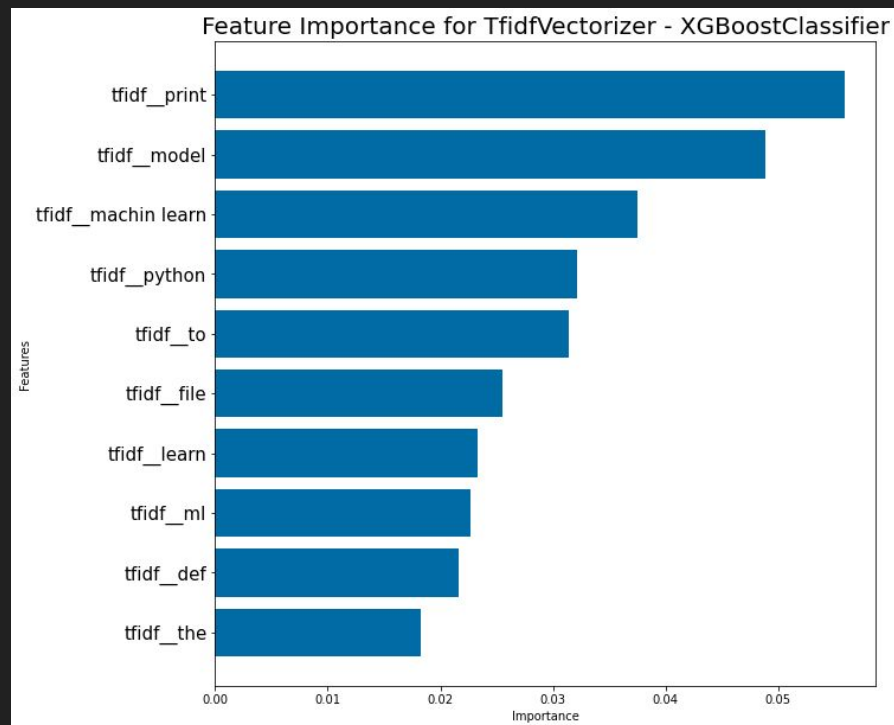
```
tfidf_xgb :  
{'ct__tfidf__max_df': 0.95,  
'ct__tfidf__max_features': 1500,  
'ct__tfidf__min_df': 0,  
'ct__tfidf__ngram_range': (1, 3),  
'ct__tfidf__stop_words': None,  
'ct__tfidf__tokenizer': Tokenizer,  
'xgb__eta': 0.1,  
'xgb__max_depth': 4,  
'xgb__tree_method': 'gpu_hist'}
```

Feature Importance

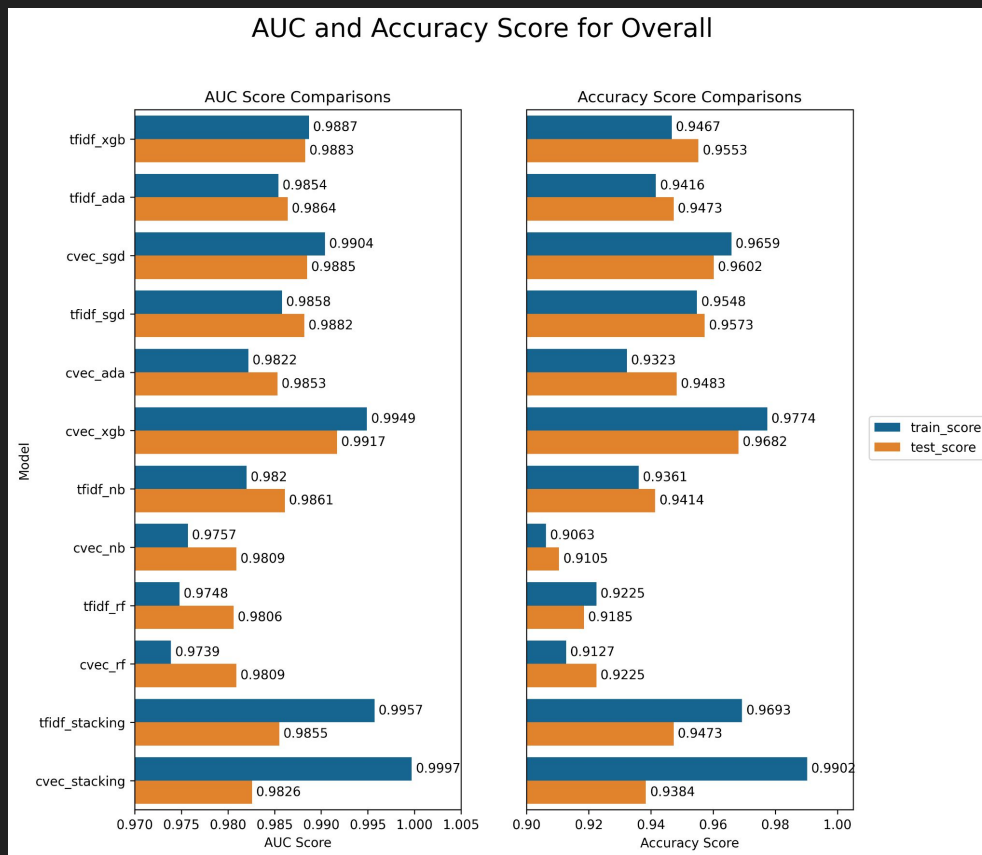
CountVectorizer



TfidfVectorizer



Overall Models Score Comparisons



Unseen Data

	future_tfidf	future_cvec
score	0.978947	0.973684
misclassification	0.021053	0.026316
sensitivity	0.968421	0.957895
specificity	0.989474	0.989474
precision	0.989247	0.989130
f1	0.978723	0.973262
auc_score	0.995346	0.993241

Conclusions & Recommendations

Conclusions

1. Both countvectorizer and tfidfvectorizer are viable methods for the modeling process.
2. Boosting algorithms and SGDClassifier exceptionally good generalisations between the train and test data.
3. Use the scoring = roc_auc for gridsearchcv

Recommendations

1. To improve the model to include the spam posts
2. To use another set of subreddits - like datascience and machinelearning

Thank you!