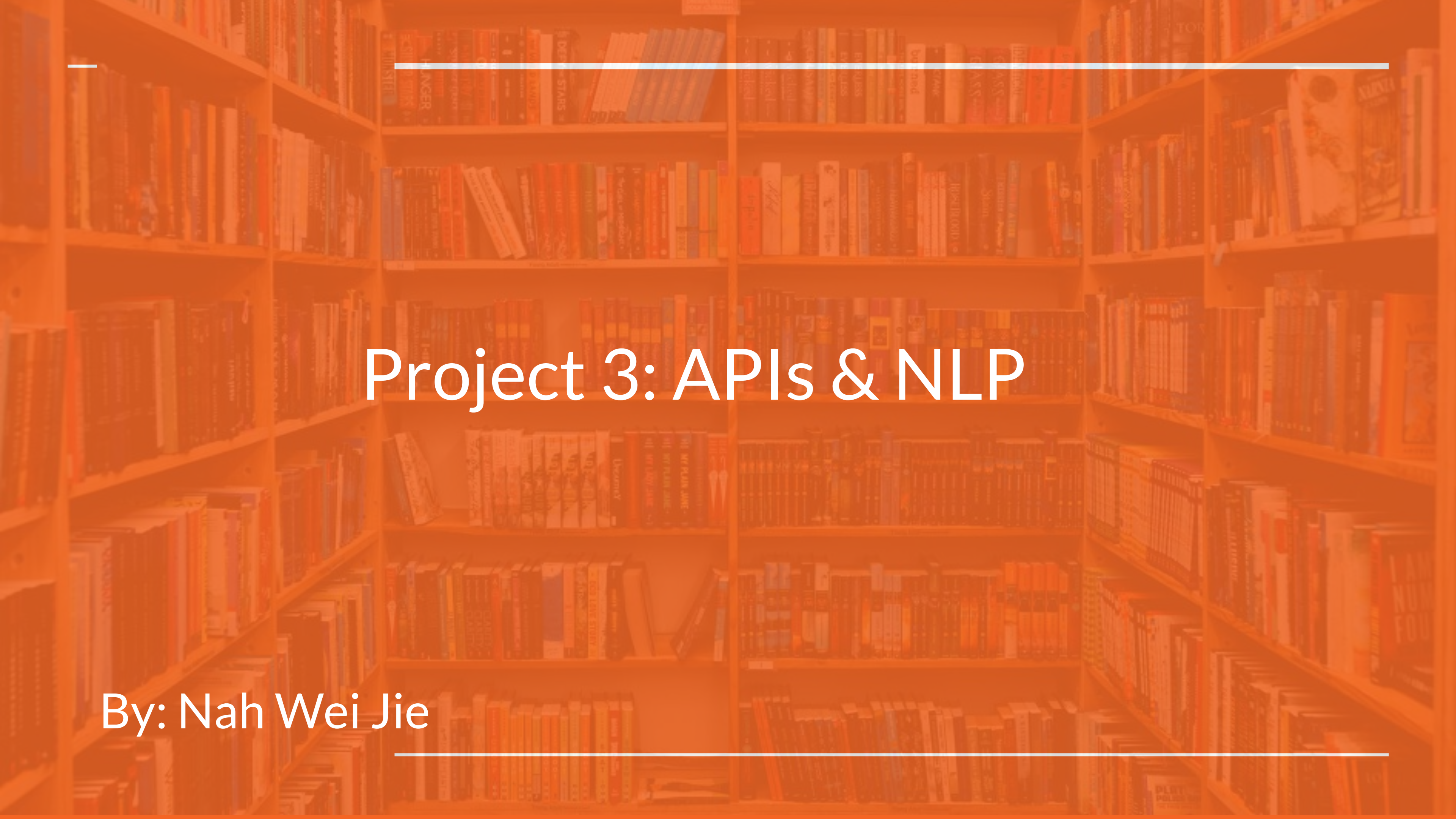


Project 3: APIs & NLP

By: Nah Wei Jie

FairPrice

Share-A-Textbook

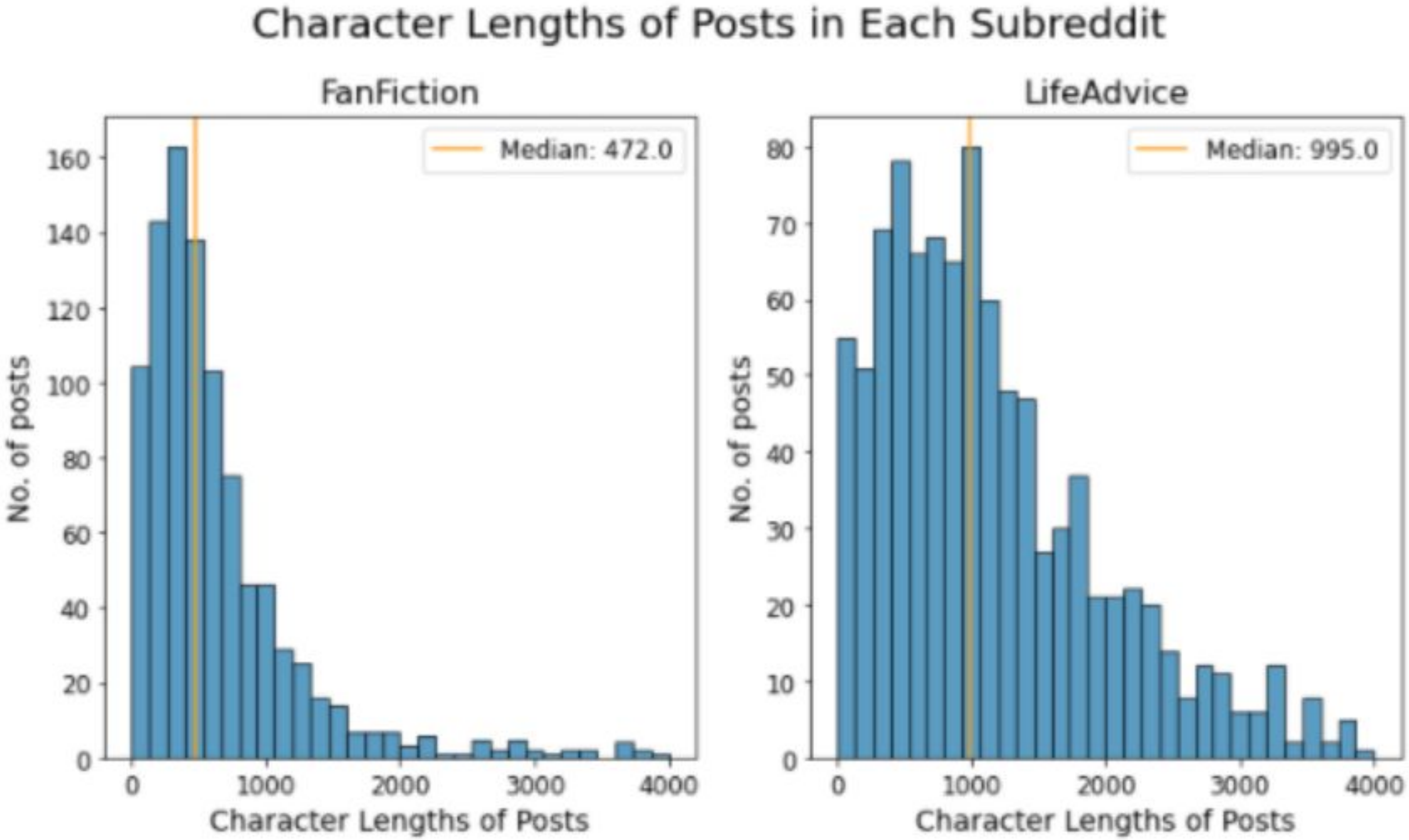
A Little Giving
A Lot More Joy
SECONDARY 4
ADDITIONAL MATHS 360
ALL ABOUT English
Textbook Secondary 2 Express
ALL ABOUT English
Textbook Secondary 2 Express
Get It Right!
English

Problem Statement:

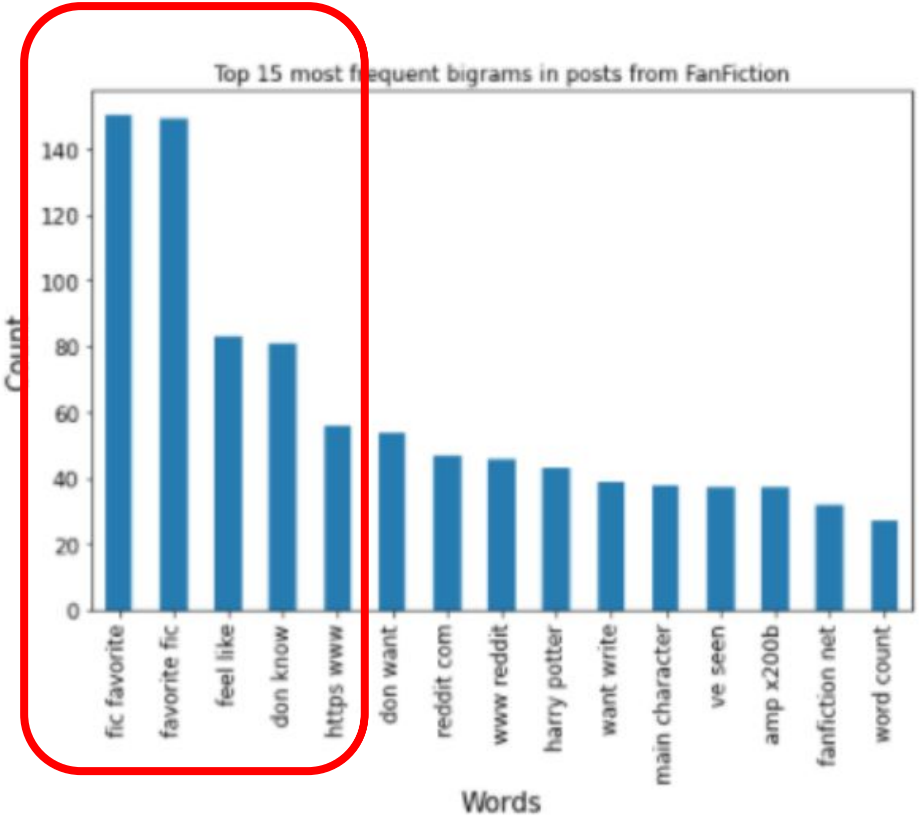
To create a text classification model to determine whether reddit posts belong to either of these two subreddits "FanFiction" or "LifeAdvice"

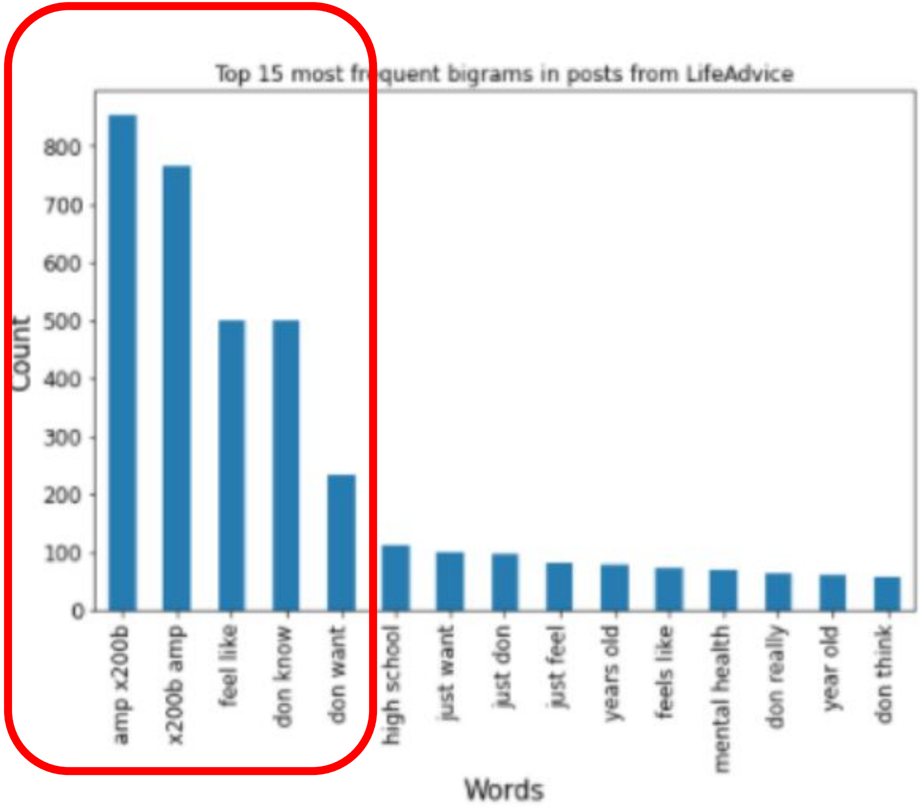
FairPrice
Share-A-Textbook

NLB | National Library
Singapore

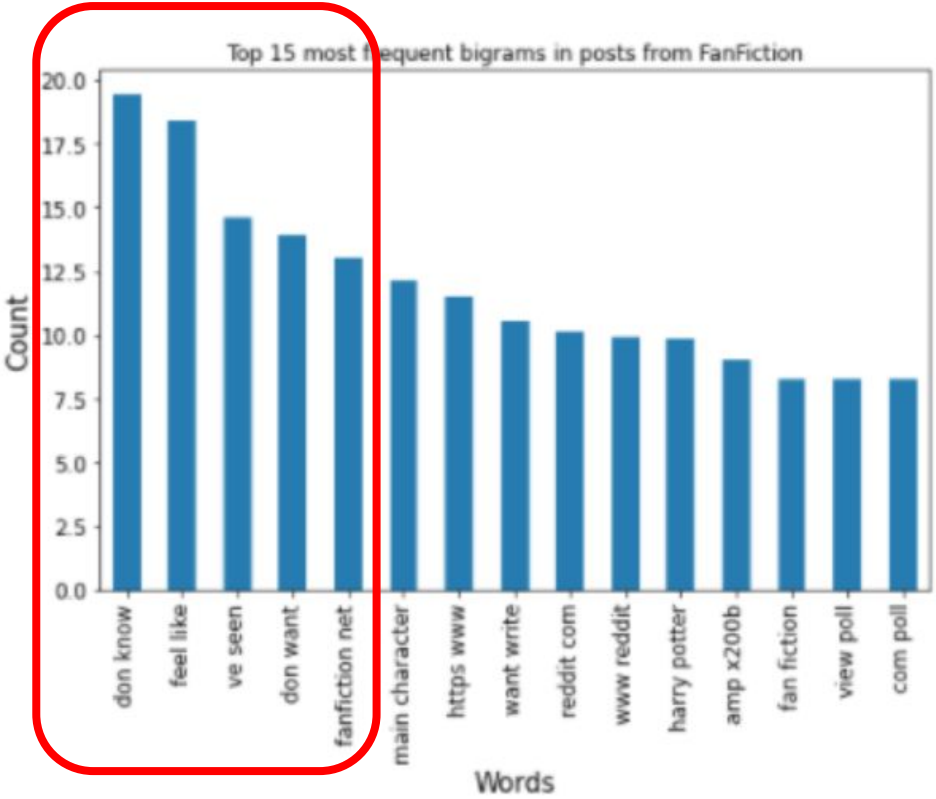

Character Lengths of Posts in Each Subreddit
FanFiction
Median: 472.0
No. of posts
0
20
40
60
80
100
120
140
160
0
1000
2000
3000
4000
Character Lengths of Posts
LifeAdvice
Median: 995.0
No. of posts
0
10
20
30
40
50
60
70
80
0
1000
2000
3000
4000
Character Lengths of Posts

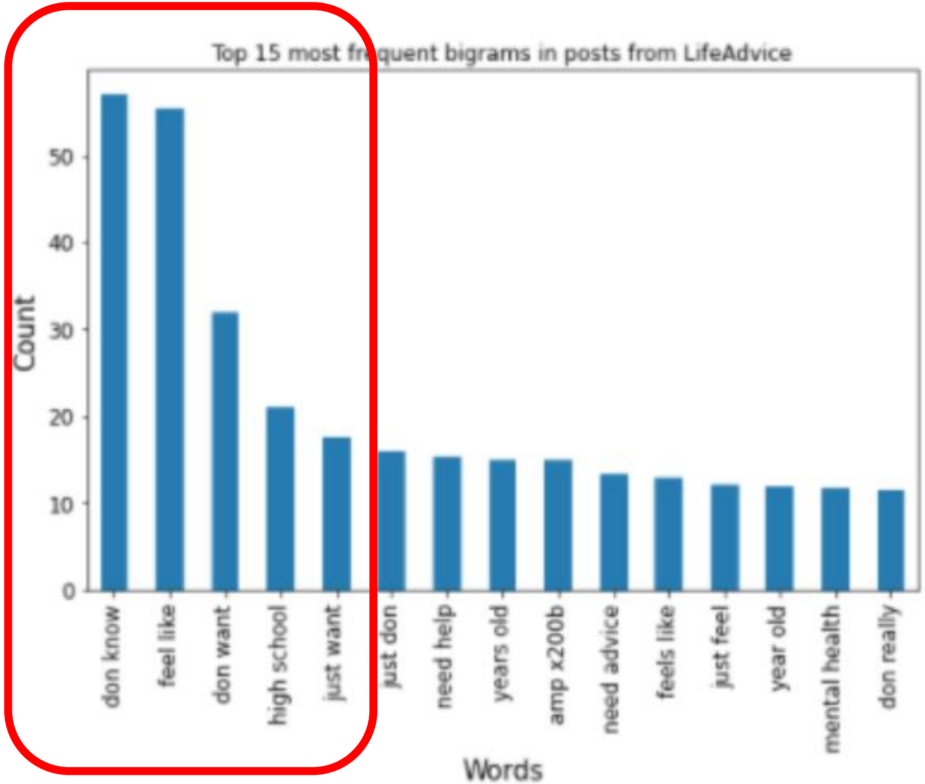
CountVectorizer on posts



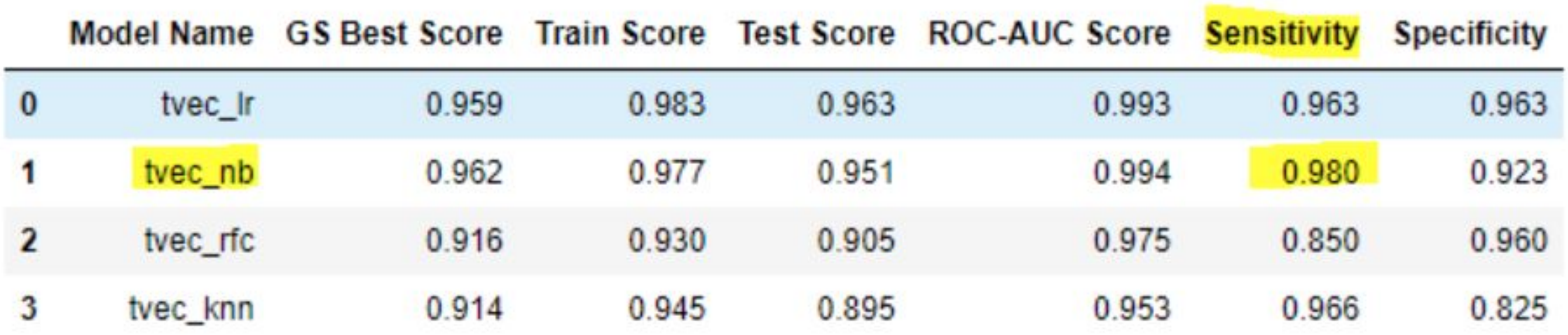


TF-IDFVectorizer on posts



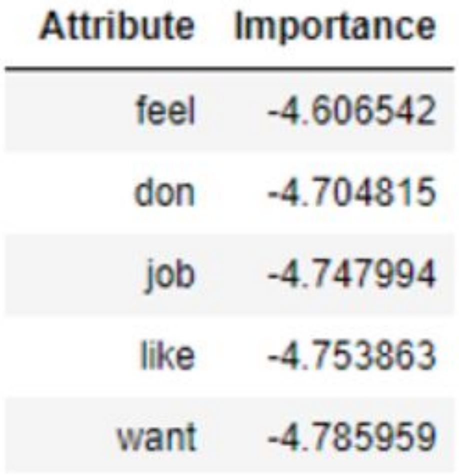


Modelling



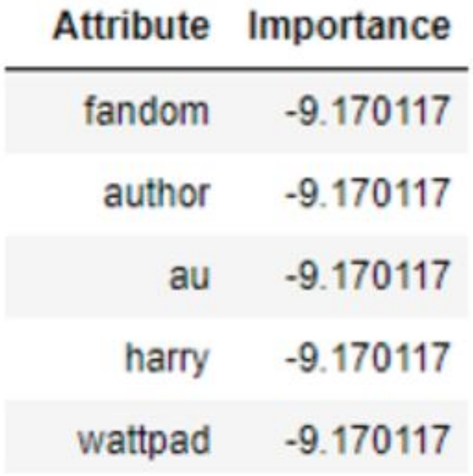
	Model Name	GS Best Score	Train Score	Test Score	ROC-AUC Score	Sensitivity	Specificity
0	tvec_lr	0.959	0.983	0.963	0.993	0.963	0.963
1	tvec_nb	0.962	0.977	0.951	0.994	0.980	0.923
2	tvec_rfc	0.916	0.930	0.905	0.975	0.850	0.960
3	tvec_knn	0.914	0.945	0.895	0.953	0.966	0.825

Top 5 most important words in determining subreddits



Attribute	Importance
feel	-4.606542
don	-4.704815
job	-4.747994
like	-4.753863
want	-4.785959

Life Advice



Attribute	Importance
fandom	-9.170117
author	-9.170117
au	-9.170117
harry	-9.170117
wattpad	-9.170117

Fan Fiction

Conclusion

Limitations

No control over books donate

(Problem of Imbalance Classes)

Dataset used to train model

(Are subreddits good proxies for excerpt for books?)