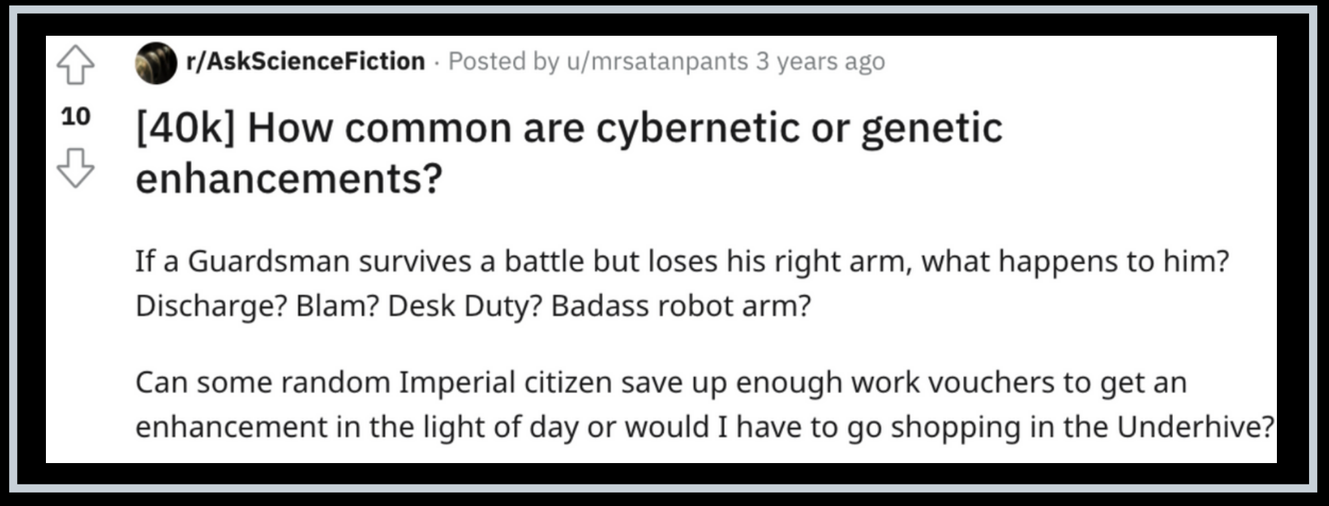
DSI 927 - Classifying Subreddits using Natural Language Processing

ABDELJALIL KABBAJ

NOVEMBER 5, 2021

Fantasy vs Sci Fi Subreddits: Which model will accurately differentiate between these two subreddits ?



What does our Data Frame look like?

- Using a custom made while loop, I was able to scrape about 10,000 posts from each subreddit, resulting in:
  - 20835 rows
  - 4 columns:
    - Subreddit: 1 for Fantasy, 0 for Sci Fi
    - Title: Title of the post
    - Selftext: Main body of the text
    - Word count: Number of words per post
  - Balance of Classes (value counts):
    - 1(Fantasy) : 0.501464
    - 0(Sci Fi) : 0.498536

Exploring our Data

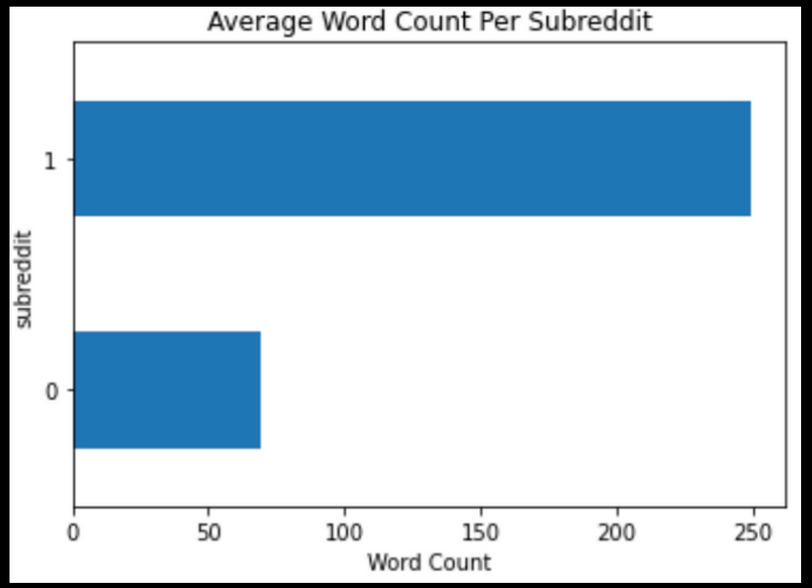


Figure 1.

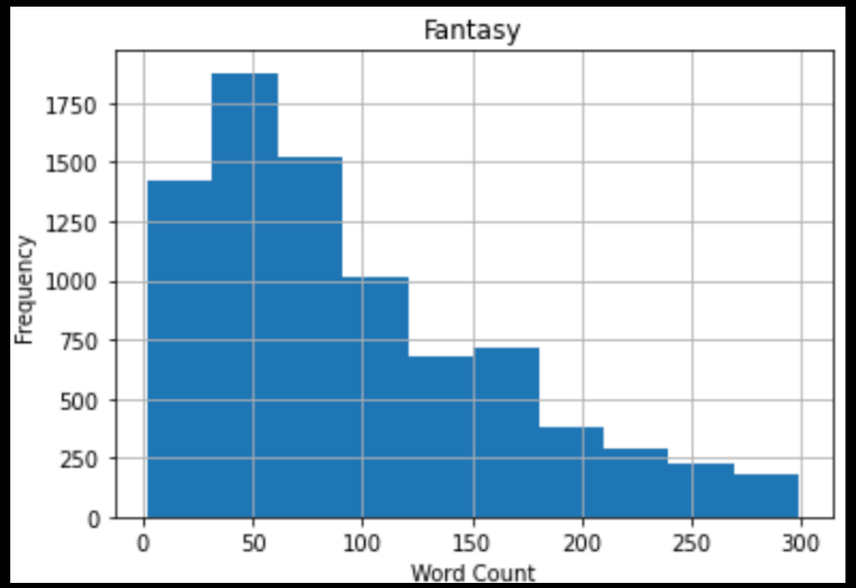


Figure 2.

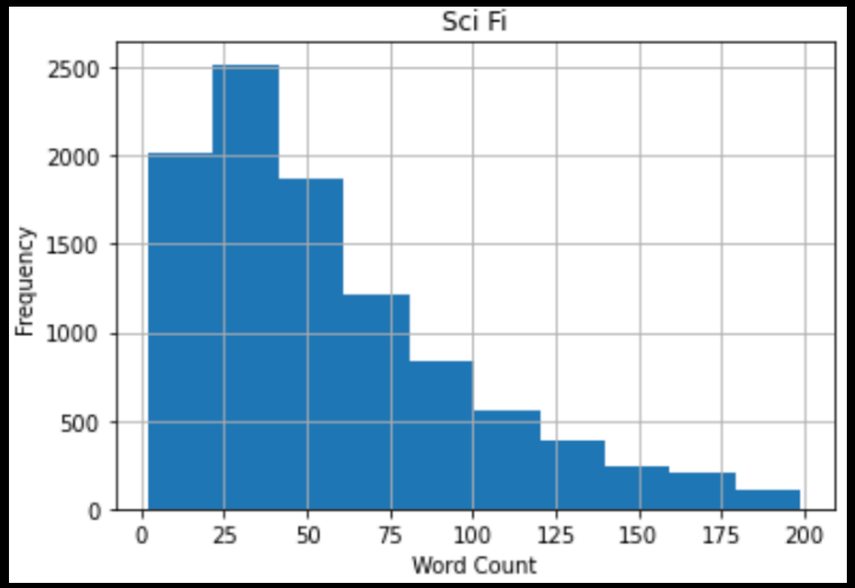


Figure 3.

Modeling our Data

- Feature as our X: Selftext
- Feature as our y: Subreddit
- Adding custom stop words to the list:
  - rfantasy, science fiction, fantasy

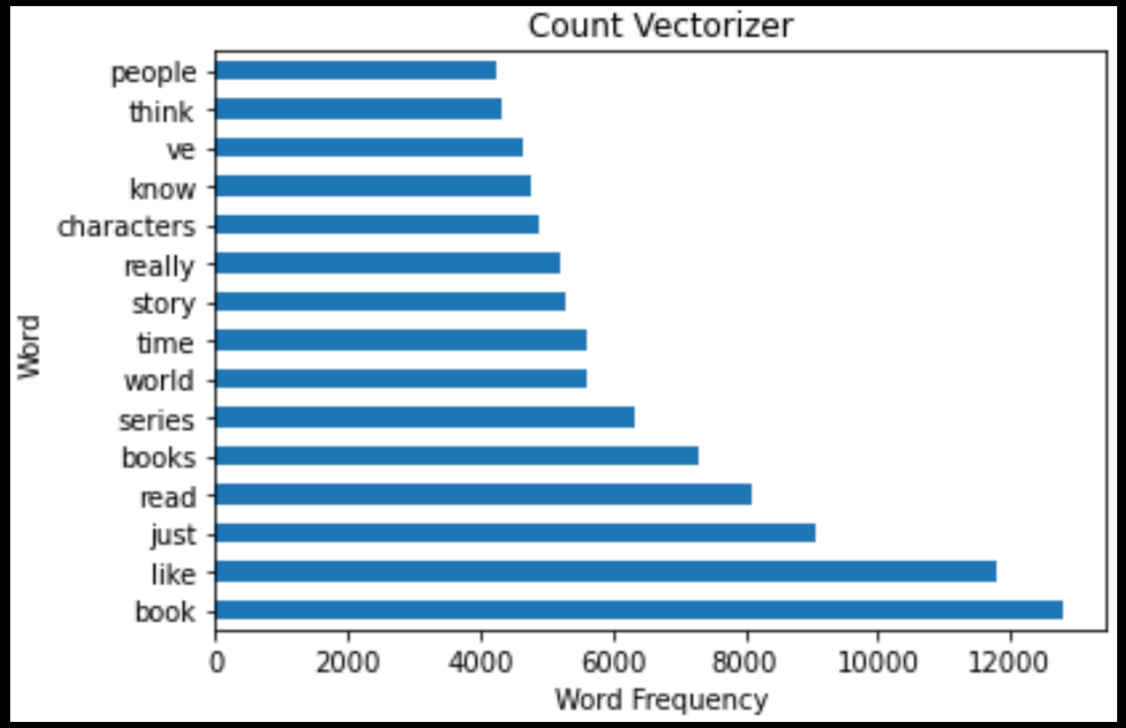


Figure 1.

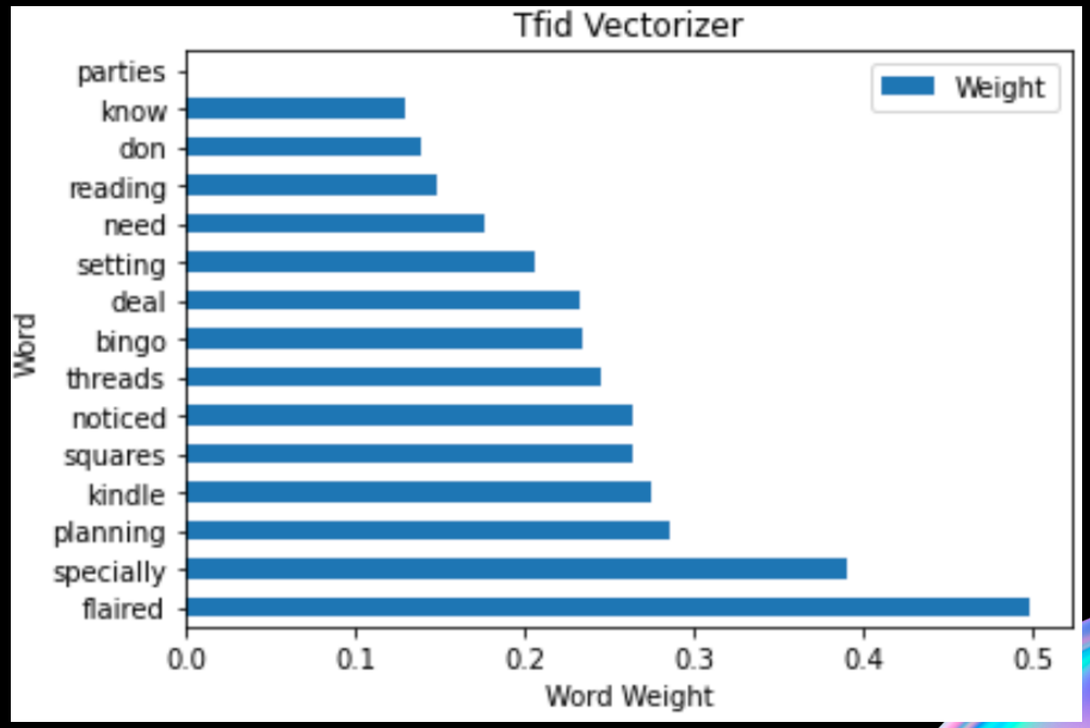
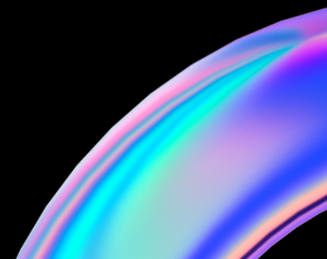


Figure 2.

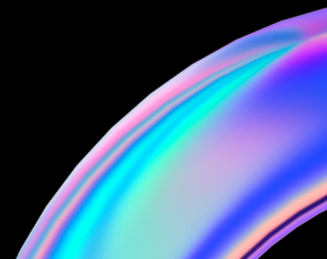
Scoring Our Models

	Train Score (CVEC)	Test Score (CVEC)	Train Score (TFID)	Test Score (TFID)
Random Forest	0.9996	0.9262	0.9996	0.9422
XGBoost Classifier	0.9477	0.9285	0.9550	0.9291
Naive Bayes	0.9553	0.9345	0.9480	0.9174



Scoring After Tuning Hyper Parameters

	Train Score (CVEC)	Test Score (CVEC)	Train Score (TFID)	Test Score (TFID)
Random Forest	0.9534	0.9306	0.9713	0.9402
XGBoost Classifier	0.9658	0.9366	0.9760	0.9356
Naive Bayes	0.9609	0.9383	0.9607	0.9295



Random Forest, XGBoost, Naive Bayes

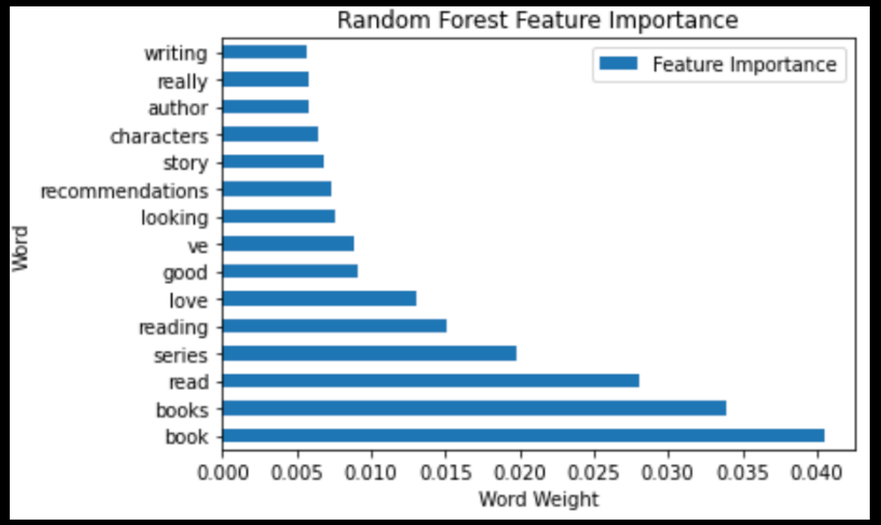


Figure 1.

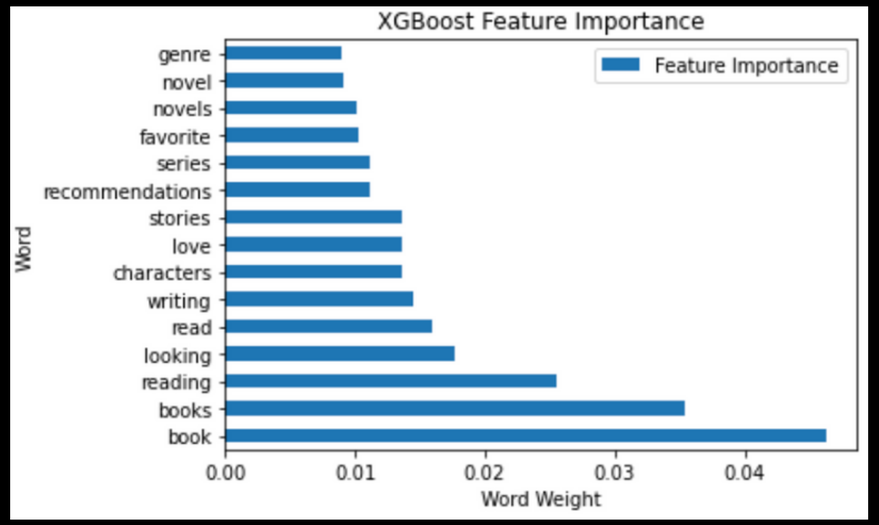


Figure 2.

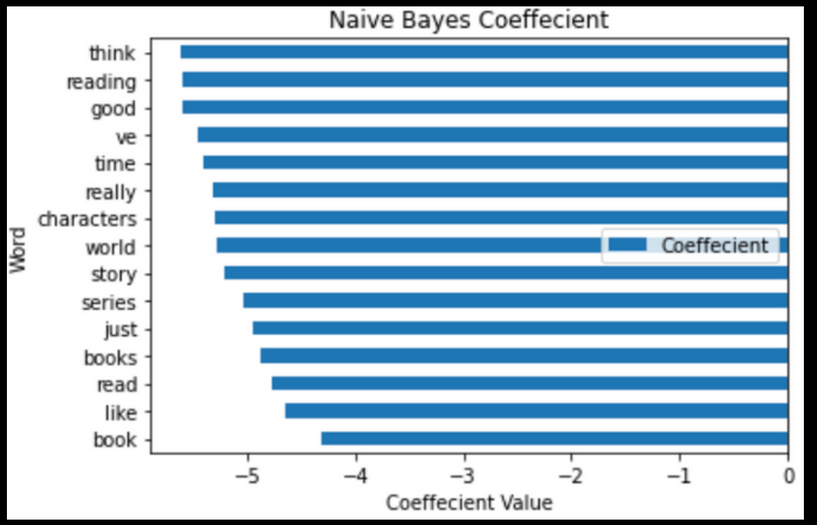


Figure 3.

Looking at Our Predictions

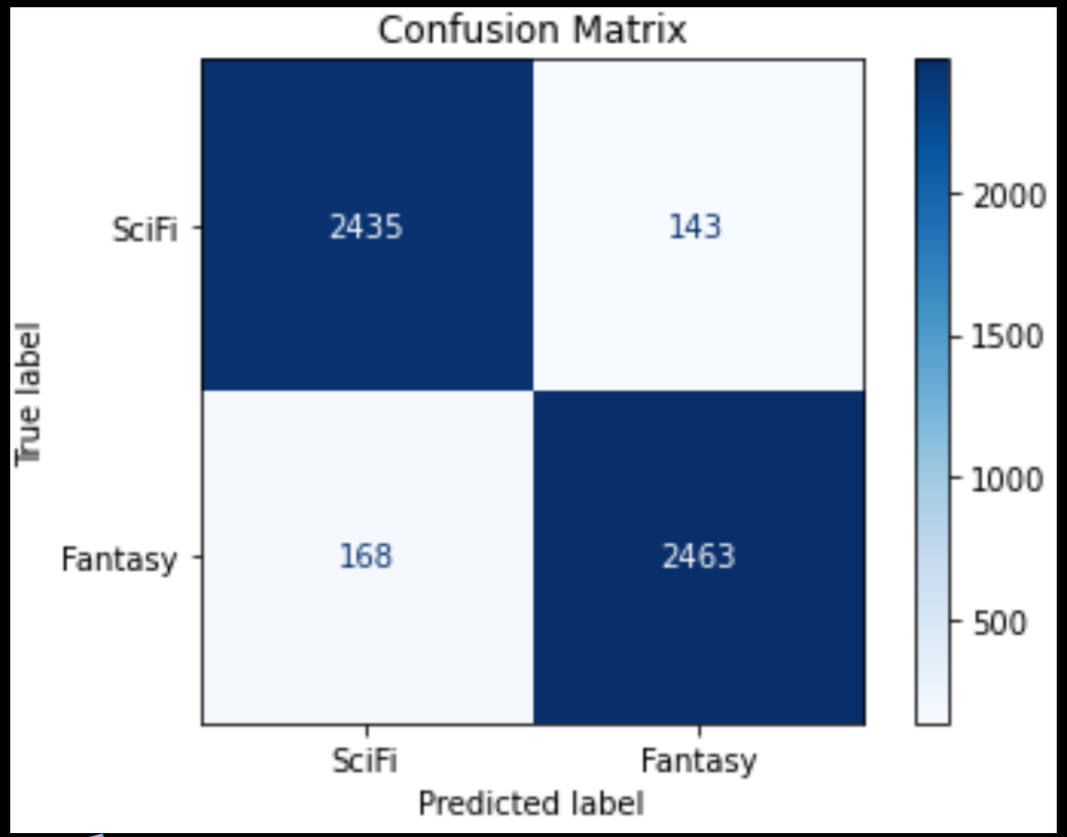


Figure 1.

Confusion Matrix done on Best Performing Model: Tuned RFC with Tfidf Vectorization

- Accuracy: What percentage of our observations did we successfully predict ?
  - 94.1%
- Sensitivity/Recall: How well is our model able to detect our subreddit in its correct class ?
  - 93.9%
- Precision: How well did our model predict Fantasy classes ?
  - 94.3%
- Specificity: How well did our model predict Sci Fi classes ?
  - 94.2%

Recommendations

- Best Performing Model:
  - Random Forest, TFIDF Vectorization
    - Train: 0.9713
    - Test: 0.9402

- After making a successful model here is what I recommend:
  - This can be taken a step further with greater computational power.
  - For instance, building a model for the entirety of the subreddits and implementing it on reddit.
  - Doing so would help users effectively post on the right subreddit.
  - Hypothetically, for instance, I could post a thread on the Fantasy subreddit and get poor replies and reddit could have the computational power to detect in which subreddit my post is most appropriate and make recommendations.
  - This will help the users realize that their post could have been more appropriate in the Sci Fi reddit.
  - Increase user commitment, engagement and satisfaction.
  - Further accommodate users (to better respond to their needs).

Thank you.