GA-DSI-123

PROJECT 3
Web APIs & NLP

Ayako Homma

March 6, 2023

Problem Statement

Subreddit moderators face challenges with overlapping contents and users between two fashion-related subreddits, r/malefashionadvice and r/femalefashionadvice.

To address this issue, we will leverage APIs and Natural Language Processing (NLP) techniquest to collect and analyze data from the two subreddits. The goal of this project is to develop a machine learning model that can accurately classify posts from each subreddit with a test accuracy of at least 0.8. By developing a model, moderators will be able to better manage their subreddits and provide more targeted content to their users.

What is Reddit?

Founded in 2005, Reddit is an American social news and discussion website. Users can share their interests and hobbies, and the posts are categorized by subject into user-created communities called subreddits.

Reddit is the **17th** most popular social media platforms with **430 million** monthly active users and **1.5 billion** monthly visits to the site.
(Semrush, published on November 2022)

Compared with other social networks, Reddit has a higher share of users in **18-29 years old, male, a high-income, and live in cities and urban areas.**
(Statista, published on August 2022)

Source: Semrush, Statista Report

Key steps in research process

1	Data Collection
2	Data Cleaning
3	Data Preprocessing & EDA
4	Data Modeling & Evaluation
5	Conclusion & Recommendations

Data Collection: Two subreddits

r/malefashionadvice

5.3m members
Created on Sep 3, 2009

- Collected **2,778** posts
- Posts created between **February 28 – March 1, 2023**

r/femalefashionadvice

3.3m members
Created on Dec 23, 2010

- Collected **2,793** posts
- Posts created between **February 28 – March 1, 2023**

Data Cleaning and Preprocessing

Data Cleaning

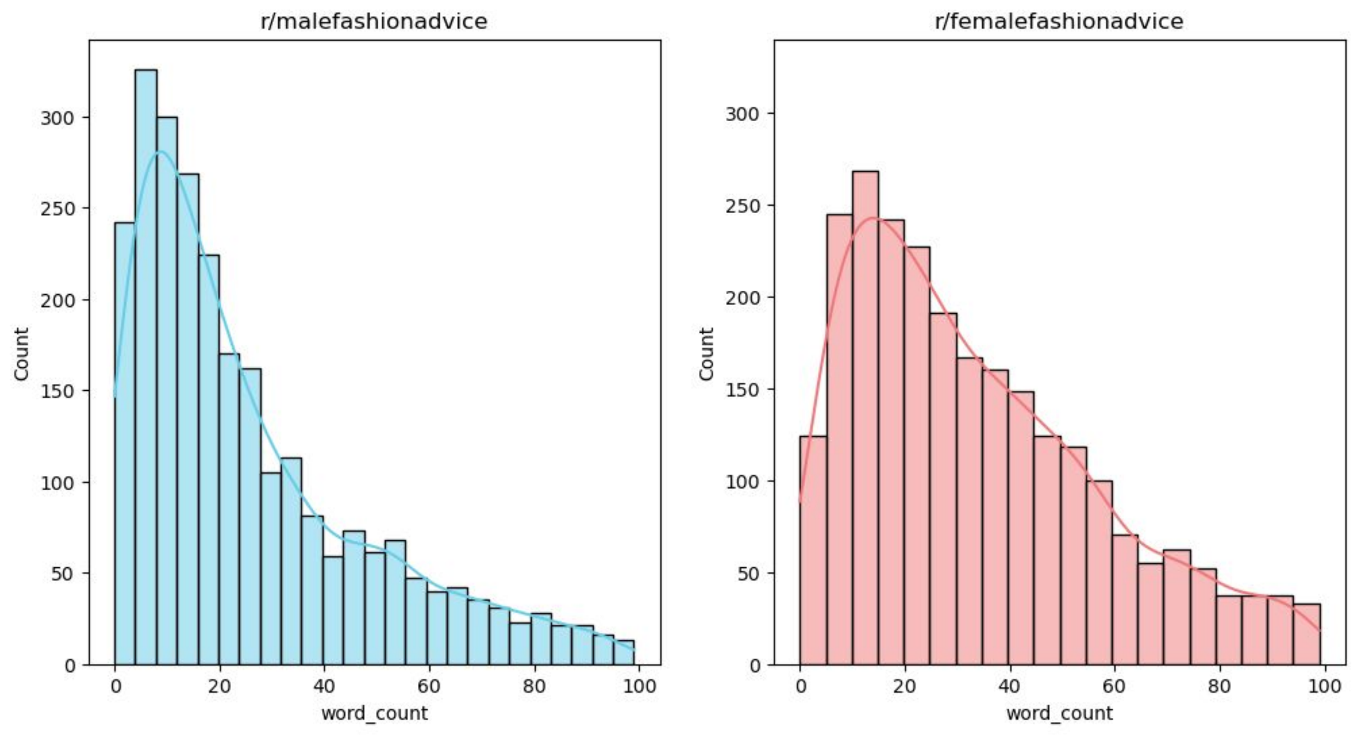
- Handled **Null values and [removed] values**.
- Dropped **duplicates posts.**

Data Preprocessing

- Removed **special characters** such as \n (new line character), &gt; (>) and &lt; (<), &amp (&) and '[^]+\.[^]+' (web link).
- Used different preprocessing methods such as **tokenization and lemmatization**.

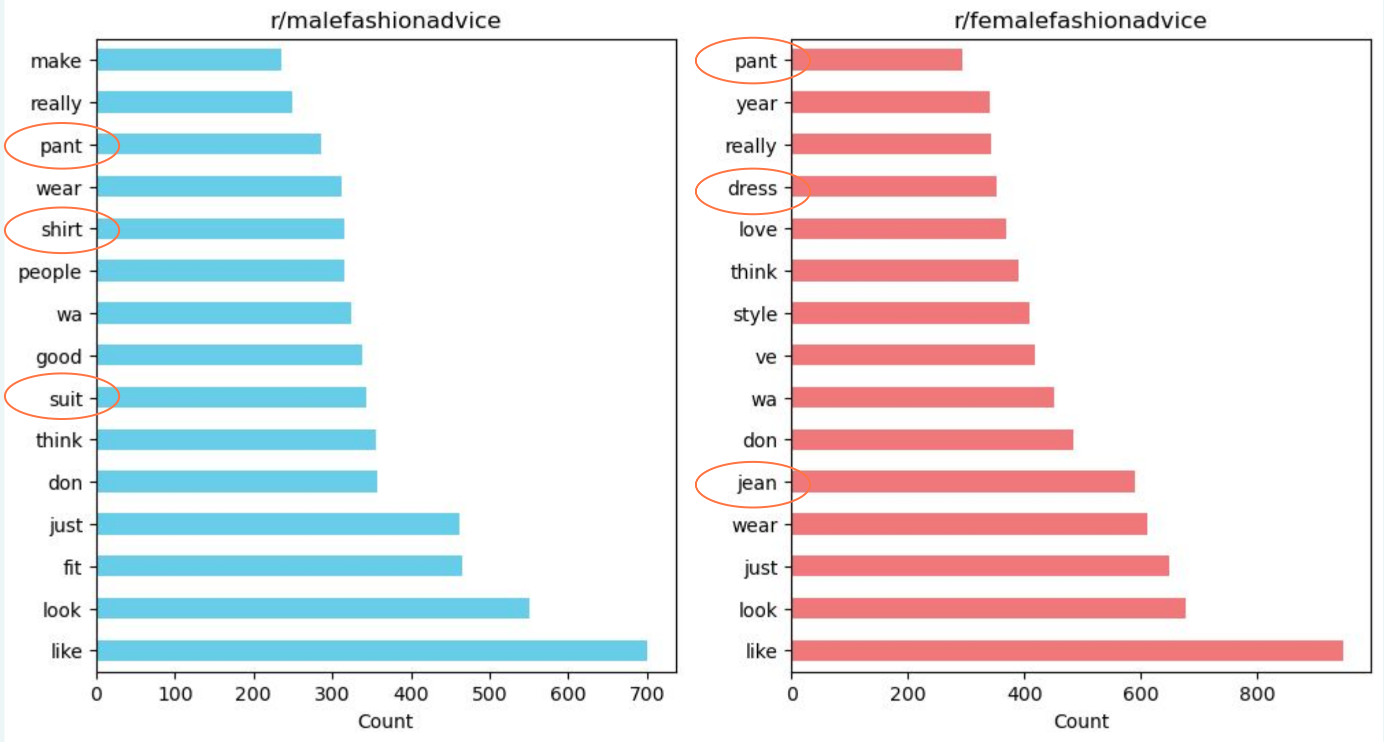
The combined clean data with 5,544 posts is used for machine learning models

Comments in r/malefashionadvice tend to have a lower word count than those in r/femalefashionadvice



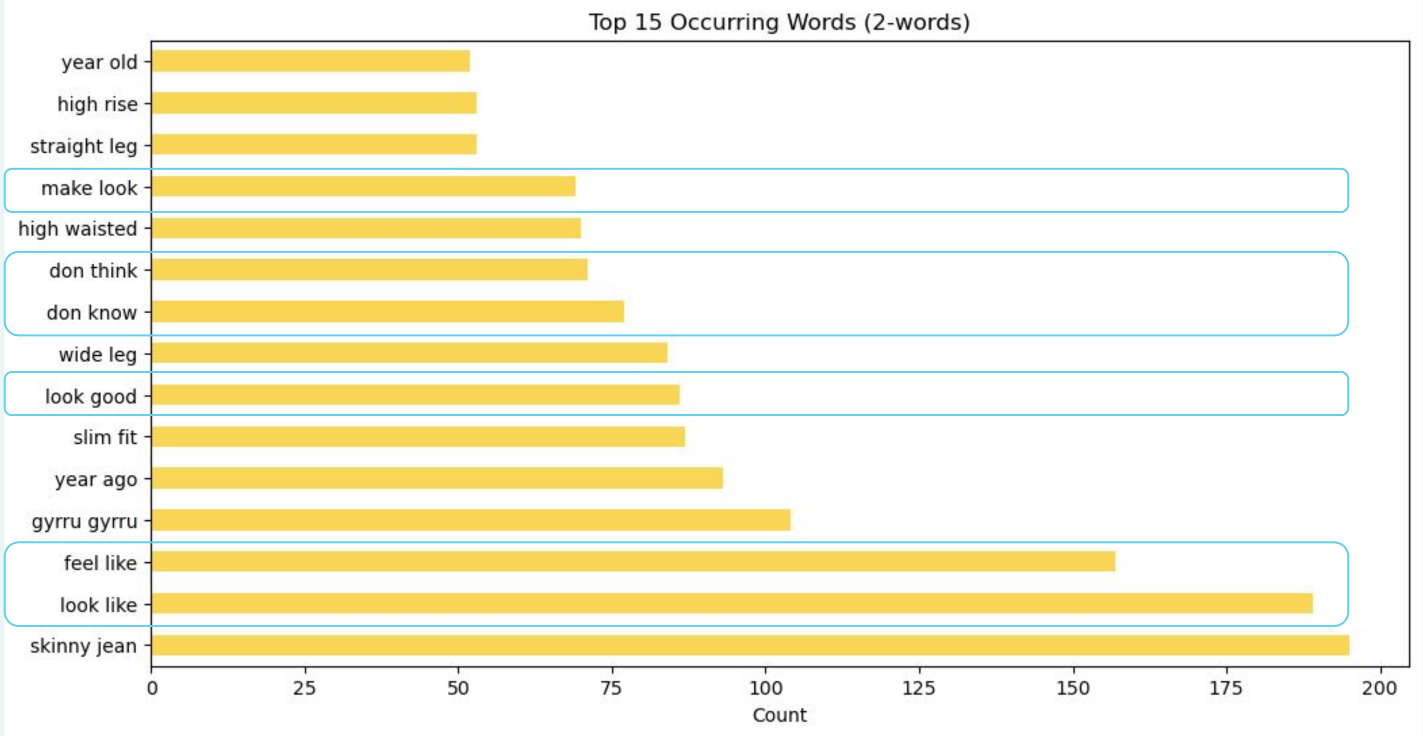
Source: Reddit (r/malefashionadvice & r/femalefashionadvice)

The most common single word occurrence of fashion items differ by subreddits



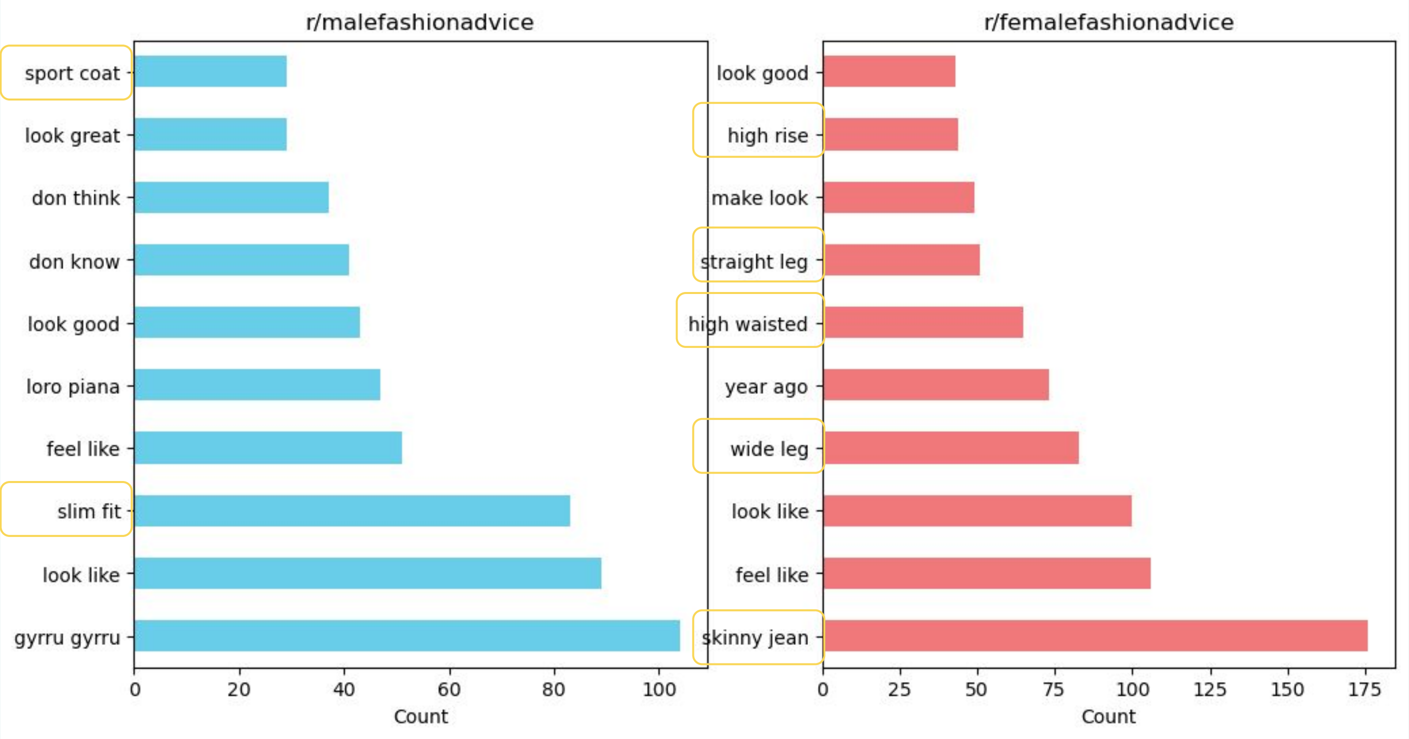
Source: Reddit (r/malefashionadvice & r/femalefashionadvice)

Users on both subreddits seek fashion advice



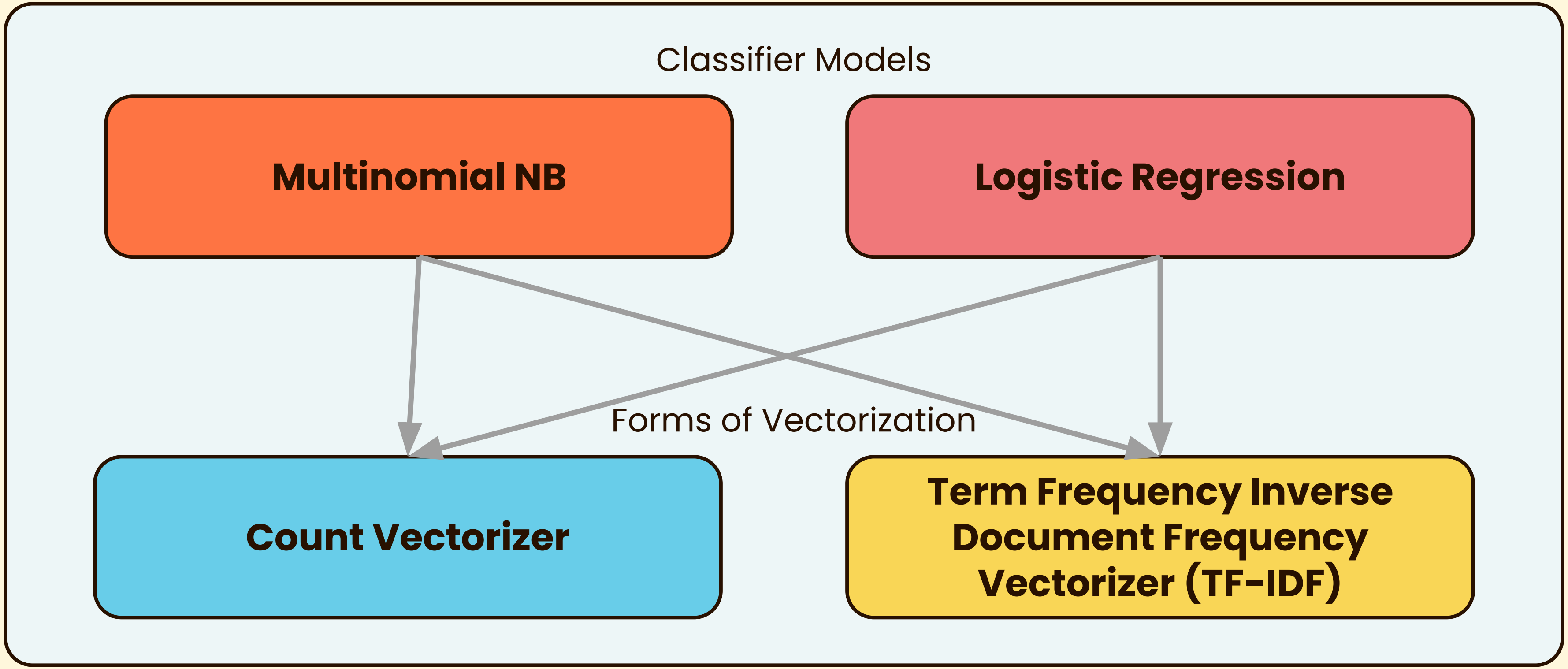
Source: Reddit (r/malefashionadvice & r/femalefashionadvice)

Differences in fashion style preferences through the most common two-word occurrences are also observed



Source: Reddit (r/malefashionadvice & r/femalefashionadvice)

Classifier Models & Forms of Vectorization



Classifier Models

Multinomial NB

Logistic Regression

Forms of Vectorization

Count Vectorizer

Term Frequency Inverse Document Frequency Vectorizer (TF-IDF)

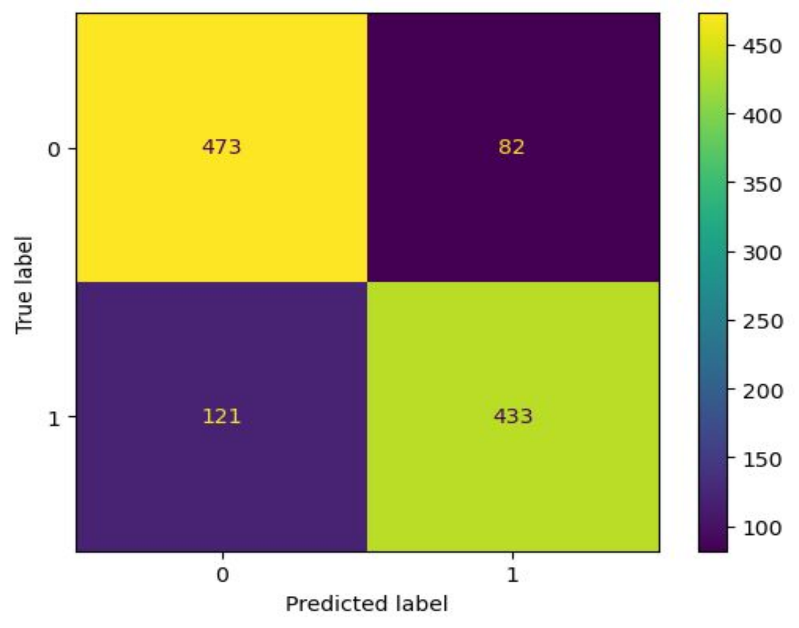
Modeling & Evaluation

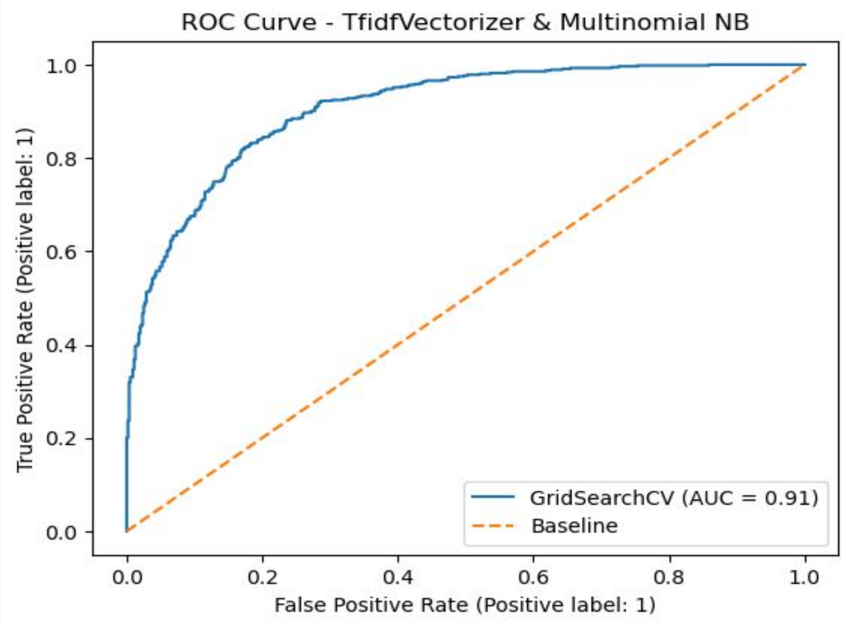
Vectorization Type	Model Type	Train Accuracy	Test Accuracy	AUC	Precision	Recall	F1-Score
Count Vectorizer	Multinomial NB	0.933	0.810	0.901	0.824	0.787	0.805
Count Vectorizer	Logistic Regression	0.963	0.787	0.872	0.761	0.838	0.797
TF-IDF Vectorizer	Multinomial NB	0.952	0.817	0.906	0.841	0.782	0.810
TF-IDF Vectorizer	Logistic Regression	0.880	0.795	0.870	0.777	0.829	0.802

Source: Reddit (r/malefashionadvice & r/femalefashionadvice)

Baseline: 0.5

The best performing model was using TfidfVectorizer & Multinomial NB





Precision: 0.841 / Recall: 0.782 / F1-Score: 0.810

ROC AUC score: 0.906

Source: Reddit (r/malefashionadvice & r/femalefashionadvice)

Baseline: 0.5

Conclusion

Based on the evaluation metrics, both the TF-IDF Vectorizer with Multinomial Naive Bayes model and the Count Vectorizer with Multinomial Naive Bayes model achieve good accuracy in classifying posts from the two subreddits.

However, the TF-IDF Vectorizer with Multinomial Naive Bayes model achieves a slightly higher test accuracy of 0.817, indicating that this model may be better at accurately classifying posts from the two subreddits.

On the other hand, the Count Vectorizer with Multinomial Naive Bayes model achieved higher recall values, which indicates that this model may be better at identifying all the posts from a particular subreddit.

Recommendations

The choice of vectorization technique depends on the specific needs and goals of the project. If the priority is to accurately classify posts from the two subreddits, then the TF-IDF Vectorizer with Multinomial Naive Bayes model may be the better choice. However, if the priority is to ensure that all posts from a particular subreddit are identified, then the Count Vectorizer with Multinomial Naive Bayes model may be more suitable.

In any case, it is recommended to conduct further analysis and fine-tuning of both models to improve the precision and recall values, particularly for the subreddit with lower recall. This could involve exploring different feature selection techniques or adjusting the hyperparameters of the models.

THANK YOU!

Any questions?