# Project 3: r/books vs r/Fantasy

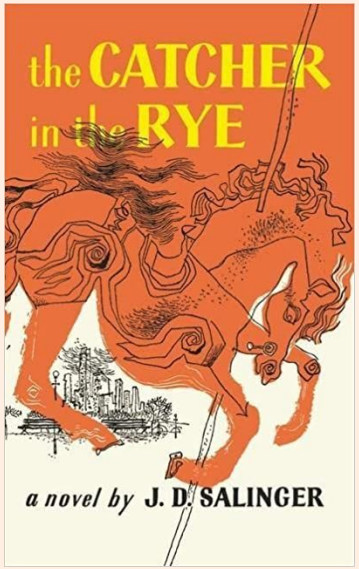
Samuel Hewitt | DSIR 523

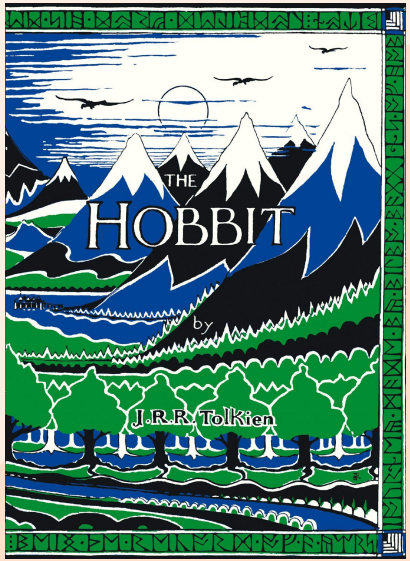
# Problem Statement

Using NLP to train a classifier on which subreddit a post came from, specifically between r/books and r/Fantasy, we want to determine the best Pipeline Classifiers that dictate a r/Fantasy post.

# Background Info

- r/books includes all things related to books, authors, genres, or publishing
- r/Fantasy includes Fantasy, SciFi, Horror, and Alt History
- Data pulled from July 2020 to June 2022

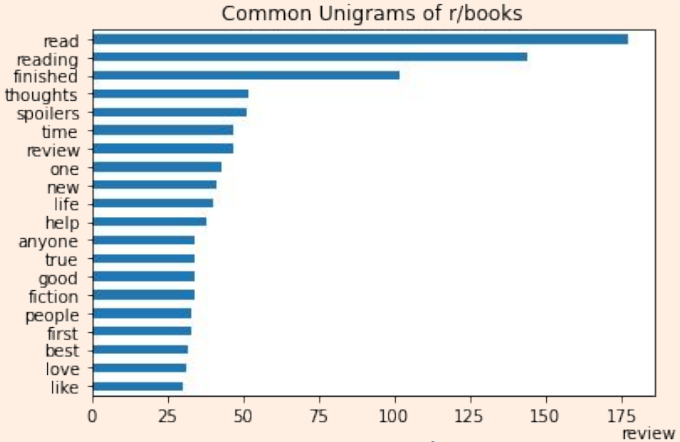


# Collecting Data

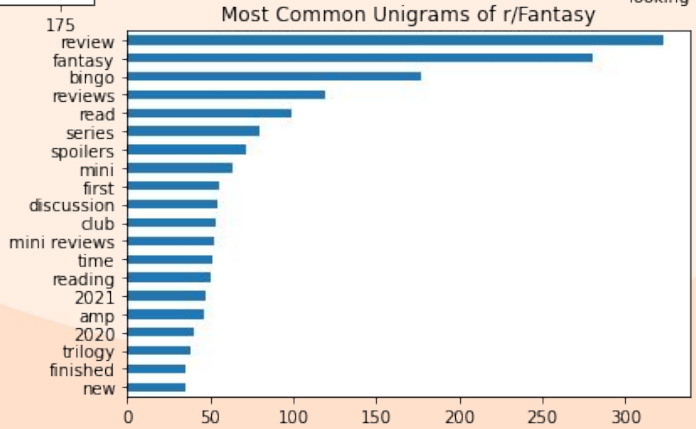
- Subreddits do not have over 100 submissions per day, resulting in duplicate posts scrapped that were dropped
- Baseline Accuracy: 51.14% of posts are from r/Fantasy

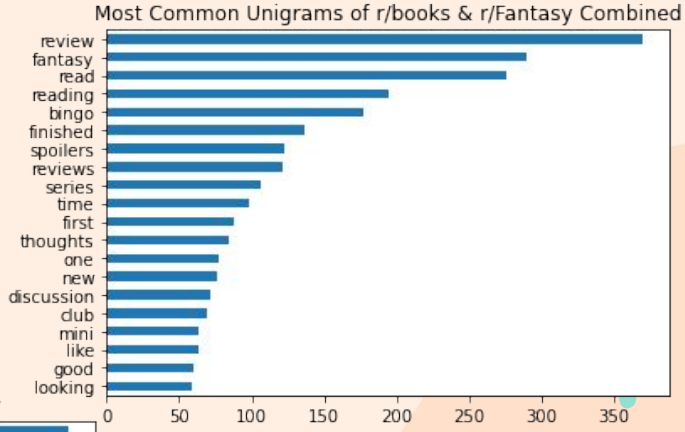
# CountVectorized Data - Unigrams




- Most prominent r/books unigrams are general updates

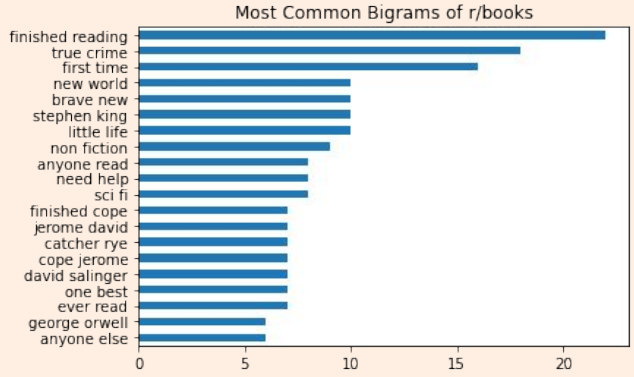
- Most prominent r/Fantasy unigrams are reviews



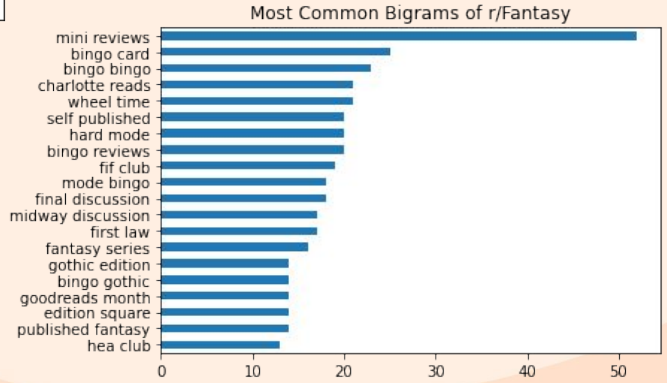

- When combined, reviews surpass general updates

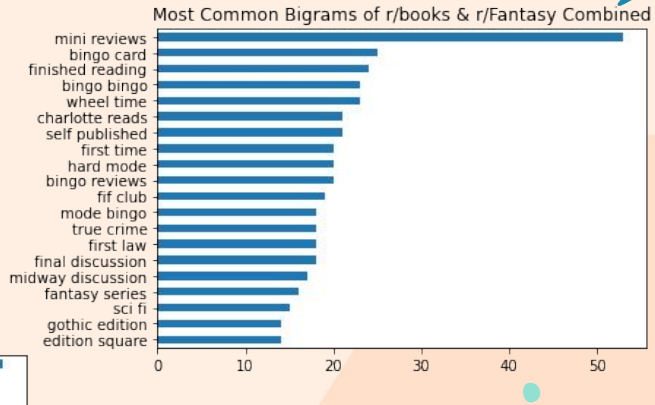
# CountVectorized Data - Bigrams




- Notable writers: Stephen King, JD Salinger, George Orwell

- Bingo Community event
- Notable series: Wheel of Time

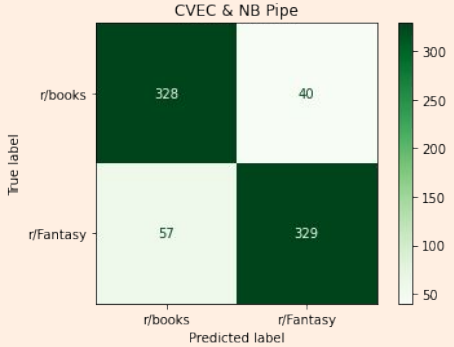


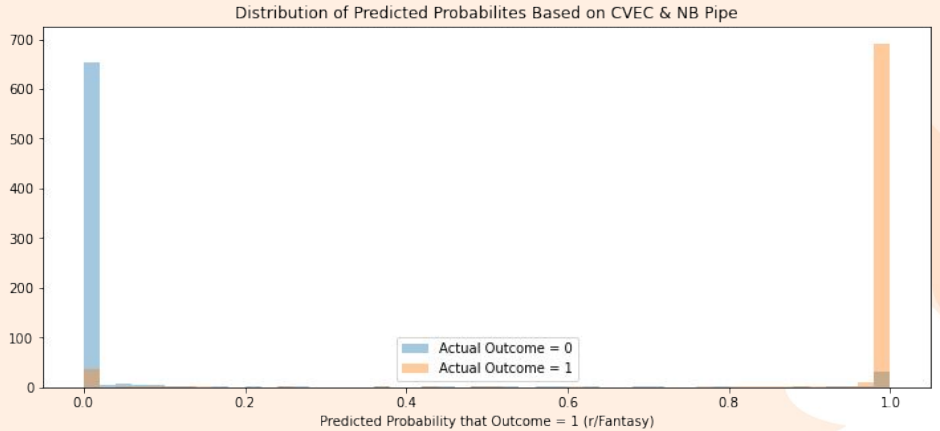




- Mini reviews and bingo lead combined bigrams

# GridSearch Models - CVEC & NB

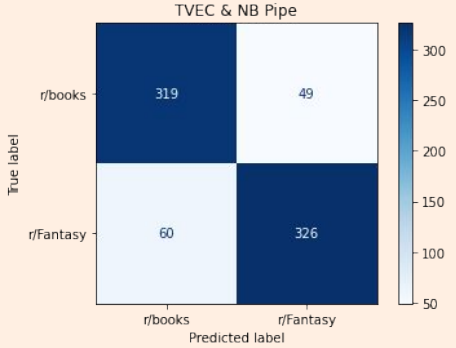
- Train Score: 0.8863334807607254
- Test Score: 0.8713527851458885
- Greater amount of absolute probability predictions vs 50/50 predictions

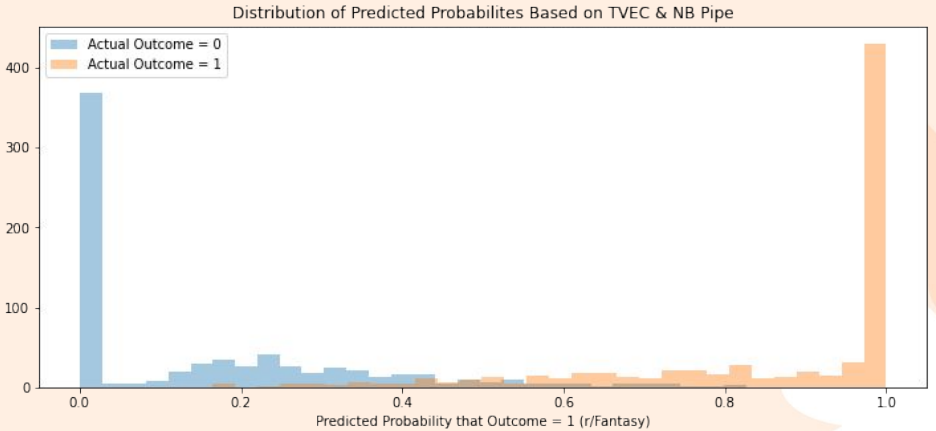


# GridSearch Models - TVEC & NB

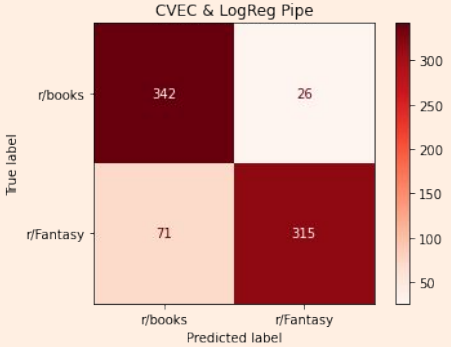
- Train Score: 0.8978328173374613
- Test Score: 0.8554376657824934
- Less absolute predictions compared to CVEC & NB model

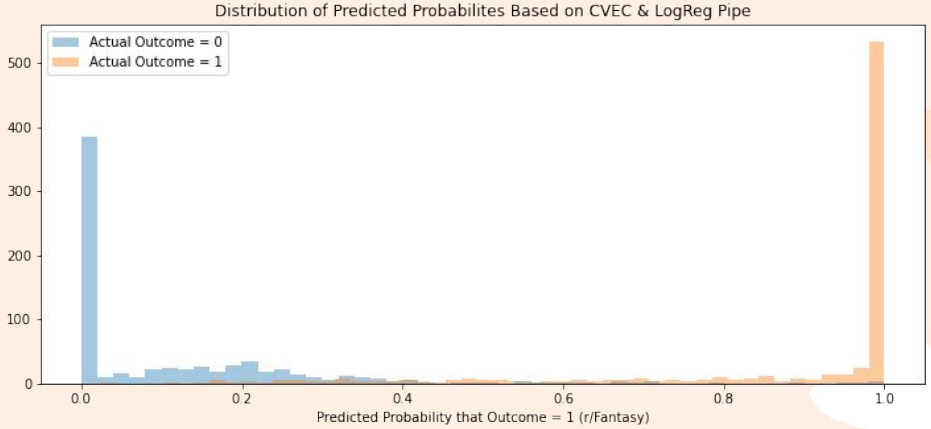


# GridSearch Models - CVEC & LogReg

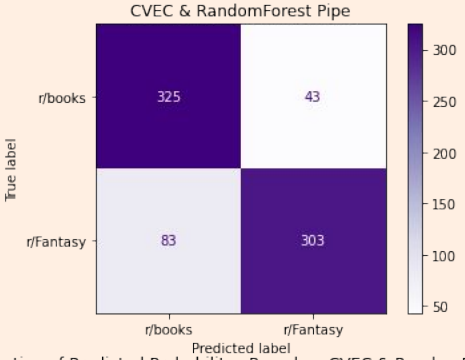
- Train Score: 0.9177355152587351
- Test Score: 0.8713527851458885
- More absolute r/Fantasy probability predictions than absolute r/books probability predictions

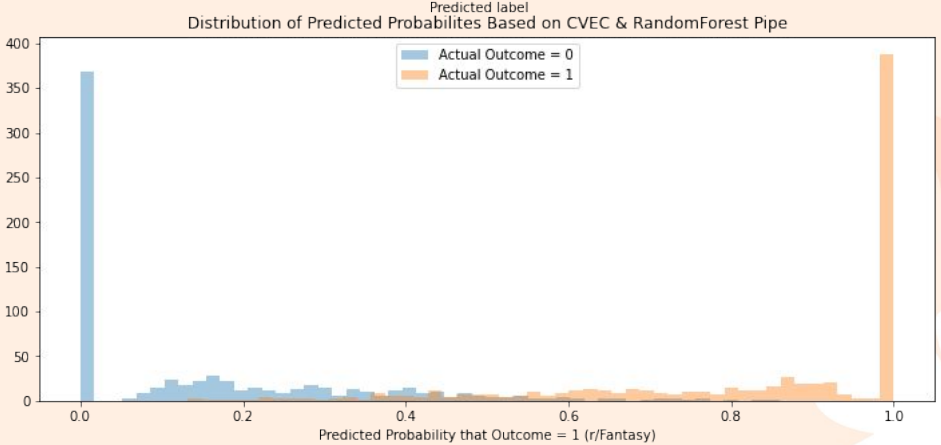


# GridSearch Models - CVEC & RandomForest

- Train Score: 0.9655019902697921
- Test Score: 0.8328912466843501
- Best train score, but weakest test score



# Conclusion/Recommendations

- GridSearching with CountVectorizer and Multinomial Naive Bayes had the best result when classifying r/Fantasy from r/books posts
  - 87.14% accuracy vs 51.14% baseline accuracy
  - Most accurate fit when comparing train and test scores
- Further Investigation/Recommendations:
  - Pull data from Horror and SciFi specific subreddits to compare to r/books data.
  - Perhaps pull data from before COVID pandemic to determine if people read more due to COVID shutdowns.

# Thanks for your time!