NLP: Bernoulli Naive Bayes

vs

Random Forest

By: Corina Lentz

Table Of Contents

- Problem Statement
- Background
- The Data
- Data Compiling, Cleaning, & Pre-Processing
- Exploratory Data Analysis
- Modeling Results
- Conclusion & Recommendations
- Resources

Problem Statement

More businesses are making the transition into online sales [1], which means an increased demand for predicting online shopping behavior, using methods such as Natural Language Processing (or NLP). The medical field is also turning to artificial intelligence to automate electronic medical and health records through NLP [2].

But which NLP model will give the most accurate predictions, given two possible outcomes? In this study we will compare the Bernoulli Naive Bayes and Random Forest NLP models, to see which one is the most accurate.

Background

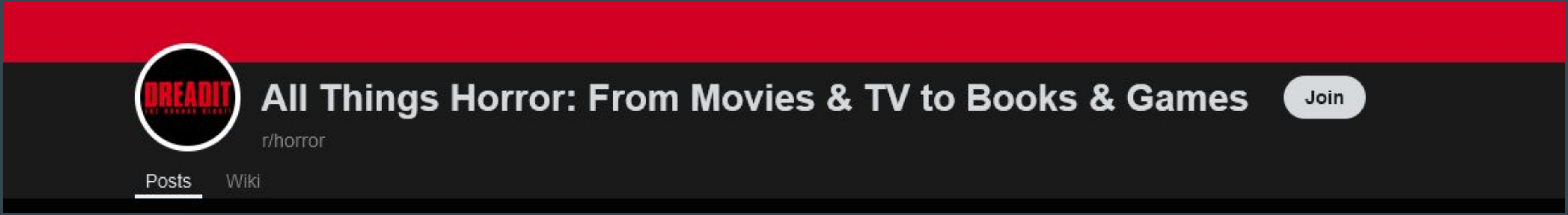
- Natural Language Processing was created in the 1950s but it wasn't until the late 1980s into the early 1990s when machine learning was created and integrated into NLP that it really became the robust language processing system that we know today [3].

- The first algorithm used for Random Forest was created in 1995, but Random Forest itself was created and trademarked in 2006 [4].

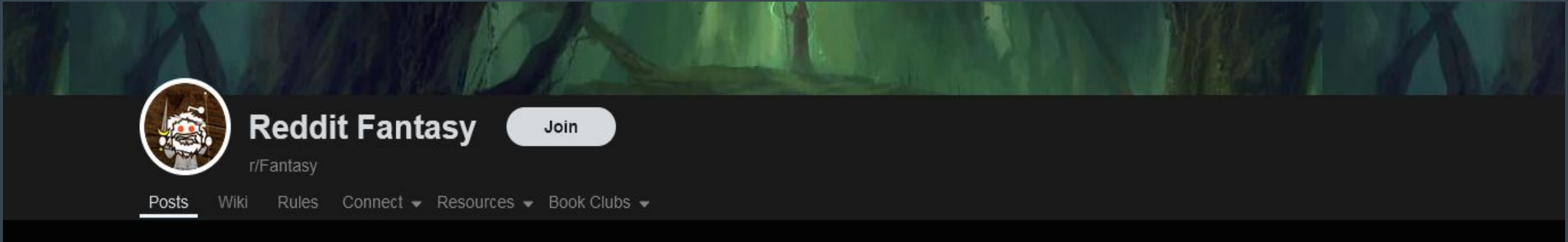
- Bernoulli Naive Bayes is based on the Bayes Theorem which was published in 1765 [5].

The Data

For this project I used Pushshift's Reddit API [6] to scrape the posts from two subreddits: r/horror [7]



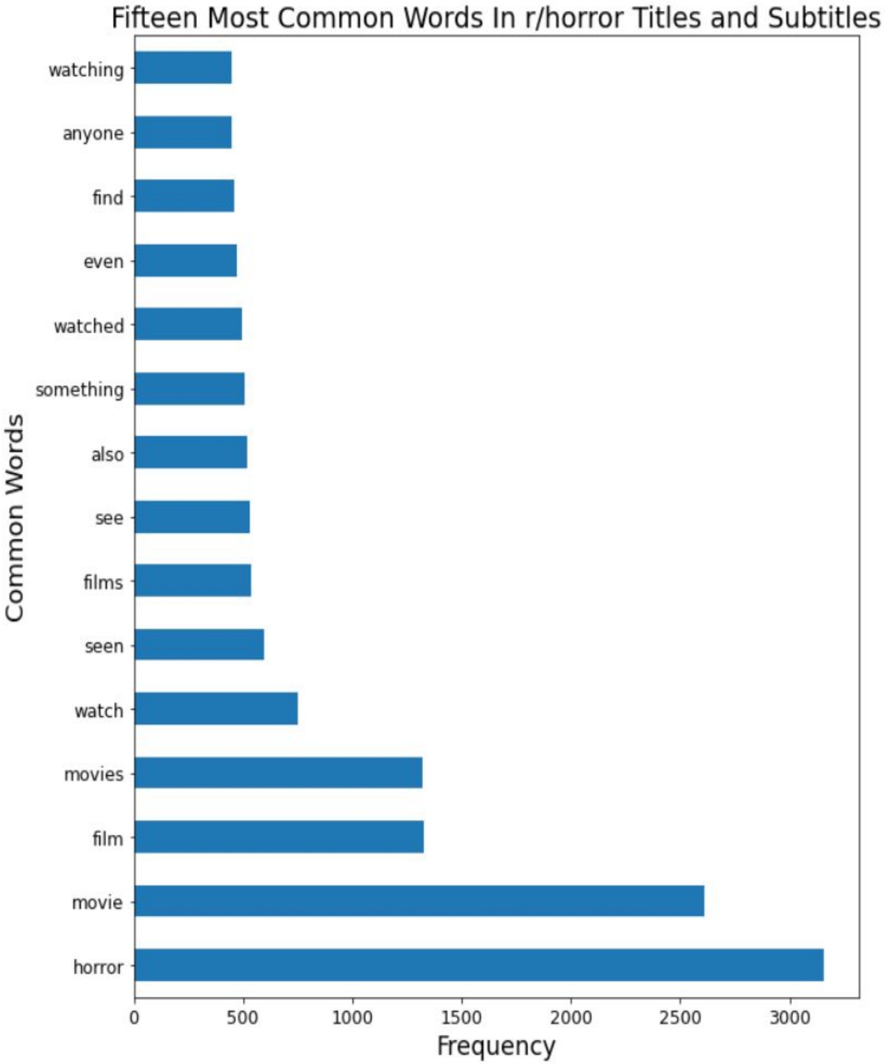
and r/Fantasy [8].

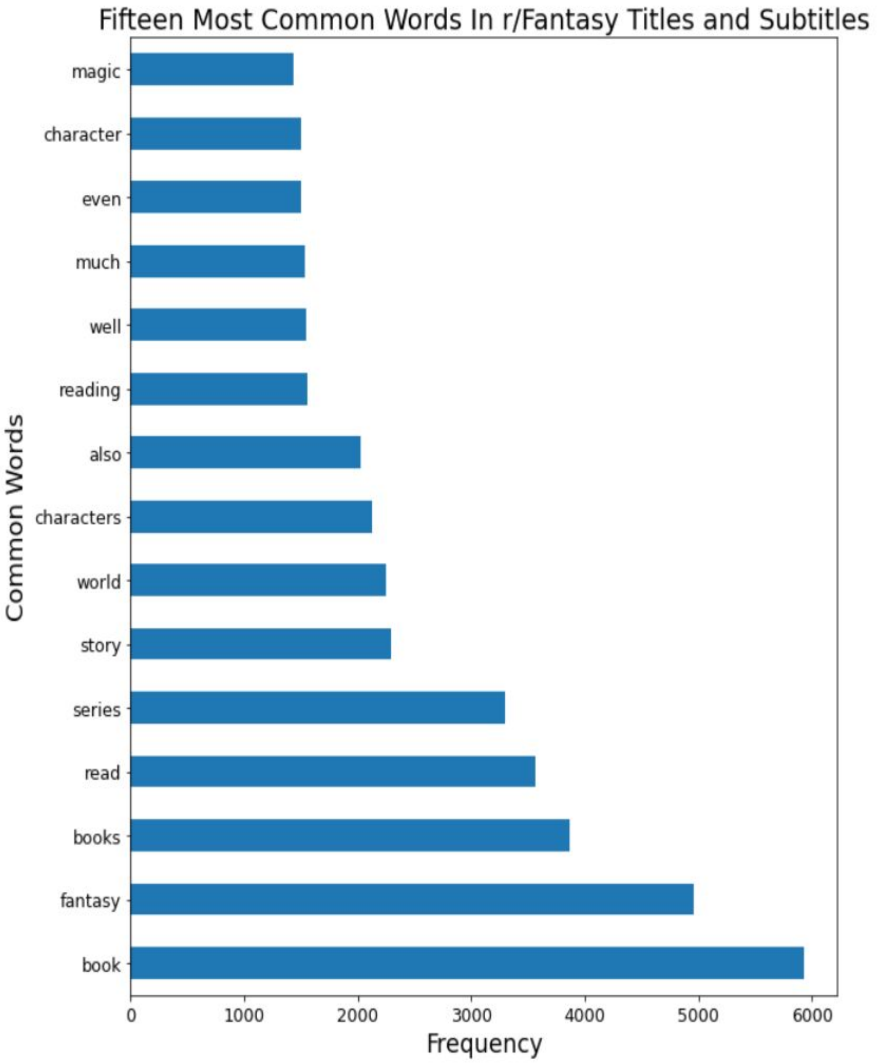


Data Compiling, Cleaning, & Pre-Processing

- I collected 5,000 posts from each subreddit for a total of 10,000 posts.
- After collecting the data I compiled the titles, subtitles, and the subreddit where the posts originally came from into a dataframe.
- I merged the text from the titles and subtitles into single strings of text so I could work with them more easily. (This also made it so I didn’t have to drop rows for posts with no subtitle.)

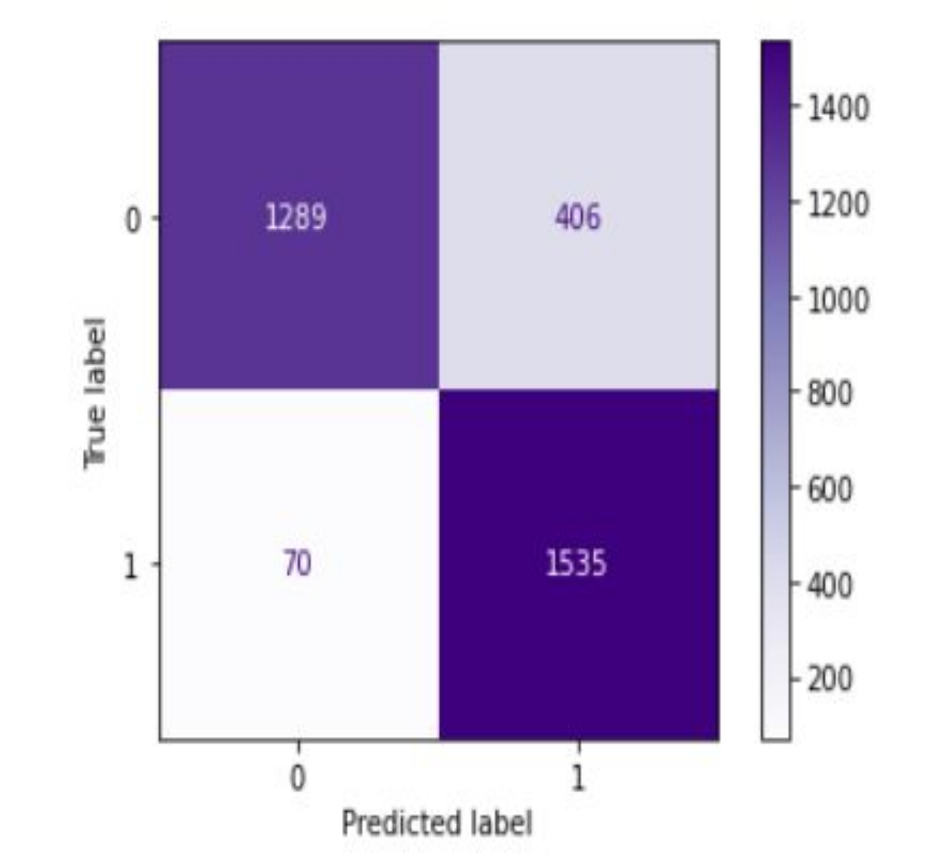
Exploratory Data Analysis


Fifteen Most Common Words In r/horror Titles and Subtitles
Common Words
watching
anyone
find
even
watched
something
also
see
films
seen
watch
movies
film
movie
horror
0
500
1000
1500
2000
2500
3000
Frequency


Fifteen Most Common Words In r/Fantasy Titles and Subtitles
Common Words
magic
character
even
much
well
reading
also
characters
world
story
series
read
books
fantasy
book
0
1000
2000
3000
4000
5000
6000
Frequency

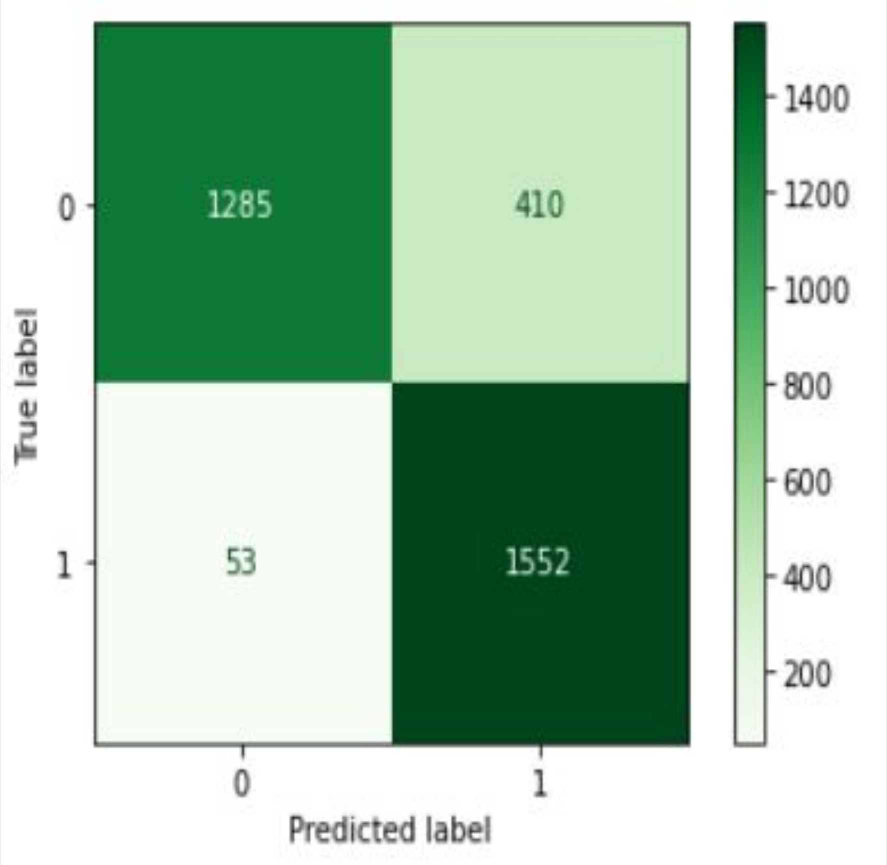
Modeling Results

Bernoulli NB Confusion Matrix



True label \ Predicted label	0	1
0	1289	406
1	70	1535

Random Forest Confusion Matrix



True label \ Predicted label	0	1
0	1285	410
1	53	1552

	Accuracy	Training Score	Testing Score	Variance
Bernoulli NB	0.8531	0.8514	0.8557	0.004265
Random Forest	0.8561	0.864	0.8596	0.004332

Conclusion & Recommendations

Although the Random Forest model was more accurate than the Bernoulli Naive Bayes model it was by a fairly slim margin. Either of these models can be used to give excellent results.

Going forward, I would recommend dedicating more time to building-out an even more robust stop words list to improve accuracy even more. I'd also recommending testing a Logistic Regression model for further comparison.

Resources

[1] https://fortune.com/2020/07/15/ecommerce-online-shopping-coronavirus-business-trends-covid/

[2] https://healthtechmagazine.net/article/2021/07/how-can-healthcare-leverage-natural-language-processing-medical-records-perfcon

[3] https://en.wikipedia.org/wiki/Natural_language_processing

[4] https://en.wikipedia.org/wiki/Random_forest

[5] https://en.wikipedia.org/wiki/Bayes'_theorem

[6] https://github.com/pushshift/api

[7] https://www.reddit.com/r/horror/

[8] https://www.reddit.com/r/Fantasy/

Questions?