# Finding Successful Features in Book Reviews

Name: Mahamuge D N Costa
Student No: 13030224
Module Name: Big Data
Assignment Component: B
Module Code: UFCF8H-15-M
Dataset Link: https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews
Notebook Link: https://colab.research.google.com/drive/12uRsM-Tvhsbo-6_b04oOusaBYwPY3asi?usp=sharing

## Table of Contents

## List of Figures

# The Problem Definition

The market for books and written material is very large and very competitive. This is in main part due to the rise of digitized books (Yahoo, 2023) as well as the continued and renewed interest of the conventional reader to go back to the parchment to get the "feel" of reading a real book. Hence, the consumer is ever willing to research their books and read the plethora of reviews and scores alike to make their choice in the book they would want to purchase. As an active member in these spaces, they serve as the backbone of the industry by providing comprehensive reviews for the books they read. However, often, the production side of these books (writers, editors, publishers etc.) may not fully understand the broad features that appeal (or not) to the masses that consume literary works (Blankenship, 2023). This invalidates those dedicated readers and their work in leaving reviews that actually speak to the content within the book and its actual feel. In order to determine this "star" power, this paper aims to investigate a natural language processing model that may be able to determine features that affect the ratings for books. Additionally, it aims to discover those features and characteristics in books that drive customers to leave positive or negative feedback on these books.

# The Dataset and Filtering

## The Dataset Analysis

The dataset was obtained through Kaggle (Bekheet, 2022) which Bekheet had filtered from the larger complete review of all products dataset (about 140 million review records) by Ni who had scraped it from Amazon (Ni, Li and McAuley, 2019). This dataset comprises of 10 attributes which describe the reviews of 212,404 distinct books. The total number of distinct reviews are 3 million and is updated regularly. There are 10 attributes in the dataset which include:

1. The Title of the Book
2. The Review score
3. The Review text

These 3 attributes are filtered from the source dataset since the assumption was made that these values do not have significance for the reviews. The excluded attributes are as follows:

1. ID
2. Price
3. User_id
4. profileName
5. review/time
6. review/summary
7. The Helpfulness of the Review

From the remaining data, filtered accordingly, exploratory data analysis (EDA) was performed. It is clear that only a handful of books have a higher amount of reviews published due to their sheer popularity be it negative or positive. This can be visualized using a plot of data showing the spread of review counts:

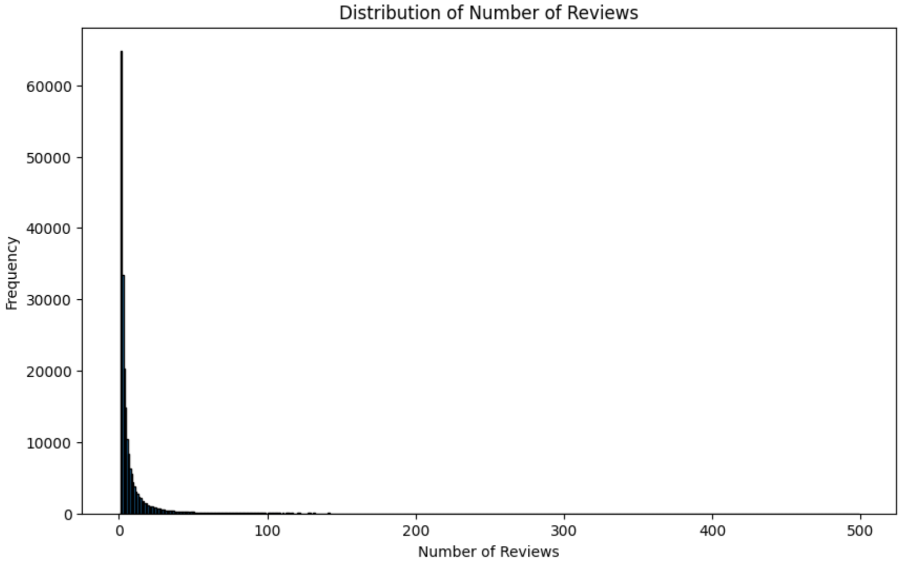


*Figure 1 Review Counts in the 0 to 500 range where frequency is the number of books*

Here it is clear that most books are receiving less than 100 and more closer to 0 reviews. This is solidified upon further investigation:

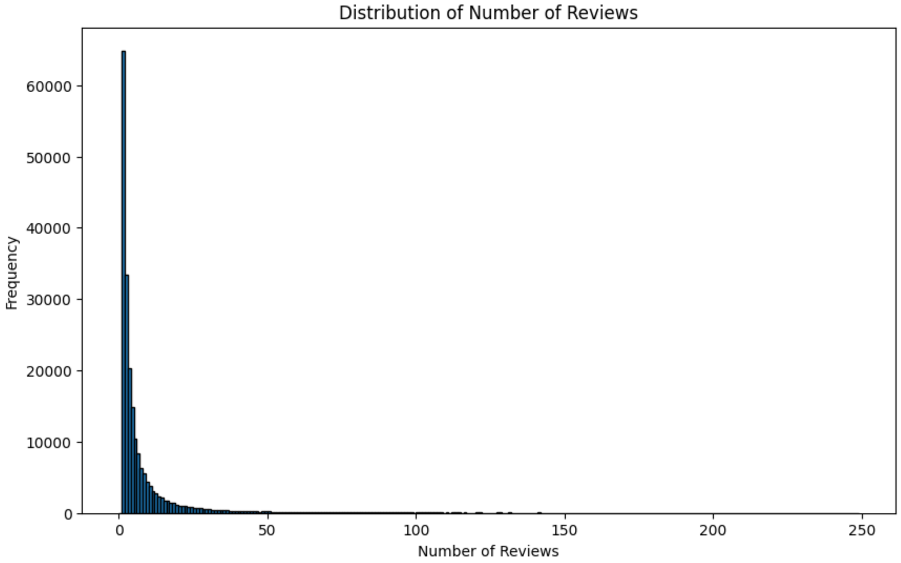


*Figure 2 Review Counts in the 0 to 250 range*

And further:

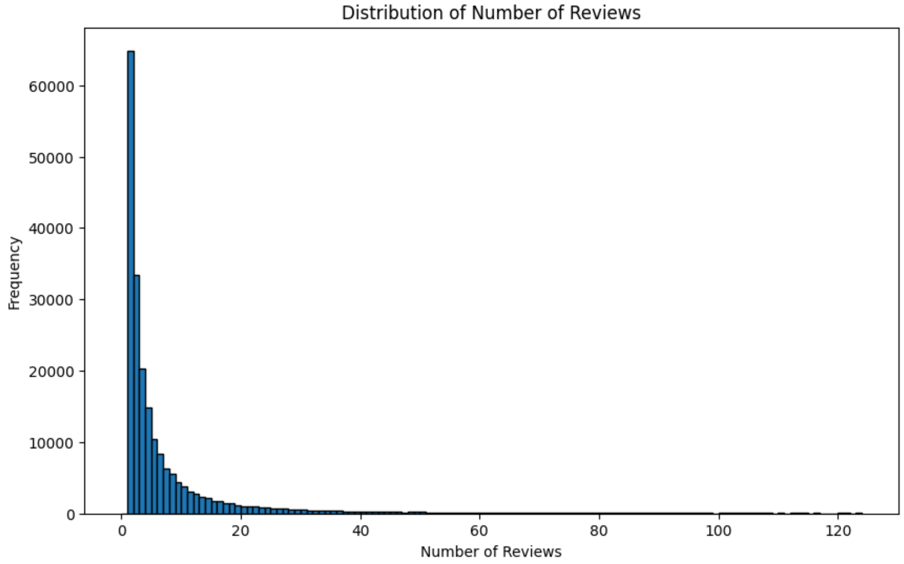


*Figure 3 Review Counts in the 0 to 125 range*

This can be due to many reasons including users who tend to simply click on the rating and leave empty the review text since that is far more convenient. This leaves the majority of books with closer to 0 reviews. Hence, in order to analyze reviews such that a minimum sufficient number of reviews is considered for each book, an arbitrary number of at least 200 was chosen. A randomized sample of reviews could have been selected for the development of this model, however, to allow for equivalent distribution of reviews with similar user groups (since specific books tend to be read by users with similar interests at a higher frequency) and to allow for equivalent spread of reviews between different books, the sample of books with numbers of reviews between 200 and 300 were selected.

Furthermore, the reviews included in the dataset were found to have missing data, punctuation, contracted words, mixed case sentences and extra spaces etc. Just to name a few anomalies.

All these attributes affect any model that needs to be trained on this data. Therefore, prior to any manipulations, this data was cleaned (Bansal, 2015). In addition to this, to avoid word variations resulting in false feature creation, the data was lemmatized. This meant that each word in each review would be reduced to their base format so that there are no variations of the same word. Finally a total of 107,791 reviews were analyzed to determine what user's deem important in book ratings.

With this data further EDA showed the significant words in reviews. The following word clouds were generated using the vectorized data of the top 10 books from the sample (top 10 selected from the aggregate average of the score) (Sharma, 2020):

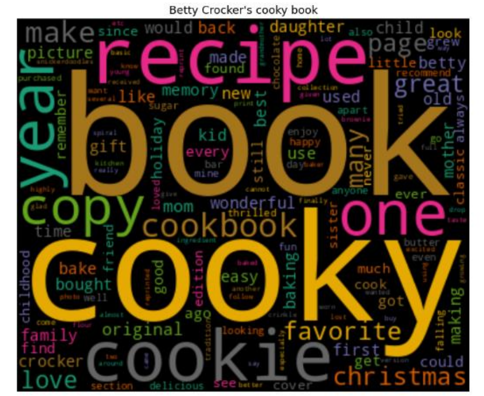


Figure 4 Top 1

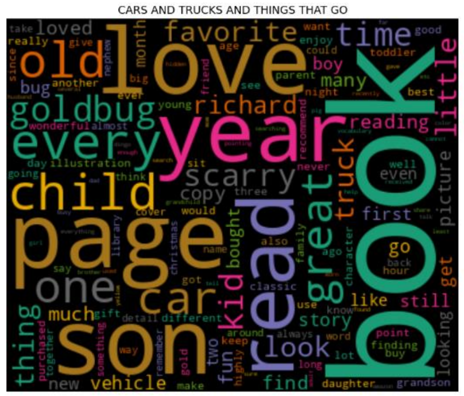


Figure 5 Top 2

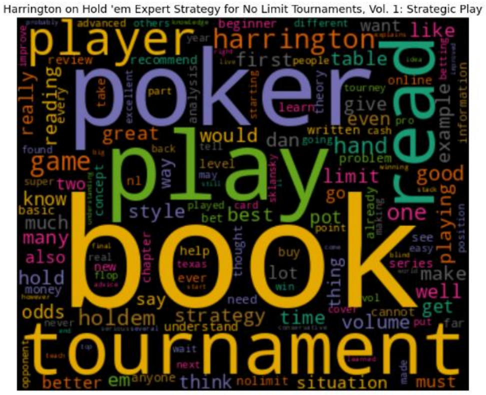


Figure 6 Top 3

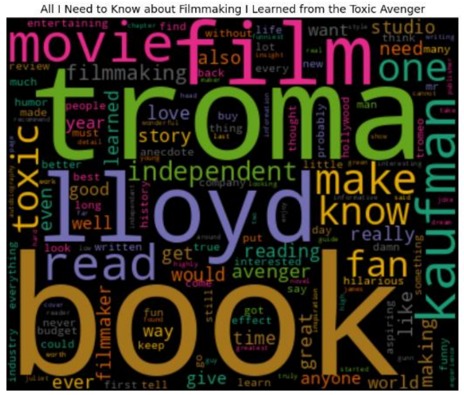


Figure 7 Top 4

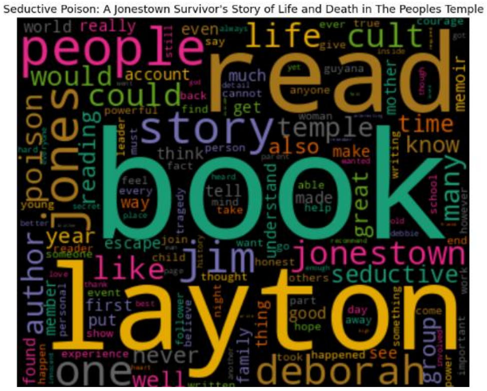


Figure 8 Top 5

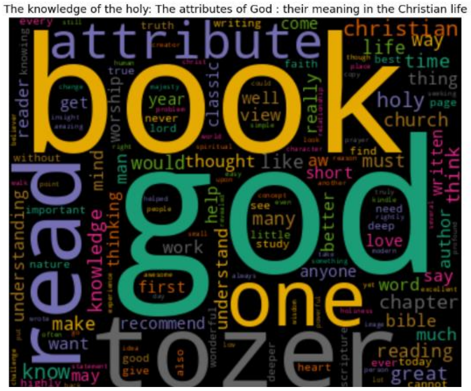


Figure 9 Top 6

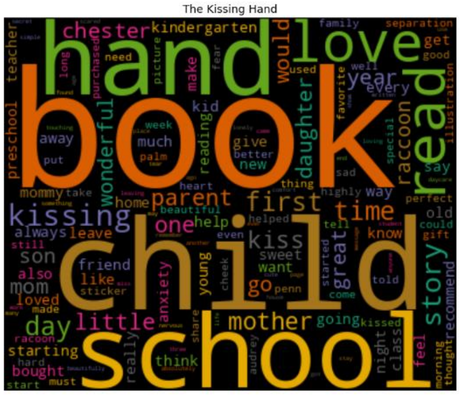


*Figure 10 Top 7*

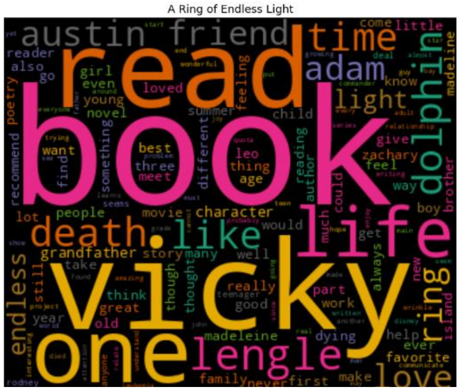


*Figure 11 Top 8*

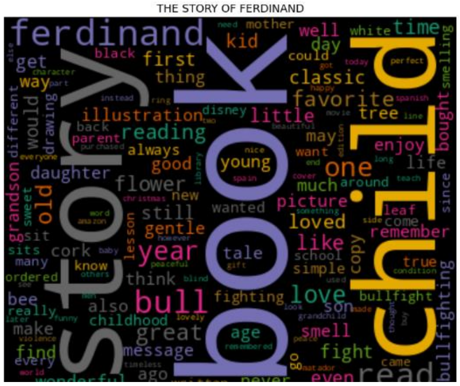


*Figure 12 Top 9*

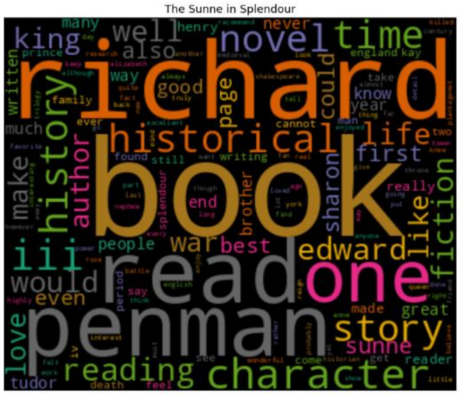


*Figure 13 Top 10*

From this analysis, it is clear that certain words like “Book” and “Read” while important, may not have feature importance relevant to the rating score of the book since majority of book reviews will have those words.

Furthermore, analyzing the reviews generated, a significant number of reviews were 5 star reviews. However, the aggregate of reviews of 1 star to 4 star allowed a comparable matching of 62,813 five star reviews to 44,978 non-five star reviews (Blankenship, 2023). This was decided to be better since one aim of this analysis was to allow for the writer to visualize what terms affected their books positively, which may be arguably more significant than the negative.

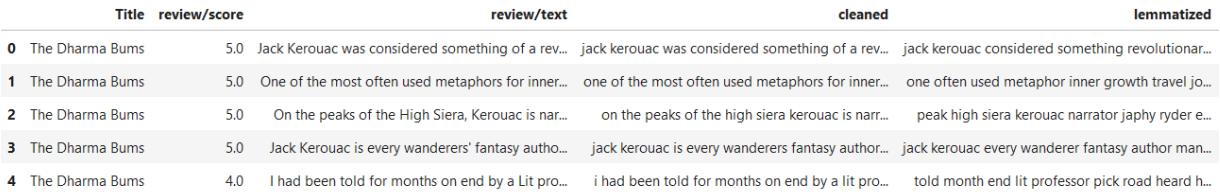
In order to visualize the true functionality of natural language processing, an appropriate classification model needs to be chosen so as to determine the possible rating of a given review.

## The Workflow

The pipeline of developing the model is as follows (Occhipinti, Rogers and Angione, 2022):

1. Obtain Raw text with classification
2. Clean the text set - removing stop words, capitalization, non-alphabetic characters, lemmatizing etc.
3. Tokenize this text.
4. Vectorizing the dataset - Converting the dataset to a numeric representation.
5. Create the machine learning model
    a. Evaluate Multiple model performance using train data
    b. Select most appropriate and best performing model
    c. Run test data on this target model to evaluate final performance
6. Evaluate performance of the final model
7. Use the model to determine which features lead to better ratings and which lead to lesser ratings.

As described, the data was tokenized and cleaned and made ready to be vectorized so as to gain an accurate feature matrix. To ensure that the dataset adhered to Natural Language Processing guidelines, the NLTK library was used for a majority of the cleaning process. With it the Lemmatizing and Stopword elimination processes were achieved:

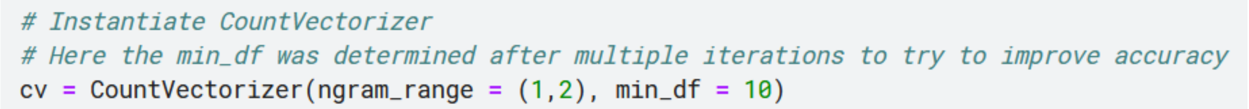


| | Title | review/score | review/text | cleaned | lemmatized |
|---|---|---|---|---|---|
| 0 | The Dharma Bums | 5.0 | Jack Kerouac was considered something of a rev... | jack kerouac was considered something of a rev... | jack kerouac considered something revolutionar... |
| 1 | The Dharma Bums | 5.0 | One of the most often used metaphors for inner... | one of the most often used metaphors for inner... | one often used metaphor inner growth travel jo... |
| 2 | The Dharma Bums | 5.0 | On the peaks of the High Siera, Kerouac is nar... | on the peaks of the high siera kerouac is narr... | peak high siera kerouac narrator japhy ryder e... |
| 3 | The Dharma Bums | 5.0 | Jack Kerouac is every wanderers' fantasy autho... | jack kerouac is every wanderers fantasy author... | jack kerouac every wanderer fantasy author man... |
| 4 | The Dharma Bums | 4.0 | I had been told for months on end by a Lit pro... | i had been told for months on end by a lit pro... | told month end lit professor pick road heard h... |

*Figure 14 Cleaned dataset*

In the vectorization stage, some anomalies were noted. Since the CountVectorizer from Scikitlearn defaults to eliminating only single instance features from the dataset, any and all words that occurred more than once were vectorized. Hence, when printing samples of these vectorized arrays, certain features that had little "English" meaning had been considered as features. Therefore, a minimal number of word occurrence for each word or pair of words was needed in order to consider the word/pair as a feature.

This was carried out by iteratively running and testing accuracy variations of a Linear Regression model on the entire dataset (with no splitting):

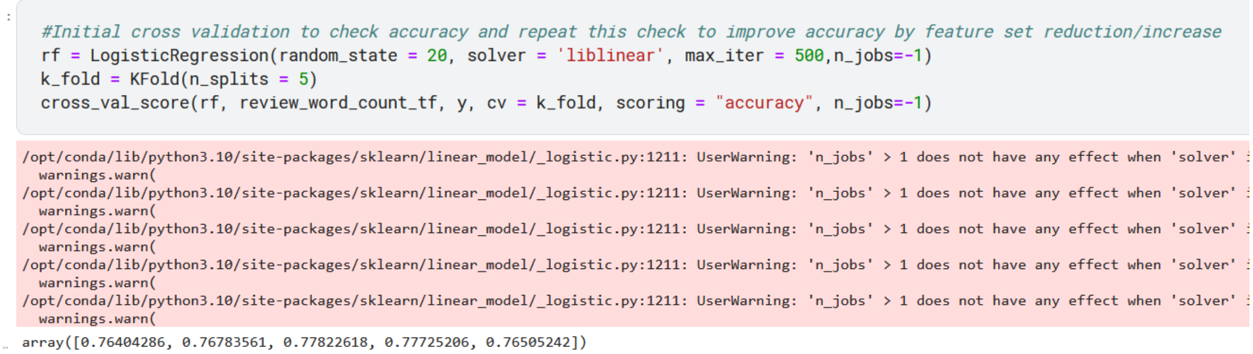


```
# Instantiate CountVectorizer
# Here the min_df was determined after multiple iterations to try to improve accuracy
cv = CountVectorizer(ngram_range = (1,2), min_df = 10)
```

*Figure 15 Adjusting min_df property for feature reduction*

Here the min_df property was continually adjusted and the Linear Regression model was tested with cross validation to obtain best performance:

As shown in Figure 15, the best performance was obtained using a min_df value of 10. This meant each feature needed to have at least 10 repetitions in the entire dataset to be considered a feature for the model.



```
#Initial cross validation to check accuracy and repeat this check to improve accuracy by feature set reduction/increase
rf = LogisticRegression(random_state = 20, solver = 'liblinear', max_iter = 500,n_jobs=-1)
k_fold = KFold(n_splits = 5)
cross_val_score(rf, review_word_count_tf, y, cv = k_fold, scoring = "accuracy", n_jobs=-1)
```

```
/opt/conda/lib/python3.10/site-packages/sklearn/linear_model/_logistic.py:1211: UserWarning: 'n_jobs' > 1 does not have any effect when 'solver'
  warnings.warn(
/opt/conda/lib/python3.10/site-packages/sklearn/linear_model/_logistic.py:1211: UserWarning: 'n_jobs' > 1 does not have any effect when 'solver'
  warnings.warn(
/opt/conda/lib/python3.10/site-packages/sklearn/linear_model/_logistic.py:1211: UserWarning: 'n_jobs' > 1 does not have any effect when 'solver'
  warnings.warn(
/opt/conda/lib/python3.10/site-packages/sklearn/linear_model/_logistic.py:1211: UserWarning: 'n_jobs' > 1 does not have any effect when 'solver'
  warnings.warn(
/opt/conda/lib/python3.10/site-packages/sklearn/linear_model/_logistic.py:1211: UserWarning: 'n_jobs' > 1 does not have any effect when 'solver'
  warnings.warn(
array([0.76404286, 0.76783561, 0.77822618, 0.77725206, 0.76505242])
```

*Figure 16 LR Cross Val for min_df adjustment*

To allow for word prominence, the TF_IDF vector transformation was used so that weighing was given to "heavier" words in the reviews. The cleaning and minimal word repetition resulted in a feature set of 100,456. Prior to this feature reduction, the feature set from vectorizing was above 3,000,000. This may have heavily affected the performance of the algorithm and single occurrence words and pairs may have affected the fit of the model as well. Additionally, both unigrams and bigrams were considered in feature generation. This was done to improve contextual representation of the text rather than just considering singular words.

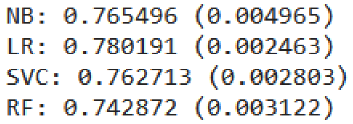
As mentioned, the review scores of each book was categorized into either "5 star" or "not 5 star". This was decided in order to allow for Binary classification and to account for the significant number of 5 star reviews which may have skewed the models. This bivariate analysis allows for better performance especially in text classification problems with its large feature set. This also allows for users of this model to more specifically understand those factors that users look for in a 5 star book. Since this model is to be developed on existing score and text data, it can be considered a supervised learning model.

## Selecting the Algorithm for Modelling

In order to select an appropriate learning algorithm for this text analysis and classification, cross validation was used. Here, the following 4 models were selected for comparison as they are often used in text classification (Occhipinti, Rogers and Angione, 2022):

1. Multinomial Naïve Bias
2. Logistic Regression
3. Linear Support Vector Classifier
4. Random Forrest

In the results obtained, the best accuracy score was obtained by Logistic Regression (0.78) and hence was selected for the model for this classification problem:



```
NB: 0.765496 (0.004965)
LR: 0.780191 (0.002463)
SVC: 0.762713 (0.002803)
RF: 0.742872 (0.003122)
```

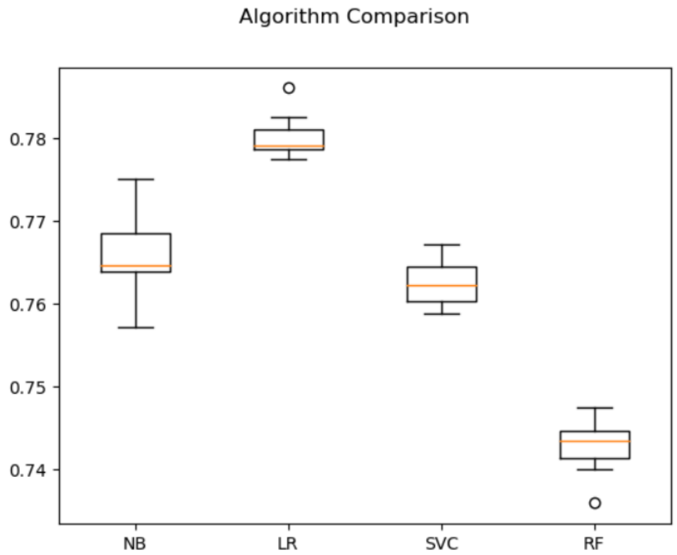


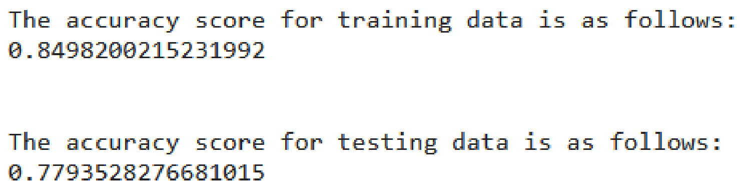
*Figure 17 Accuracy scores of algorithms considered*

In addition to these results, the training duration for the Logistic Regression model was only surpassed by the Naïve Bias. Therefore, with the higher accuracy and lower training duration, the choice is Linear Regression.

## Training the Model

The data was split into 75% training and 25% testing data and was fit. The performance of the model is as follows.

Using linear regression, the accuracy scores for training and testing are:

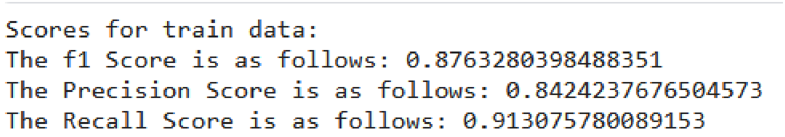


```
The accuracy score for training data is as follows:
0.8498200215231992


The accuracy score for testing data is as follows:
0.7793528276681015
```

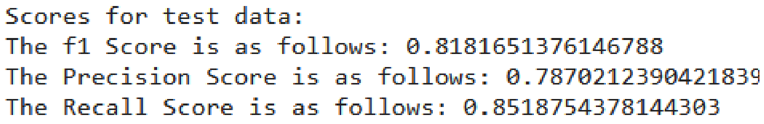
*Figure 18 Accuracy of LR*

Here it is evident that the model is performing as expected with barely 6% deviation between the training and testing set. Furthermore, when evaluating the precision, recall and the F1 score:



```
Scores for train data:
The f1 Score is as follows: 0.8763280398488351
The Precision Score is as follows: 0.8424237676504573
The Recall Score is as follows: 0.913075780089153
```

*Figure 20 F1, Precision and Recall for LR Training*



```
Scores for test data:
The f1 Score is as follows: 0.8181651376146788
The Precision Score is as follows: 0.7870212390421839
The Recall Score is as follows: 0.8518754378144303
```

*Figure 19 F1, Precision and Recall for LR Testing*

Similar low deviations from training and testing is noted. In natural language processing, especially when processing user generated text data, higher accuracy and performance scores are difficult to achieve considering the sheer variation of language and typing methodologies utilized by users. Hence, these scores can be considered to be sufficient for our model. Additionally, the parent dataset of 3 million records could not be used due to its large size and computational complexity. These scores may improve with a larger dataset, however, that is not guaranteed since it is still trained on user generated text and not standardized text.

## High Importance Features

One of the main goals of this endeavor was to allow users to understand which features of books are appealing to users and which should be avoided. Using the feature names generated in the vectorization phase, and the coefficient scores of the Linear Regression algorithm, the 20 most and least important features are as follows:

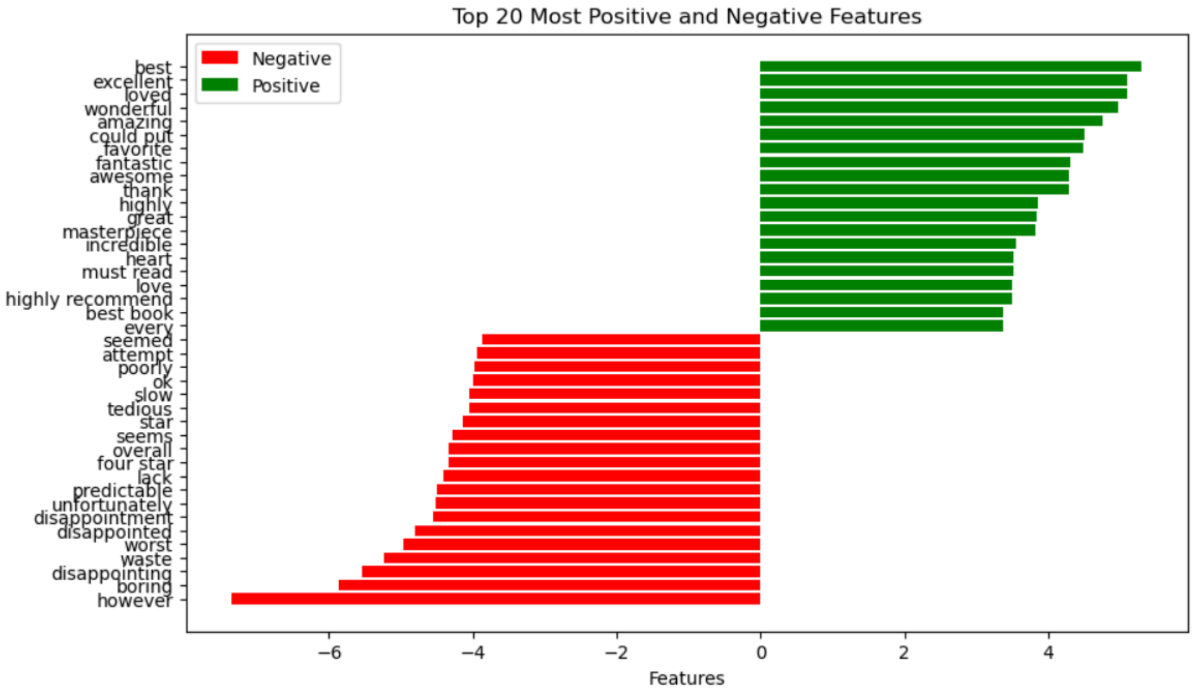


*Figure 21 Feature Importance*

Here it is clearly seen that certain word combinations affect the score given by users. With this evaluation, users may utilize this model to determine and predict the features that are to illicit positive reception of their work. Additionally, users on the business end of book production (publishers, vendors etc.) may use this model to determine effective strategies for marketing and sales to ensure that their books are successful.

## Less Suitable Model

While it maybe commonly used for classification problems and while it is known to perform effectively with these problems, Random Forests may not be suitable to be used here. In actual testing, Random Forest was the most computationally expensive of all the models evaluated. Coupled with its lower accuracy, it is clear that it is not suitable for this operation. Additionally, in Random Forests, with this high count of dimensions and high number of individual records mean that training and fitting the model takes considerable amount of time and resources. This could be arguably limited by limiting the growth of the tree but that would lead to decreasing in its accuracy.

# Conclusion

As stated in the problem definition, this model maybe utilized to determine features of high importance in books. This may allow writers to gain an understanding of attributes that are needed for successful storytelling and to understand trends in the industry to maximize sales. On the other hand, vendors and publishers can use this model to determine the most successful elements of books and use that for strategies to maximize profits (such as marketing, target demographics etc.).

The model heavily depends on the pre-processing of the data provided. If the data is not processed sufficiently, the accuracy of all models evaluated drops significantly. Additionally, the curse of dimensionality comes into play with the multitude of features due to vectorization of unclean text data.

# Word Count

2170 Words.

# References

Bansal, S. (2015) *6 Practices to enhance the performance of a Text Classification Model* [Document]. Available at: https://www.analyticsvidhya.com/blog/2015/10/6-practices-enhance-performance-text-classification-model/ [Accessed 20 Augustus 2023].

Bekheet, M. (2022) *Amazon Books Reviews* [Online]. Available from: https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews [Accessed 20 Augustus 2023].

Blankenship, L. (2023) *New research from Hallie Cho investigates the relationship between quantitative star ratings, qualitative text reviews, and product demand* [Document]. Available at: https://business.vanderbilt.edu/news/2023/03/10/how-written-product-reviews-influence-consumer-impressions-of-star-ratings/ [Accessed 20 Augustus 2023].

Ni, J., Li, J. and McAuley, J. (2019) Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China, 2019. Association for Computational Linguistics.

Occhipinti, A., Rogers, L. and Angione, C. (2022) A pipeline and comparative study of 12 machine learning models for text classification. *Expert Systems with Applications* [Dcoument]. 201, p.117193. Available at: https://www.sciencedirect.com/science/article/pii/S0957417422005802 [Accessed 20 Augustus 2023].

Sharma, A. (2020) *A Beginner's Guide to Exploratory Data Analysis (EDA) on Text Data (Amazon Case Study)* [Web Page]. Available at: https://www.analyticsvidhya.com/blog/2020/04/beginners-guide-exploratory-data-analysis-text-data/ [Accessed 20 Augustus 2023].

Yahoo. (2023) *Global Books Market to 2030: Growing Number of Readers Worldwide and the Rise of E-books fuel the Sector* [Web Page]. Yahoo Available at: https://finance.yahoo.com/news/global-books-market-2030-growing-160800759.html [Accessed 20 Augustus 2023].