



IMDB Movie Reviews with Ratings
Final White Paper

DSC680 – Applied Data Science
Project – Milestone 3
Winter 2021
Manish Kalkar

Final White Paper

1. Background

Customer reviews are increasingly available online for a wide range of products and services. They supplement other information provided by electronic storefronts such as product descriptions, reviews from experts, and personalized advice generated by automated recommendation systems. While researchers have demonstrated the benefits of the presence of customer reviews to an online retailer, a largely uninvestigated issue is what makes customer reviews helpful to a consumer in the process of making a purchase decision. Drawing on the paradigm of search and experience goods from information economics, we develop and test a model of customer review helpfulness. An analysis of 50000 unique movie reviews on IMDB across wide range of movies indicated that review extremity, review depth, and overall rating affect the perceived helpfulness of the review. Overall rating moderates the effect of review extremity on the helpfulness of the review. For movie experience, reviews with extreme ratings are less helpful than reviews with moderate ratings. For all movie reviews, review depth has a positive effect on the helpfulness of the review. Review depth has a greater positive effect on the helpfulness of the review for the search of the movie. We discuss the implications of our findings in terms of overall ratings for both theory and practice.

This is a dataset for sentiment classification containing reviews by the movie viewers and corresponding ratings. Preprocessed Stanford IMDB Movie Review dataset is in the CSV format with sentiment and rating columns.

2. Business Problem

The dataset contains the IMDB Movie Reviews. Going through all the reviews can be very tedious because of reviews captured online in open text fields. The opinions expressed in free text fields somehow need to be categorized in the review forums. The data mining / text analysis needs to be done in order to identify the review category. The categorized review then can be utilized for review management system. The Classification of individual reviews also determine rating given by the viewer based on individual reviews. It enables the IMDB to develop more loyal viewership, thereby increase in business, brand value and profits.

3. The Data / Dataset

Dataset - <https://www.kaggle.com/nisargchodavadiya/imdb-movie-reviews-with-ratings-50k>

Preprocessed Stanford IMDB movie review dataset of 50000 unique movie reviews in CSV format with sentiment and rating column also to predict ratings. The dataset includes basic sentiment and rating columns for each IMDB Movie Review.

4. Methods

- Exploratory Data Analysis
- Data Preparation
 - Feature Selection
 - Feature Engineering
- Sentiment Analysis
- Model Selection
- Model Evaluation
- Results / Outcome
- Conclusion

5. Exploratory Data Analysis

The goal is to perform the analysis of the IMDB Movie Reviews along with the Sentiment Analysis of the reviews and build model to predict the Overall Rating of the Movies. Review and Rating are the only relevant features / columns in the dataset.

Dimensionality & Feature Selection / Reduction

List of features available and their categorization relevant vs. irrelevant

1. Review - relevant feature – selected for the analysis and modeling
2. Rating - relevant feature – selected for the analysis and modeling
3. Sentiment - relevant feature – selected for the analysis only

Interestingly Rating 1 has got the highest number of counts followed by 10, 8 and 4. The reviews seem to have Ratings that are distributed almost evenly across from the happiest movie watchers with Rating 10 has Count of 9731 are almost the same as unhappiest movie watchers with Rating 1 that has count of 10122.

Based on the count, Rating 1 has got the highest percentages followed by 10, 8 and 4. The reviews seem to have Ratings that are distributed evenly across from the happiest movie watchers with Rating 10 are about 20%, which are almost the same as unhappiest movie watchers with Rating 1 are 19.46%.

	Review	Rating	Sentiment
0	Kurt Russell's chameleon-like performance, cou...	10	1
1	It was extremely low budget(it some scenes it ...	8	1
2	James Cagney is best known for his tough chara...	8	1
3	Following the brilliant "Goyôkiba" (aka. "Hanz...	8	1
4	One of the last classics of the French New Wav...	10	1
...
49995	(spoiler) it could be the one the worst movie ...	4	0
49996	So, you've seen the Romero movies, yes? And yo...	1	0
49997	Just listen to the Broadway cast album and to ...	3	0
49998	I have been a fan of the Carpenters for a long...	3	0
49999	Set in 1945, Skenbart follows a failed Swedish...	1	0

50000 rows × 3 columns

Feature Engineering

Below are the steps taken to further engineer the features to make them ready for the analysis.

- Address any missing data issues
- The free text field “Review” was further processed by text cleaning:
 - Made text lowercase
 - Removed text in square brackets
 - Removed links
 - Removed special characters
 - Removed words containing numbers
 - Removed punctuation
- Split the data into 80% - Trained Data and 20% Test Data
- Applied TFIDF Vectorizer to transform the text into feature vectors that can be used as input to estimator

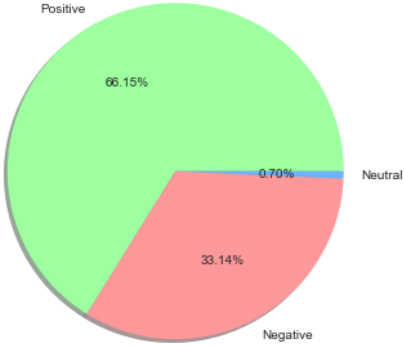
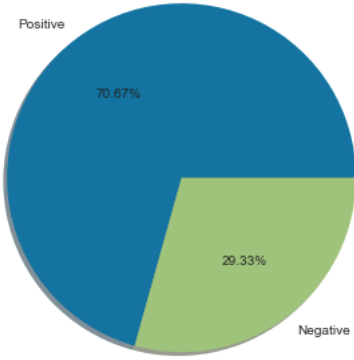
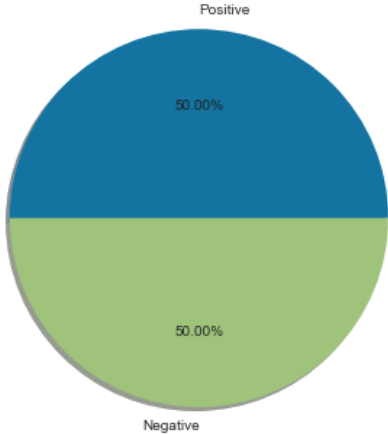
Sentiment Analysis

Added / renamed below features / columns to accommodate sentiment data

- Added - sentiment
- Added - sentimentScore
- Renamed - sentiment → OriginalSentiment

Sentiment Analysis has been performed based on below listed methods:

- VADER
- Textblob
- Original Sentiment score given in the dataset

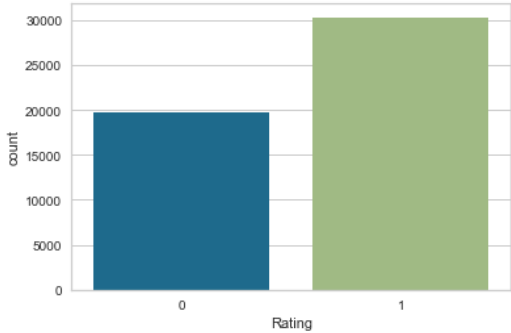
VADER	textblob Blobber and NaiveBayesAnalyzer	Based on Original Sentiment Score given the dataset						
Measure Sentiment	Measure Sentiment	Measure Sentiment						
-- sentiment - pos - Positive, neg - Negative, neu – Neutral -- sentimentScore - 2 for Positive, 0 for Negative and 0 for Neutral	-- sentiment - pos - Positive, neg – Negative -- sentimentScore - 1 for Positive, 0 for Negative -- Neutral and Negative categories are combined and considered as Negative Review category.	<table border="1"> <tr> <td>Rating</td><td>10 - 5</td><td>4 - 1</td></tr> <tr> <td>Score</td><td>1</td><td>0</td></tr> </table> O-R: Overall Rating S-S: sentimentScore Sentiment 1 - Positive Sentiment 0 - Negative	Rating	10 - 5	4 - 1	Score	1	0
Rating	10 - 5	4 - 1						
Score	1	0						
Results	Results	Results						
								
Positive Reviews - 66.15% Negative Reviews - 33.14% Neutral Reviews – 0.70%	Positive Reviews - 70.67% Negative Reviews - 29.33%	Positive Reviews - 50% Negative Reviews - 50%						

6. Graphical Analysis

A. Plot the Total Number of Ratings (Counts and Percentages) of the IMDB Reviews



B. Simplified Ratings for Modeling

Ratings - Counts	Ratings - Percentages
<pre># Display value counts for Rating df_feedback.Rating.value_counts()</pre> <pre>1 30331 0 19669</pre>	<pre># Calculate percentages of each Rating of the IMDB Movie review df_feedback['Rating'].value_counts(normalize=True)*100</pre> <pre>1 60.662 0 39.338</pre>
 <p>A bar chart with 'Rating' on the x-axis and 'count' on the y-axis. The y-axis ranges from 0 to 30,000 with increments of 5,000. There are two bars: a blue bar for Rating 0 with a count of approximately 19,669, and a green bar for Rating 1 with a count of approximately 30,331.</p>	

7. Modeling

Model Selection

Ratings were further simplified for modeling by re-assigning the values as below

- Rating 4 and above = 1
- Rating 3 and below = 0

With about 60% of positive reviews, below types of models were selected to run:

1. Multinomial Naive Bayes
2. Random Forest
3. Linear SVC
4. Logistic Regression
5. XGB Classifier
6. Logistic Regression Model using class_weight
7. Neural Network Classifier with Keras

Model Evaluation – Results / Outcome

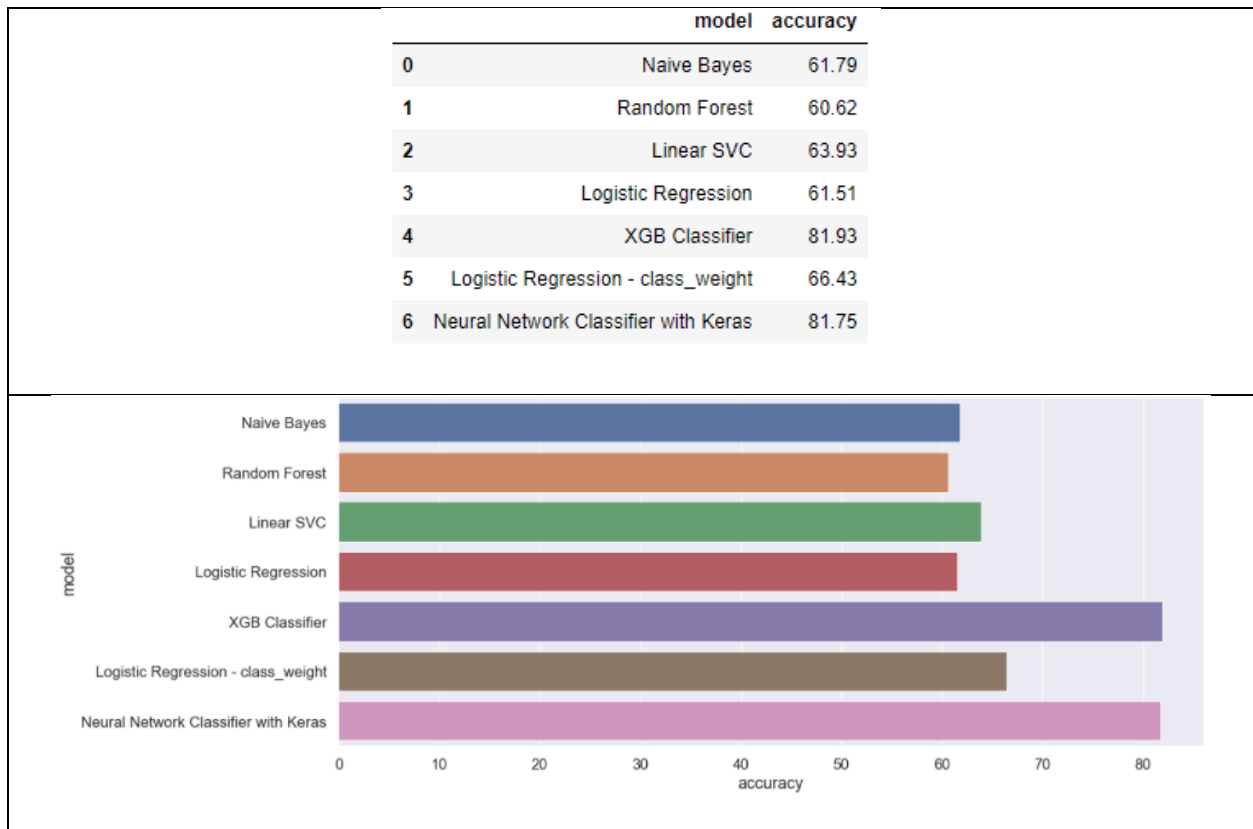
The accuracy of the XGB Classifier and Neural Network with Keras shows best results as below:

- XGB Classifier - 81.93%
- Neural Network Classifier with Keras - 81.75%

Performance Metric of each selected model

Model	Accuracy (%)	f1 – Score (Weighted avg)	Precision (Weighted avg)	Recall (Weighted avg)
Multinomial Naive Bayes	61.79	0.48	0.74	0.62
Random Forest	60.62	0.46	0.37	0.61
Linear SVC	63.93	0.54	0.70	0.64
Logistic Regression	61.51	0.48	0.76	0.62
XGB Classifier	81.93	0.82	0.82	0.82
Logistic Regression (cw)	66.43	0.61	0.68	0.66
Neural Network - Keras	81.75	0.82	0.82	0.82

Model Evaluation – Plot the Results



Confusion Matrix of each selected model

TFIDF Vectorizer resulted into below result:

```
# TFIDF Vectorizer

tv = TfidfVectorizer(min_df=0,max_df=1,use_idf=True,ngram_range=(1,2), stop_words='english')

tfidf_x_train = tv.fit_transform(x_train)
tfidf_x_test = tv.transform(x_test)

print('tfidf_x_train:',tfidf_x_train.shape)
print('tfidf_x_test:',tfidf_x_test.shape)

tfidf_x_train: (40000, 2048395)
tfidf_x_test: (10000, 2048395)
```

Multinomial Naive Bayes	Random Forest	Linear SVC	Neural Network - Keras																																				
 <table><caption>MultinomialNB Confusion Matrix</caption><tr><th></th><th>Predicted Class 0</th><th>Predicted Class 1</th></tr><tr><th>True Class 0</th><td>125</td><td>3813</td></tr><tr><th>True Class 1</th><td>8</td><td>6054</td></tr></table>		Predicted Class 0	Predicted Class 1	True Class 0	125	3813	True Class 1	8	6054	 <table><caption>RandomForestClassifier Confusion Matrix</caption><tr><th></th><th>Predicted Class 0</th><th>Predicted Class 1</th></tr><tr><th>True Class 0</th><td>0</td><td>3938</td></tr><tr><th>True Class 1</th><td>0</td><td>6062</td></tr></table>		Predicted Class 0	Predicted Class 1	True Class 0	0	3938	True Class 1	0	6062	 <table><caption>LinearSVC Confusion Matrix</caption><tr><th></th><th>Predicted Class 0</th><th>Predicted Class 1</th></tr><tr><th>True Class 0</th><td>431</td><td>3507</td></tr><tr><th>True Class 1</th><td>100</td><td>5962</td></tr></table>		Predicted Class 0	Predicted Class 1	True Class 0	431	3507	True Class 1	100	5962	 <table><tr><th></th><th>Predicted Class 0</th><th>Predicted Class 1</th></tr><tr><th>True Class 0</th><td>2993</td><td>945</td></tr><tr><th>True Class 1</th><td>880</td><td>5182</td></tr></table>		Predicted Class 0	Predicted Class 1	True Class 0	2993	945	True Class 1	880	5182
	Predicted Class 0	Predicted Class 1																																					
True Class 0	125	3813																																					
True Class 1	8	6054																																					
	Predicted Class 0	Predicted Class 1																																					
True Class 0	0	3938																																					
True Class 1	0	6062																																					
	Predicted Class 0	Predicted Class 1																																					
True Class 0	431	3507																																					
True Class 1	100	5962																																					
	Predicted Class 0	Predicted Class 1																																					
True Class 0	2993	945																																					
True Class 1	880	5182																																					
Logistic Regression	XGB Classifier	Logistic Regression (cw)																																					
 <table><caption>LogisticRegression Confusion Matrix</caption><tr><th></th><th>Predicted Class 0</th><th>Predicted Class 1</th></tr><tr><th>True Class 0</th><td>90</td><td>3848</td></tr><tr><th>True Class 1</th><td>1</td><td>6061</td></tr></table>		Predicted Class 0	Predicted Class 1	True Class 0	90	3848	True Class 1	1	6061	 <table><tr><th></th><th>Predicted Class 0</th><th>Predicted Class 1</th></tr><tr><th>True Class 0</th><td>2913</td><td>1025</td></tr><tr><th>True Class 1</th><td>782</td><td>5280</td></tr></table>		Predicted Class 0	Predicted Class 1	True Class 0	2913	1025	True Class 1	782	5280	 <table><caption>LogisticRegression (cw) Confusion Matrix</caption><tr><th></th><th>Predicted Class 0</th><th>Predicted Class 1</th></tr><tr><th>True Class 0</th><td>957</td><td>2981</td></tr><tr><th>True Class 1</th><td>376</td><td>5686</td></tr></table>		Predicted Class 0	Predicted Class 1	True Class 0	957	2981	True Class 1	376	5686										
	Predicted Class 0	Predicted Class 1																																					
True Class 0	90	3848																																					
True Class 1	1	6061																																					
	Predicted Class 0	Predicted Class 1																																					
True Class 0	2913	1025																																					
True Class 1	782	5280																																					
	Predicted Class 0	Predicted Class 1																																					
True Class 0	957	2981																																					
True Class 1	376	5686																																					

7. Assumptions

- Description of the IMDB Movie Review correctly align with the rating and the sentiment given by the reviewer
- In terms of the fairness, all Movie Reviewers have actually watched those movies before submitting the review and giving the rating.

8. Limitations

- The dataset with only 50000 reviews is comparatively smaller subset looking at the overall IMDB Movie Reviews that may be out there in the central database.
- Rating scale of 1 To 10 is too wide to perform the analysis.

9. Conclusion

Why Are Reviews Important?

IMDB Movie Review It is now owned and operated by IMDb.com, Inc., a subsidiary of Amazon. As of June 2021, the database contained some 8 million titles (including television episodes) and 10.4 million person records. Additionally, the site had 83 million registered users. As IMDB continues to see rapid growth, reviewers are increasingly making decisions with the help of one of the most powerful tools: IMDB reviews. Reviews aren't just beneficial for movie lovers trying to find the perfect match based on their interests. They're one of the most effective ways for you to boost your brand's conversion, credibility, and overall online presence. If you have very few reviews—or if the reviews, you do have are negative—you're less likely to convince movie watchers that your movie beats the competition. Reviews are everywhere, from Amazon to Facebook to Google to Yelp and beyond, and for good reason. Research shows that 84 percent of shoppers trust online reviews as much as a personal recommendation, and 91 percent of shoppers occasionally or regularly read online reviews. Why? Because they create trust and add transparency to the purchasing experience, so consumers are more willing to buy.

10. Challenges

- We observed that we had positive reviews almost the same as negative reviews, which makes little harder to train the model and predict accuracy. Additionally, the rating scale of 1 To 10 is too wide to perform the analysis.
- It was crucial to try various different models to better judge the reviews available into relatively smaller subset of the overall IMDB Movie Review database. Most of the tried models provided the results with relatively lower accuracy levels except XGB Classifier and Neural Network with Keras.

11. Implementation Plan

Below six steps were considered while implementing the effort end to end:

- Step 1: Define Problem Statement
- Step 2: Data Collection
- Step 3: Data Cleaning
- Step 4: Data Analysis and Exploration
- Step 5: Data Modelling
- Step 6: Optimization and Deployment

12. Ethical Considerations – Human Bias

Although online reviews text mining can often legally and reasonably proceed without formal ethics approvals, it's recommended to look into improving ethical standards when it comes to human biases. How many of the IMDB Movie Reviews in the available dataset are unbiased and with the right sentiments, especially when they were posted online as free texts. It would be interesting to analyze the movie reviews based on their ratings and the sentiments.

13. References

- **Dataset:** <https://www.kaggle.com/nisargchodavadiya/imdb-movie-reviews-with-ratings-50k>
- <https://towardsdatascience.com/benefits-and-ethical-challenges-in-data-science-compas-and-smart-meters-da549dacd7cd>
- https://help.imdb.com/article/imdb/discover-watch/what-to-watch-faq/GPZ2RSPB3CPVL86Z?ref=helpms_ih_gi_siteindex#
- <https://en.wikipedia.org/wiki/IMDb>