Melissa Results

Abstract: Paragraph synopsis of research

Introduction: Description of your project

Data: Description of the data used

Methodology: Description of technology used

Results: Summary of key findings

Conclusion: Restate what problem or question your project addresses

**Notes:**

**Getting the data:**

* Read in Hugging Face E-book reviews
* 1% sample due to the size of the data- about 50K reviews
* Labeled the reviews as being positive (4-5 stars), neutral (3 stars), negative (1-2 stars)

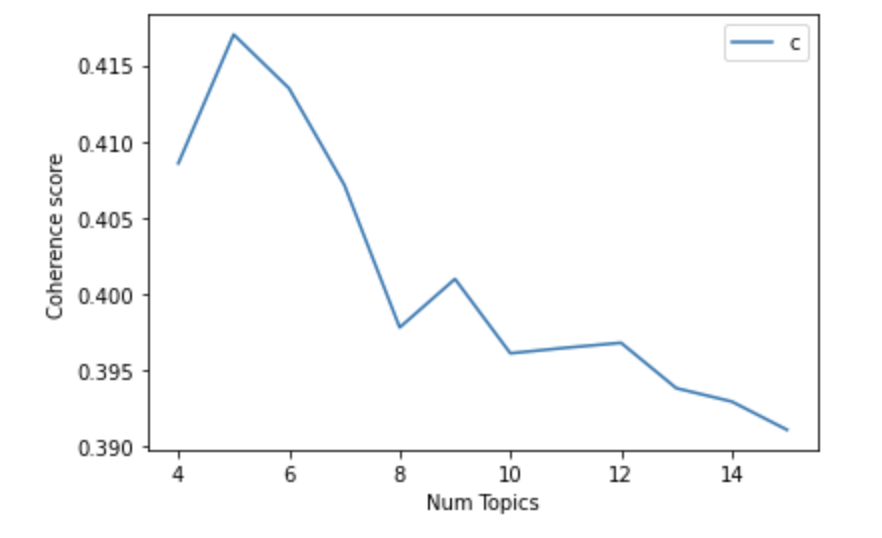
**Preprocessing the data:**

* Cleaning- removed anything not a letter or number
* Removed words less than 3 characters long
* Lemmatized words

**Part 1- LDA Topic Modeling**

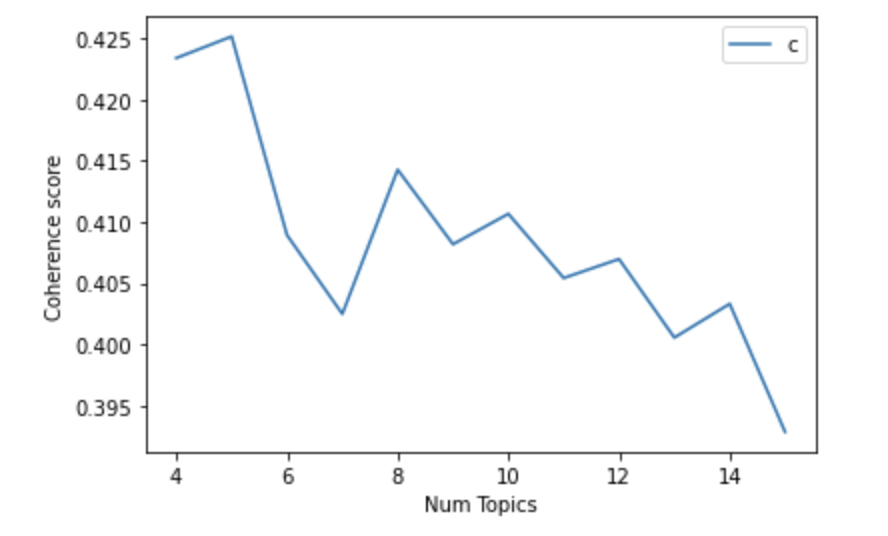
* Plotted coherence scores for 4-15 topics

1. ***Model 1***
   * Unigrams: no\_below = 500, no\_above = 0.75
   * Best = 5 topics
   * Coherence score = 0.41701197299887516



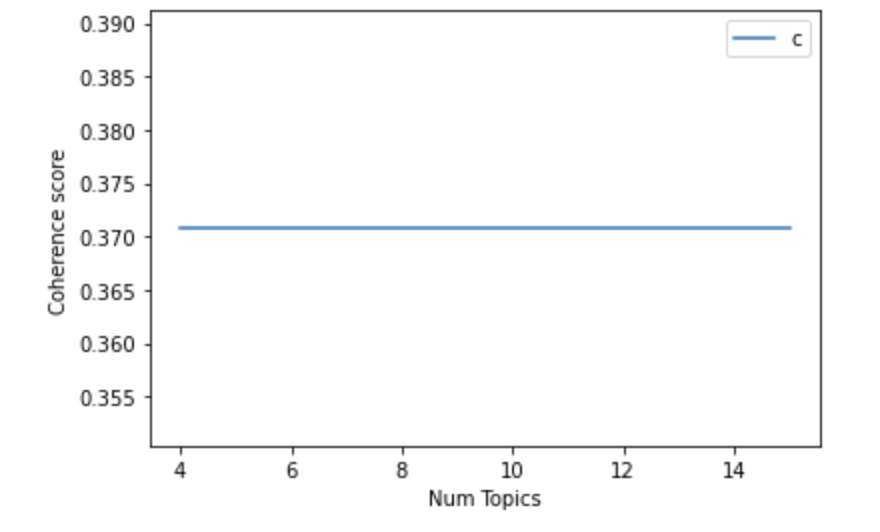
1. ***Model 2***

* no\_below = 500, no\_above = 0.50
* Best = 5 topics
* Coherence score = 0.4251618105898782



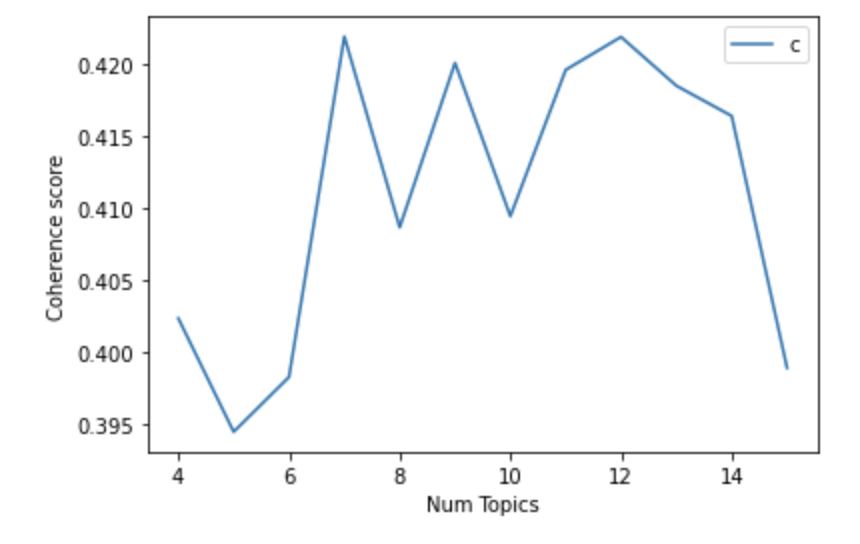
1. ***Model 3***

* no\_below = 5000, no\_above = 0.50
* Terrible- flat line for coherence score for each number of topics



1. ***Model 4***

* No\_above = 0.75
* Best = 7 topics
* Coherence score = 0.42195141623906346
* Selected this one because although the coherence score was slightly lower than model 2, the topics made more sense when considering the top words.
* Some of the top words associated with each topic include:
  + Topic 1: kindle, easy, idea, use, work
  + Topic 2: character, novel, plot, interesting, action, mystery
  + Topic 3: story, love, romance, heart, character
  + Topic 4: book, one, would, think, know, time
  + Topic 5: life, people, child, world, god, history, time
  + Topic 6: family, year, man, woman, friend, mother
  + Topic 7: book, series, one, next, loved, wait, recommend



**Part 2- Custom Embeddings Creation**

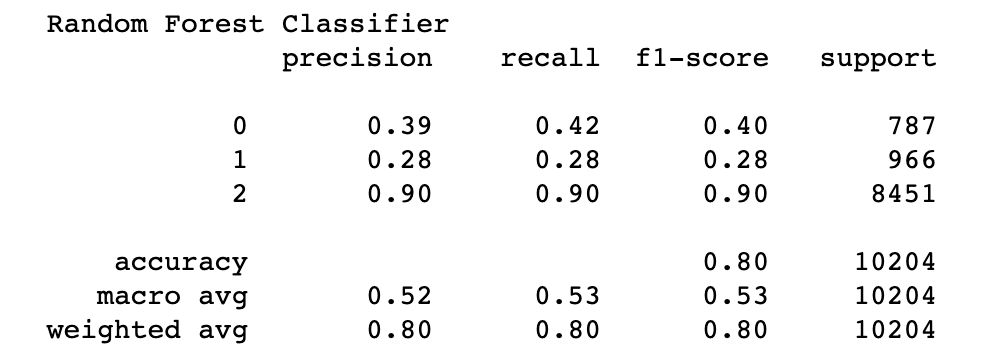
* Created 300-D custom embeddings based on our reviews
* Debated using Amazon-product reviews embeddings but was not sure if it would be specific enough to e-books, so decided to create our own
* Used 300D embeddings to collect lots of dimension on each review

**Part 3- Classification Modeling**

* Predict the review category based on the probability of each topic, and also the embeddings
* Created a dataframe that added the topic probabilities for every review from model 4, plus the 300D embeddings for every review, to the original reviews dataset.
* Used topic probabilities and the embeddings to predict the review rating category (positive, neutral, negative)
* Tried 3 different model architectures- random forest, KNN classification, and multilayer perceptron network (fully connected neural network).
* Ran grid search on every model to find the optimal parameters
* Got classification report for every model

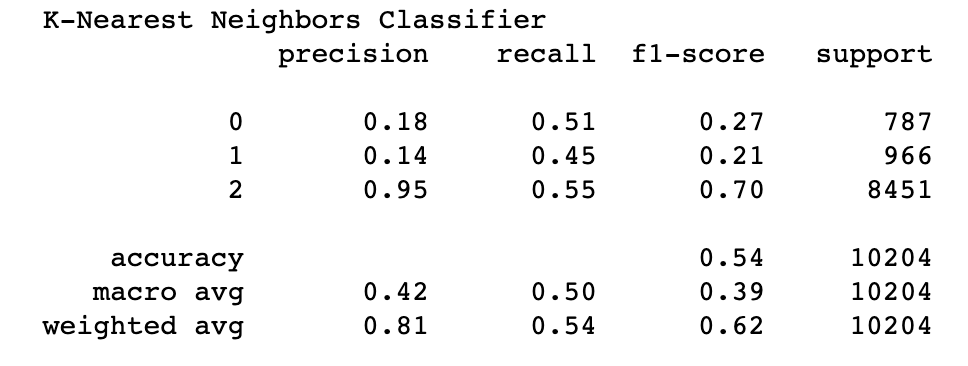
1. ***Random Forest***

* Best model training accuracy- 0.9401
* Pamraters- balanced class weight, max depth = None, 100 estimators



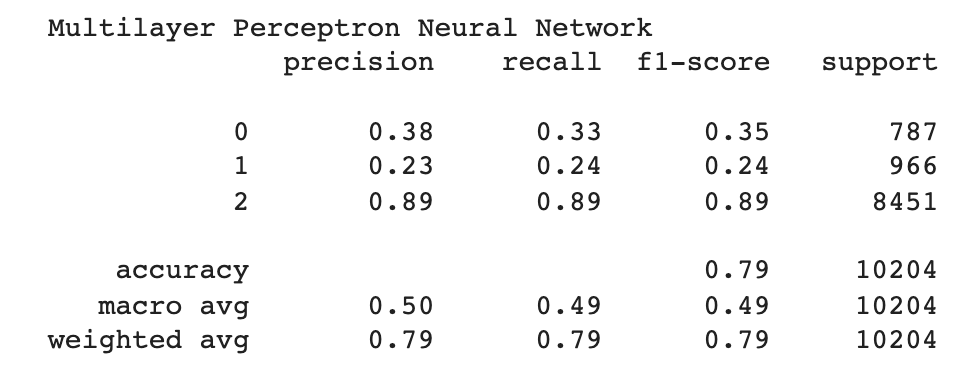
1. ***KNN Classification***

* Best model training accuracy- 0.8430
* 5 neighbors



1. ***Multilayer Perceptron Neural Network***

* Best model training accuracy- 0.9160
* Input nodes = 307 (topic probabilities and embeddings)
* 3 hidden layers, first with 500 hidden nodes, second with 500 hidden nodes, third with 100 nodes
* Relu activation



**Melissa Sections for Paper**

Research question:

Can topic probabilities and custom embeddings be used together to predict review rating category (positive, neutral, negative)?

Hypothesis:

The probability of each review containing each topic, along with embeddings to describe each review, will be strong predictors of review rating as evidenced by good model performance.

**Methodology**

**Topic Modeling Methodology**

Topic modeling is a type of unsupervised machine learning that can be used to detect recurrent themes in a body of text. This has applications in classifying text, finding similarities between texts, and in recommendation systems. An algorithm commonly used for this purpose is the Latent Dirichlet Allocation (LDA). Defined from the class 9 slides, LDA is ‘a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of latent topics. Each observed word originates from a topic that we do not directly observe. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities.’ This method can be thought of as similar to clustering, except instead of numerical features defining the clusters, the clusters are defined by tokens (words) [link 1]. The goal is to find the hidden, or latent, structure within the text. This latent structure in the text is defined by a distribution of topics and distribution of words in each topic. This distribution of distributions is known as a dirichlet. Once the dirichlet is discovered, the topics will be allocated to the documents, and the words of the documents will be allocated to the topics. LDA believes that each word in each document comes from a topic, and the topics are selected from a per-document distribution over topics. From the lecture slides and [link 2], the conditional probabilities determined by LDA are as below:

**p(topic t | document d) [link 2]**

* The proportion of words in a document *d* that are assigned to topic *t*
* Tries to capture how many words belong to the topic *t* for a given document *d*

**p(word w | topic t)**

* Proportion of assignments to topic *t* over all documents that come from this word *w*
* Tries to capture how many documents are in topic *t* because of word *w*

**p(word w with topic t) = p(topic t| document d) \* p(word w| topic t)**

* Probability a word *w* belongs to topic *t*

To evaluate the performance of LDA in defining the latent topics, the coherence of each topic can be calculated. For a single topic, the coherence score measures the degree of semantic similarity between high scoring words in the topic. The c\_v coherence of the model can also be calculated to gauge the overall quality of the model This helps to understand whether or not the topics are cohesive. In addition, the perplexity score can also be computed, which is another indicator of how good the entire LDA model is. To improve model performance, there are various parameters that can be tuned, including number of latent topics, the learning decay, as well as the alpha (document-topic density) and beta (topic-word density). High values of alpha result in documents being composed of more topics, while low values of topics result in documents being composed of fewer topics. High values of Beta result in topics composed of large numbers of words from the entire corpus, while low values of Beta result in topics composed of few numbers of words from the entire corpus[link 3].

* Class slides- <https://courseworks2.columbia.edu/courses/135369/files/folder/slides?preview=12617512>
* Link1- <https://medium.com/analytics-vidhya/topic-modeling-using-lda-and-gibbs-sampling-explained-49d49b3d1045>
* Link 2- <https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2>
* Link 3- <https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/>

**Embeddings Methodology**

From the class slides, Word embeddings are learned representations for text where words that have the same meaning have the same representation. Words are represented as a vector of real numbers, typically in dimensions of 50, 100 or 300. These vectors enable comparisons of word similarities, machine translation, named entity recognitions, relation extraction, sentiment analysis, and coreference resolution. The dimensions used to describe a text can be used to relate it to other texts in however many dimensions the embeddings occupy. Texts that have similar vectors are thought to be more similar than those with less similar vectors. A common way to create these embeddings is to use Word2Vec. Word2Vec is a two-layer neural network that takes a corpus as an input and outputs a vector representation of the corpus, according to the vector dimensions specified.

* Class slides- <https://courseworks2.columbia.edu/courses/135369/files/folder/slides?preview=12299521>

**Topic Modeling and Embeddings Classification Model Methodology**

Representing each review by the probability of containing each topic, along with the 300 dimensions of embeddings, provided an opportunity for multiclass classification through supervised machine learning. Each review assigned a star rating indicating the reviewer’s satisfaction, from 1 to 5, with 1 being least satisfied and 5 being most satisfied. Categorizing each review into a rating category of positive (4-5 star rating), neutral (3 star rating) or negative (1-2 star rating) simplified the task from a 5-category classification to a 3-category classification. These categories were also deemed more meaningful than five separate categories, where it can be difficult to determine the recurrent difference between ratings separated by one star. These three rating categories, positive, neutral and negative, were used as the dependent variable. The topic probabilities and embeddings used to describe each review were used as the independent variables.

Another factor to consider during the classification was the imbalance of the data. The data was heavily skewed in favor of positive reviews (nearly 34,000), with much smaller numbers of neutral and negative reviews (about 4,000 and 3,000 respectively). To avoid bias in the model, the training data was upsampled using SMOTE (Synthetic Minority Oversampling Technique) to provide the model with equal counts of each review category on which to train. This upsampling generated synthetic data of the neutral and negative review categories, making the number of reviews in each category in the training data equal, to just under 34,000 each.

Three types of classification algorithms were used, including Random Forest, K-Nearest Neighbors, and Multilayer Perceptron Networks. A brief description of each classification model’s methodology is below:

* Random Forest Classifier is an ensemble method that includes multiple decision trees, built on a bootstrapped sample of the training data. Each tree is composed of internal nodes, which are split, and leaf nodes, which are terminal nodes and do not split. Each time a split in one of these decision trees is made, a random subset of the total features are selected, with the split using one of those features in the subset. The split is greedy, meaning that the decision for each split is based on the potential increase in purity for that specific step, and not for future splits. This process repeats for all trees, with the goal of increasing the purity of samples in each node with each split. Some of the tuning parameters for Random Forest include the number of trees (called estimators) in the Forest, the maximum depth of each tree, the criteria to motivate or prevent a split (either minimum observations required to perform the split, or minimum observations permitted in a leaf node), maximum number of leaf nodes, and splitting criteria.
* Random forest slides- <https://courseworks2.columbia.edu/courses/135307/files/folder/Slides?preview=12214979>
* K-Nearest Neighbors Classifier is a simple algorithm where observations are classified based on the classifications of their most similar k neighbors. The similarity to other observations in the training data is calculated by a distance measurement, such as euclidean distance. Once the k nearest neighbors to an observation are determined, the majority vote of these neighbors is taken to assign the category to the observation in question. There is only one tuning parameter, that is, the number of k neighbors to consider when determining the majority vote.
* The Multilayer Perceptron Network Classifier is a type of neural network. These networks can learn non-linear relationships and discover other relationships that other models may miss. However, they require a large number of parameters, including the number of hidden layers, hidden nodes in the hidden layers, and iterations. These networks are composed of an input layer, which takes in data flattened into a vector, hidden layer(s), and an output layer. The nodes in the hidden layer(s) use nonlinear activation functions. These networks are fully connected, meaning that every node in a given layer is connected to all the nodes in the next layer. The strength of the connection between two nodes is called a weight, which is initially randomly assigned. The data is fed forward through the network, where the input values are multiplied by these initial weights and then summed. Each layer also includes biases which are constants, and adjusts the output along with the weighted sums of the nodes. A technique called backpropagation is then used to adjust these weights to improve model performance. Some of the tuning parameters for these networks include the number of hidden nodes in each hidden layer, the learning rate, the batch size of training data, and the number of epochs (number of times the algorithm will work through the entirety of the training data).

**Results**

**Topic Modeling Results**

As an alternate representation of the reviews, each product review can be represented by topic probabilities. After removing non alphanumeric characters, removing words comprised of fewer than three characters, removing custom stopwords, and performing lemmatization, Latent Dirichlet Allocation (LDA) was used to define the number of recurrent topics found within the body of the reviews. Multiple iterations of the LDA algorithm were run to test different possibilities. The LDA algorithm provides parameters for minimum and maximum token thresholds. These were used to ensure that only meaningful tokens, which appeared a reasonable number of times within all of the reviews, were used to define the topics. Different values for the token parameters were applied, including:

| Iteration | Tokenization | Maximum Token Threshold\* | Minimum Token Threshold\*\* | Ideal Topic Count\*\*\* | C\_v Coherence Score | C\_v Plot by topic |
| --- | --- | --- | --- | --- | --- | --- |
| Model 1 | Unigram | 500 | 0.75 | 5 | 0.4170 | Appendix |
| Model 2 | Unigram | 500 | 0.50 | 5 | 0.4252 |  |
| Model 3 | Unigram | 5000 | 0.50 | n/a | 0.3710 |  |
| Model 4 | Unigram | n/a | 0.75 | 7 | 0.4220 |  |

*\*Exclude tokens that appear in more than x reviews*

*\*Exclude tokens that appear in less than x% of total reviews*

*\*\*\*Topic counts of 4 through 15 were tested*

Although Model 2 produced the highest coherence score, which indicates similarity among words associated with each topic, Model 4 produced the most interpretable results. The topics produced by Model 2 were not as clearly differentiable as those produced by Model 4.

For Model 2, some of the top words associated with each topic include:

* Topic 1: series, love, character, next, wait
* Topic 2: time, know, think, make, going
* Topic 3: story, character, little, romance, part
* Topic 4: life, love, family, friend, real
* Topic 5: recommend, author, interesting, reader, easy

For Model 4, some of the top words associated with each topic include:

* Topic 1: kindle, easy, idea, use, work
* Topic 2: character, novel, plot, interesting, action, mystery
* Topic 3: story, love, romance, heart, character
* Topic 4: book, one, would, think, know, time
* Topic 5: life, people, child, world, god, history, time
* Topic 6: family, year, man, woman, friend, mother
* Topic 7: book, series, one, next, loved, wait, recommend

From the associated words, it might be reasonable to infer that from the topics produced by Model 4, the first might be related to (Kindle) e-books and their ease of use, the second might be related to the plot of action or mystery books, the third might be related to romance stories, the fifth might be related to philosophy, the sixth might be related to family, and the seventh might be related to series of books (as opposed to single book releases). It is not immediately clear what is being described in topic 4.

Compared to the topics produced by Model 2, those from Model 4 are clearer. Topic 1 in Model 2 might be related to series of books, topic 3 might be related to romance stories, and topic 5 might be related to books that reviewers would recommend to others. It is not immediately clear what is being described in topics 2 and 4. This further emphasized that in machine learning, it is important to consider the interpretability of models in addition to their objective results. Objectively, one could say that Model 2 had a slightly better performance than Model 4, based on coherence score. However, interpretability of the topics produced by Model 4 seems clearer. This same intuition guided the number of topics tested in the models, where anything less than 4 might not capture topic variability and anything greater than 15 might be too specific to certain reviews.

**Embeddings Results**

Embeddings were intended to serve as multidimensional representations of each review. To create the embeddings, there are multiple parameters that can be customized, including the minimum count in the corpus per token, the embeddings vector size and neighboring word window size. For the embeddings created here, the minimum count in the corpus for each token was 1, the embeddings vector size was 300, and the neighboring word window size was 5. Together, these parameters intended to extract information about each review, and how it relates to the other reviews, in 300 dimensions. Using pre-trained embeddings related to product reviews were considered, but there was concern that these might not be specific enough to either e-books or even Amazon purchases. For this reason, it was determined that training custom embeddings would provide the most relevant information.

**Topic Modeling and Embeddings Classification Model Results**

Using the dataset structure with the topic probabilities and the embeddings dimensions as independent variables and the review rating categories as the dependent variable, three different classification modeling approaches were tried. Random Forest, K-Nearest Neighbors, and Multilayer Perceptron Networks were tuned, trained and cross-validated. The optimal models found from GridSearchCV were also applied to the test sets to gauge their performances.

Below are the training set accuracy, test set accuracy, and test set weighted f1 scores of each of the three models. The optimal parameters for each, discovered during GridSearchCV, are also noted.

| Model | Optimal Model Parameters | Training Set Accuracy | Test Set Accuracy | Test Set Weighted F1 Score |
| --- | --- | --- | --- | --- |
| Random Forest | 100 estimators, no max tree depth | 0.9401 | 0.8012 | 0.8023 |
| K-Nearest Neighbors | 5 neighbors | 0.8430 | 0.5387 | 0.6184 |
| Multilayer Perceptron Network | 3 hidden layers- hidden nodes per layer (500, 500, 100) | 0.9160 | 0.7876 | 0.7864 |

From the results above, the Random Forest had the strongest performance among all of the models. The Multilayer Perceptron Network is the second best performing model, not too much of a difference from the Random Forest. The K-Nearest Neighbors model performed the worst out of the three models tested. The decline in accuracy from training set to test set in all of the models is indicative of overfitting to the training data. This occurs when a model learns the training set too well and fails to generalize to new and unseen data, such as that of the test set. It seems that the K-Nearest Neighbors model may have the worst degree of overfitting due to the large drop from training set accuracy to test set accuracy. The weighted f1 scores for the Random Forest and Multilayer Perceptron Network are fairly similar, indicating that these models both have adequate balances of precision (true positives/actual results) and recall (true positives/model predictions) when exposed to test data. Although the class disparity in the training data was resampled using SMOTE, the high weighted f1 sores for these models indicate that these models are able to recognize instances of the different review rating clases in the test set, which contains mostly positive review ratings.

On the contrary, the K-Nearest Neighbors model displays the worst test set performance in regards to both accuracy and weighted f1 score.

Due to these results, to predict the review rating category of new reviews, it would be best to use the Random Forest model. This model performs well in terms of accuracy and does a good job of recognizing instances of each class, despite their differences in frequency. To use this Random Forest model, new review data would have to first be processed in the same manner that the data used to train the model was- by considering the topic probabilities of each of the seven topics found using LDA in each new review, and transforming the text of the new reviews into embeddings of the same dimensions.