**Final Report**

Cameron Gallien

DSE6212OM\_DSE6311OM\_SU2023S5\_Text and Image Mining

Merrimack University

Prof. Peter Salemi

August 23, 2023

**Executive Summary**

This paper outlines a project aimed at developing an accurate sentiment analysis model tailored to the needs of the ABC Company's website. The company is adding a movie discussion and review section to its site and seeks an algorithm that assigns sentiment scores on a scale of 0 to 1 to enhance user experience. These sentiment scores will categorize reviews from "very negative" (close to 0) to "very positive" (close to 1), enabling users to sort and prioritize reviews based on their sentiment.

To achieve this, a dataset of 50,000 labeled movie reviews was provided by the ABC Company. The goal was to create a predictive model capable of distinguishing positive from negative reviews. The dataset was meticulously prepared, including steps to clean text data, split reviews into training and testing sets, and employ techniques like tokenization, stopword removal, stemming, and TF-IDF matrix creation.

Four different models were tested: Naïve Bayes, Lasso classification, Long Short-Term Memory (LSTM) network, and Convolutional Neural Network (CNN). The Naïve Bayes model underwent refinement using trigrams, resulting in a 57.1% accuracy and an improved area under the ROC curve of 0.85. The Lasso classification model achieved an accuracy of 86.2% post hyperparameter tuning, with an impressive AUC of 0.92. The LSTM model yielded an accuracy of 85.1%, and the CNN model displayed 84% accuracy, though with little overfitting.

Considering the unique business need, the Lasso classification model emerged as the most suitable choice. Its high accuracy and AUC values of 86.2% and 0.92 respectively, make it a reliable tool for the ABC Company's website. The model's ability to generate sentiment scores ranging from 0 to 1 aligns perfectly with the company's requirement. Users will be able to sort reviews based on their sentiment scores, allowing them to access critical or positive opinions as needed to aid in movie-related decisions.

However, considering the highest accuracy achieved was 86.2%, it is suggested that future efforts be directed toward developing even more accurate models. This could involve collaboration with other companies or using larger review datasets for training. Despite the success of the recommended model, there is room for improvement, particularly to enhance user experience and decision-making on the website, as an accuracy rate of 86.2% might still leave room for misclassification.

**Approach & Data**

I approached this project with one goal, and that was to create the most accurate model possible for predicting movie review sentiment scores. To aid in creating this model, the ABC Company provided 50,000 movie reviews that were assigned a sentiment of “positive” or “negative.” The sentiment of each review was used to teach the model which words included in movie reviews generally correlate to a specific sentiment. These reviews would be utilized to train the models on what to look for in both positive and negative movie reviews.

The reviews were split into training and testing sets, so that we would have a separate set of reviews that the models have not been trained on. This test set would be used to simulate how the model would perform on future movie reviews uploaded to the site. From there, the models were analyzed to determine which model produced the most accurate results.

Before any models were created, we needed to start with the basic data processing steps that were laid out in the analytic plan and preliminary results reports. This consisted of first cleaning the movie review data to eliminate punctuation, unnecessary characters, and numbers that would not provide much value to our movie review model. From there, the data was separated into the training and testing sets that were mentioned above. The subsets that were chosen were 37,186 (75%) reviews for training, and 12,396 (25%) for testing.

From there, further data processing steps were taken. A tokenization step was utilized to separate each word into an individual token that the model could then analyze. This tokenization was done at the word level to capture the meaning of individual words, and to make the model more computationally efficient. Tokens were originally limited to 500 per review, as there were not many reviews with that many words, however after doing some quick tests of the models utilized, this was later increased to 1000. In addition, varying lengths of n-grams were experimented with on the models used in the preliminary analysis.

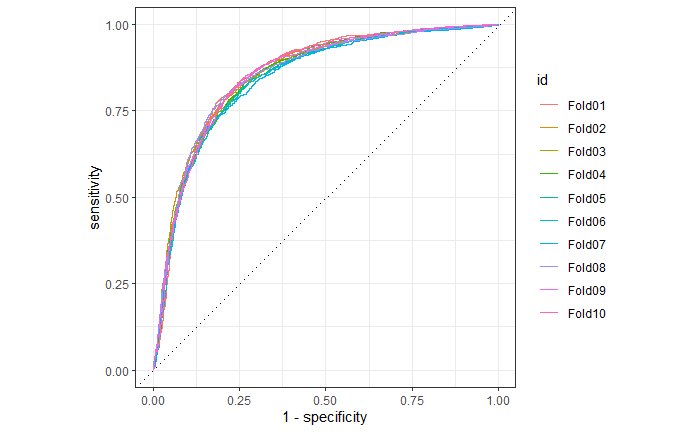
I also removed stopwords from the movie reviews. Stopwords consist of common words that contain very little informational meaning (Hvitfeldt & Silge, 2022). Examples include “a,” “an,” “of,” and “the.” These words are used regardless of a movie review’s sentiment and will be deleted from the data as they add no substantial value to our analysis. Different stopword lists were experimented with during my final analysis, and through this I determined that the “Snowball” pre-made stopword list produced the greatest results among those that were tested. This was the same stopword list that was used in my preliminary analysis, so this did not have any impact on the updated versions of those models.

Stemming techniques were then implemented in order to reduce words to their base or “stem” which would allow different variations of the same word to be considered as the same. Finally, all of these words were placed in a TF-IDF matrix, which would store the text data numerically, and take into account the frequency of each word. This allows the model to highlight terms that are important within the reviews, and downplay the common terms that are used frequently.

From there it was time to decide which models to experiment with. I once again utilized the Naïve Bayes and Lasso classification models from my preliminary analyses, with some slight tweaks made to attempt to achieve a better model. In addition, a Long Short-Term Memory (LSTM) network and a Convolutional Neural Network (CNN) were also utilized, with varying degrees of success. All four of these models were compared to find the most accurate model possible for implementation on the ABC Company’s new website.

**Findings & Model Evaluation**

The first model chosen was the Naïve Bayes model. The naïve bayes model was chosen for its ability to properly analyze a large number of features (Hvitfeld & Silge, 2022). In my preliminary analyses this was by far the least accurate model I was able to produce. With that being said, I experimented with bigrams and trigrams to see if they could improve upon the original results. Trigrams ended up giving us the best results, however they were still not optimal. By using trigrams instead of unigrams, I was able to increase the model’s accuracy from 52.7% to 57.1%, and there was a slight improvement with the ROC curve as well:



The area under the ROC curve was improved from .833 to .85, which showed further evidence that the model had improved. That being said, a mean accuracy of 57.1% is still not high enough to warrant being considered a reliable tool for movie viewers.

The next model utilized was the Lasso classification model. The Lasso classification model was chosen because of its ability to penalize features, a value that can be customized (Hvitfeld & Silge, 2022). The ability to penalize features was explored in my preliminary analysis, where it was determined that the optimal penalty for the model was 0.0007880463.

This model with the customized penalty produced the best results of those utilized in my preliminary analyses, with an average accuracy of 86.2%. The ROC curve also showed considerable improvement from the Naïve Bayes model, as the area under the curve was .92.

A graph with a line

Description automatically generated

When utilizing bigrams and trigrams on this model the accuracy decreased, unlike the Naïve Bayes model. The trigram variation of this model produced an accuracy of only 64.8%, and the area under the ROC curve was .71. Since I was not able to produce any greater results for this model, the original from the preliminary results was kept as is for evaluation among the final models.

Two additional deep learning models were created for this project as well. The first was an LSTM model. This specific model was chosen because according to Hvitfeldt & Silge (2022), LSTMs are “capable of learning long-range dependencies and broader context.” The LSTM model consisted of an embedding layer, a pre-made LSTM layer, and a dense layer which would create the output of the model. The original LSTM model that I created had an overall test accuracy of 85.1%. While many deep learning models have problems with overfitting, that was not the case with this model as the history plot shows below:

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generated

The accuracy generally increases in both the validation and training sets as the epoch increases. This is a good sign for the overall effectiveness of the model and proves that it is working as expected. I was a little underwhelmed with the results, so I also attempted to utilize weight freezing to see if that would provide any better results. For this exercise I used the GloVe word embeddings, which encapsulates semantic relationships between words to aid in the model’s understanding of each word. This did not provide better results however, as the accuracy of this model was only 69.1%, a big drop off from the original LSTM model. This could explained by a difference between the context that the word embeddings were derived from compared to the context of the available movie reviews.

Since the LSTM model was not able to out-perform the Lasso model, I wanted to try another. The last model that was utilized was a Convolutional Neural Network. This model was chosen because one of its key features is being able to capture specific local patterns in text data (Hvitfedlt & Silge, 2022). The architecture of this network consists of an embedding layer, a one-dimensional convolution layer, a one-dimensional max pooling layer, a dense hidden layer, and a dense output layer. The history plot that this model created originally made me very optimistic for this model:

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generated

Unfortunately, while this model produces a history plot showing no overfitting, the overall accuracy results were slightly underwhelming. This model produced only an 84% accuracy, still lower than the previous model, and the Lasso classification model that were previously utilized.

**Recommendations**

After reviewing all of the created models, it is evident by the results of the model testing that the Lasso classification model from the preliminary report produced the greatest accuracy out of all models shown. With a test accuracy of 86.2%, and an area under the ROC curve of .92, there is more than enough evidence to show that this is the most successful algorithm to be used on the ABC Company website.

While most of the analysis thus far has been on the accuracy of predicting positive or negative, this prediction is based on an overall probability of each review falling into one of the two categories. This probability is what would be used for our sentiment score, as the probabilities range from 0 to 1 just as the requested sentiment scores would. Seeing as a very high probability of a positive review would most likely point to a very positive review, this algorithm can be seamlessly integrated onto the company website to produce the sentiment score.

However, with our best model only having an overall success rate of 86.2%, it is recommended that efforts be directed toward refining the model even further. Collaborations with external entities or expanding the training dataset could potentially yield more accurate sentiment prediction, reinforcing the website's usefulness in aiding user decision-making. While I stand by the Lasso classification model as a suitable solution for the ABC Company’s need, there may still be room for improvement to make the website the go-to place for movie reviews.

**References**

Hvitfeldt, E., & Silge, J. (2022). Supervised machine learning for text analysis in R. CRC Press, Taylor & Francis Group. Retrieved Aug 4, 2023, from https://smltar.com/.