Sentiment Analysis for Reviews

Final Project

Aleah Bobb

Professor Peter Salemi

25 August 2023

Table of Contents

Executive Summary

The ABC Company has provided a dataset in hopes of adding a page to their website where potential users can discuss movies and post their reviews. The company aims to enhance their website by allowing users to discuss and review movies. Their objective is to create an algorithm that provides sentiment scores from 0 to 1. Users can then choose to view critical or positive reviews based on the sentiment scores. A sentiment analysis algorithm is developed to assign sentiment scores from 0 to 1 to movie reviews, which can in turn enable reviewers to sort and filter reviews and make informed decisions. The dataset in question contains 50,000 movie reviews from the users, where each review has been labeled as a 'positive' or 'negative' sentiment.

Machine learning techniques are utilized to create the necessary algorithm for ABC Company's business goal. Two machine learning models are compared and contrasted in order to achieve the best algorithm for the sentiment analysis. Preprocessing steps are first conducted in order for the models to run smoothly. The preprocessing steps ensures the sentiment values are converted to readable categories, which are positive and negative. It also ensures that any empty and duplicate reviews, URLs, non-alphanumeric characters, and digits are removed. Tokenization, removing stop words, handling negations and stem words, and applying TF-IDF are all necessary for the machine learning models to learn the provided dataset. One model requires normalizing the predictor columns while the other requires one-hot encoding.

The two machine learning models being compared and contrasted are the Naive Bayes and Convolutional Neural Network models. The Naive Bayes is chosen for this task since the model is known to be efficient for text classification and handling sparse data. It assumes conditional independence of features given sentiment. The Convolutional Neural Network model is the second model chosen for the task. This model learns hierarchical features from any text data and it is great for capturing local and global dependencies in language. It also automates feature learning, which in turn reduces manual feature engineering. Furthermore, it provides effective regularization, which avoids overfitting the text data.

After running the two machine learning models, one gains insights in each model's performance. The Naive Bayes model has a low accuracy and receiver operating characteristic, a metric used to evaluate the performance of a binary classification model, of .500. This suggests the model's poor performance since the model's predictions are not significantly better than random guessing. On the other hand, the Convolutional Neural Network model has a higher accuracy and receiver operating characteristic of 0.762. This indicates reasonable accurate predictions and well-calibrated predicted outcomes, which demonstrates reliable outcomes for the sentiment analysis.

While the Naive Bayes model is a simple choice for an sentiment analysis algorithm, it has limits to capturing complex relationships and nuanced contexts. The Convolutional Neural Network model is a better choice for ABC Company's business goal. The data provided is a large data set, which are great for running CNN models. Convolutional Neural Network models

capture hierarchical patterns and both local and global dependencies within text data. This model can achieve high accuracy with the proper tuning. This model is recommended for the company since it is better suited for capturing nuanced relationships within the text data.

With the Convolutional Neural Network model, the company can integrate this model into workflow for real-time or batch sentiment analysis. The algorithm can input reviews for automated sentiment predictions and provide real-time predictions, analyze trends, and insights. Input reviews for automated sentiment predictions. The company can then create visualizations of sentiment trends and popular movies. The company can utilize their algorithm to enhance their customers' decision-making and tailor business strategies based on insights. ABC Company can leverage the Convolutional Neural Network model to enhance customer insights, customers' informed decisions, and tailored strategies. ABC Company can use the model to have access to real-time sentiment analysis and trends, which improves user experiences and future business decisions.

Approach & Data

The Overall Approach

The ABC Company is adding a page to their website where potential users can discuss movies and post their reviews. In addition to movie reviews, the company would like to have an algorithm that automatically provides a sentiment score on a scale from 0 to 1. A sentiment score close to "0" would represent a "very negative" sentiment and a score close to "1" would represent a "very positive" sentiment. With this capability, the users can sort the reviews based on the sentiment score. A user may want to see very critical reviews in order to decide whether or not to rent or purchase a movie. On the flip side, a user may want to see very positive reviews.

The objective is to create an algorithm of a sentiment analysis is to assign sentiment scores, the supervised classification problem, to movie reviews on a scale from 0 to 1. The scale indicates the level of positivity or negativity. The sentiment score will enable users to sort and filter positive and negative reviews. The capability to sort through reviews this way will help users make informed decisions about whether or not to purchase or rent a movie.

The Available Data

The ABC Company has provided a dataset containing 50,000 movie reviews from the users. Each review has been labeled as a 'positive' or 'negative' sentiment.

The Feature Engineering Steps

In order to create an algorithm of the sentiment analysis, feature engineering steps needs to be the first step. After importing the given data, the sentiment values in the sentiment column are converted to readable categories by mapping the 1s to "positive" and 0s to "negative" using

indexing. Moreover, the column is converted to a factor to treat it as a categorical variable for the Naive Bayes model. For the Convolution Neural Network model, the indexing of the sentiment column is the same. All rows with empty review content, URLs in the review content, and duplicate rows are removed. Non-alphanumeric characters and digits are also removed. Moreover, all processed text are converted to lowercase.

The data is then split into training and testing sets. The data is tokenized and the stop words are removed. Tokenizing the text into individual words or "tokens" allows the algorithm to comprehend and analyze the text at a granular level. Stop words such as "the", "is", "and" will be removed since these common words do not carry significance to the sentiment information. Since negations can significantly impact this analysis, they need to be handled to capture appropriate subsequent words such as "not good", "very good", etc. Afterwards, stemming is applied in order to reduce words to their root form. This step captures word variations so words like "movies" "movie" will not be considered two different words. Another preprocessing step that is also applied to the data is the TF-IDF or Term Frequency-Inverse Document Frequency, which assigns weights to words based on their significance in a review. Words that occur more frequently in a specific review but less frequently in the overall corpus will receive higher weights. All predictor columns are normalized. Word embeddings is another technique that represent words as dense numerical vectors that capture semantic relationships, which is later applied to the data during the Convolution Neural Network modeling. Similar preprocessing steps are followed for the Convolution Neural Network model, such as tokenization and calculating the one-hot encoding sequence.

These steps help transform the raw review text data into a format suitable for modeling, where the text is tokenized, preprocessed, and transformed into numerical features through TF-IDF. These steps are necessary since the goal is to prepare the data in a way that captures the most relevant information from the text.

The Machine Learning Methods

The first machine learning method for the sentiment analysis is the Naive Bayes model. This model is very efficient for text classification tasks, which is great for a large data set like the dataset ABC Company has provided. Naive Bayes assumes that the features or words are conditionally independent given the class label, which is the sentiment. This model is also great for handling very sparse text data due to its probabilistic nature.

The Convolutional Neural Network model is the second machine learning method for the sentiment analysis. CNNs can learn hierarchical features. This includes low level features, like words, and high level features such as phrases and sentence structures. The CCN aids in capturing the hierarchical nature of language. CNN can automatically learn automatically learn relevant features from sentiment data, which in turn can reduce the necessity for manual feature engineering. Furthermore, CNNs are great for regularization and avoiding overfitting the data. This ensures the model generalizes well to the data.

Detailed Findings & Evaluation

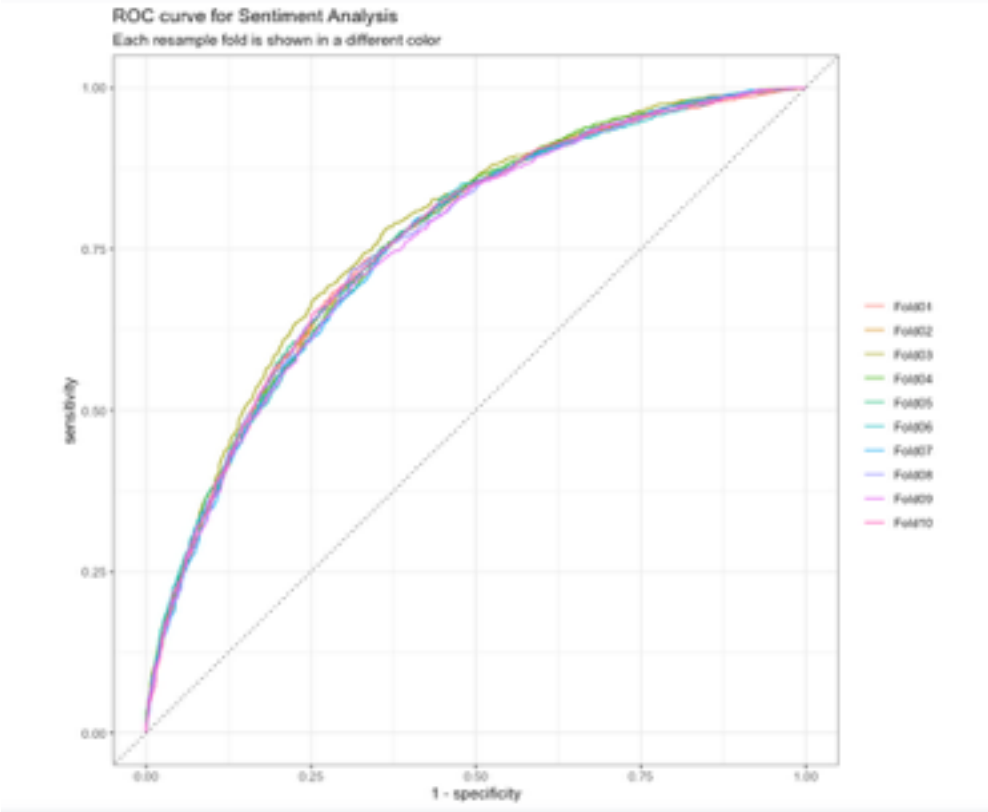
Naive Bayes Model

Based on Figure 1, both accuracy and roc_auc values are very close to 0.5, which suggests that the Naive Bayes model is not performing much better than random guessing. An accuracy of 0.5 indicates that the model's predictions are essentially random, and the roc_auc value of 0.5 indicates that the model's ability to discriminate between the classes is not strong. Figure 2 represents the ROC curve for Naive Bayes model, with each resample fold shown in different colors.

Figure One: Naive Bayes Model's Metrics

.metric	.estimator	mean	n	std_err	.config
<chr>	<chr>	<dbl>	<int>	<dbl>	<fctr>
accuracy	binary	0.5001600	10	0.0032956567	Preprocessor1_Model1
roc_auc	binary	0.5003424	10	0.0001107214	Preprocessor1_Model1

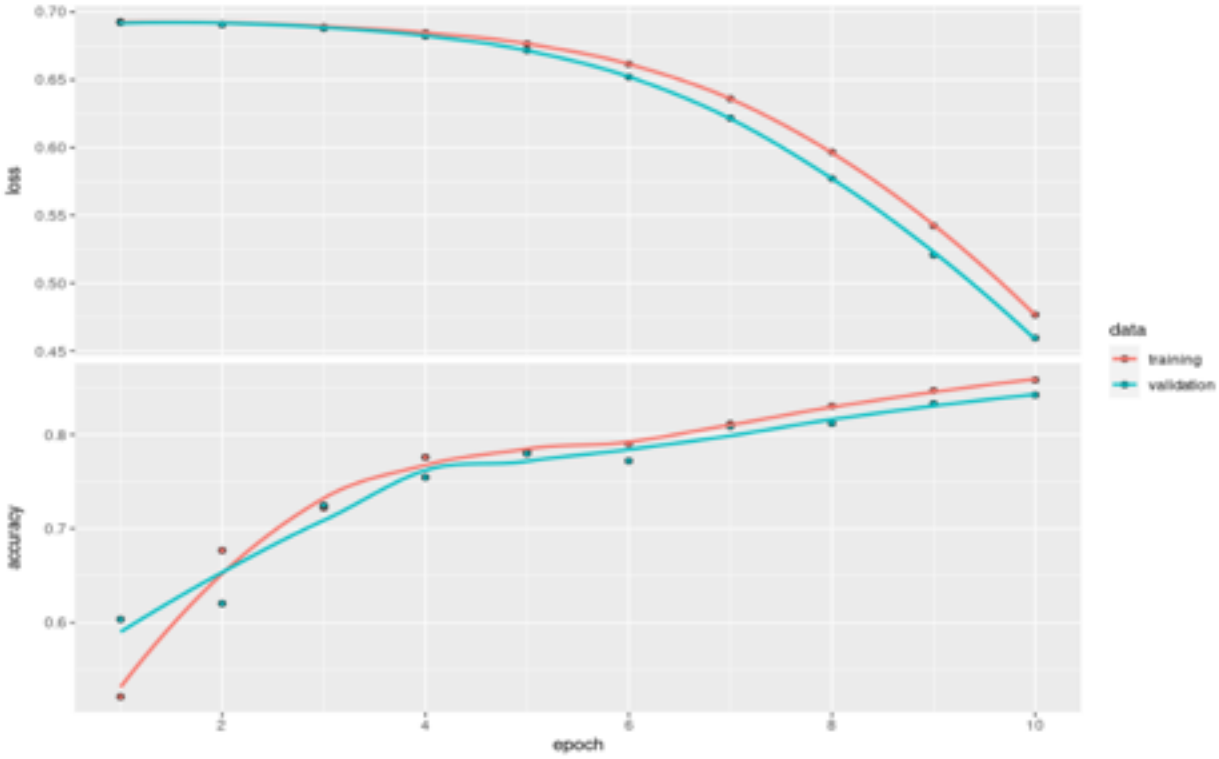
Figure 2: ROC Curve for Sentiment Analysis



Convolution Neural Network (CNN)

The average accuracy value of CNN model is 0.7615, which indicates the model correctly classifies approximately 76.15% of the data points in the validation set. In other words, the model's predictions are accurate for about 76.15% of the cases when compared to the actual ground truth values in the validation set. Figure 3 demonstrates that the model's predicted probabilities are well-calibrated, which suggests they accurately reflect the true probabilities of positive outcomes. In conclusion, the Convolutional Neural Network model has a higher accuracy rate than the Naive Bayes model.

Figure 3: The CNN's Calibration Plot



Recommendations

The Naive Bayes is a simple and interpretable algorithm, making it easy to understand the classification process. It is efficient and can handle a large number of features or words. It works well with text data and can capture simple relationships between words and sentiments. However, the model may not capture complex relationships in the data and could struggle with nuanced contexts. As shown in Figure 1, the performance suffered since it demonstrated to not be the best model for ABC Company's data set.

Convolutional Neural Networks (CNNs) are capable of learning hierarchical patterns from data, which can be beneficial for analyzing text sequences. It can also capture local and

global dependencies in the text, allowing them to learn more complex relationships. With the appropriate hyperparameters and training, CNNs have the potential to achieve high accuracy on sentiment analysis tasks. The data provided is a very large data set, where CNNs require a larger amount of data to effectively train. On the other hand, CNNs can be prone to overfitting if not properly regularized. They also may require more effort for model tuning and interpretation.

The Convolutional Neural Network model is recommended for ABC Company's business goal, which is to achieve higher accuracy in sentiment prediction and capture more nuanced relationships between words and sentiments. The Convolutional Neural Network model should be integrated into the business workflow to provide real-time or batch sentiment analysis on any incoming data. The model can better assist in understanding customer reviews, identifying trends, and making informed business decisions.

The model is designed to analyze text sequences, making it well-suited for sentiment analysis. By inputting customer reviews into the model, it can automatically predict whether the sentiment of the review is positive or negative. Since it has a higher accuracy than the naive bayes model, it is better suited for more accurate insights about customer sentiment and trends. The model can be integrated into the company's software, allowing for automated sentiment analysis of new customer reviews as they come in. Moreover, it can also provide real-time predictions of the customers' sentiments. The model can be utilized to aggregate these predictions to generate insights on sentiment trends, popular movies, and areas of concern. Furthermore, the company can be able to then create visualizations and reports that showcase sentiment trends and insights. ABC Company can then present the data in an accessible format that aids decision-makers in understanding customer sentiment at a glance. By leveraging the Convolutional Neural Network model for sentiment analysis, the ABC Company can gain deeper insights into customer sentiment, make informed decisions, and tailor their strategies to better meet customer needs and expectations.

References

Hvitfeldt, E., and Silge, J. *Supervised Machine Learning for Text Analysis in R*. (First Edition). CRC Press.