1 Abstract

This study evaluated the performance of logistic regression and multiclass regression against Decision Trees in the classification of textual data. Utilizing the IMDB Reviews data set and the 20 News Groups data set, we implemented logistic regression and multiclass regression models, respectively, and compared their performance with Decision Trees after conducting data preprocessing. The logistic regression model outperformed Decision Trees on the IMDB data with a 0.94 AUROC compared to 0.74, while multiclass regression achieved 73% accuracy compared to XX%, demonstrating its superiority in cases of linear relationships and less complex data structures, consistent with findings from related literature.

2 Introduction

The assignment explores different methods to classify textual data, first by implementing logistic regression and multiclass regression, and then evaluating the models performance against Decision Trees. Each regression model used a different data set. Logistic regression was implemented using the IMDB Reviews data set provided by Stanford University. Multiclass regression was implemented using the 20 News Groups data set provided by Scikit Learn. Both data sets required various data preprocessing methods such as simple linear regression and mutual information to filter out irrelevant words. The logistic regression implementation had a .94 AUROC whereas Decision Trees had a .74 AUROC when applied to the IMDB movie data. The multiclass regression implementation had a 73% accuracy whereas Decision Trees had a XX% accuracy. This aligns with the findings presented in Wilhelm Grzesiak's paper "Classification of Daily Body Weight Gains in Beef Calves Using Decision Trees, Artificial Neural Networks, and Logistic Regression" where he found Logistic Regression to perform better than Decision Trees in situations with linear relationships and less complex data structures. This is particularly relevant in our case, as the IMDB Reviews and 20 News Groups data sets, being textual, often exhibit linear associations between words and sentiments or topics. Furthermore, logistic regression's strength in handling large datasets and providing probabilistic outputs makes it a suitable choice for text classification, as demonstrated by its higher AUROC scores compared to Decision Trees in our study.

3 Datasets

The IMDB Reviews dataset contains movie review documents with their ratings. Each document in the labeledBowTest.feat dataset begins with a rating out of 10, followed by a dictionary where each key is a word index corresponding to a row number in the imbd.vocab dataset, and the values indicate the word's frequency in the document. We processed the data by mapping features (words) to their indices and then determining each feature's document frequency to filter out rare (words in less than 1% of the documents) and common words (words in more than 50% of the documents). This filtering reduced the initial 89,527 features to 1,744 relevant ones, enabling our model to concentrate on the most informative words for analysis.

The 20 News Groups dataset consists of about 20,000 documents from 20 different newsgroups, each labelled according to its respective newsgroup, forming a multiclass classification problem. The 'target' array indicates the category label for each document, while 'target_names' provides the names of the newsgroups corresponding to these labels. We refined the dataset by selecting five categories: 'sci.med', 'rec.motorcycles', 'rec.sport.baseball', 'sci.space', and 'talk.politics.mideast', and removed headers, footers, and quotes to streamlines relevant data. Textual data was converted into numerical format using a Count Vectorizer, which excluded common English stop words and filtered out rare and common terms. We then used Mutual Information to identify the top 150 most relevant words for the document classification task. This feature selection process resulted in a more concise and targeted dataset.

Our first data set had a perfectly balanced class distribution of 'Good' and 'Bad' movie reviews, seen in Figure 11. This was determined based off the document review scores, such that any score

less than or equal to 4 was considered 'Bad', and anything greater than or equal to 7 was considered 'Good'. Our second data set also has a very balanced class distribution of documents between the 20 different classes, as seen in Figure 12. Balanced datasets are preferable as they tend to lead to better, more generalizable models since the model has an equal opportunity to learn from all classes.

4 Results

4.0.1 Top 20 Features from the Simple Linear Regression on IMDB Data

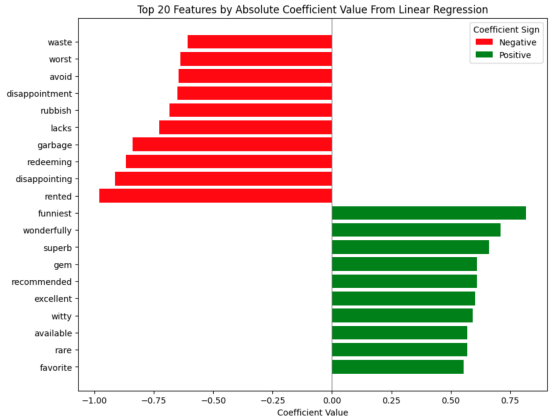

Figure 1: Top 20 IMBD Data Features Identified by the Simple Linear Regression Model

Our initial experiment involved plotting the top 20 features with the most positive and negative regression coefficients based on their absolute values. They were derived from a pool of the top 200 features identified through simple linear regression analysis of the IMDB dataset. Among these features, the ten most positive ones were 'favorite', 'rare', 'available', 'witty', 'excellent', 'recommended', 'gem', 'superb', 'wonderfully', and 'funniest'. Their corresponding coefficient values ranged from 0.56 to 0.82. Conversely, the ten most negative features were 'waste', 'worst', 'avoid', 'disappointment', 'rubbish', 'lacks', 'garbage', 'redeeming', 'disappointing', and 'rented', with coefficient values ranging approximately from -0.61 to -0.98. These features highlight the words most seen in positive and negative movie reviews from the IMBD dataset, allowing us to classify future good and bad movie reviews based on the frequency of these words.

4.0.2 Predicting Sentiment Score Using Linear Regression

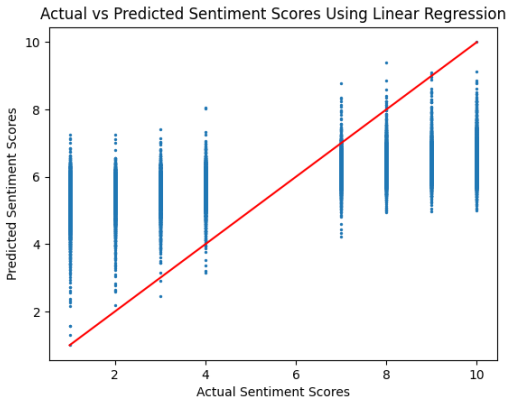

Figure 2: Actual vs Predicted Sentiment Scores Using Linear Regression

The scatter plot indicates that the linear regression model tends to overestimate sentiment scores for lower actual values and underestimate them for higher actual values, as evidenced by the clustering of points below the red line at higher actual scores and above the line at lower scores. This pattern suggests a lack of fit for extreme values. The normalization of predicted scores to a 1-10 scale has adjusted the original model outputs, which ranged from approximately -15 to 20, allowing for a more direct comparison with the actual scores that were already on this scale. However, the normalization does not change the relative accuracy of the predictions, only the scale on which they are presented.

4.0.3 Logistic and Multiclass Regression Convergence

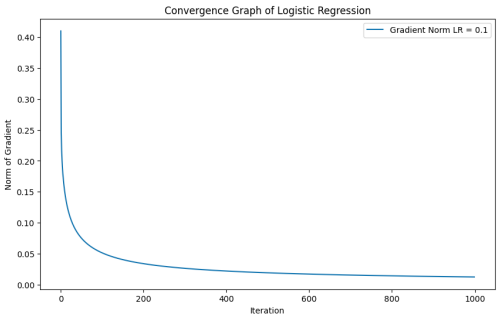

Figure 3: Logistic Regression Convergence Plot With Learning Rate 0.1

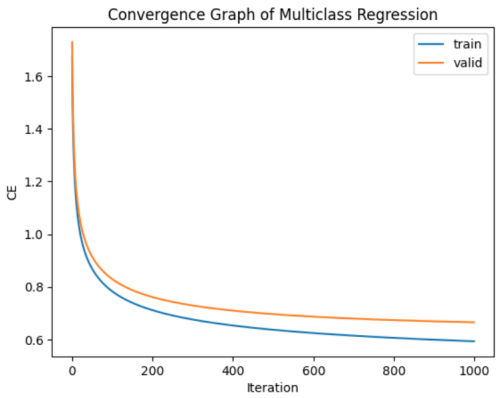

Figure 4: Multiclass Regression Convergence Plot With Learning Rate 0.0005

After choosing learning rates, we constructed convergence plots for our logistic and multiclass regression models. In Figure 2, the logistic regression model with a learning rate of 0.1 converges rapidly; this suggests that the rate is effective for the dataset without causing overshooting of the cost function's minimum. Figure 3 shows a multiclass regression model with a lower learning rate of 0.0005, demonstrating good convergence for training and validation. Further, the graph indicates a low risk of overfitting as our validation and training data converge in a similar pattern. The learning rates were determined through experimentation and chosen based on which yielded the best accuracy without converging too early. Our logistic regression model has a gradient norm of 0.002247

and an accuracy of 91%. The steep decline in the gradient norm for logistic regression indicates a successful choice of learning rate. While our multiclass regression model has a lower test accuracy of 73.27%, with a more gradual decline in convergence. Overall, the learning rates seem well-tuned for their respective models and datasets, balancing the speed of convergence with the accuracy of the models.

4.0.4 ROC of Logistic Regression and DT on IMDB Test Data

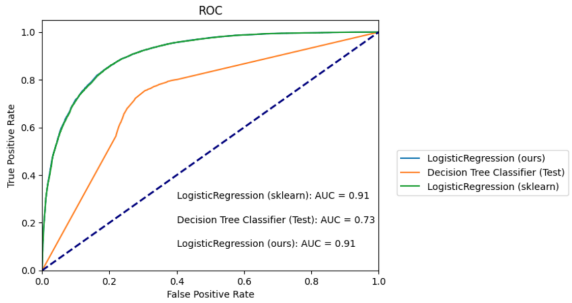

Figure 5: ROC of Logisitc Regression and DT on IMDB Test Data

Figure 4 displays the ROC curves of sklearn-logistic regression and sklearn-DT compared to our logistic regression model on the IMDB dataset. Our logistic model has a test accuracy of 82.8% and an AUC value of 0.91, performing on par with the sklearn-logistic regression model. This comparison indicates the effectiveness of our model. Further, an AUC of 0.91 is quite high, suggesting that our model can effectively distinguish between the positive and negative classes. The sklearn-DT classifier has an AUC of 0.74, indicating that it has a more moderate predictive ability but is not as effective as the logistic regression models. The higher AUC value associated with our logistic model implies that it is more reliable to make predictions on this particular dataset and is significantly better at distinguishing between the positive and negative classes in the dataset.

4.0.5 AUROC of Logisitic Regression and DT on Varying Training-Testing Data Splits

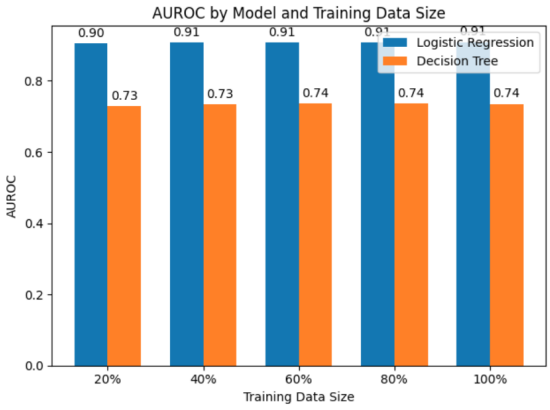

Figure 6: AUROC of Logisitic Regression and DT on Varying Training-Testing Data Splits

The bar plot compares the performance of logistic regression and decision tree models across various training data sizes using the AUROC metric. Consistently across all training data sizes, logistic regression outperforms the decision tree model with AUROC values ranging from 0.90-0.91. In contrast, the decision tree model's AUROC remains around 0.73-0.74, significantly lower than the

logistic regression's. This suggests that the decision tree may be less capable of generalizing from the training data to the test data. Further, the logistic regression model's performance does not seem to be sensitive to the training data size as it maintains a high AUROC across all subsets. The decision tree's performance also appears stable across different training data sizes, but it consistently achieves a lower AUROC score, indicating that increasing the training data size does not improve its ability to differentiate between classes. This analysis shows that logistic regression is a better model for this particular dataset and task, as it consistently provides stronger predictions.

4.0.6 Classification Accuracies of Multiclass Regression and DT on Test Data

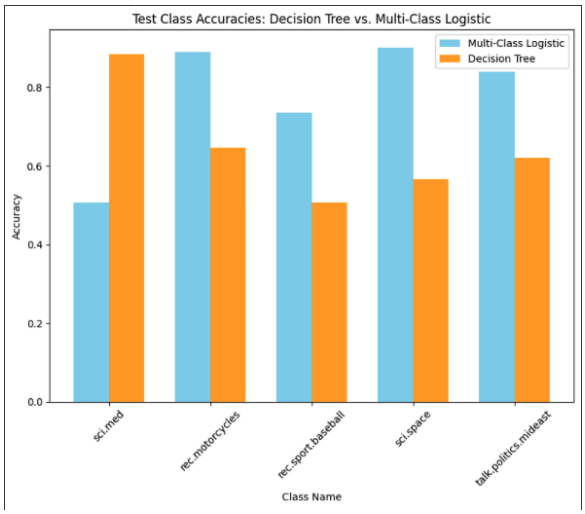

Figure 7: Accuracy's per class used for 2nd Data Set

The comparative bar graph shows the test class accuracies for two distinct models: the Decision Tree (DT) and the Multi-Class Logistic Regression. Across five different categories—'sci.med', 'rec.motorcycles', 'rec.sport.baseball', 'sci.space', and 'talk.politics.mideast'. The graph reveals significant variance in performance between the two models. The DT model has better accuracy in 'rec.motorcycles' and 'sci.space', with very strong peformance in 'sci.space'. Conversely, the Multi-Class Logistic Regression model outperforms the DT in 'sci.med', with closer margins across the other categories.

4.0.7 Top 20 IMBD Data Features Identified by Logistic Regression Model

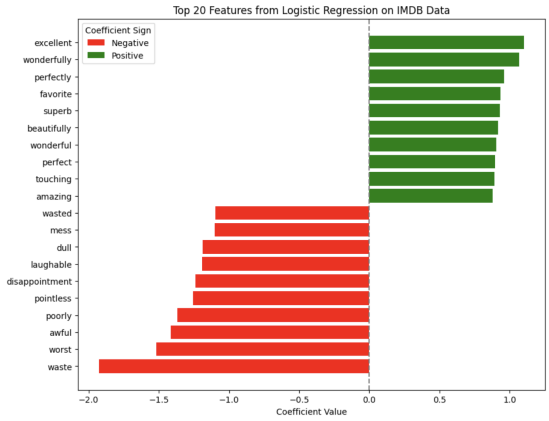

Figure 8: Top 20 features from Logistic Regression model

The bar chart visualizes the top 20 influential words from a logistic regression model applied to IMDB data for sentiment analysis. Green bars indicate words positively correlated with positive reviews, such as "excellent" and "perfect," while red bars show words like "waste" and "worst" that predict negative sentiment. The length of the bars denotes the strength of the correlation: the longer the bar, the stronger its impact on the model's predictions.

4.0.8 Heatmap of Top 5 Most Positive Features In the Multiclass Classification from the 20 Newsgroup Datasets

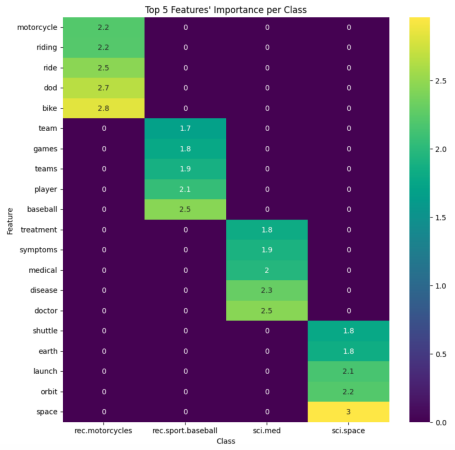

Figure 9: Heatmap of Top 5 Most Positive Features In the Multiclass Classification

The heatmap displays the top 5 most influential features for the five classes derived from the 20 Newsgroup dataset. Each row represents a feature, and each column corresponds to a class. The colours represent different levels of importance for each feature in the class, with the more yellow colours indicating higher importance. Each class has 5 key features most relevant for classification. The top 5 features for comp.sys.mac.hardware are 'centris', 'se', 'simms', 'mac', and 'apple'. For misc.forsale, they are 'interested', 'shipping', 'sell', 'offer', and 'sale'. For rec.motorcycles, the features are 'ride', 'bikes', 'motorcycle', 'bike', and 'dod'. In rec.sport.baseball, they are 'player', 'players', 'teams', 'team', and 'baseball'. Lastly, the talk.politics.mideast features are 'jews', 'serdar', 'turkish', 'israeli', and 'israel'. The heatmap effectively shows that the selected features for

each class are highly relevant and class-specific, which likely contributes to the strong performance of the classifier on the dataset.

4.0.9 Enhancing Multi-Class Classification by Considering Top-k Accuracy Scoring

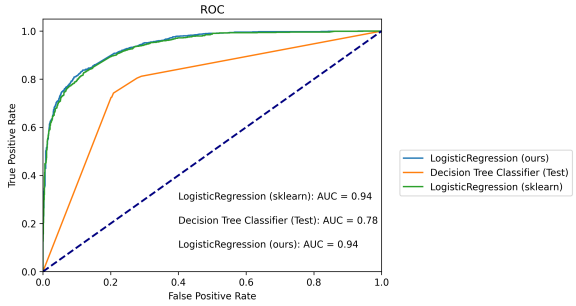

Figure 10: Top 3 classes AUROC

The ROC curve displays the performance of two Logistic Regression models—one custom and one from scikit-learn—with identical AUC scores of 0.94, and a Decision Tree Classifier with an AUC of 0.78. Both Logistic Regression models demonstrate equivalent predictive capabilities, significantly outperforming the Decision Tree Classifier in distinguishing between the aggregated positive classes (0, 1, 4) and the rest.

5 Discussion and Conclusion

This project has effectively demonstrated the superiority of logistic and multiclass regression models over Decision Trees in the context of text classification, using the IMDB Reviews and 20 News Groups datasets. The key takeaways include the impressive performance of logistic regression, with a 0.94 AUROC in the IMDB dataset, and the multiclass regression's 73% accuracy in the 20 News Groups dataset.

Our investigation also highlighted the importance of feature selection in text classification. The top features identified, such as 'excellent', 'perfect', 'waste', and 'worst', for the IMDB data, and the top 5 features in the multiclass classification from the 20 News Group dataset, all make logical sense. They align well with the sentiments or topics they represent, indicating the models' ability to accurately capture and utilize key textual elements.

In future research on text classification, there are several interesting areas to explore. First, it's a good idea to look at different types of machine learning models. So far, we've used models like logistic and multiclass regression, but trying out others, such as Neural Networks or Support Vector Machines, might give us different or better results. Another area to work on is improving how we process and understand the text data. We can do this by using advanced techniques like TF-IDF, which helps us figure out which words in the text are most important, or by using word embeddings like Word2Vec, which help us understand the meaning of words in context.

In conclusion, this project has not only shown the effectiveness of logistic and multiclass regression in text classification but also opened avenues for further research to enhance model performance and understanding.

6 Appendix

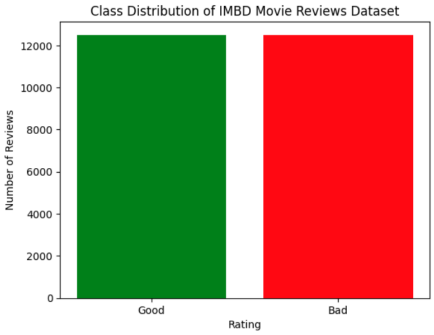

Figure 11: Class Distribution of Dataset 1

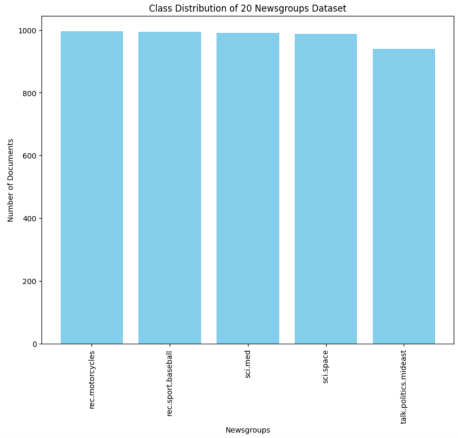

Figure 12: Class Distribution of Dataset 2

References

[1] Wilhelm Grzesiak (2023). Classification of Daily Body Weight Gains in Beef Calves Using Decision Trees, Artificial Neural Networks, and Logistic Regression *MDPI*, Article number: 20309. Available at: https://www.mdpi.com/2076-2615/13/12/1956