Classification of Textual Data: COMP 551 Winter 2024 Assignment 2

Emile Riberdy (260985547), Shubham Vashisth (261155310), Mohaddeseh Yaghoubpour (261094819)

February 27, 2024

Abstract

This study implements Simple Logistic Regression and Multiclass Logistic Regression models and compares their performance with Decision Trees to elucidate their classification accuracies across a spectrum of training set sizes. Our comprehensive evaluation demonstrates that Simple Logistic Regression consistently surpasses Decision Trees classifiers across all training sizes, and Multiclass Logistic Regression outperforms Decision Trees at larger training set sizes. Intriguingly, at smaller training set sizes, Decision Trees exhibit a small superior performance to Multiclass Logistic Regression. Both Multiclass Logistic Regression and Decision Trees demonstrate a trend of increasing accuracy with the expansion of training sets up to 70%, beyond which the gains either plateau or decline, underscoring the pivotal influence of training set size on model performance. Notably, the Logistic Regression model maintains an impressive AUROC score exceeding 0.9 across varying training sizes, showcasing its robustness and efficiency. These insights advocate for a nuanced selection of models based on training set size and underscore the superiority of Logistic Regression in maintaining consistent performance, thereby guiding future research towards optimizing machine learning model selection and application.

1 Introduction

This project embarks on a sophisticated exploration of text classification and sentiment analysis, harnessing the capabilities of advanced machine learning models. It utilizes two distinct datasets: IMDB Reviews and 20 News Groups. The IMDB dataset, comprising movie reviews with binary sentiment labels, serves as a gateway to discerning consumer sentiment patterns. Meanwhile, the 20 News Groups dataset provides a comprehensive platform for text classification, spanning a wide array of topics, which facilitates the development of robust information retrieval and topic categorization techniques. These datasets have been instrumental in previous analytical studies like [1], [2], and [3].

Our methodology involves a rigorous preprocessing regime and a strategic feature selection process to optimally condition the datasets for analysis. The IMDB Reviews benefitted from a combination of stopword elimination and frequency-based filtering, augmented by Simple Linear Regression to pinpoint features with predictive potency. In parallel, the 20 News Groups dataset was subjected to a congruent preprocessing workflow, employing Mutual Information to isolate features emblematic of specific newsgroup topics. Our manual implementation of Logistic and Multiclass classifiers, derived from classroom learnings, was complemented by the deployment of a Decision Trees classifier from the SkLearn Python library for comparative analysis. The resultant insights reveal Logistic Regression's dominance over Decision Trees in sentiment classification within the IMDB corpus, thereby accentuating the critical role of feature selection in amplifying model precision. It is important to note that Decision Trees were slightly better than Multiclass Logistic Regression in terms of accuracy in small raining set size. However, as the training set size increased, Multiclass Logistic Regression performed better.

2 Datasets

2.1 IMDB Reviews

The IMDB Reviews dataset is a binary sentiment classification dataset consisting of textual movie reviews that have been labeled as either positive or negative. This dataset is instrumental in developing models that can accurately assess sentiment in textual data.

In the preprocessing stage, reviews were extracted from text files, with their sentiment labels inferred from the directory structure, either 'pos' or 'neg'. Each review's rating was derived from its filename, which also served as a unique identifier to ensure orderly data processing. We refined the dataset by filtering out words that were too infrequent (appearing in less than 1% of the reviews) or too common (appearing in more than 50% of the reviews), reducing dimensionality and emphasizing words with potential predictive power. Additionally, stopwords were removed using the Natural Language Toolkit (nltk) to eliminate high-frequency but low-information words from the feature set.

Feature selection was informed by implementing a Simple Linear Regression to evaluate the correlation of words with the review ratings. This process identified words whose presence strongly correlates with the sentiments expressed in the reviews, focusing on those with the highest absolute regression coefficients. This judicious approach to feature selection was designed to equip our classification models with a nuanced understanding of sentiment-laden language, striking a balance between ignoring noise and capturing meaningful patterns within the reviews.

2.2 20 News Groups Analysis

The 20 Newsgroups dataset serves as a cornerstone in text-based machine learning tasks, such as text classification and clustering, encompassing a collection of approximately 20,000 documents across 20 varied newsgroups. In our exploratory study, we honed in on a subset of documents from 'comp.graphics', 'sci.med', 'soc.religion.christian', 'talk.politics.mideast', and 'rec.sport.hockey'.

Initial preprocessing involved the removal of headers, footers, and quotes from the raw text, thereby directing our classifiers' attention to the essential content. Moreover, we standardized the formatting by converting newline and tab characters into spaces.

Employing scikit-learn's CountVectorizer, we transformed the textual data into a token count matrix. Subsequent filtering based on term frequency ensued, keeping only those words present in more than 1% and fewer than 50% of the documents, to discard the outliers in term frequency.

Our selection of features was informed by the Mutual Information (MI) metric, which quantifies variable dependency. This led to a prioritization of features, with the MI informing us of each term's relevance to the class labels. The results delineates the top 50 features with the highest MI scores, thus shedding light on the terms most indicative of each class.

This methodical approach to analysis was instrumental in the creation of a training dataset, an amalgam of these paramount features across categories, matched with their corresponding labels. Such a refined dataset is crucial for the training of classification models, offering a concentrated depiction of the dataset that highlights the most salient features for discerning the topics within the newsgroups.

3 Results

3.1 Top Features from Simple Linear Regression on IMDB Data

The analysis of the IMDB reviews dataset via simple linear regression has yielded a set of features with significant positive and negative correlations to movie review sentiments. Figure 1 illustrates the top 20 features identified. The horizontal bar plot distinctly marks the 10 most positive and 10 most negative features, arrayed according to their regression coefficients. Words like "funniest" and "wonderfully" show the strongest positive correlation, indicating a strong predictive power for positive sentiment. Conversely, words such as "waste" and "worst" have the highest negative coefficients, suggesting a robust association with negative reviews. This visualization encapsulates the duality inherent in sentiment analysis, highlighting the linguistic indicators most influential in determining the polarity of a review.

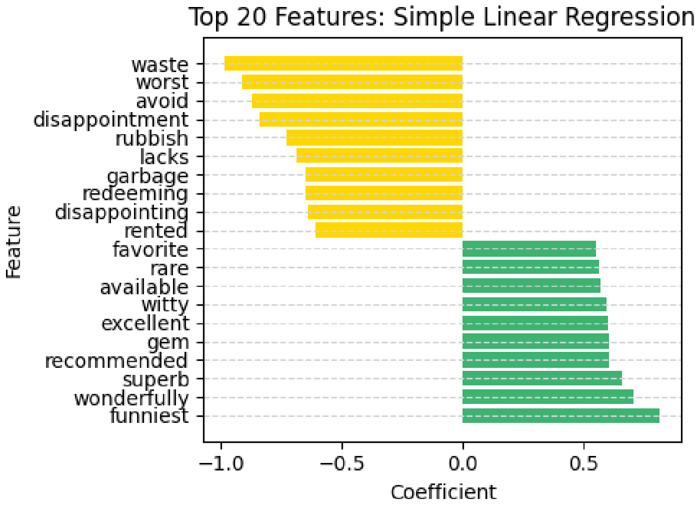

Figure 1: Top 20 features from simple linear regression on the IMDB dataset. Positive features are represented with green bars, while negative features are in yellow.

3.2 Convergence of Logistic and Multiclass Regression Models

The convergence behaviors of logistic and multiclass regression models were meticulously analyzed utilizing the Cross Entropy loss function. For the logistic regression model tailored to the IMDB dataset, a learning rate of $\alpha = 0.01$ was judiciously selected. Depicted in Figure 2, the training and validation losses manifest a substantial decline, signaling a swift attainment of predictive precision for the task of sentiment analysis.

In contrast, the multiclass regression model, when applied to the intricate 20 Newsgroups dataset, necessitated a more conservative learning rate of $\alpha = 0.0001$. The depicted convergence in Figure 3 unveils a consistent diminution in loss, heralding an effective adjustment to the multifaceted nature of topic categorization inherent in the dataset.

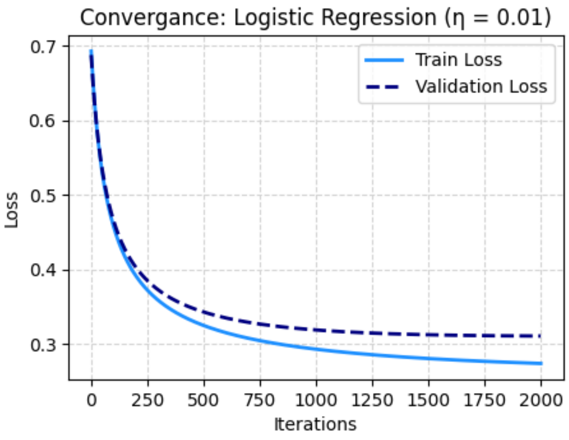

Figure 2: Training and validation loss for logistic regression on IMDB data, illustrating significant convergence with learning rate $\alpha = 0.01$.

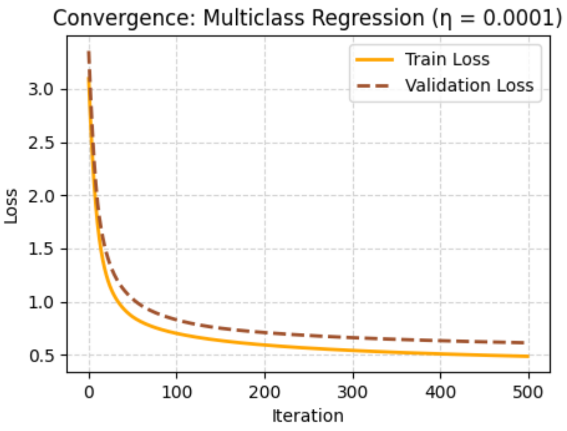

Figure 3: Training and validation loss for multiclass regression on 20 Newsgroups data, showing steady convergence with learning rate $\alpha = 0.0001$.

These convergence patterns emphasize the pivotal role of an optimal learning rate for each model, which is quintessential to foster expeditious learning while maintaining stability. The selected learning rates of 0.01 for logistic regression and 0.0001 for multiclass regression corroborate that both models have reached a state of convergence efficaciously, with the validation loss mirroring the training loss, thereby endorsing successful generalization to unseen data.

3.3 ROC Curves for Logistic Regression and Decision Trees on IMDB Test Data

For the evaluation of our binary classification models on the IMDB test data, Receiver Operating Characteristic (ROC) curves were utilized to visually represent the performance of both the Logistic Regression model and the Decision Trees classifier.

As illustrated in 4, the Logistic Regression model achieved an AUROC of 0.94, signifying an excellent discrimination capability between positive and negative sentiment reviews. The model's ROC curve, depicted in blue, shows a high true positive rate across approximately all levels of the false positive rate, which indicates strong performance.

In contrast, the Decision Trees classifier exhibited an AUROC of 0.75. The corresponding ROC curve is represented in red and demonstrates lower performance compared to the Logistic Regression model. This is particularly evident in the middle range of the false positive rate, where the curve for the Decision Trees model lies significantly below the Logistic Regression curve, indicating fewer true positives for the same level of false positives.

The comparison between the two models reveals that Logistic Regression has a superior capability to distinguish between the classes in the IMDB dataset. This is likely due to the model's linear decision boundary, which may better capture the nuances in the dataset as opposed to the hierarchical, piecewise decision boundaries used by Decision Trees.

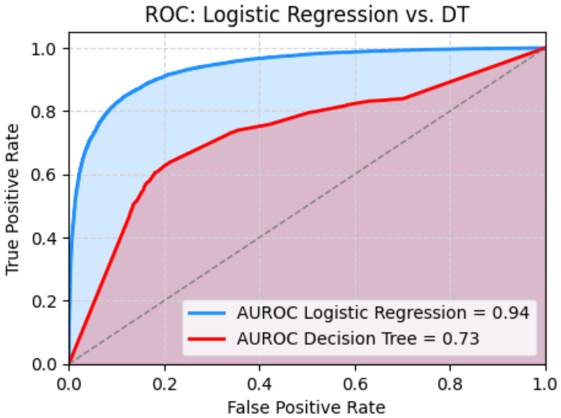

Figure 4: ROC curves comparing Logistic Regression and Decision Trees classifiers on the IMDB test dataset. The Logistic Regression model demonstrates superior performance with an AUROC of 0.94, compared to the Decision Tree's AUROC of 0.75.

3.4 AUROC of Logistic Regression and Decision Trees Across Training Sizes

We assessed the performance of Logistic Regression and Decision Trees classifiers on the IMDB test dataset by computing the Area Under the Receiver Operating Characteristic (AUROC) across various proportions of training data. The AUROC metric provides an aggregate measure of performance across all classification thresholds, with 1 indicating a perfect model and 0 signifying no discriminative ability.

As indicated in Figure 5, the classifiers were trained on subsets of the IMDB dataset, with training sizes ranging from 10% to 90% for both Logistic Regression and Decision Trees.

The Logistic Regression model exhibited robust performance, maintaining an AUROC score above 0.9 for all training set sizes, indicating its effectiveness in discriminating between positive and negative reviews. The Decision Trees classifier showed lower AUROC values, ranging from 0.66 to 0.73, suggesting that it has less predictive power compared to the Logistic Regression model.

The stability of the Logistic Regression model's AUROC scores suggests that it is less sensitive to the size of the training data and can maintain high performance even with less data. In contrast, the Decision Tree's performance varied more with the size of the training set, indicating potential overfitting or sensitivity to the specific training data used. The analysis highlights the importance of choosing the appropriate model for binary classification tasks and demonstrates the effectiveness of Logistic Regression in handling the nuances of natural language data for sentiment analysis.

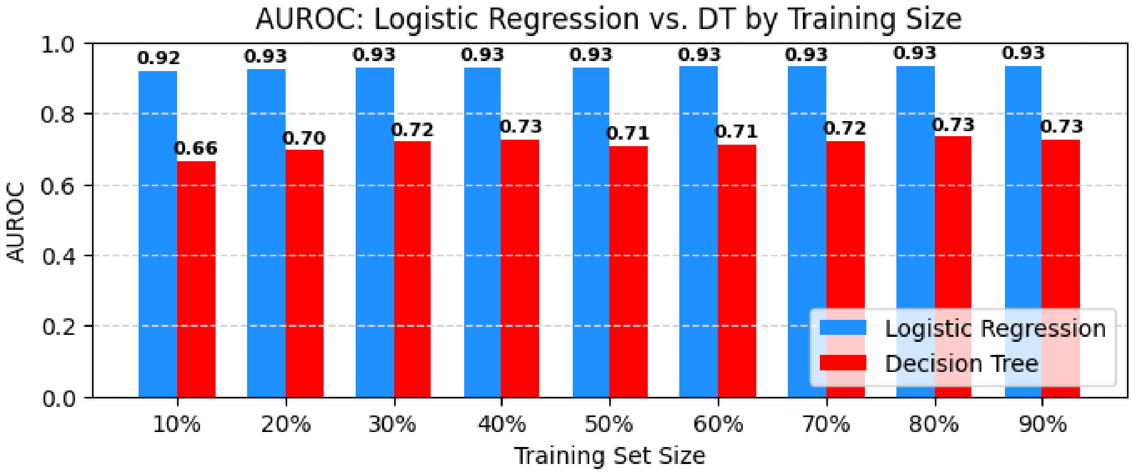

Figure 5: AUROC scores for Logistic Regression and Decision Trees classifiers as a function of training set size. Logistic Regression consistently outperforms Decision Trees, with its AUROC scores remaining above 0.9 across all training sizes.

3.5 Classification Accuracies of Multiclass Regression and Decision Trees

In this section, we evaluate the classification accuracies of the Multiclass Logistic Regression model and the Decision Trees classifier across different sizes of the training set. The training sizes examined were 20%, 40%, 60%, 80%, and 100% of the available data and the evaluation metric used is the accuracy on the test set.

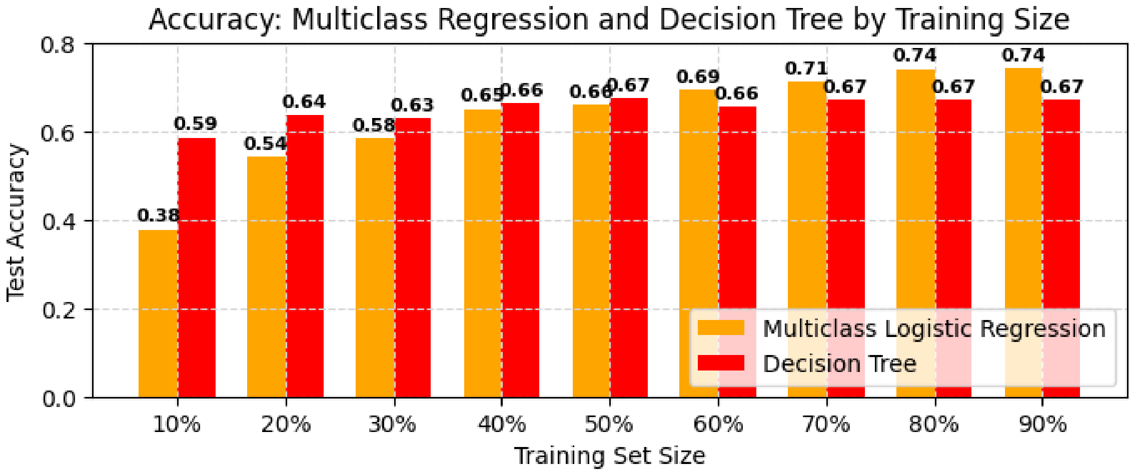

Figure 6: Test accuracy comparison between Multiclass Logistic Regression and Decision Trees models across various training set sizes. The Multiclass Logistic Regression model demonstrates higher accuracy across all training sizes, with peak performance at a training set size of 90%.

The Multiclass Logistic Regression model consistently outperformed the Decision Trees classifier across larger training set sizes. However, Decision Trees had a better accuracy in lower training set sizes. The test accuracy of the Multiclass Logistic Regression model was highest at a training set size of 90%, reaching an accuracy of approximately 76.4%. This indicates that the model was able to utilize the larger volume of data to effectively learn and make predictions on the test set.

The Decision Trees classifier achieved its highest test accuracy with an 80% training set size, obtaining an accuracy close to 67%. However, the model did not show significant improvement with an increase in the training set size beyond this point, which may suggest a tendency towards overfitting, where additional training data does not translate to better generalization.

It is worth noting that both models exhibited an increase in test accuracy as the training set size was augmented from 10% to 70%. However, while the accuracy of the Multiclass Logistic Regression model began to stabilize beyond a 70% training set size, the Decision Trees model's accuracy saw a slight decrease, highlighting the former's robustness and the latter's potential overfitting issues.

The comparative test accuracies are visualized in the bar plot shown in Figure 6. The bars colored in orange represent the Multiclass Logistic Regression model, and those in red represent the Decision Trees classifier. The observed trends and differences in performance are indicative of the inherent strengths and weaknesses of the two models in handling multiclass classification tasks on the dataset under study.

3.6 Top Features from Logistic Regression on IMDB Dataset

We trained a Logistic Regression model on the IMDB dataset to classify movie reviews as positive or negative. After fitting the model, we extracted the logistic regression coefficients corresponding to each word feature. These coefficients represent the strength and direction of the association between the feature and the likelihood of a positive review outcome. A positive coefficient suggests that the presence of the word is associated with a positive review, while a negative coefficient suggests an association with a negative review.

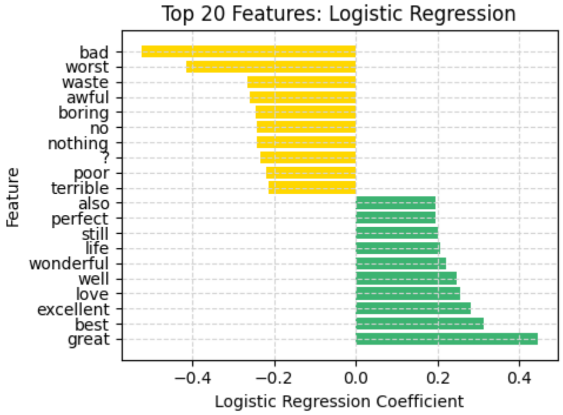

Figure 7: Top 20 features from Logistic Regression on the IMDB data, showing the 10 most positive and 10 most negative words by their logistic regression coefficients.

The bar plot in Figure 10 illustrates the top 20 features determined by the model. The x-axis represents the logistic regression coefficients, while the y-axis lists the feature names, which are specific words from the reviews. The top 10 features with positive coefficients are indicative of a positive sentiment, whereas the bottom 10 features with negative coefficients are indicative of a negative sentiment.

Comparing Figures 1 and 10 reveals a considerable overlap in positive and negative words identified as significant by both models. Notably, words such as "excellent", "great", and "worst" appear in the lists of top features for both simple linear regression and logistic regression. This overlap corroborates the effectiveness of both models in discerning key predictors of sentiment in movie reviews.

3.7 Feature Importance Heatmap for Multi-class Classification on News Groups

In multi-class classification tasks, understanding the relevance of different features for each class is crucial for interpreting the model's decisions. In our study, we extracted the top features for each class by analyzing the weights of the trained Multiclass Logistic Regression model. The heatmap in Figure 8 represents the importance of each feature across the five different newsgroup classes used in our analysis: 'comp.graphics', 'rec.sport.hockey', 'sci.med', 'soc.religion.christian', and 'talk.politics.mideast'.

Each row in the heatmap corresponds to a feature, and each column corresponds to a class. The color intensity in the heatmap reflects the normalized weight of each feature, indicating the level of influence that feature has on classifying a document into the corresponding category. Darker shades represent a higher importance, signifying a stronger association between the feature and the class.

The features identified as most indicative of each class were as follows: 'computer', 'hardware', 'file', 'graphics', and '3d' were strongly associated with 'comp.graphics'; 'game', 'ranger', 'player', 'leafs', and 'toronto' with

’rec.sport.hockey’; ’doctor’, ’available’, ’desease’, ’med’. and ’effect’ with ’sci.med’; ’christians’, ’bible, ’god’, ’christian’, and ’heaven’ with ’soc.religion.christian’; and ’israeli’, ’peace’, ’serdar’,’israel’, and ’country’ with ’talk.politics.mideast’. These features are intuitively representative of the topics commonly discussed in the corresponding newsgroups.

The heatmap visualization serves as a useful tool for interpreting the model and offers an informative way to present the results to both technical and non-technical audiences, allowing for a better understanding of the model’s behavior.

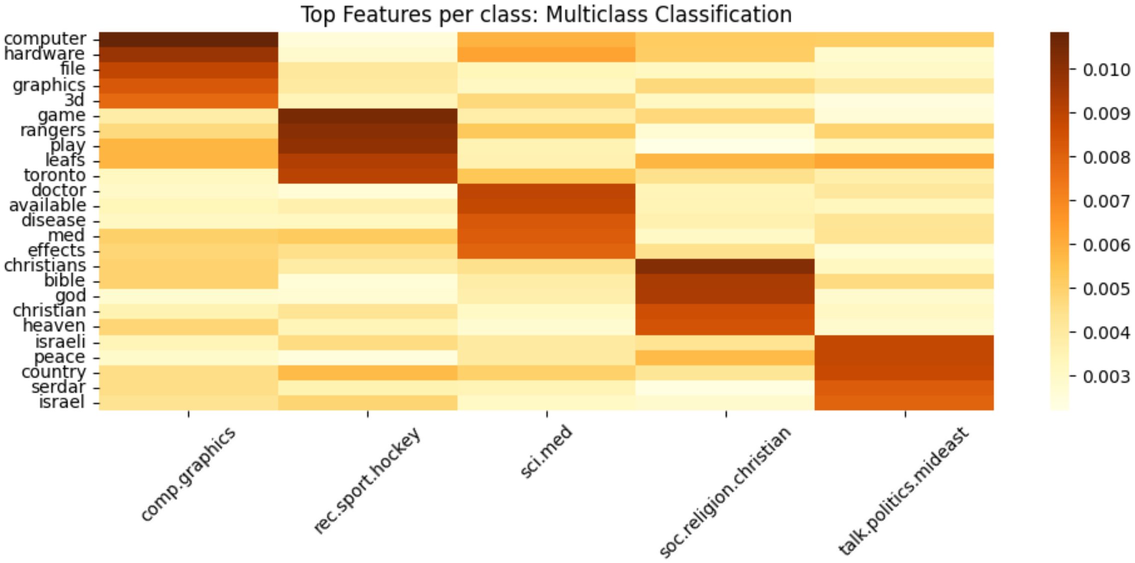

Figure 8: Feature importance heatmap displaying the top five features for each class in the multi-class classification task on the selected 20-news group dataset. The intensity of the color represents the normalized weight of each feature, providing a clear visualization of the most influential terms for each newsgroup category.

3.8 Additional experiment

To further explore the subject, we conducted two additional experiments.

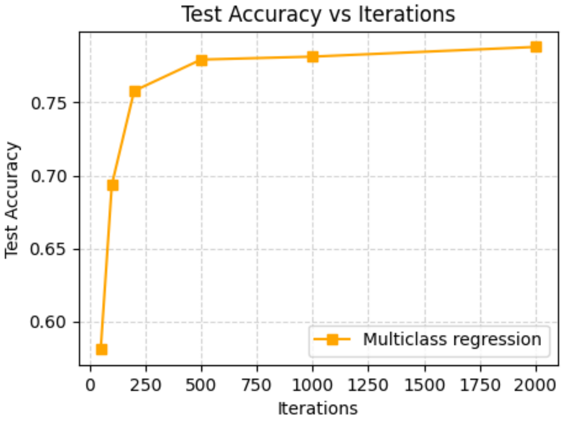

Figure 9: Test Accuracy vs Iterations for a Multinomial Logistic Regression model.

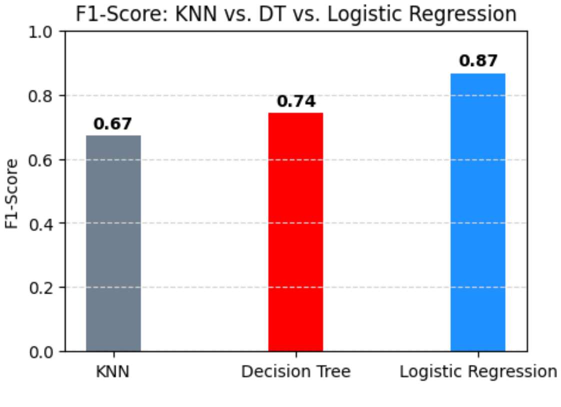

Figure 10: Comparing F1-score of KNN, Decision Trees and Logistic Regression.

3.8.1 Test Accuracy vs Iterations for a Multinomial Logistic Regression model

In this experiment, we compared the F1-Score of KNN, Decision Tree, and Logistic Regression on the IMDB dataset. As indicated in Figure **??**The F1-Score, which is the harmonic mean of precision and recall, serves as a measure to evaluate the performance of these binary classifiers. Our findings show that Logistic Regression achieved the highest F1-Score of 0.87, indicating a robust performance in classifying the sentiment of movie reviews. The Decision Trees classifier recorded an F1-Score of 0.74, which, while lower than Logistic Regression, still demonstrates a reasonable level of accuracy. KNN lagged with an F1-Score of 0.67, suggesting it may be less suited for the high-dimensional feature space of text data. These results highlight the effectiveness of Logistic Regression in processing and classifying textual data, particularly for sentiment analysis tasks within the domain of natural language processing.

3.8.2 Test Accuracy vs Iterations for a Multinomial Logistic Regression model

Second, the Multinomial Logistic Regression model's accuracy was scrutinized by varying the number of iterations during the training process. 20 News Groups dataset was split into training and validation sets with a 50% test size, ensuring reproducibility and randomness in the split. The model was then trained multiple times using the implemented Multiclass Logistic Regression algorithm with iterations ranging from 50 to 2000. It was observed that the test accuracy improved significantly up to 200 iterations, after which the increase in accuracy plateaued, reaching stability at approximately 76% for 1000 iterations and beyond (9). This experiment highlights the importance of choosing an optimal number of iterations for the training process to balance between computational efficiency and model accuracy.

3.9 Conclusion

The exploration of classification accuracies for Simple Logistic Regression, Multiclass Logistic Regression and Decision Trees across various training set sizes has yielded insightful outcomes. The Simple Logistic Regression model demonstrated superior performance over the Decision Trees classifier across all evaluated training set sizes. The Logistic Regression model exhibited robust performance, maintaining an AUROC score above 0.9 for all training set sizes.

Both Multiclass Logistic Regression and Decision Trees models exhibited a trend of increasing test accuracy with larger training set sizes up to 70%. However, while the Multiclass Logistic Regression model's accuracy began to plateau beyond a 70% training set size, the Decision Trees model's accuracy experienced a slight decline, indicating potential overfitting issues. These findings underscore the importance of model selection and training set size in achieving optimal classification performance.

In terms of future work, investigating additional machine learning models and employing advanced feature selection techniques could offer new insights and potentially enhance classification accuracy. Furthermore, exploring the application of these models in other domains or with varied datasets may provide a broader understanding of their generalizability and effectiveness.

4 Statement of Contributions

Each team member actively participated in all aspects of the assignment, utilizing in-person meetings, online discussions, Slack, and shared platforms like Colab, GitHub and Overleaf. However, there was a slight emphasis on specific areas: Shubham took the lead in Logistic Regression, Emile in Multiclass Regression, and Mohaddeseh in the report and analysis.

References

[1] Saeed Mian Qaisar. Sentiment analysis of imdb movie reviews using long short-term memory. In *2020 2nd International Conference on Computer and Information Sciences (ICCIS)*, pages 1–4, 2020.

[2] Vrushang Patel, Sheela Ramanna, Ketan Kotecha, and Rahee Walambe. Short text classification with tolerance-based soft computing method. *Algorithms*, 15(8), 2022.

[3] Yongheng Mu and Yun Wu. Multimodal movie recommendation system using deep learning. *Mathematics*, 11(4), 2023.