COMP 551 A2: *Classification of Textual Data*

Rebecca Mizrahi (260975001) rebecca.mizrahi@mail.mcgill.ca
Software Engineering

Amélie Barsoum (290988658) amelie.barsoum@mail.mcgill.ca
Software Engineering

Samantha Handal (260983914) samantha.handal@mail.mcgill.ca
Software Engineering

1. Abstract

In this report, we engage in experimentation to compare the performance of logistic regression and multiclass regression models against Decision Trees in the classification of textual data. We preprocess, implement, and evaluate these models on two distinct datasets: the IMDB movie reviews and a subset of the 20 Newsgroups dataset. Our findings reveal that logistic regression outperforms Decision Trees in binary classification tasks with an accuracy of 80% on IMDB data and an AUROC of 0.89. In the realm of multiclass classification, our custom multiclass regression model demonstrates superior accuracy over Decision Trees, achieving 73.6% on the 20 Newsgroups data. Further experiments regarding the impact of training data size on model performance indicate that logistic regression maintains high accuracy across smaller and larger data proportions, highlighting its robustness.

2. Introduction

In machine learning, specifically in natural language processing (NLP), the classification of textual data is an important task task, enabling a wide array of applications. This report delves into our implementation and evaluation of logistic regression (LR) and multiclass regression models from scratch. Our evaluation and experimentation is centered on two benchmark datasets: the IMDb movie reviews for sentiment analysis and a subset of 5 of the 20 Newsgroups dataset for topic classification. The IMDb dataset has a binary classification nature, and allows us to perform sentiment analysis, meaning that a movie can be classified into positive sentiment or negative sentiment, based on the movie's reviews. Conversely, the 20 Newsgroups dataset, with its multi-class labeling, allows for more complex classification tasks, such as categorizing documents into distinct topics.

Logistic regression and multiclass regression helps us with a general understanding of linear classifiers in NLP tasks, as they are powerful tools for understanding the significance of features (words) in textual data. Implementing these models taught us about training, feature selection and evaluation. The comparison of their performance vis a vis decision trees also shed light on how much more effective they are than DTs for textual data.

Researchers have already explored the efficacy of LR in sentiment analysis tasks, particularly with the IMDb dataset (Tripathi et al., 2020), and LR gave the best validation AUC

of nearly 96%. Similarly, multiclass regression has been shown to effective on textual data (Joshi et al., 2021), reporting a high F1-score of 0.86, which will likely translate to our 20 News groups dataset.

3. Datasets

We use two datasets for textual data classification: the IMDb movie reviews dataset for sentiment analysis and a subset of the 20 Newsgroups dataset for topic classification.

The IMDb dataset consists of movie reviews, used for binary sentiment classification. Our preprocessing involves standardizing text by removing punctuation, converting to lowercase, and filtering words based on their frequency across documents to filter out words that appear in less than 1% of the documents and words that appear in more than 50% of the documents, which are the rare and "stopwords". This filtration eliminates rare and common words, narrowing down the feature space to more meaningful data for analysis. The dataset's reviews are vectorized into a binary feature representation, indicating the presence or absence of these filtered words. Ratings extracted from filenames categorize reviews as positive (ratings above 5) or negative, which facilitates our sentiment analysis.

With the 20 Newsgroups Dataset, for multi-class classification, we select five distinct categories from the 20 Newsgroups dataset: 'comp.graphics', 'misc.forsale', 'rec.sport.baseball', 'sci.med', and 'talk.politics.guns'. The preprocessing steps include removing non-essential sections (headers, footers, quotes) and applying a document frequency filter to refine the feature set. The vectorization of this dataset focuses on creating features based on mutual information scores, selecting the most relevant words for class distinction. This method prepares the dataset for multiclass regression analysis, to categorize the data into their distinct topics.

4. Results

4.1 Top Features for IMDb Sentiment Classification Determined by Logistic and Linear Regression

Figure 1 reflects the 10 words that are most indicative of a positive review, as well as the 10 indicative of a negative review, based on weights assigned by a Linear Regression model. Figure 2 shows the same idea of figure 1, except executed with Logistic Regression rather than Linear. Simply through observation, one can see that this is most likely accurate; the positive words reflect positive sentiments, like enjoyable and excellent, and give interesting insight on the type of movies that receive positive reviews, such as "unusual". Similarly, the highly negative-weighted words are evidently negative as well. The features selected as holding the most weight via logistic regression interestingly do not totally overlap with those selected by linear regression. One can argue, based on human understanding of the words, that it selects better features.

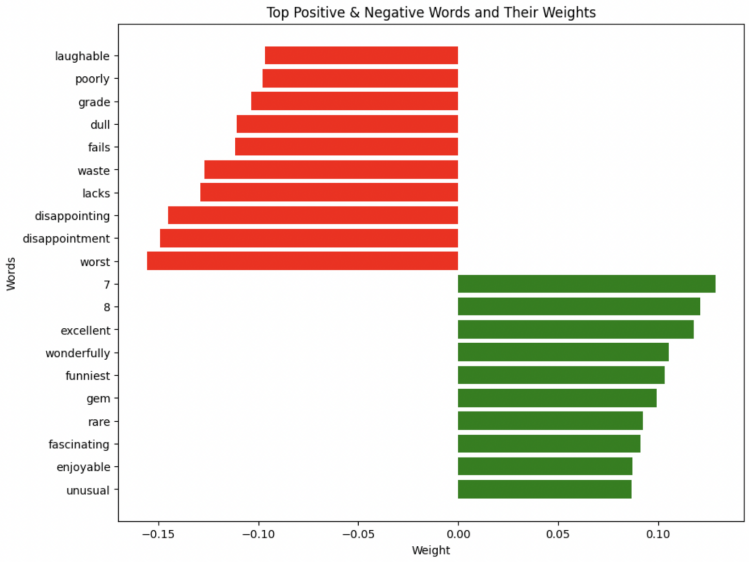

Figure 1: Top features based on linear regression coefficients.

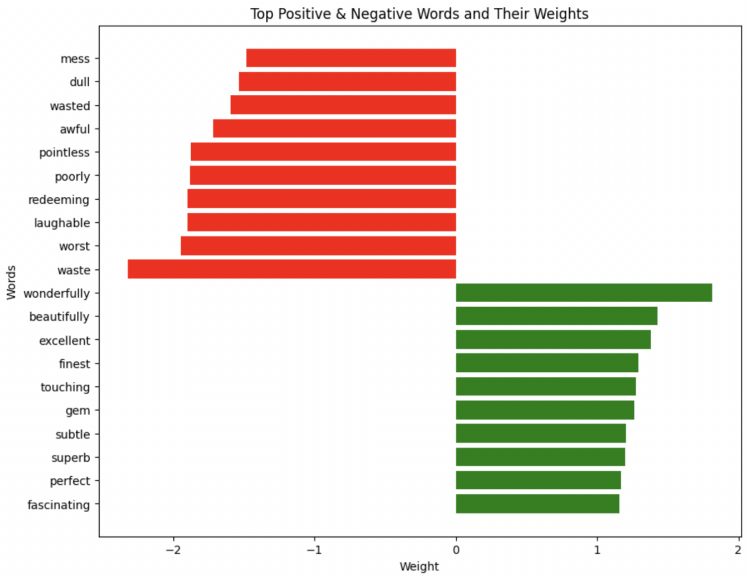

Figure 2: Top features based on logistic regression coefficients.

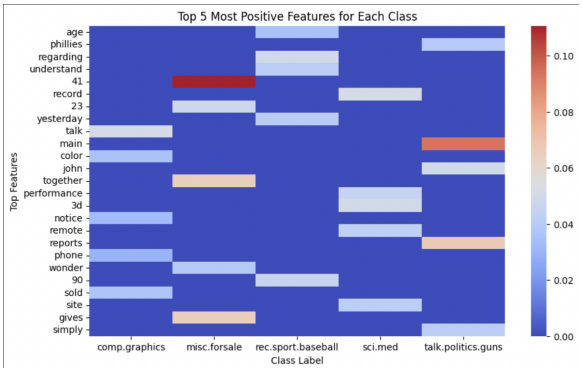

Figure 3: Heat Map of 5 Top Features for 5 News Groups

4.2 Model Implementation and Classification Performance

4.2.1 Binary Classification Performance on IMDB Reviews

In the binary classification task, we constructed a logistic regression model from to analyze the IMDB reviews dataset, aiming to make decisions between positive and negative movie reviews. After preprocessing the text data and training the model, we obtained an accuracy of 81% on the training dataset. When the model was applied to the test dataset, it maintained a high accuracy level of 80%, suggesting that our logistic regression model was effective in its predictive capacity and demonstrated minimal overfitting.

4.2.2 Multi-Class Classification Efficacy on the 20 Newsgroups Dataset

For the multi-class classification challenge, we made a Multiclass Linear Regression model to tackle the 20 Newsgroups dataset, which involved categorizing texts into one of five distinct topics. Each word was examined as a feature with MultiClass Linear Regression in order to determine which features correlate to which News Group. The 5 most positive, or more relevant, words/features were selected for viewing; Figure 3 shows a heatmap which represents this. The model's training involved one-hot encoding of the target labels and fitting the model to the training data. The subsequent testing phase gave us an accuracy of 74%, indicating that the model was reasonably successful in navigating the multi-class classification problem, though with expectedly less accuracy than the binary classification task due to the increased complexity of the problem.

Finally, Figures 9 and 10 show the training loss of the logistic and the multi-class linear regression respectively. While the loss of these two models do not converge together exactly, they both show a similar pattern, and each of their losses converge to a specific unmoving value after adequate iterations.

4.3 Comparative ROC Analysis for IMDb Sentiment Prediction

Figure 2 depicts ROC curves for the execution of Decision Tree and Logistic Regression models for classification of the IMDB data set. For extra information, the ROC curve of the scikit implementation of Logistic Regression was also tested. It functioned basically indentically to our implementation of Logistic Regression. The plot shows that while the

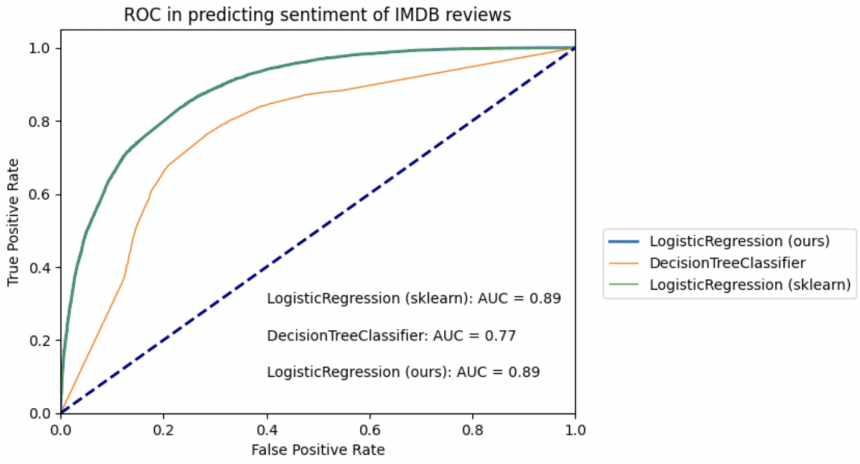

Figure 4: ROC curves of Logistic Regression and sklearn Decision Tree & Logistic Regression on the IMDB data

decision tree achieved a AUC of 0.77, the curve achieved from logistic regression surpassed it with a 0.89 area. It is possible that further careful adjusting of the decision tree model may allow it to reach a greater AUC, but we can conclude that logistic regression generally performs better with the IMDB dataset.

4.4 Accuracy Multiclass Regression vs. Decision Trees

In evaluating the performance of multiclass classification models, we use the 20 Newsgroups dataset subset, with five categories for classification. We implemented a multiclass linear regression model from scratch and also applied a decision tree classifier from sklearn for comparison. The multiclass linear regression model was trained on the processed feature matrix from the 20 Newsgroups dataset. The decision tree classifier was also trained on this dataset but used the original class labels instead of the one-hot encoded format required for multiclass regression. The accuracy of our multiclass linear regression model on the test dataset was approximately 73.6%, which indicates a relatively high performance level for classifying text data into one of several categories. This performance metric suggests that our model effectively captured the distinctions between different topics in the dataset. In contrast, the decision tree classifier gave an accuracy of approximately 63%, which is lower than that of the multiclass regression model. This difference in performance could be attributed to the inherent characteristics of decision trees of possible overfitting, especially when dealing with high-dimensional feature spaces, most likely common with textual data. The multiclass linear regression model outperformed the decision tree classifier on the 20 Newsgroups dataset, shown by the accuracy. This result shows the potential of linear models in handling complex classification tasks involving multiple classes,and shows the importance of choosing the right algorithm for text classification problems.

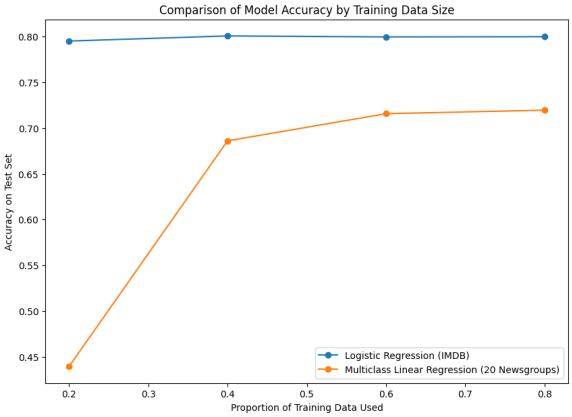

Figure 5: Comparison of Model Accuracy by Training Data Size

4.5 Data Size Impact on Model Accuracy

In this experiment, we investigated how the amount of training data affects the performance of two models: Logistic Regression for the IMDB dataset and Multiclass Linear Regression for the 20 Newsgroups dataset. The models were trained on varying proportions of the data: 20%, 40%, 60%, and 80%. We used a systematic approach where both models were trained and evaluated on subsets of the training data by setting a fixed random seed. For the IMDB dataset, the Logistic Regression model was evaluated based on accuracy, treating the sentiment classification task as binary. For the 20 Newsgroups dataset, the Multiclass Linear Regression model's accuracy was tested on the multi-class task, with the target labels one-hot encoded. The plot in Figure 5 illustrates the accuracy of each model on the test set as a function of the proportion of training data used. Results show that the Logistic Regression model on the IMDB dataset maintains a high level of accuracy across all data sizes, indicating robustness and its ability to perform well even with little data. In contrast, the Multiclass Linear Regression model for the 20 Newsgroups dataset shows a clear trend of increasing accuracy with more training data, suggesting that the complexity of the multi-class task benefits from larger datasets. For Logistic Regression on IMDB, he accuracy was consistently high, with a slight variation as the training data size increased. However, for Multiclass Linear Regression on 20 Newsgroups, the accuracy improved significantly from 20% to 60% of data usage, then plateaued, suggesting that beyond a certain point, additional data does not contribute to significant improvements in performance. The experiment showed us that Logistic Regression for binary classification tasks is less sensitive to the size of the training data than Multiclass Linear Regression for multi-class tasks. However, both models show improvement in performance with more data, up to a certain threshold. This insight important to understanding the data requirements of different models.

In the same vein, we aimed to evaluate how the size of the training data affects the performance of Logistic Regression and Decision Tree classifiers on the IMDB dataset. We measured performance in terms of AUROC (Area Under the Receiver Operating Characteristic curve) and accuracy, considering the same portions of the dataset as previously. We divided the IMDB dataset into different training sizes and used a fixed test set for

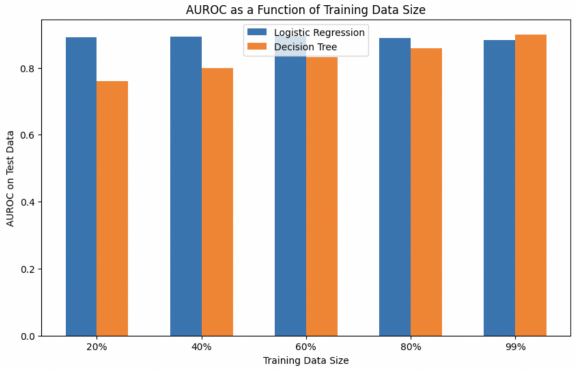

Figure 6: AUROC as a Function of Training Data Size

Figure 7: Accuracy as a Function of Training Data Size

evaluation. For each portion, we trained both Logistic Regression and Decision Tree models. Logistic Regression's probability estimates were used to calculate the AUROC, and the class predictions were used to find the accuracy. The Decision Tree's class predictions served both purposes.The bar plots show the performance metrics for both models across different training sizes: in Figure 4 6, the Logistic Regression model outperformed the DT across all training data sizes. The AUROC values are higher for Logistic Regression, indicating a better ability to sort positive and negative classes. In Figure 5 7, the accuracy trends were similar to AUROC, with Logistic Regression again achieving higher accuracy across all portions of training data compared to the DTs. The results show once again the robustness of Logistic Regression in handling binary classification tasks such as here sentiment analysis, as it had superior performance across all levels of training data availability. The DT's lower performance might be due to overfitting, especially when the training data is limited.

4.6 Compare with LASSO and Ridge Regression

We decided that implementing LASSO and Ridge regression would be beneficial to strengthen our analysis and findings related to Logistic Regression and the IMDb dataset. Ridge Regression adds "L2" regularization to the model, penalizing large coefficients. In contrast, LASSO regression adds "L1" regularization, which can set some coefficients to zero, effec-

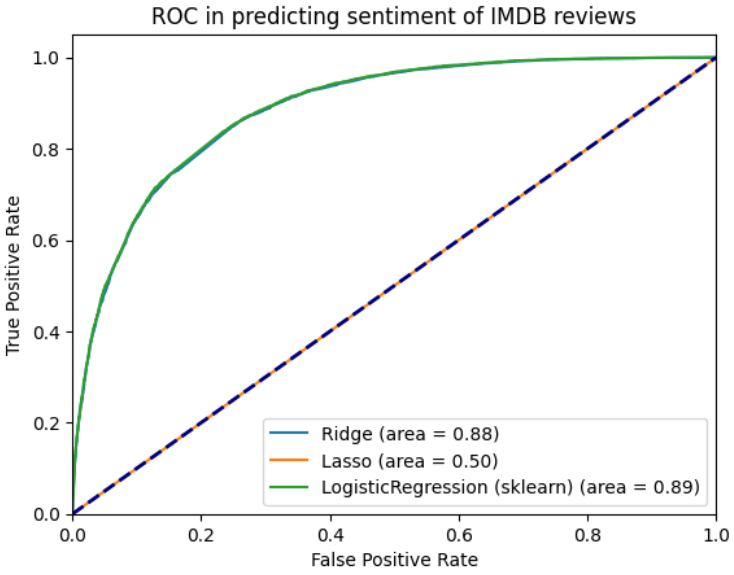

Figure 8: ROC in predicting sentiment of IMDb reviews

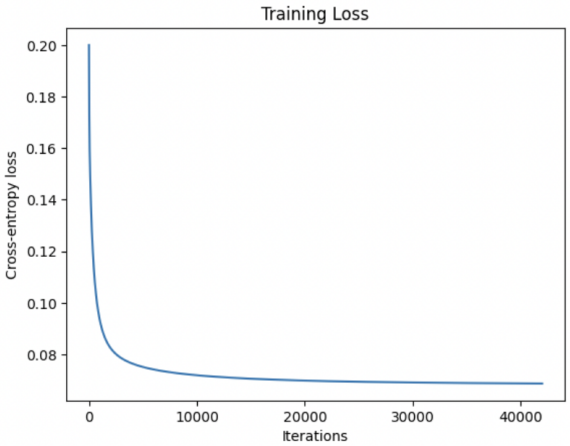

Figure 9: Multiclass Linear Regression Training Loss as a Function of Iterations

tively performing feature selection. We first began by importing the models from scikit-learn, We then calculated the MSE and accuracy for Ridge and LASSO, which were 0.15 and 0.8, and 0.25 and 0.5 respectively. We plotted these results' ROC curves against Logistic Regression's and still see that Logistic regression performs the best, strengthening our findings that is it ideal for binary classification tasks with textual data.

5. Discussion and Conclusion

This project embarked on an exploratory journey to evaluate and compare the efficacy of logistic regression, linear regression, and multiclass linear regression models in the context of

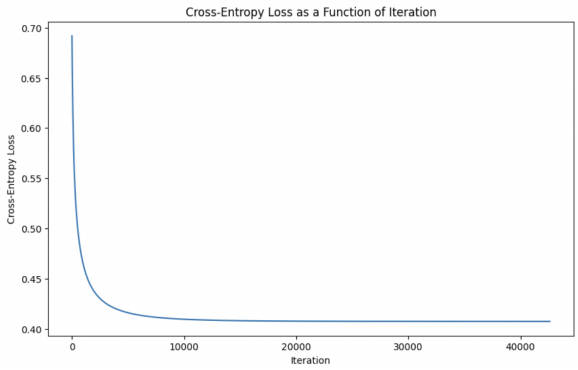

Figure 10: Logistic Regression Training Loss as a Function of Iterations

textual data classification. Through experimentation on two distinct datasets—the IMDB movie reviews for binary sentiment analysis and a subset of the 20 Newsgroups dataset for multiclass topic categorization—we have gained valuable insights into the strengths and limitations of these models, which have been backed up by our research efforts.

We conclude that logistic regression is a robust model for binary classification tasks, demonstrating superior performance in terms of accuracy and AUROC when benchmarked against Decision Trees on the IMDB dataset. Its success can be attributed to its ability to model the probability of class membership, thus making it particularly well-suited for binary classification problems where understanding the uncertainty of predictions is crucial. The top features identified by logistic regression intuitively align with positive and negative sentiments, underscoring the model's capability to discern key textual cues that distinguish between sentiments.

In contrast, linear regression, while not traditionally used for classification tasks, provided an interesting perspective when applied to the IMDB dataset. The model's performance, in terms of identifying the most positive and negative words, suggests that linear models can capture meaningful patterns in data. However, the gradient checking and monitoring of cross-entropy loss highlighted the challenges in directly applying linear regression to classification problems, especially in ensuring the correctness of gradient computations and the appropriateness of the loss function.

Multiclass linear regression, tailored for the more complex task of categorizing texts into multiple topics, showcased quite accurate results on the 20 Newsgroups dataset. Despite the inherent challenges of multiclass classification, the model achieved a high accuracy, affirming its potential for applications requiring fine-grained categorization. The exploration of the impact of training data size further revealed that while logistic regression maintains high accuracy across varying data sizes, indicating its robustness, multiclass linear regression benefits from larger datasets, suggesting a trade-off between model complexity and data requirement.

The comparative analysis of model performance highlighted an essential aspect of machine learning: the balance between model complexity and the availability of data. Logistic regression's consistent performance across different data sizes underscores its efficiency and robustness, making it a preferred choice for binary classification tasks with limited data.

On the other hand, the nuanced improvement in accuracy observed with multiclass linear regression as the training data size increases points to the model's dependency on sufficient data to effectively capture the complexity of multiclass problems.

In conclusion, our study of logistic regression, multiclass regression, and Decision Trees on two textual datasets has given results explaining the models' performance and utility. Our experimentation and analysis lead to key takeaways: logistic regression stands out for its power in binary classification tasks, aligning with the literature of Tripathi et al., 2020 on sentiment analysis. The multiclass regression model also performs well, particularly when considering the complexity of the task of categorizing documents into multiple classes. The top features identified by the models are intuitively relevant to their respective sentiments and topics, reinforcing the models' validity. Moreover, the additional experiments with LASSO and Ridge regression further validate logistic regression's supremacy in handling binary classification of textual data. For future research, we suggest investigating ensemble methods, neural network approaches, and delving deeper into hyperparameter optimization to potentially enhance model performance further.

Building upon the findings of this study, future investigations could delve into several promising directions:

1. **Ensemble Methods**: Exploring ensemble techniques that combine predictions from multiple models to improve overall accuracy and robustness. 2. **Deep Learning Approaches**: Experimenting with neural network architectures, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), for more sophisticated feature extraction and classification capabilities. 3. **Hyperparameter Optimization**: Utilizing techniques such as grid search and random search to fine-tune model parameters and achieve optimal performance.

The interplay between model complexity, data size, and feature selection continues to be a fertile ground for research, promising to unveil more nuanced understandings and innovative solutions in the classification of textual data.

6. Statement of Contributions

For this project, all three team members first agreed on our list of experiments, including the extra ones we would be conducting. Once decided, Rebecca started working on task 1 with the IMDb dataset, while Samantha started working on task 2 with the 20 Newsgroups dataset, which involved the data acquisition and preprocessing for each dataset respectively, as well as the Logistic Regression and Multiclass regression model implementations. Amélie, Samantha and Rebecca then divided amongst them the implementations for the suite of experiments for task 3. Amélie also commented and adjusted the formatting of our Notebook so that it was ready for submission. All three team members then created an outline for the report together, and wrote the sections on the parts they implemented.

References

S. Tripathi, R. Mehrotra, V. Bansal and S. Upadhyay, "Analyzing Sentiment using IMDb Dataset," 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN), Bhimtal, India, 2020, pp. 30-33, doi: 10.1109/CICN49253.2020.9242570. keywords: Radio frequency;Sentiment analysis;Forestry;Motion pictures;Regression tree anal-

ysis;Random forests;Logistics;Sentiment Analysis;IMDb reviews dataset;Bag of Words;Logistic Regression;Naive Bayes;Decision Tree;Random Forest,

S. Joshi and E. Abdelfattah, "Multi-Class Text Classification Using Machine Learning Models for Online Drug Reviews," 2021 IEEE World AI IoT Congress (AIIoT), Seattle, WA, USA, 2021, pp. 0262-0267, doi: 10.1109/AIIoT52608.2021.9454250. keywords: Drugs;Medical conditions;Text categorization;Static VAr compensators;Support vector machine classification;Vegetation;Predictive models;Drug Review;Machine Learning;Text Classification;Disease Classification,

arunmohan_003, "Perturbation test On Logistic Regression", 2021, https://www.kaggle.com/code/arunmoh
test-on-logistic-regression