Assignment2COMP551- Group 19

Ridwanur Rahman, François Pagé, Isabella Chiaravallotti

February 28, 2024

1 Abstract

For this experiment, we tested logistic regression for sentiment prediction of movie reviews (IMDB dataset), and multiclass regression for categorization of news documents (20 news groups dataset). We also compared our models against the scikit-learn decision tree implementation. While preparing these datasets for investigated the importance of feature selection and found that we can greatly reduce the dimensionality of our dataset by carefully selecting features. We found that for the IMDB dataset, both logistic regression and decision tree performed well with AUROC calculated as 0.99 and 1.0, respectively.

For the 20 newsgroups dataset, we found that both multiclass regression and decision tree performed relatively poorly, although decision tree performed slightly better than multiclass regression. We found the validation accuracy of the multiclass regression to be 45.0%. Subsequently, we found the training and validation accuracy for scikit-learn decision tree to be 53.24% and 50.15% respectively.

2 Introduction

In the age of social media and continuous news cycles, analysis of text-based data is becoming increasingly pertinent. In this experiment we performed sentiment prediction and class prediction on two text-based datasets(the IMDB movie review dataset and the 20 newsgroups dataset) using logistic regression and multi-class regression. Both datasets consist of text files and their corresponding sentiment (positive or negative) or category(alt.atheism, comp.graphics, rec.sport.baseball,sci.med and talk.politics,misc). More on these datasets can be found below in the "Datasets" section.

As natural language processing technology has advanced, these datasets have been utilized to test different methods and hypotheses. For example, Shakut et. al (2020) used the IMDB movie review dataset to train a neural network that can be used to achieve "opinion mining" and to identify positive and negative words. The IMDB set has become a classical dataset for comparing different machine learning methods for sentiment analysis and emotion identification, demonstrated by Basa (2023), Dahir (2023), and Palomo (2023) to name a few. Authors found that when comparing models, some are able to perform better than others on the task of sentiment analysis. The above authors found support vector machine, logistic regression, and transformer based models, respectively, as the top performers. The authors also note that using a term-frequency method for feature selection tends to be better than the traditional bag-of-words method.

Similarly, the 20 newsgroup dataset serves as a good basis for investigating different machine learning techniques. Recently, Saigal (2020) used the dataset to test different implementations of the support vector machine model. The authors elaborated on how the SVM model can be used to categorize text, noting trade-offs between computational time and model performance. Another experiment conducted by Raj (2022) did a similar investigation in which they assessed different clustering methods for news group classification. They found that clustering based on key words was effective for identifying the correct news class and identify co-clustering as ideal (as opposed to hierarchical clustering).

The literature indicates that these two datasets serve as a good benchmark for investigating machine learning models that can perform sentiment analysis and categorization.

3 Datasets

For this experiment we looked at two datasets, the IMDB Reviews dataset and the 20 newsgroups dataset (Maas, 2011; Scikit-learn). The IMDB dataset consists of movie reviews (text files), the sentiment about the review (either positive or negative) and the numerical review (on a scale of 1 to 10). 50,000 reviews were collected, and split evenly between training and testing. The 20 newsgroups set includes 18000 posts on 20 different topics. For our experiment, we selected 5 topics. The training and testing sets are separated by a specific data.

4 Results

For task 1 we conducted data pre-processing. Pre-processing involved organizing the data into a usable format, and selecting the most important features for modeling. Both sets are very large and filtering features was necessary to reduce noise and

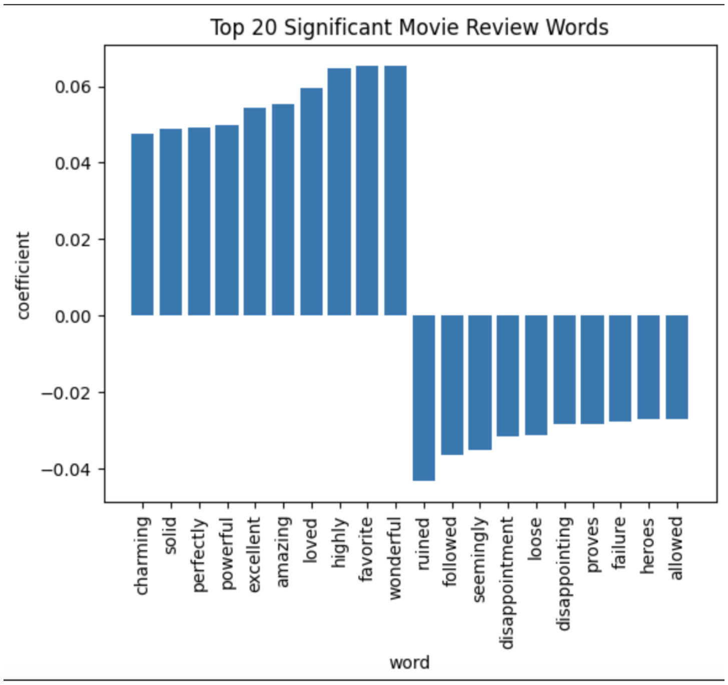

Figure 1: This bar chart show the twenty most significant words according to our linear regression on the IMDB reviews data. We can see that the words with the most positive coefficients are associated with positive reviews and the words with the most negative coefficients are associated with negative reviews

reduce computational time.

For the IMDB Reviews dataset, we began by loading the text files for the training and testing set. Then, we utilized CountVectorizer(scikit-learn) to turn the text files into a matrix of word counts, quantifying how many times each word present in the set was used for each review. Transforming the data also included filtering. We removed stop words, or commonly used words, as well as rare words as these words will have negligible impact on the model. Next, we defined and trained a linear regression model to get an idea of which of the remaining words would be most important for our model. We took the absolute values of the linear regression coefficients to identify which words had a very negative or very positive correlation with the sentiment of the review (1 being positive and 0 being negative). We took the top 500 features with the greatest absolute linear regression coefficient to be used in our model. The words with the 10 most negative and 10 most positive coefficients were plotted (Figure 1). We can see that these words meet expectations when it comes to describing a positive or negative review. For example, some of the positive words are ”charming”, ”excellent”, and ”perfectly” while some of the negative words are ”ruined”,”disappointment”, and ”failure”.

Next, for the 20 newsgroup set we chose 5 news groups:’alt.atheism’,’comp.graphics’,’rec.sport.baseball’,’sci.med’, and ’talk.politics.misc’. We also converted the next data into a matrix of word counts while filtering for stop words and rare words. The groups were re-coded from their name to a value from 0:4.

Before selecting features, we visualize the data using a principal components analysis (PCA) (Figure 2). In short, a PCA provides a summary of the data by plotting the features that explain the most variance. The observations in the dataset will cluster according to the principal components. We were able to confirm that our news groups clustered separately,indicating that we should be able to clearly distinguish them based on the words used in each observation.

Next, we chose which features to use for our model. To determine feature importance, we calculated correlations separately on each group. We achieved this by transforming the response variable to 1 for the first group and 0 for the other four groups. Then we transformed the response variable to 1 for the second group (group 1) and 0 for the other four groups, and so on. We identified the features (words) most highly correlated with each news group. Then we took the 5 most correlated words from each group, and determined their correlation with the other four groups. This allowed us to create a heat map showing how each selected word correlates with each news group (Figure 3). These 25 words were selected as the features for our model.

Task 2 involved implementing the two models, Logistic Regression and Multiclass Regression, and Task 3 involved using our implementations to run experiments, evaluate accuracies and compare them.

Figure 4 shows cross-entropy loss for logistic regression. The learning rate was 0.6, epsilon was 0.005 and maximum iterations was set to 10000. We also monitored cross entropy as a function of iteration (Figure 6). We can see that loss is lower when we add bias (addbias : true) and test loss was similar to training loss. Due to the lack of a spike in the cross entropy on the validation set across iterations, we can conclude that over-fitting wasn’t yet reached and therefore, a more accurate model would be possible given more compute time.

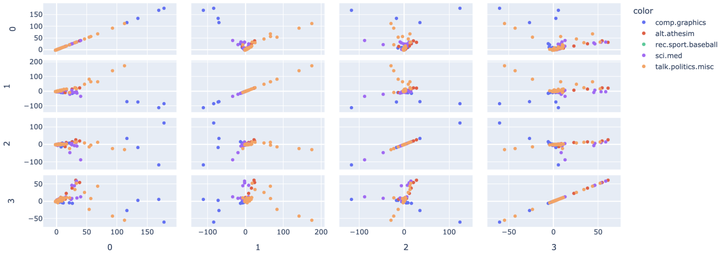

Figure 2: This figure shows a principle components analysis performed on the 20 News Group Dataset. The PCA allows us to see how the different categories cluster together according to the features that explain the most variance

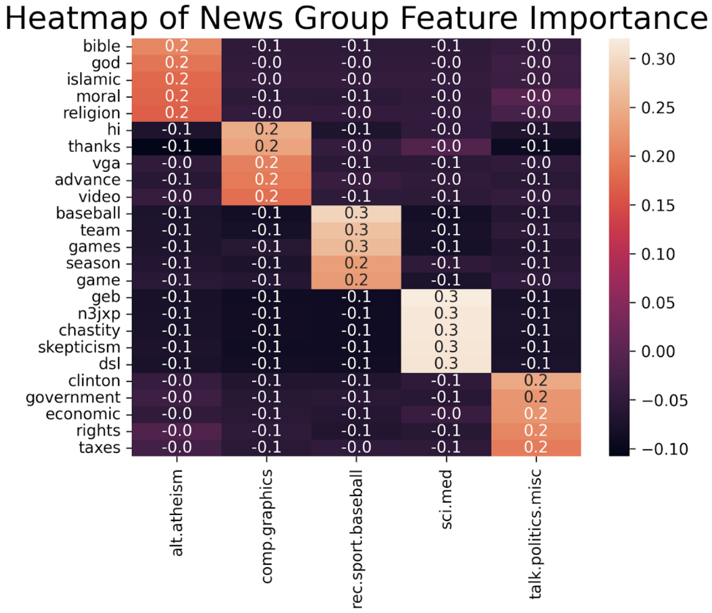

Figure 3: This figure depicts a heat map for the selected features in the 20 News Group Dataset

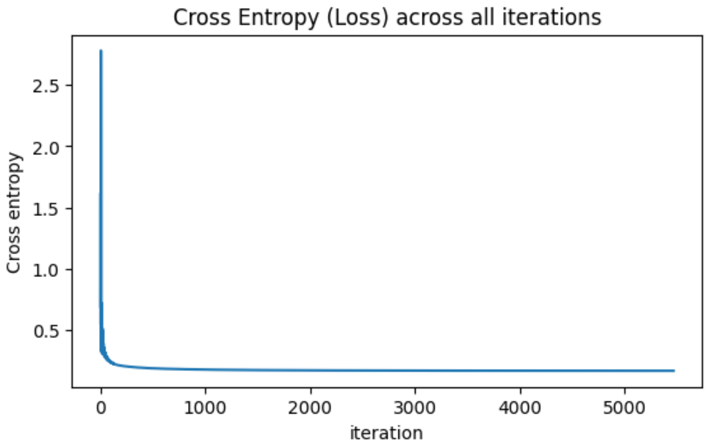

Figure 4: cross entropy loss for all iterations for logistic regression

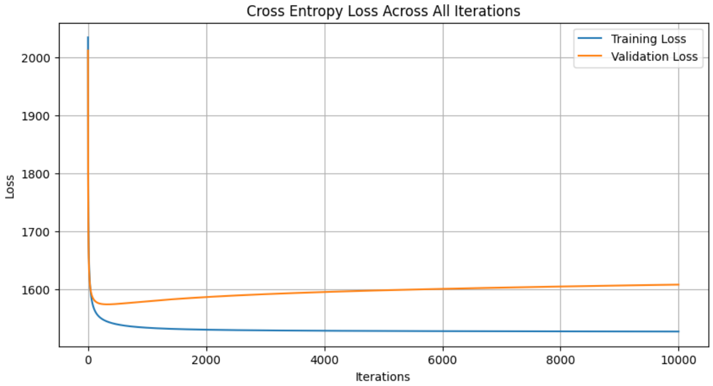

Figure 5: cross entropy loss for all iterations for multiclass regression

We observed the features with 10 highest and 10 lowest coefficients to understand which features had the greatest impact on our logistic regression model. These features and their coefficients can be seen in Figure 7.

We compared our logistic regression model to the scikit-learn decision tree (Figure 8). The two models performed similarly, with the AUROC for decision tree being 1.0 and the AUROC for our logistic regression being 0.99. We assessed the impact of training set size of accuracy for decision tree and logistic regression. We found that training set size had a greater impact on logistic regression than it did for decision tree, seen in Figure 10.

Figure 5 shows cross-entropy loss for Multiclass regression. The learning rate was 0.005, epsilon was 0.0001 and maximum iterations was set to 10000. The plot reveals a rapid decline in both training and validation loss, which quickly plateaus, indicating that the model is converging to a solution. The close proximity of the training and validation loss curves throughout the training process suggests that the model is generalizing well, as there is no significant divergence indicative of over-fitting. Given the absence of over-fitting and the stabilized loss values, it is reasonable to infer that extending the number of iterations beyond 10,000 is unlikely to yield substantial improvements in model performance.

We found the validation accuracy of the Multiclass regression to be 45.0%. Subsequently, we found the training and validation accuracy for scikit-learn Decision Tree to be 53.24% and 50.15% respectively. We then compared our Multiclass regression model to the scikit-learn Decision Tree (Figure 8). As the training set size increases, both models exhibit a slight improvement in accuracy, indicating a positive correlation between the amount of training data and model performance. Notably, the Decision Tree model consistently outperforms the Multiclass Regression model across all training set sizes. This performance difference could be attributed to the Decision Tree's ability to capture non-linear patterns within the data, which may not be as effectively represented by the linear boundaries of the Multiclass Regression model.

5 Discussion and Conclusion

The investigation into sentiment prediction and document classification presented distinct outcomes for logistic regression, multiclass regression, and decision tree models. Logistic regression exhibited robust performance on the IMDB dataset, paralleling decision tree results with high AUROC scores. Contrarily, the multiclass regression model demonstrated moderate efficacy on the 20 newsgroups dataset with a validation accuracy of 45.0%, slightly lagging behind the decision tree's per-

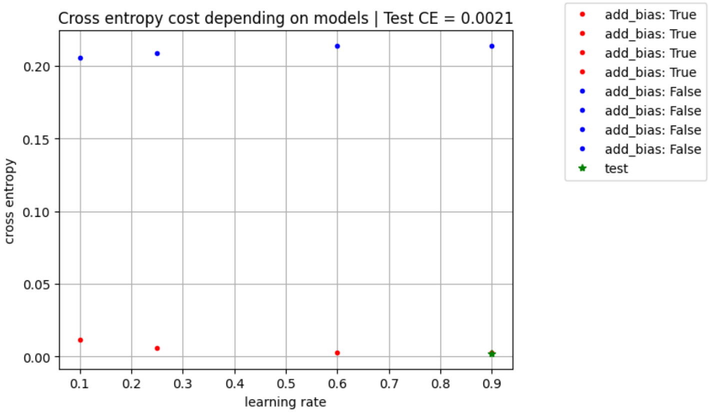

Figure 6: The model has constant hyper parameters: Max iteration = 10000, epsilon = 0.001. As can be seen on the plot, a constant bias offset is needed to accurately model the data.

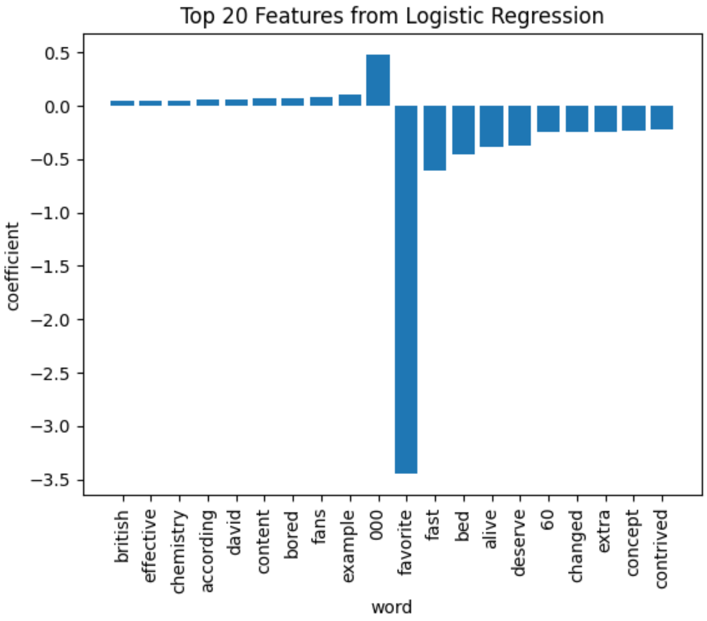

Figure 7: top 20 features from Logistic Regression

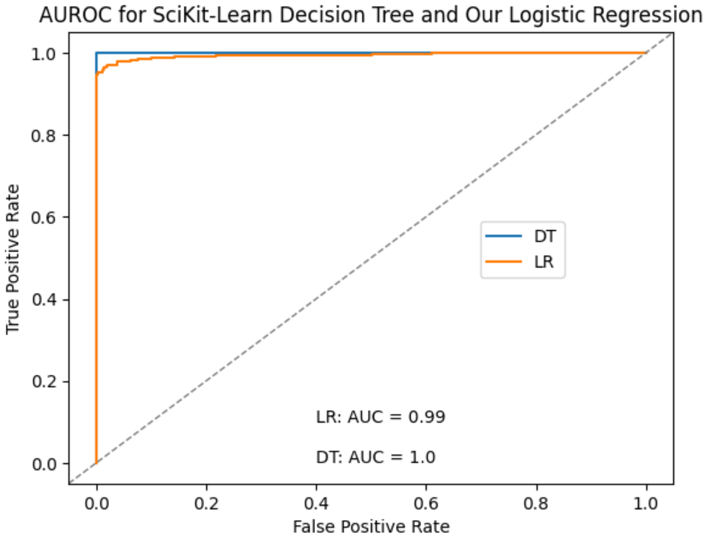

Figure 8: AUROC for the scikit-learn decision tree versus our logistic regressionn

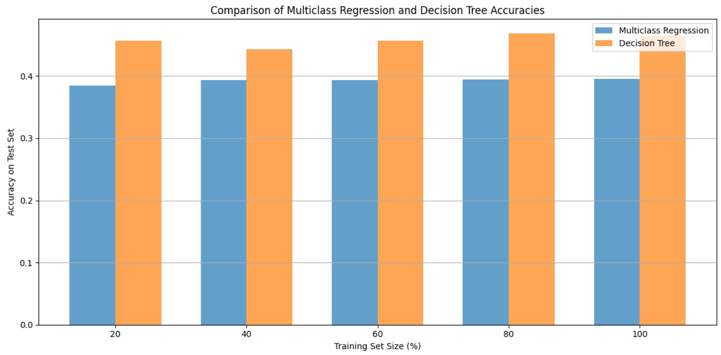

Figure 9: bar plot comparison of model accuracies for different training set sizes Decision Tree and Multiclass Regression

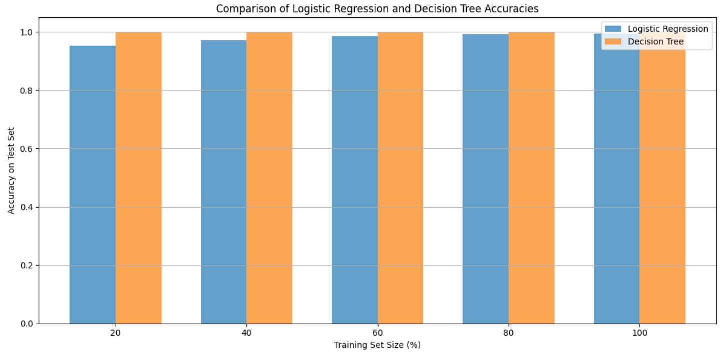

Figure 10: bar plot comparison of model accuracies for different training set sizes for logistic regression and decision tree

formance. Incremental training data bolstered model accuracy, indicating the value of expansive training sets for enhancing model predictions. Nevertheless, the decision tree consistently outstripped the multiclass regression, potentially due to its proficiency in navigating the data's non-linear characteristics. Cross-entropy analysis underscored the absence of overfitting and the convergence of models, suggesting optimization of current iterations rather than extended compute time. Collectively, these insights affirm the applicability of these models to text data analysis while also highlighting the inherent advantages of decision trees in handling complex patterns.

6 Statement of Contributions

Isabella conducted the pre-processing, ran experiments, and contributed to the report. Francois implemented logistic regression, ran experiments on it and contributed to the report. Ridwanur implemented multiclass regression, ran experiments on it and contributed to the report.

7 References

Başa, Selen Nazli, and Muhammet Sinan Basarslan. ”Sentiment Analysis Using Machine Learning Techniques on IMDB Dataset.” 2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). IEEE, 2023.

Dahir, Ubaid Mohamed, and Faisal Kevin Alkindy. ”Utilizing machine learning for sentiment analysis of IMDB movie review data.” International Journal of Engineering Trends and Technology 71.5 (2023): 18-26.

Maas, Andrew L.,Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Palomo, Beatriz Alejandra Bosques, et al. ”Sentiment Analysis of IMDB Movie Reviews Using Deep Learning Techniques.” International Congress on Information and Communication Technology. Singapore: Springer Nature Singapore, 2023.

Raj, A., Susan, S. (2022). Clustering Analysis for Newsgroup Classification. In: Bhateja, V., Khin Wee, L., Lin, J.CW., Satapathy, S.C., Rajesh, T.M. (eds) Data Engineering and Intelligent Computing. Lecture Notes in Networks and Systems, vol 446. Springer, Singapore. https://doi.org/10.1007/978-981-19-1559-8$_2$8

Saigal, P., Khanna, V. Multi-category news classification using Support Vector Machine based classifiers. SN Appl. Sci. 2, 458 (2020). https://doi.org/10.1007/s42452-020-2266-6

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

Shaukat, Zeeshan, et al. ”Sentiment analysis on IMDB using lexicon and neural networks.” SN Applied Sciences 2 (2020): 1-10.