HW1 Report

Yifei Gan

1 Introduction

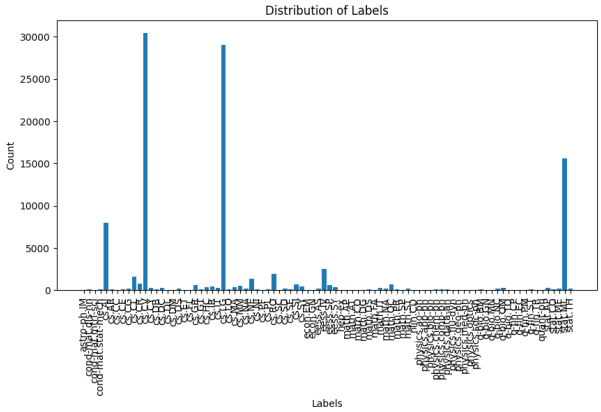

Figure 1: Distribution of Labels

This dataset contains 88 labels and here's the distribution. Five different models have been used to train. Each of the micro and macro F1 scores are provided, along with the validation classification report. I use optuna to find the best parameters for each model.

2 Naive Bayes

Metric	Precision	Recall	F1-Score	Support
Micro Avg	0.80	0.67	0.73	15353
Macro Avg	0.11	0.06	0.06	15353
Weighted Avg	0.72	0.67	0.68	15353
Samples Avg	0.86	0.76	0.77	15353

Table 1: Naive Bayes

Micro F1 Score: 0.7298355520751761
Macro F1 Score: 0.06243270237959762
Running time: 20.2 seconds.

3 Linear SVC

Metric	Precision	Recall	F1-Score	Support
Micro Avg	0.83	0.71	0.76	15353
Macro Avg	0.78	0.35	0.45	15353
Weighted Avg	0.81	0.71	0.74	15353
Samples Avg	0.87	0.79	0.79	15353

Table 2: Linear SVC

Micro F1: 0.7614123936132798
Macro F1: 0.44832968113393307
Running time: 43.8 seconds.

4 Logistic Regression

Metric	Precision	Recall	F1-Score	Support
Micro Avg	0.82	0.69	0.75	15353
Macro Avg	0.39	0.16	0.20	15353
Weighted Avg	0.78	0.69	0.72	15353
Samples Avg	0.86	0.78	0.78	15353

Table 3: Updated performance metrics for the classification model.

Micro F1: 0.7517519643236356
Macro F1: 0.20467415361754257
Running time: 38.0 seconds.

5 Random Forest

Metric	Precision	Recall	F1-Score	Support
Micro Avg	0.91	0.74	0.82	15353
Macro Avg	0.87	0.32	0.43	15353
Weighted Avg	0.92	0.74	0.79	15353
Samples Avg	0.93	0.81	0.84	15353

Table 4: Updated performance metrics for the classification model.

Micro F1: 0.8162366949305431
Macro F1: 0.4330467158500335
Running time: 33 minuets 23 seconds.

6 KNN

Metric	Precision	Recall	F1-Score	Support
Micro Avg	0.78	0.64	0.70	15353
Macro Avg	0.24	0.08	0.11	15353
Weighted Avg	0.71	0.64	0.66	15353
Samples Avg	0.83	0.73	0.74	15353

Table 5: Updated performance metrics for the classification model.

Micro F1: 0.702172741964446
Macro F1: 0.11061668921111507
Running time: 32 minuets 27.4 seconds

7 Comparison

For this task, the features were created using a CountVectorizer with unigram and bigram representations. By utilizing unigrams (single words), the models could capture essential keywords, while bigrams allowed the representation of short contextual phrases. The n-gram approach was particularly useful for distinguishing between technical terms and sequences that provided context crucial for topic differentiation in arXiv paper abstracts. Additionally, I performed some basic text preprocessing, such as lowercasing, removing punctuation, and eliminating stopwords, which reduced noise and improved the overall quality of the features.
Across all the models evaluated, the micro and macro F1 scores presented distinct differences, which can be traced back to the imbalance in label frequency and the specific characteristics of each model. The micro F1 score, which prioritizes overall accuracy across all samples, was consistently higher than the macro F1 score, which calculates the average performance per class without accounting for label frequency. For instance, the Linear SVC model achieved a micro F1 score of 0.76, while its macro F1 score was only 0.45. The substantial gap indicates that while the model was effective at predicting the more common labels, it struggled with the rare ones, highlighting a performance imbalance.

7.1 Naive Bayes

This model had the lowest macro F1 score (0.06) and a relatively lower micro F1 score (0.73). Naive Bayes' assumptions about feature independence and simple probabilistic framework seem insufficient for dealing with the complexity and contextual interdependencies in the abstract texts. The low macro F1 score reflects the model's poor ability to predict rare labels, potentially due to its tendency to overfit to the more frequently occurring terms, neglecting the minor but important ones.

7.2 Linear SVC

Linear SVC showed a higher micro F1 score (0.76) and an improved macro F1 score (0.45) compared to Naive Bayes. The use of n-grams improved the model's ability to understand short phrases, which contributed to better classification of documents with technical jargon. However, the macro F1 score suggests it still struggled to balance predictions across all labels effectively, with frequent labels overshadowing rare ones.

7.3 Logistic Regression

Logistic Regression achieved a micro F1 score of 0.75 and a macro F1 score of 0.20. The model performed decently but faced difficulties distinguishing less frequent classes, resulting in a low macro F1 score. The lack of non-linear decision boundaries likely hindered its ability to capture some of the complex relationships inherent in the features, despite the use of bigrams.

7.4 Random Forest Classifier

Random Forest achieved the highest micro F1 score of 0.82 and a macro F1 score of 0.43. This model benefitted from its ensemble nature, allowing it to effectively capture interactions between features. The higher macro F1 score compared to other models indicates better performance on less frequent labels, though the disparity between micro and macro F1 still points to the challenges posed by the imbalance in the dataset.

7.5 K-Nearest Neighbors (KNN)

KNN performed the worst in terms of micro F1 score (0.11) and had a macro F1 score of 0.70. KNN's reliance on calculating distances in a high-dimensional space made it more sensitive to irrelevant features and noise, which, coupled with the large number of labels, led to suboptimal performance, especially for rare labels.

Given the results, classifying all 88 labels may not be the most efficient approach, primarily due to the highly imbalanced nature of the dataset. The low macro F1 scores across all models, such as Naive Bayes (0.06) and KNN (0.11), illustrate the difficulty in accurately predicting rare labels.

Even the best-performing Random Forest model showed a macro F1 score of 0.43, highlighting significant limitations. A more focused approach that targets a subset of more frequent labels or employs specialized techniques like hierarchical classification or label grouping might yield better results.