220 Assignment 3

Shriya Sravani Y
UCSC
sy4@ucsc.edu

1 Introduction

In this report, I will be processing data, engineering features and training models to predict which category or categories an arXiv paper's abstract belongs in. This is a multi-label classification problem.

2 Dataset Preprocessing

The given dataset is a json file named arxiv-data.json. It contains 3 columns,

- title: The title of the paper.
- abstract: A brief summary of the paper, serving as the primary input feature (X).
- terms: A list of labels indicating the categories the paper belongs to (y).

The dataset consists of 51774 rows.

The training set is 70% on the dataset(36241 rows). The validation set is 15% and test set is 15% of the dataset.

This dataset exhibits a class imbalance as it can be seen in the graph given below.

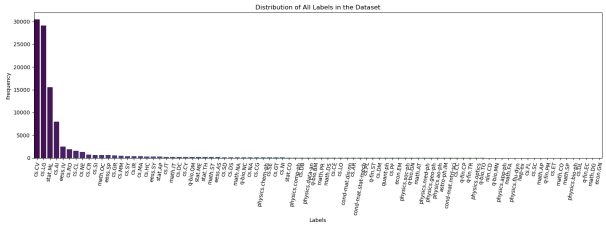

Figure 1: Distribution of Review Scores

The major classes are as follows:

Category	Frequency
cs	78,689
stat	16,402
eess	3,654
math	1,581
q-bio	578
physics	542
cond-mat	146
q-fin	138
econ	49
quant-ph	47
astro-ph	31
nlin	25
hep-ex	20

Table 1: Major categories and their frequencies in the dataset.

The top 10 frequent classes are displayed below:

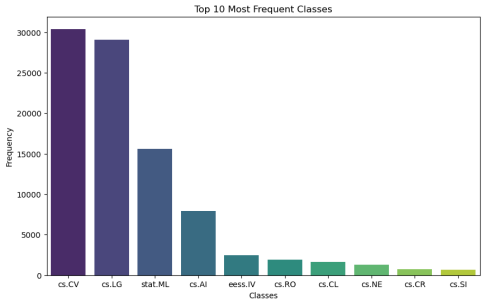

Figure 2: Distribution of Review Scores

The labels are encoded using a multi-label binarizer.

3 Training and Feature Engineering

3.1 Naive Bayes model

This model uses the TF-IDf vectorization with a max_features of 5000 and stop_words in english. The micro average precision score is 0.83, recall is 0.64 and the F1 score is 0.72. The macro F1 score is 0.03.

Class	Precision	Recall	F1-Score	Support
cs.CV	0.93	0.92	0.93	4585
stat.ML	0.66	0.68	0.67	2347
Accuracy			0.50	15353
Macro Avg	0.06	0.03	0.04	15353
Weighted Avg	0.71	0.64	0.65	15353

Table 2: Classification Report(Naive Bayes

- **Validation Micro F1:** 0.7229
- **Validation Macro F1:** 0.0355
- **Training Time:** 0.76 seconds
- **Inference Time:** 0.30 seconds

3.2 Logistic Regression

This model encodes labels using the multi-label binarizer and uses the tf-idf word embedding method with max_features = 5000.

The micro F1 score for this model is 0.74 and the macro F1 score is 0.07. It has a better micro F1 score as compared to the Naive Bayes showing that it performs better on individual classes.

Class	Precision	Recall	F1-Score	Support
cs.CV	0.95	0.93	0.94	4585
stat.ML	0.71	0.65	0.68	2347
Accuracy			0.51	15353
Macro Avg	0.13	0.04	0.07	15353
Weighted Avg	0.75	0.64	0.67	15353

Table 3: Classification Report(Logistic Regression

- **Validation Micro F1:** 0.7350
- **Validation Macro F1:** 0.0722
- **Training Time:** 7.42 seconds
- **Inference Time:** 0.06 seconds

The micro F1 score considers all instances across all classes equally, irrespective of class imbalance. It focuses on the overall correctness of predictions. A score of 0.7350 shows the model performs reasonably well in predicting labels when considering all classes, weighted by their frequency in the dataset.

3.3 Decision Tree

A score of 0.70 suggests the model has reasonable overall performance when considering the dataset as a whole.It is able to capture the model's performance across all samples proportionally to their frequency.

- **Validation Micro F1:** 0.7064
- **Validation Macro F1:** 0.2463
- **Training Time:** 222.63 seconds
- **Inference Time:** 0.12 seconds

Class	Precision	Recall	F1-Score	Support
cs.CV	0.92	0.88	0.90	4585
stat.ML	0.56	0.81	0.66	2347
Accuracy			0.44	15353
Macro Avg	0.39	0.21	0.25	15353
Weighted Avg	0.72	0.66	0.67	15353

Table 4: Classification Report(Decision Tree)

3.4 SVM

This model uses the LinearSVC and Tf-Idf vectorizer.

Class	Precision	Recall	F1-Score	Support
cs.CV	0.95	0.93	0.94	4585
stat.ML	0.69	0.66	0.68	2347
Accuracy			0.45	15353
Macro Avg	0.74	0.28	0.37	15353
Weighted Avg	0.81	0.69	0.72	15353

Table 5: Classification Report(LinearSVC)

- **Validation Micro F1:** 0.7559
- **Validation Macro F1:** 0.3684
- **Training Time:** 11.80 seconds
- **Inference Time:** 0.06 seconds

A score of 0.7559 indicates the model performs reasonably well on a per-instance basis, weighted by the frequency of all classes. It is sensitive to classes with low precision or recall, as seen here for stat.ML. At 45.61%, the validation accuracy aligns with the weighted performance across all classes.

The model has an efficient training and inference times.

3.5 Random Forest

It creates multiple decision trees during training by bootstrapping from the dataset.

It combines predictions by majority voting (for classification). Each tree is independent of the others.

Class	Precision	Recall	F1-Score	Support
cs.CV	0.94	0.95	0.95	4585
stat.ML	0.78	0.73	0.76	2347
Accuracy			0.62	15353
Macro Avg	0.87	0.32	0.43	15353
Weighted Avg	0.93	0.73	0.78	15353

Table 6: Classification Report(Random Forest)

- **Validation Micro F1:** 0.8136
- **Validation Macro F1:** 0.4319
- **Training Time:** 51.52 seconds
- **Inference Time:** 0.94 seconds

A score of 0.81 indicates that the model performs well in terms of the overall correctness of its predictions, weighted by the number of samples. A score of 0.4319 indicates that the model's performance varies significantly between classes.

The training process took 51.52 seconds, which is relatively longer than simpler models. This is likely due to the complexity of the model, such as the number of features. The inference process required 0.94 seconds to predict the labels for the entire validation set. This is efficient given the size of the dataset (15,353 samples).

4 Overall Performance

Metric	NB	LR	DT	SVM	RT
Micro F1 Score	0.72	0.73	0.70	0.75	0.81
Macro F1 Score	0.03	0.05	0.24	0.38	0.43
Training Time (s)	0.76	7.42	222.63	11.8	51.52
Inference Time (s)	0.30	0.06	0.12	0.06	0.94

Table 7: Performance Metrics for Different Classifiers

5 Best Model

The random forest gives the best performance on the validation set so I ran it on the test set. It is run on tf-idf vector and gives a result similar to the validation set values.

Class	Precision	Recall	F1-Score	Support
cs.CV	0.95	0.95	0.95	4530.0
math.SP	0.00	0.00	0.00	1.0
Accuracy			0.61	
Macro Avg	0.84	0.29	0.41	
Weighted Avg	0.93	0.73	0.78	

Table 8: Test Classification Report (Random Forest)

- **Training Time:** 24.74 seconds
- **Inference Time:** 0.54 seconds

The accuracy on the test set (61%) is very close to that of the validation set (62%), indicating that the model generalizes reasonably well. However, accuracy is not an ideal metric for this imbalanced dataset as it disproportionately reflects performance on majority classes.

Micro F1 is slightly lower on the test set (80%) than on the validation set (81%). The macro F1 score drops from 43% on validation to 41% on the test set. This metric is more sensitive to imbalances, and the decline indicates that the model is struggling with minority classes (math.SP) in the unseen test data.

6 Analysis

Overall the **Random Forest** classifier has performed the best in terms of the micro and macro F1 scores.

Micro F1 score measures how well the models perform overall. The scores are relatively high across all models, with Random Forest (RF) achieving the best score of **0.81**, indicating that it handles the dominant classes well. Naive Bayes (NB) and Decision Trees (DT) perform slightly worse due to their simplicity and sensitivity to class imbalances.

Macro F1 score highlights the model's performance on underrepresented classes. Random Forest (RF) and SVM perform the best, scoring **0.43** and 0.38, respectively. This indicates that they handle minority classes better than other models.

Impact of Features: The use of **TF-IDF** for feature extraction positively impacted all models, especially SVM and RF, because these algorithms benefit from high-dimensional and sparse feature spaces. TF-IDF captures the importance of words across abstracts while reducing noise.

SVM and Random Forest classifiers are more robust to noise and class imbalance, which positively influenced both micro and macro F1 scores.

As for classifying all 88 labels, it may not be worth as most of the classes are underrepresented. Though the micro F1 scores remain reasonable, the macro F1 scores are low for few of the models, indicating poor performance on minority classes.