# Question 1

1. Download the data here: TRDataChallenge2023.zip. This zip file contains a single text file in JSON Lines format.
2. Programmatically load the data into your preferred analytical environment.
3. Report the number of documents, postures, and paragraphs in the dataset.

**Number of Documents: 18000**

**Number of Postures: 27659**

**Number of Paragraphs: 542169**

4. Describe the data and any aspects relevant to the modeling task in Question 2.

The provided dataset with each line representing a document containing metadata and textual information.

- **Document Metadata:**
  - `documentId`: Unique identifier for each document.
  - `postures`: A list of procedural postures associated with the document. The procedural postures define the case's legal status or phase (e.g., “On Appeal”).
  - `sections`: Each document has a list of sections, which themselves contain:
    - `headtext`: Section header text, which often categorizes the content (e.g., “Background”).
    - `paragraphs`: A list of paragraphs containing the main content of each section.
- The dataset comprises 18,000 judicial documents labeled with procedural postures and organized into over 542,000 paragraphs, displaying a dense and structured format typical of legal documents.

- There are 194 unique labels. Statistics below:

```
Label Distribution (Sorted by Frequency):
                                          Label  Count
0                                     On Appeal   4942
1                                Appellate Review   4652
2                               Motion to Dismiss   1449
3                 Review of Administrative Decision   1395
4                       Motion for Attorney's Fees    609
..                                             ...    ...
189  Motion to Serve Additional Discovery Requests      1
190                Motion to Extend Claims Bar Date      1
191             Petition for Divorce or Dissolution      1
192        Declinatory Exception of Improper Venue      1
193                Motion to Vacate Summary Judgment      1
```

- The top two labels, "On Appeal" (4942 instances) and "Appellate Review" (4652 instances), dominate the dataset. These labels represent common procedural postures in judicial opinions, which is expected given the nature of legal documents and the frequency of appeals in the legal process.

- Many labels (such as "Motion to Serve Additional Discovery Requests," "Petition for Divorce or Dissolution," and others) appear only once in the dataset. This indicates that these postures are rarely encountered, which poses challenges for any predictive modeling efforts.

- Challenges in Predicting Low-Frequency Labels: Predicting low-frequency labels presents several challenges: Insufficient Training Data, Models typically require a minimum number of examples to learn effectively. The imbalanced nature of the dataset can lead to model bias. For low-frequency labels, the model may memorize instances instead of generalizing patterns.

| Threshold N | Classes Above N | Documents Reserved | Classes Reserved % | Documents Reserved % |
|---|---|---|---|---|
| 1 | 145 | 17028 | 74.7 | 99.7 |
| 2 | 124 | 16986 | 63.9 | 99.5 |
| 3 | 103 | 16923 | 53.1 | 99.1 |
| 4 | 86 | 16855 | 44.3 | 98.7 |
| 6 | 74 | 16792 | 38.1 | 98.3 |
| 8 | 68 | 16749 | 35.1 | 98.1 |
| 10 | 62 | 16693 | 32.0 | 97.8 |
| 20 | 50 | 16515 | 25.8 | 96.7 |
| 30 | 41 | 16293 | 21.1 | 95.4 |
| 50 | 27 | 15742 | 13.9 | 92.2 |
| 100 | 14 | 14797 | 7.2 | 86.6 |
| 1000 | 4 | 12438 | 2.1 | 72.8 |

- Choosing a threshold N=20 for label frequency strikes a balance between retaining sufficient classes and minimizing noise in the model.

Count before removing labels with fewer than 10 instances: 17077

Count after removing labels with fewer than 20 instances: 16515

Percentage of data removed: 3.29%

Percentage of data retained: 96.71%

Number of unique classes remaining after filtering: 50

---

## Question 2

Our business partners would like to automate the labeling of judicial opinions with procedural postures. You are tasked with conducting an initial exploration to assess feasibility.

1. Based on the provided dataset, build a model that achieves the desired automation.

- **We used logistic regression to solve this classification problem.**
- **We Applied train, test split. (0.2 percent test). Size of training dataset: 13212. Size of testing dataset: 3303. Total dataset size: 16515**
- **Data filtering techniques are applied and tf-idf is used for the process of transforming text data into a numerical format.**

2. Analyze your model's performance.

The hyperparameter I choose is max_iteration=100 for logistic regression due to time limitation.

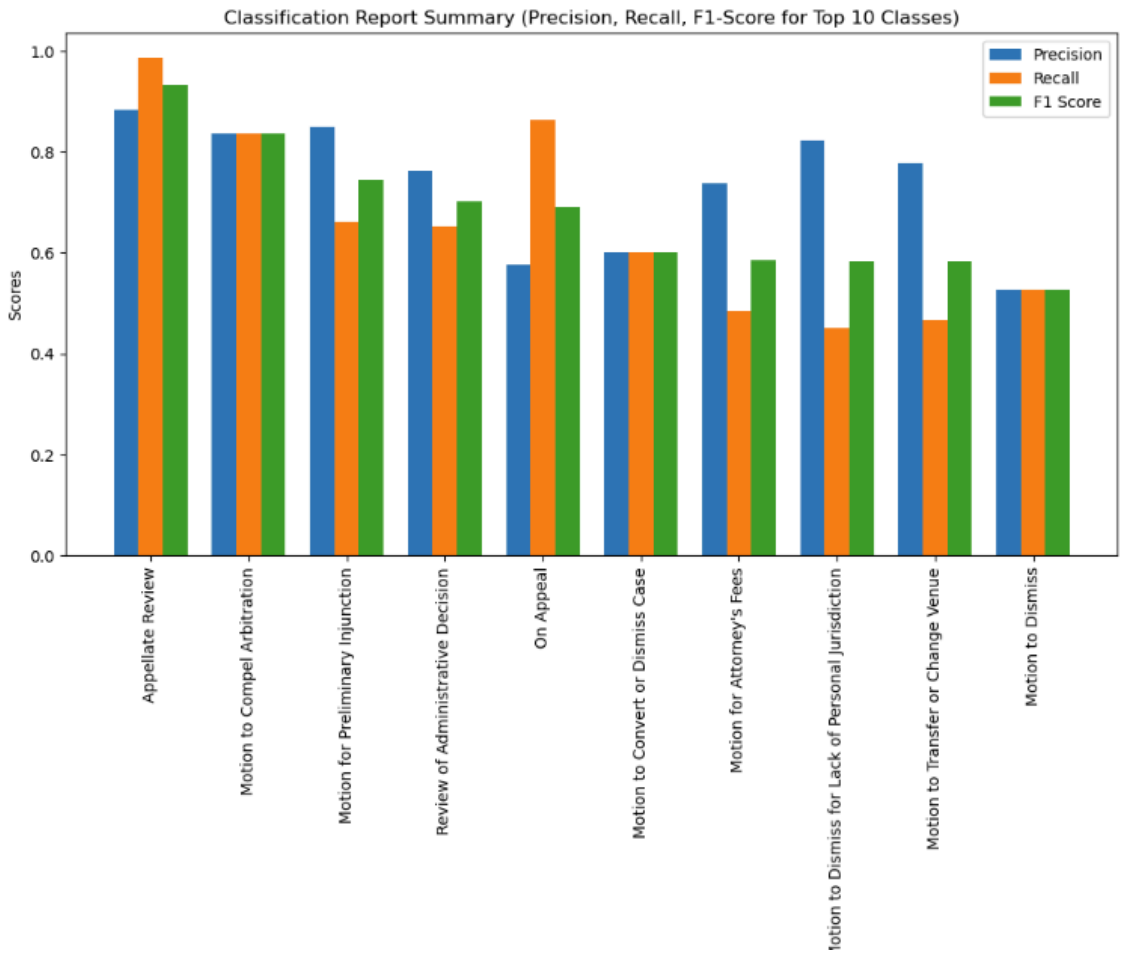
```
  n_iter_i = _check_optimize_result(
Label Accuracy: 0.6979

Classification Report:
```

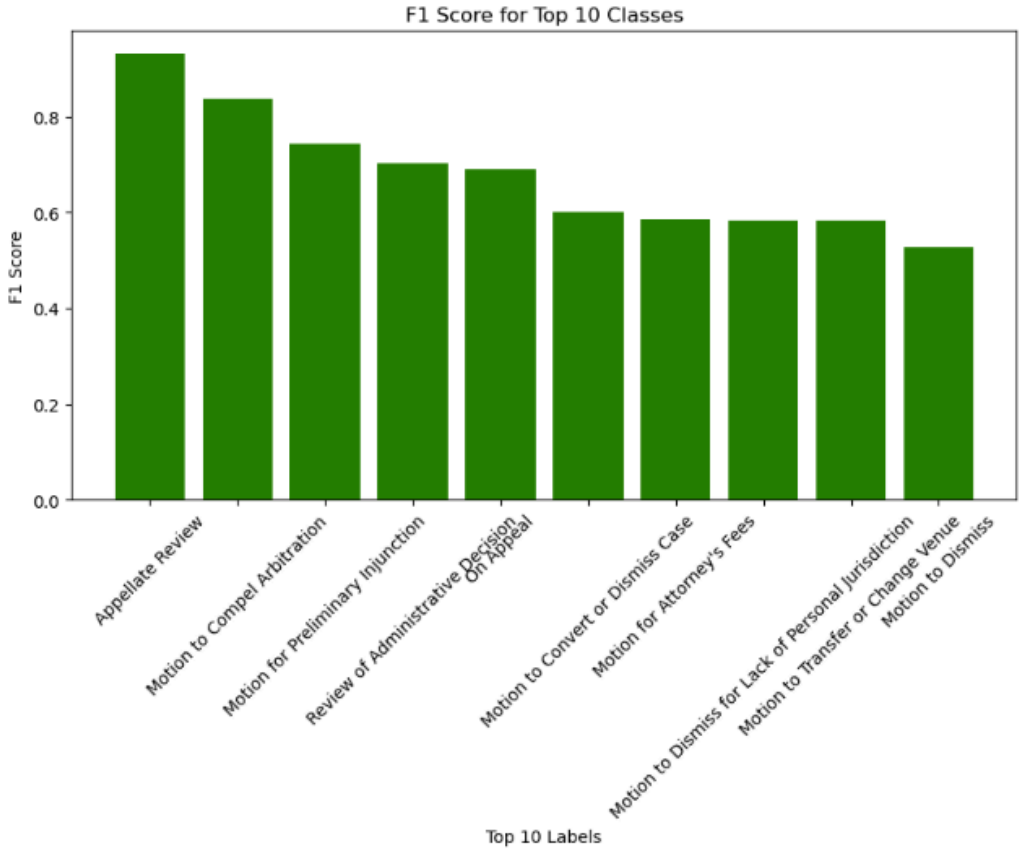
| | precision | recall | f1-score |
|---|---|---|---|
| Appellate Review | 0.88 | 0.99 | 0.93 |
| Application to Vacate Arbitration Award | 1.00 | 0.27 | 0.43 |
| Certified Question | 0.00 | 0.00 | 0.00 |
| Juvenile Delinquency Proceeding | 0.00 | 0.00 | 0.00 |
| Motion for Additional Discovery | 0.00 | 0.00 | 0.00 |
| Motion for Attorney's Fees | 0.74 | 0.48 | 0.58 |
| Motion for Contempt | 0.67 | 0.11 | 0.18 |
| Motion for Continuance | 0.00 | 0.00 | 0.00 |
| Motion for Costs | 0.00 | 0.00 | 0.00 |
| Motion for Default Judgment/Order of Default | 0.75 | 0.22 | 0.34 |
| Motion for Extension of Time | 0.00 | 0.00 | 0.00 |
| Motion for Involuntary Dismissal | 0.00 | 0.00 | 0.00 |
| Motion for Judgment as a Matter of Law (JMOL)/Directed Verdict | 0.50 | 0.08 | 0.13 |
| Motion for New Trial | 1.00 | 0.04 | 0.07 |
| Motion for Permanent Injunction | 0.00 | 0.00 | 0.00 |
| Motion for Preliminary Injunction | 0.85 | 0.66 | 0.74 |
| Motion for Protective Order | 0.00 | 0.00 | 0.00 |
| Motion for Reconsideration | 1.00 | 0.09 | 0.17 |
| Motion for Rehearing | 0.00 | 0.00 | 0.00 |
| Motion for Relief from Order or Judgment | 0.00 | 0.00 | 0.00 |
| ... | | | |
| accuracy | | | 0.70 |
| macro avg | 0.27 | 0.16 | 0.18 |
| weighted avg | 0.65 | 0.70 | 0.65 |

The classification results indicate a logistic regression model that achieves an overall accuracy of 0.6979%, with good performance for the dominant class "Appellate Review," but poor predictions (precision, recall, f1-score of 0.00) for many minority classes, leading to warnings about undefined metrics. The model did not converge within the default iteration limit, suggesting the need for increased max_iter or feature scaling. To address class imbalance and enhance performance, strategies such as resampling, employing different algorithms, and feature engineering should be considered, alongside cross-validation for more reliable performance evaluation.

## Visualization of trained results:

- confusion metrics for top 10 classes:


Classification Report Summary (Precision, Recall, F1-Score for Top 10 Classes)
Precision
Recall
F1 Score
Scores
1.0
0.8
0.6
0.4
0.2
0.0
Appellate Review
Motion to Compel Arbitration
Motion for Preliminary Injunction
Review of Administrative Decision
On Appeal
Motion to Convert or Dismiss Case
Motion for Attorney's Fees
lotion to Dismiss for Lack of Personal Jurisdiction
Motion to Transfer or Change Venue
Motion to Dismiss




- As you can see, applicated review had highest F1 score on test set, while motion to dismiss has lowest F1 score on testset.

- For analyze usage, we also have Sample Inference Instances (Top 10 Classes): Displays a table of individual predictions, limited to examples from the top 10 classes to show specific instances of correct and incorrect predictions, and Label Distribution for

Top 10 Classes in test set, for future inspections.

```
Sample Inference Instances (Top 10 Classes):
                                                Text  \
0  Gerald Schram appeals from his conviction for ...
1  In a hybrid proceeding pursuant to CPLR articl...
2  This is a civil-rights case, brought under the...
3  In 1999, Robert A. E. Hall, Jr. purchased prop...
4  Order, Supreme Court, Bronx County (Mark Fried...
5  Willie Mathers Newton appeals a judgment and s...
6  Order, Supreme Court, New York County (Debra A...
7  Appellant Melissa Everly appeals an order of t...
8  Appellants, the plaintiffs below, appeal the t...
9  On January 27, 2017—seven days after taking th...

                         True Label                   Predicted Label
0                  Appellate Review                  Appellate Review
1                         On Appeal                         On Appeal
2                             Other                         On Appeal
3                             Other                         On Appeal
4                  Motion to Dismiss                 Motion to Dismiss
5                  Appellate Review                  Appellate Review
6                             Other                         On Appeal
7                         On Appeal                         On Appeal
8                         On Appeal                 Motion to Dismiss
9  Motion for Preliminary Injunction  Motion for Preliminary Injunction
```

## 3. Make a recommendation to the business on the feasibility of this task.

- The substantial class imbalance, many labels receiving little to no predictive accuracy, raises concerns about the model's ability to provide reliable insights across all categories.
- To improve the performance: https://huggingface.co/nlpaueb/legal-bert-base-uncased We may use specifically designed for legal text. This model has been fine-tuned on legal datasets, providing a robust starting point for our task. Leveraging Legal-BERT allows us to benefit from its understanding of legal language and concepts, which is likely to improve our model's accuracy and generalization when predicting procedural postures.

- If the primary objective is to classify and analyze predominant legal scenarios, we may switch to legal bert with better accuracy. However, it is essential to approach the results generated by **AI with caution, especially in legal area**. Human annotators must be engaged to verify and validate these findings, ensuring accuracy and reliability in the

interpretation of legal postures. **This collaborative approach between AI and human expertise will enhance the overall effectiveness** in automated systems.
- Recommendations for the business would include a strategic focus on addressing class imbalance and possibly investing in alternative modeling techniques or data enrichment strategies to enhance the model's performance across all categories.

**Note**: This is an exploratory model, and state-of-the-art results are not expected. A thorough evaluation of the model's performance is more valuable than high accuracy.

---

# Question 3

Regardless of the feasibility determination in Question 2, assume that stakeholders are satisfied with the initial results and wish to proceed. You are asked to outline the next steps, and you may choose either to:

- **(a)** Conduct additional experiments to improve or validate the model, or
- **(b)** Begin the process to deploy this model in production.

Either choice is valid. Describe the next steps with sufficient detail to be accessible to both technical and non-technical audiences.

1. Explain and justify each recommended step.
2. Identify potential challenges that may need to be addressed to ensure success.

**I choose (a)**
To improve the performance, potential experiments to try:
- Incorporate More Data: If available, we will try to augment our training dataset with additional judicial opinions or legal documents. This will help the model learn from a broader range of examples and improve its ability to generalize to unseen cases.

- https://huggingface.co/nlpaueb/legal-bert-base-uncased We may use specifically designed for legal text. This model has been fine-tuned on legal datasets, providing a robust starting point for our task. Leveraging Legal-BERT allows us to benefit from its understanding of legal language and concepts, which is likely to improve our model's accuracy and generalization when predicting procedural postures. For future development, it is recommended to explore the potential of leveraging Legal BERT, a model specifically designed for legal text processing. The `**legal_bert.py**` script includes my skeleton codes, and the training pipeline has already been completed. Although time

limitations, it presents an opportunity for subsequent projects. Utilizing Legal BERT may enhance classification accuracy and better handle the complexities of legal language.

- Evaluate Additional Model Architectures: In addition to fine-tuning Legal-BERT, we will explore other model architectures like DistilBERT or RoBERTa, which might yield better results.

- Cross-Validation: Implementing k-fold cross-validation will allow us to better assess the model's performance and reduce the risk of overfitting.

- Address Low-Frequency Labels: We may consider techniques such as data augmentation for these classes or using specialized loss functions that give more weight to underrepresented labels.