Methodology for the Categorization of Allegations, Dispositions, and Repercussions in Cases of Misconduct

The categorization of misconduct dispositions was completed through basic regex. However, the classification of misconduct allegations poses unique challenges, primarily arising from reporting discrepancies, data imbalance, and the presence of multiple categories within a single allegation.

To overcome these challenges, a comprehensive approach was adopted, encompassing the following key steps:

1. **Category Definition**: Categories for misconduct dispositions and repercussions were meticulously defined through manual scrutiny of the data, with a focus on identifying underlying patterns of misconduct.
2. **Text Clustering Techniques**: Advanced clustering techniques, such as Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM), were employed to refine categories. However, certain clusters lacking real-world relevance were merged with assumed categories to enhance the accuracy of the classification.
3. **Labeling Data**: A subset of 500 randomly sampled unique allegation and allegation description pairs, representing a substantial portion of the total dataset, were manually labeled with 41 categories to create a labeled dataset for model training.

To address the challenge of classifying allegations, three models were implemented: a zero-shot multi-label model based on regular expressions applied to keyword stems, a few-shot multi-label Support Vector Machine model, and a few-shot multi-label Random Forest model. The selection of the best-performing model was based on average accuracy and F1 scores.

The zero-shot multi-label model achieved an impressive average accuracy score of 98.48% and an average F1 score of 80.00. The best-performing multi-label Support Vector Machine model, with a linear kernel and a cost of 10, exhibited an average accuracy score of 97.69% and an average F1 score of 44.31. In comparison, the optimal multi-label Random Forest model, with 50 estimators and no maximum depth, displayed an average accuracy score of 97.46% and an average F1 score of 21.50.

The findings indicate that the keyword model, based on regular expressions applied to keyword stems, outperforms both the Support Vector Machine and Random Forest models in accurately classifying misconduct allegations. Consequently, this model was selected for predicting classifications for all misconduct allegations, providing a robust and efficient approach to categorization in cases of misconduct.