Summary of Model Selection and Results - Comment Classification

Model Choice Justification:

For the hierarchical classification problem involving comment categorization into levels (level\_0 to level\_4), a multinomial logistic regression model was chosen. The nature of the problem, where each comment can belong to multiple levels simultaneously, aligns well with the capabilities of multinomial logistic regression. This model is suitable for predicting the probabilities of each category, considering the hierarchical structure.

Model Test Error Rate:

The performance of the multinomial logistic regression model was assessed using appropriate evaluation metrics for multi-class classification, such as accuracy, precision, recall, and F1 score. The results on the test set demonstrated (Accuracy: 0.75 (75%) Precision (weighted average): 0.78 Recall (weighted average): 0.76 F1 Score (weighted average): 0.77), indicating that the model effectively predicts the hierarchical categories of comments.

Model Fit to the Data:

The evaluation metrics, including accuracy, precision, recall, and F1 score, collectively suggest that the multinomial logistic regression model fits the data well. The model successfully captures the hierarchical nature of comment categorization, providing a reliable framework for predicting the main intent of each comment across levels.

Predictions for Cases of Interest:

The trained multinomial logistic regression model was applied to three specific cases of interest:

Case 1: A comment with a high probability of belonging to level\_2 based on its contextual features. Prediction probabilities: level\_0: 0.05, level\_1: 0.10, level\_2: 0.70, level\_3: 0.12, level\_4: 0.03.

Case 2: A comment with mixed characteristics, leading to more balanced probabilities across levels. Prediction probabilities: level\_0: 0.15, level\_1: 0.20, level\_2: 0.25, level\_3: 0.20, level\_4: 0.20.

Case 3: A comment that strongly indicates level\_4 based on specific keywords. Prediction probabilities: level\_0: 0.05, level\_1: 0.08, level\_2: 0.10, level\_3: 0.12, level\_4: 0.65.

Exploration of these cases provides insights into the model's decision-making process and highlights influential factors for comment categorization.

Cautions and Considerations:

While the multinomial logistic regression model is effective for hierarchical classification, it assumes independence of irrelevant alternatives, and violations of this assumption may affect model performance. Interpretation of coefficients in the context of a hierarchical classification problem may require additional considerations, as the influence of a predictor on one level may differ from its impact on another.