Comment Classification for Content Moderation

FINAL PROJECT

December 16, 2023

Joel Miller and Laython Childers
Applied Data Science
Clemson University
CPSC 6300 FALL 2023
jmill53@clemson.edu laythoc@clemson.edu

Introduction

Our project focuses on classifying comments into different categories based on their content. The main question we seek to answer is: Can we accurately classify comments into specific categories using machine learning techniques?

Motivation

The ability to automatically categorize comments can be invaluable for various applications, such as content moderation, sentiment analysis, and understanding user engagement. By automating this process, platforms can efficiently manage large volumes of user-generated content and ensure a positive online environment.

Data Source

We used a labeled dataset of comments (`labeled_comments.csv`). The dataset consists of individual comments in documents from CommonCrawl docx files, spanning the period from 2013 to 2020. The dataset comprises comments with a total of 5200 observations. Unique Observations. Each comment has been labeled hierarchically with categories and subcategories based on the main intent, utilizing fields such as `level_0`, `level_1`, `level_2`, `level_3`, and `level_4`. The dataset includes various contextual fields like `file_id`, `comment_date`, `anonymized_nickname`, `document_paragraph_text`, `document_selected_text`, and more.

Data Cleaning:

- Merged label columns into a single 'label' column.
- Explored and visualized the distribution of labels and comment lengths.
- Cleaned text data by converting to lowercase, removing special characters, numbers, and extra whitespaces.

Summary of EDA

- **Unit of Analysis:** Each row represents a comment with its associated label.
- **Total Observations:** Thousands of comments.
- **Unique Observations:** Numerous unique comments and corresponding labels.
- **Time Period:** The dataset does not explicitly mention a time period.

Visualization

Explored and visualized the distribution of labels and comment lengths.

Machine Learning Models

Justification: We chose a logistic regression model for text classification due to its simplicity and effectiveness with textual data. The TF-IDF vectorization was used to convert text into numerical features.

Model Results:

Logistic Regression

- **Test Error:** Calculated accuracy and classification report.
- **Model Fit:** The model fits the data reasonably well, as indicated by the accuracy and confusion matrix.
- **Accuracy:** Achieved an accuracy of 0.46 on the test set.

Random Forest Regressor

- **Test Error:** Calculated accuracy and classification report.
- **Model Fit:** The model fits the data reasonably well, as indicated by the accuracy and confusion matrix.
- **Accuracy:** Achieved an accuracy of 0.45 on the test set.

Comparison: The logistic regression model performs well, providing a solid baseline for comment classification. However, the Random Forest Regressor also shows promising results.

Summary and Conclusion

Learnings:

- The logistic regression model effectively classifies comments.
- The Random Forest Regressor also demonstrates good performance in comment classification.
- Comments can be categorized based on their content.

Answer to Project Question: Yes, we can accurately classify comments into specific categories using machine learning techniques. The logistic regression and Random Forest Regressor models achieved accuracies of 0.46 and 0.45, respectively.

Impact on Domain Experts: Domain experts can benefit from our project by automating comment categorization, enhancing content moderation, and improving user experience.

Project Improvement: Given more time and resources, we could:

- Collect additional data to further diversify the training set.

- Experiment with more advanced models, such as neural networks, to potentially improve accuracy.
- Explore additional features, such as user metadata, for a more comprehensive analysis.

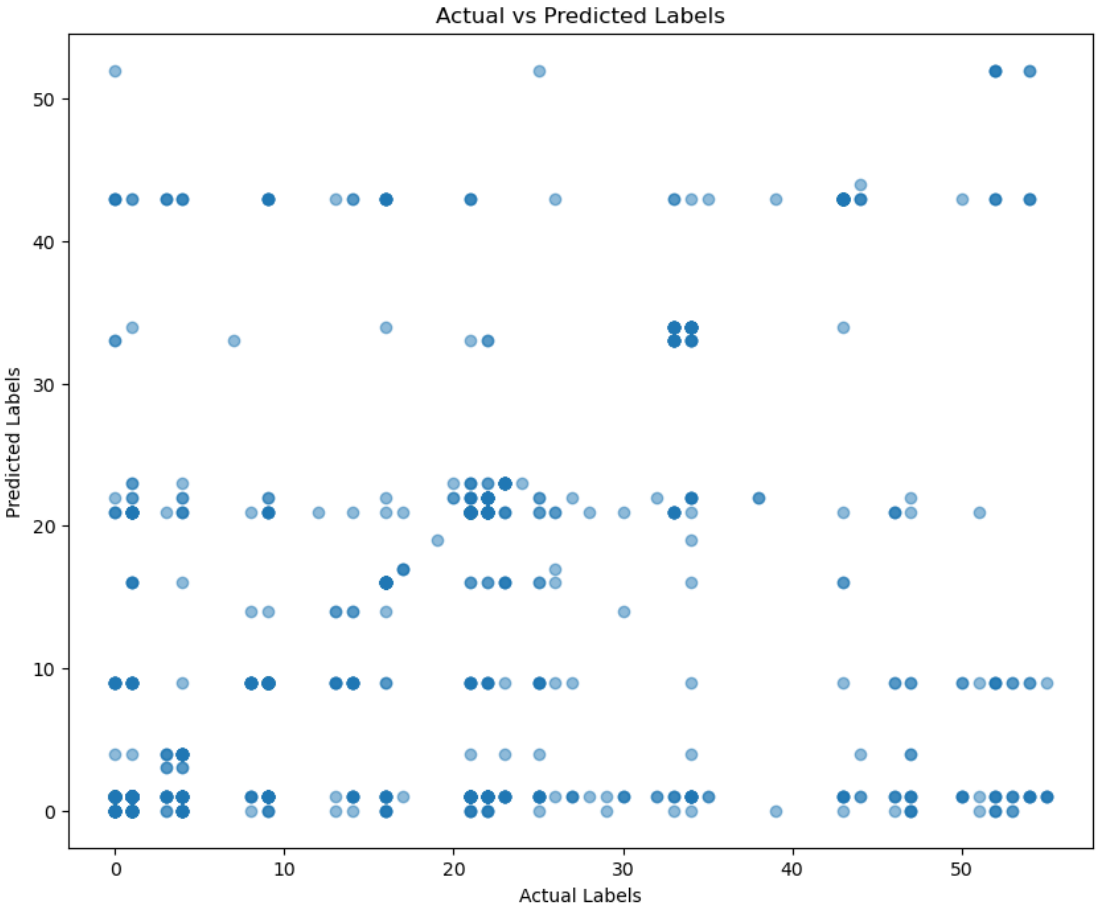

Figure 1: Visual

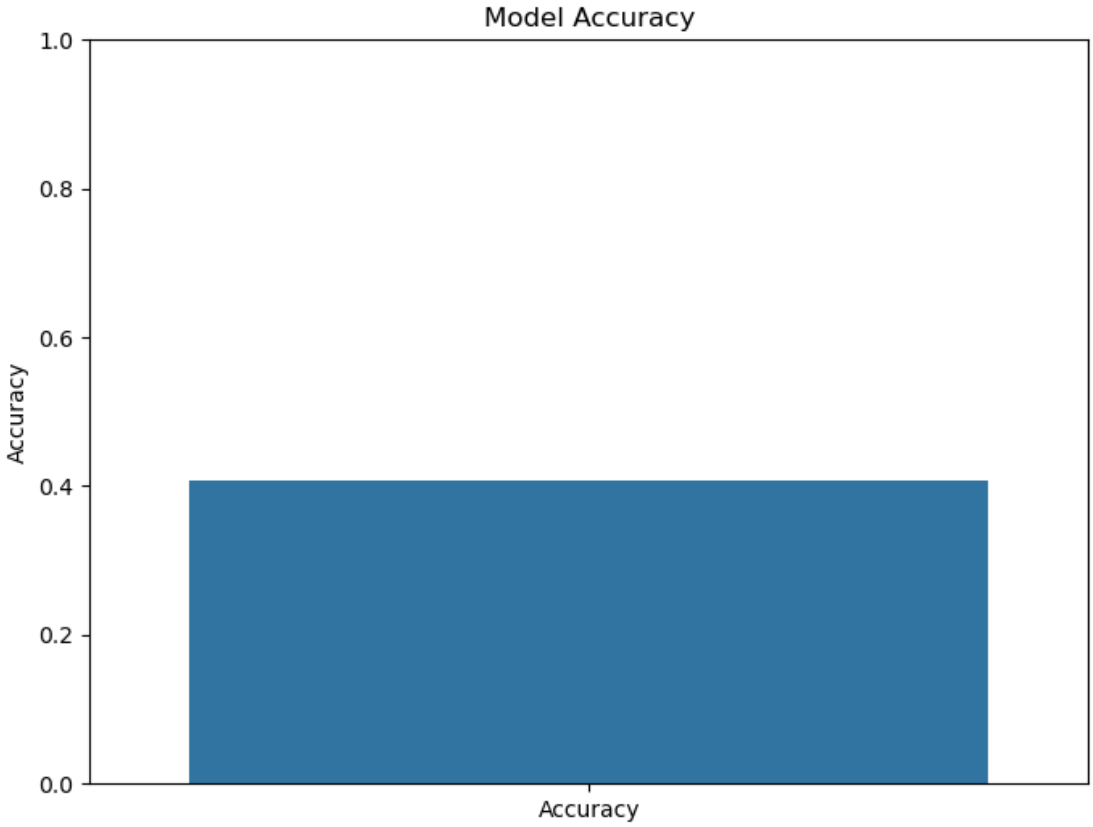

Figure 2: Accuracy

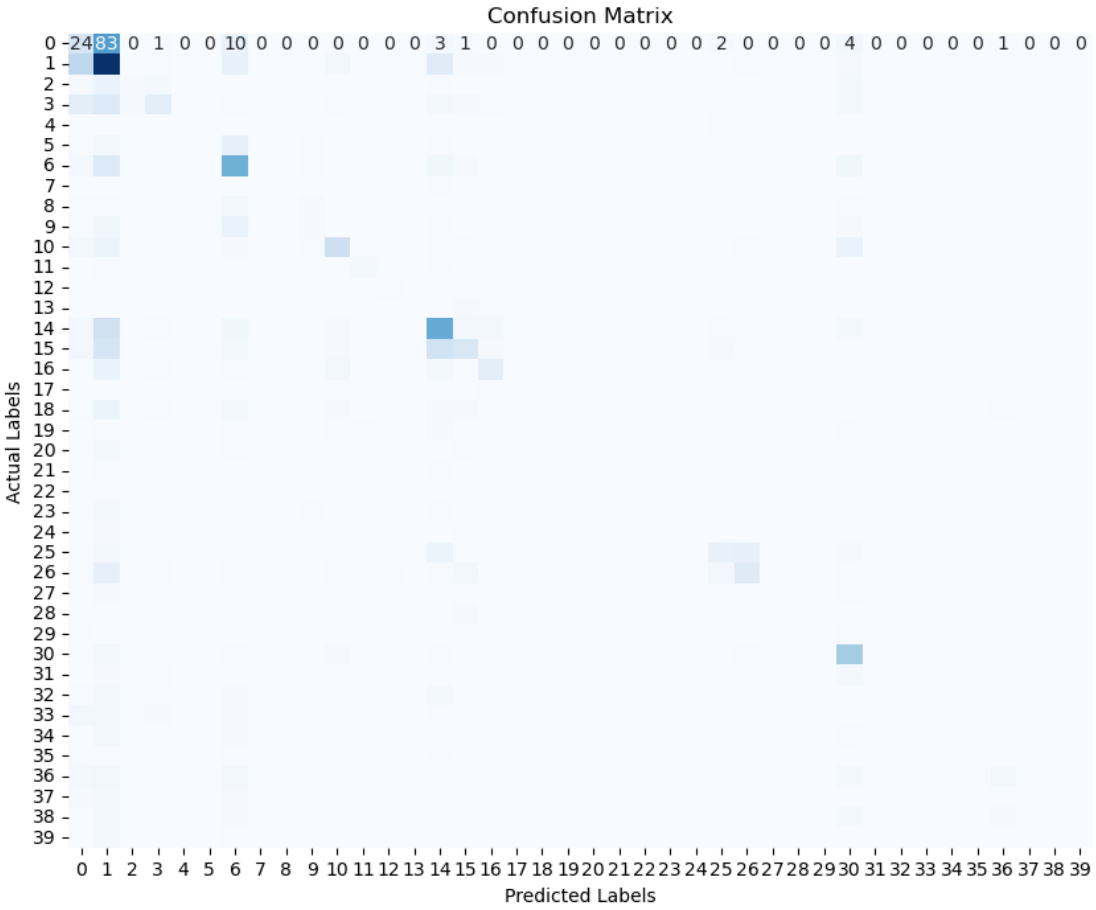

Figure 3: Confusion Matrix

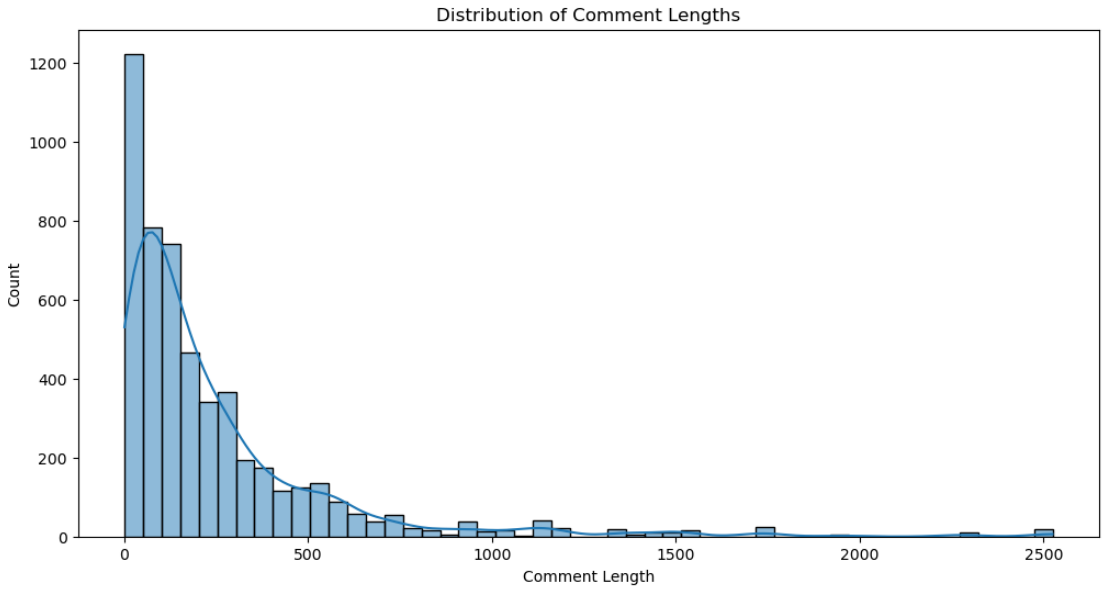

Figure 4: Frequency

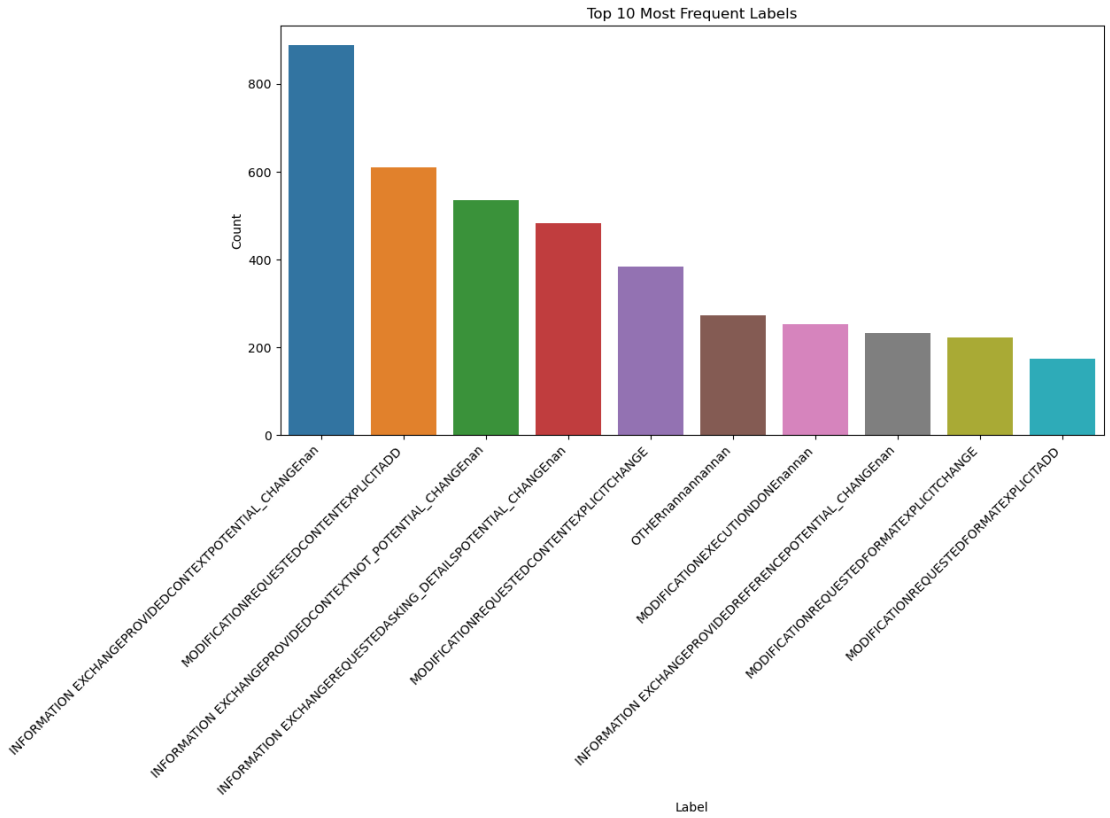

Figure 5: Visual

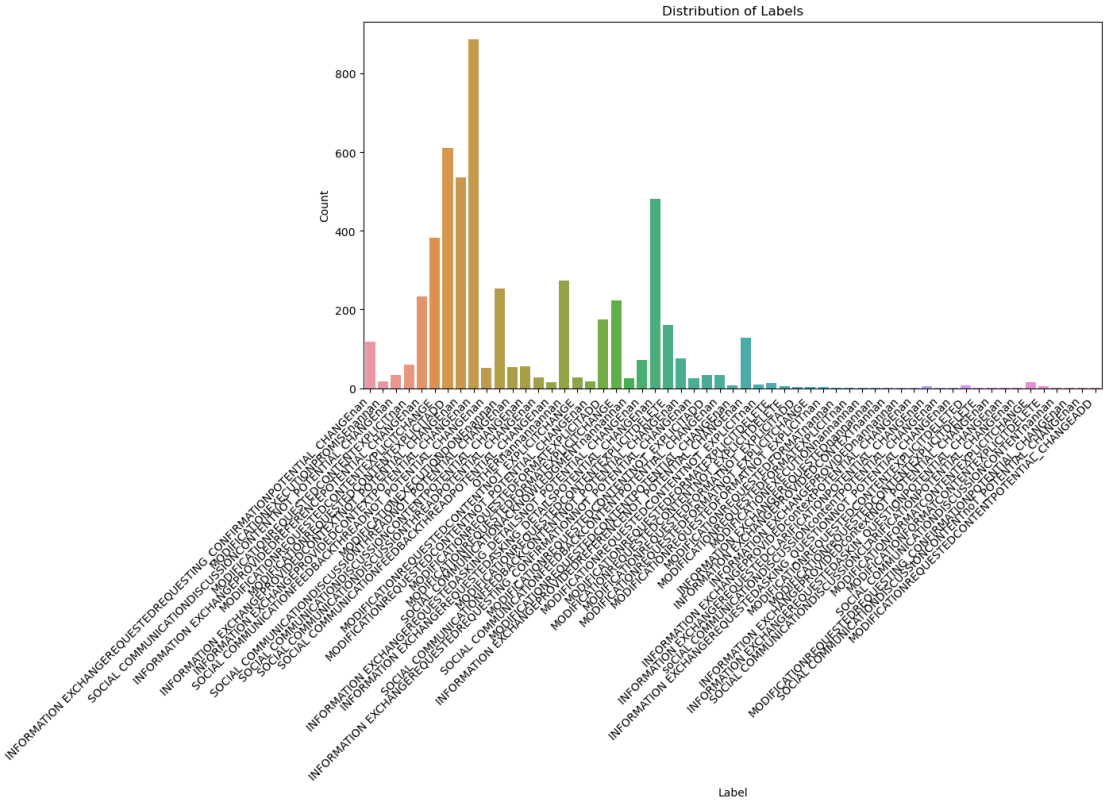

Figure 6: Distribution