Classifying Toxic Comments on the Jigsaw Toxic Comment Dataset

Gustavo Lindenberg Pacheco
Insper
Email: gustavolp1@al.insper.edu.br

I. Introduction

This study uses the Jigsaw Toxic Comment Classification Challenge dataset. The dataset's business purpose is to improve content moderation on user-generated platforms such as Wikipedia. It includes over 150,000 comments, each labeled as toxic or non-toxic. The task is to classify comments as toxic or non-toxic.

The dataset is publicly available and was provided by Kaggle as part of the 2017 competition *Toxic Comment Classification Challenge* [0]. This dataset plays a key role in the development of machine learning models to moderate harmful content in online communities automatically.

II. Classification Pipeline

The classification process involves several key steps:

1) **Text Pre-processing**: Comments were cleaned by removing punctuation, converting text to lowercase, and removing numbers and non-alphabetic characters.
2) **Tokenization**: Each comment was split into individual words (tokens).
3) **Lemmatization**: Words were reduced to their base forms using the lemmatizer from WordNet.
4) **Vectorization**: Comments were transformed into numerical data using TF-IDF (Term Frequency-Inverse Document Frequency), specifically using the top 5000 most frequent words.
5) **Modeling**: Logistic Regression was used as the classification model, trained on 80% of the data and tested on the remaining 20%.

III. Evaluation

The classifier was evaluated using the balanced accuracy score to account for class imbalance. The balanced accuracy achieved is 80.66%.

A. Word Importance

Using the coefficients from the Logistic Regression model, the top 10 words most predictive of toxic comments were identified. Figure 1 shows the most important words and their respective coefficients.

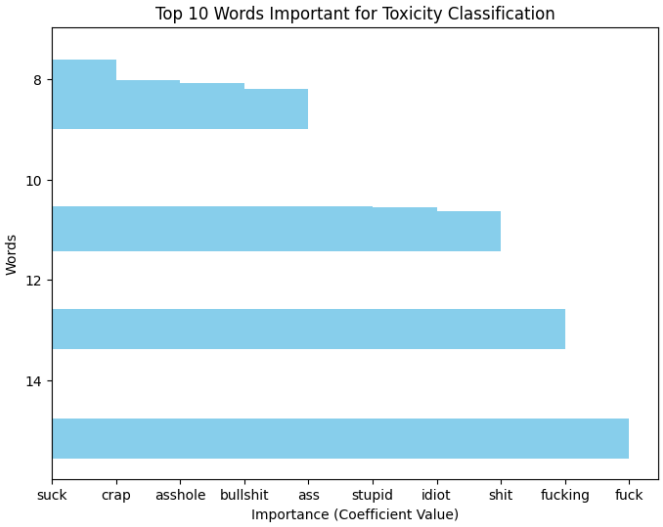

Fig. 1: Top 10 Words Important for Toxic Classification

B. Confusion Matrix

The confusion matrix in Figure 2 provides a clear indication of the classifier's performance in distinguishing between toxic and non-toxic comments. It shows the counts of true positives, false positives, true negatives, and false negatives.

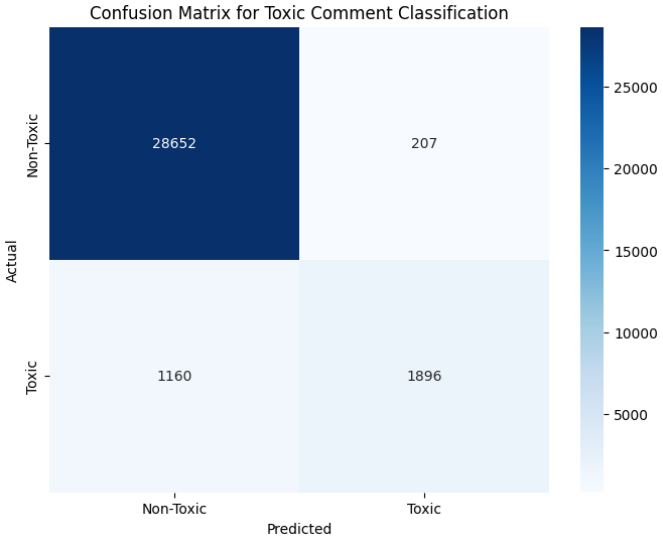

Fig. 2: Confusion Matrix for Toxic Comment Classification

C. Class Distribution in the Training Data

Figure 3 shows the class distribution in the training data. Non-toxic comments are much more frequent than toxic comments, indicating a class imbalance.

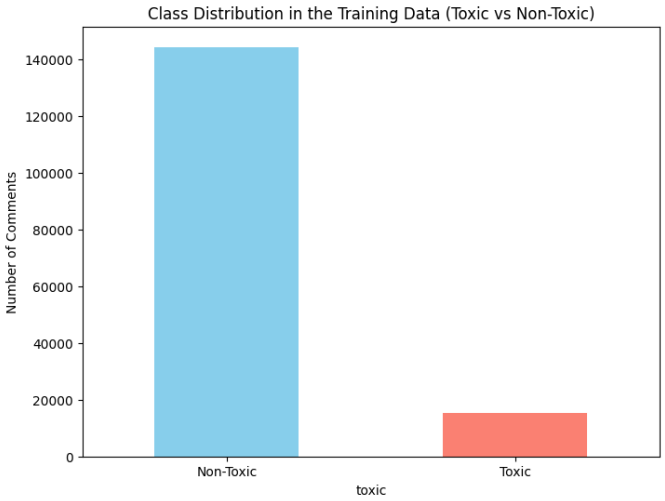

Fig. 3: Class Distribution in the Training Data (Toxic vs Non-Toxic)

IV. Dataset Size

The dataset contains over 150,000 comments, which is considered sufficient for this task. Considering the size and balance of the dataset, no downsampling was necessary.

V. Topic Modeling

Latent Dirichlet Allocation (LDA) was applied to identify five distinct topics in the comments. The toxicity rates for each topic are as follows:

- **Topic 0 (24.22% toxicity)**: Discussions about Wikipedia editing and moderation, which usually have a higher toxicity rate due to disagreements over edits.
- **Topic 1 (3.86% toxicity)**: Conversations about the quality of articles and sources. These are usually casual discussions with lower toxicity.
- **Topic 2 (3.12% toxicity)**: Conversations about Wikipedia article deletion and copyright. Low toxicity.
- **Topic 3 (2.15% toxicity)**: Generally casual and polite discussions, often expressing gratitude. This topic has the lowest toxicity rate.
- **Topic 4 (8.42% toxicity)**: Discussions around sociopolitical matters and identity topics, which can be polarizing and highlight moderate toxicity.

These insights suggest that certain topics (such as edit disputes or sensitive social themes) are more prone to toxic behavior, while more helpful and casual conversations have much lower toxicity rates.

VI. Conclusion

The toxic comment classification task was successfully implemented using Logistic Regression. The classifier achieved a balanced accuracy score of 80.66%. Word importance analysis shows that the most offensive terms are the strongest predictors of toxicity. Additionally, topic modeling revealed that certain topics, such as discussions about Wikipedia editing and sociopolitical issues, tend to have higher toxicity rates.

References

J. Sorensen, J. Elliott, L. Dixon, M. McDonald, Nithum, and W. Cukierski, "Toxic Comment Classification Challenge," Kaggle, 2017. [Online]. Available: https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge