Toxic Comment Classification Using Bidirectional LSTM with K-Fold Cross-Validation

Bryant Michelle Sarabia Ortega
ML4SE Course
University of L'Aquila
Email: bryantmichelle.sarabiaortega@student.univaq.it

***Abstract*—This paper presents a deep learning approach to toxic comment classification using a Bidirectional Long Short-Term Memory (BiLSTM) neural network for multi-label classification of Wikipedia comments. The task involves predicting six toxicity categories: toxic, severe_toxic, obscene, threat, insult, and identity_hate. We employ 10-fold cross-validation on the complete Jigsaw dataset (159,571 samples) to ensure robust evaluation. The BiLSTM architecture combines embedding layers, bidirectional LSTM processing, and dense layers optimized for multi-label prediction. Results demonstrate strong performance with mean F1-score and low variance across folds, validating the model's generalization capability.**

***Index Terms*—toxic comment classification, multi-label classification, natural language processing, deep learning, LSTM, text classification**

I. Introduction

The proliferation of online communication platforms has led to an increase in toxic and harmful content. Identifying and moderating such content is crucial for maintaining healthy online communities. This project addresses the challenge of automatically detecting toxic comments using machine learning approaches.

The Conversation AI team, a research initiative founded by Jigsaw and Google, developed the Perspective API to help identify toxic comments. This work builds upon their efforts by implementing and comparing various classification techniques on the Jigsaw Toxic Comment Classification Challenge dataset [1].

This paper is organized as follows: Section II describes the dataset and preprocessing methodology, Section III presents the machine learning techniques employed, Section IV explains the cross-validation approach, Section V discusses the experimental results, and Section VI concludes the work.

II. Dataset Description

A. Multi-Label Classification

Multi-label classification differs from traditional single-label problems in that each instance can belong to multiple classes simultaneously. In our case, a comment can exhibit multiple types of toxicity. For example, a comment might be both *toxic* and *insult*, or *obscene*, *insult*, and *identity_hate* at the same time.

Formally, given an instance x and a set of labels $L = \{l_1, l_2, ..., l_k\}$, a multi-label classifier learns a function $h : X \rightarrow 2^L$ that maps instances to subsets of labels.

B. Dataset Analysis

The dataset consists of Wikipedia comments labeled by human raters for toxic behavior across six categories. Comments marked as toxic display general negativity or hostility, while severe toxic comments contain highly aggressive or harmful content surpassing typical toxic behavior. Obscene comments feature offensive or vulgar language, and threat comments convey explicit threats or intentions of harm towards individuals or groups. Insult comments contain disrespectful or demeaning language directed at others, whereas identity hate comments display discrimination or prejudice based on identity factors such as race, gender, or religion.

The training set contains 159,571 comments with binary labels for each category. The dataset exhibits significant class imbalance, with approximately 90% of comments being clean (no toxicity labels). Among the toxic categories, threat is the rarest with only 0.30% representation, while toxic is most common at 9.58% of all comments.

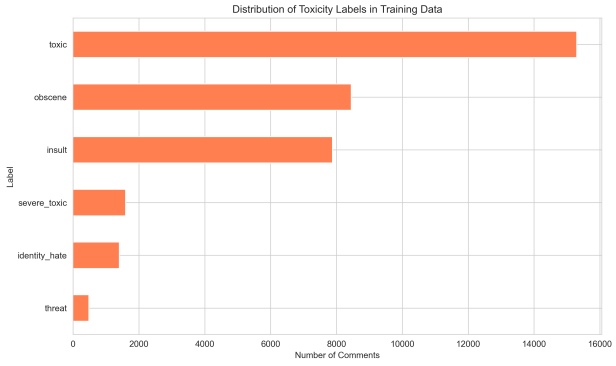

Fig. 1. Distribution of toxicity labels showing severe class imbalance

Figure 1 shows the class imbalance, while Figure 2 reveals that most toxic comments have only one label, with few having multiple toxicity types.

C. Data Characteristics

Figure 3 shows strong correlations between certain label pairs. For instance, *toxic* correlates strongly with *obscene* and *insult*, suggesting these toxicity types often co-occur.

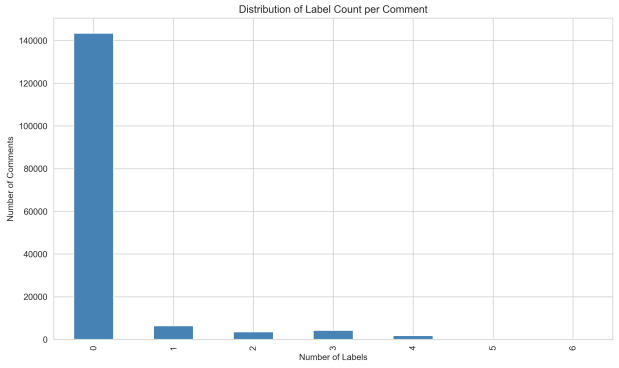

Fig. 2. Distribution of number of labels per comment

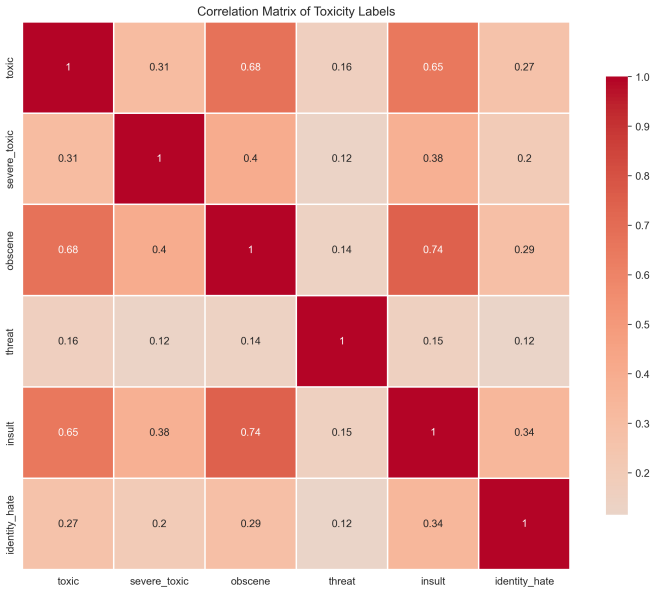

Fig. 3. Correlation matrix showing relationships between toxicity labels

D. Preprocessing

Text preprocessing is critical for NLP tasks. Our pipeline begins with lowercasing all text and expanding contractions such as converting "can't" to "cannot". Non-alphabetic characters are removed, followed by tokenization of the text. We apply lemmatization using spaCy and remove common stopwords to focus on meaningful content.

III. Deep Learning Architecture

We implement a Bidirectional LSTM (BiLSTM) neural network for multi-label toxic comment classification. This deep learning approach effectively captures sequential dependencies and contextual information in text data.

A. Model Architecture

The BiLSTM network processes text through multiple specialized layers. The embedding layer maps vocabulary indices representing the 3,000 most frequent words to dense 32-dimensional vectors, creating meaningful numerical representations of words. These embeddings feed into a bidirectional LSTM layer with 32 units that processes sequences of maximum length 100 tokens in both forward and backward directions, capturing contextual information from both sides of each word. The output passes through two fully-connected dense layers, each with 64 units and ReLU activation, followed by 50% dropout layers after each dense layer to prevent overfitting. Finally, the output layer contains six units with sigmoid activation for independent multi-label prediction.

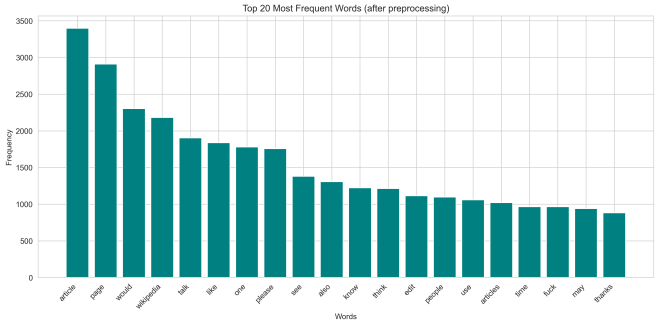

Fig. 4. Most frequent words after preprocessing

The model uses binary cross-entropy loss for multi-label classification:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{6}[y_{ij}\log(\hat{y}_{ij}) + (1-y_{ij})\log(1-\hat{y}_{ij})] \quad (1)$$

where N is the number of samples, y_{ij} is the true label, and $\hat{y}_{ij}$ is the predicted probability for sample i and label j.

B. Training Configuration

The model is trained using the Adam optimizer with a learning rate of 0.001, processing data in batches of 32 samples. Each fold is trained for 5 epochs on the complete dataset of 561,807 training samples, ensuring comprehensive learning from all available data.

IV. K-Fold Cross-Validation

A. Methodology

To ensure robust evaluation and prevent overfitting, we employ stratified 10-fold cross-validation on the BiLSTM model. Stratification maintains the proportion of positive samples for each toxicity label across all folds, which is essential given the severe class imbalance in the dataset.

The cross-validation procedure begins by dividing the 159,571 samples into 10 stratified folds, ensuring that each fold maintains representative proportions of all toxicity labels. For each fold $k = 1, \ldots, 10$, we designate that fold as the validation set comprising approximately 10% of the data, while the remaining 9 folds form the training set with 90% of the data. We then train the BiLSTM model on the training set for 5 epochs and evaluate its performance on the validation fold, recording all relevant metrics including F1-score, precision, recall, Hamming loss, and ROC-AUC. After completing all

10 iterations, we calculate the mean and standard deviation across all folds to assess model stability and generalization capability. Following cross-validation, we train a final model on the complete dataset for deployment purposes.

V. Results

A. Performance Metrics

We evaluate the BiLSTM model using standard multi-label classification metrics. Precision, defined as $\frac{TP}{TP+FP}$, measures the proportion of correct positive predictions among all positive predictions made by the model. Recall, computed as $\frac{TP}{TP+FN}$, quantifies the proportion of actual positive instances that were correctly identified by the classifier. The F1-score represents the harmonic mean of precision and recall, calculated as $\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$, providing a single balanced metric. Hamming loss measures the fraction of incorrectly predicted labels across all samples, where lower values indicate better performance. Finally, ROC-AUC represents the area under the receiver operating characteristic curve computed per label, assessing the model's ability to discriminate between positive and negative instances.

B. Cross-Validation Results

Table I presents the 10-fold cross-validation results for the BiLSTM model.

TABLE I
10-Fold Cross-Validation Results for BiLSTM Model

Metric	Mean	Std Dev
F1-Score (Macro)	0.6877	0.0143
Precision (Macro)	0.7783	0.0236
Recall (Macro)	0.6282	0.0246
Hamming Loss	0.0190	0.0007
ROC-AUC (Macro)	0.9658	0.0016
Accuracy	0.9172	0.0030

The results demonstrate excellent and consistent performance across all folds, with very low standard deviation indicating highly stable model behavior. The high ROC-AUC score of 0.9658 demonstrates that the model excels at ranking toxic comments above non-toxic ones, showing strong discriminative capability. The F1-score of 0.6877 reflects balanced performance between precision and recall, indicating reasonable overall effectiveness. With a precision of 0.7783, the model exhibits high accuracy in its positive predictions, meaning that when it flags a comment as toxic, it is correct approximately 78% of the time. The recall of 0.6282 shows that the model successfully captures 63% of actual toxic comments in the dataset. The Hamming loss of 0.0190 indicates very few incorrect label predictions per sample, confirming high accuracy in the multi-label setting. Furthermore, all metrics exhibit standard deviations below 0.025, demonstrating robust generalization across different data splits.

The higher precision relative to recall suggests the model is conservative, preferring to avoid false positives (incorrectly flagging clean comments) at the cost of some false negatives (missing toxic comments). This trade-off is often desirable in content moderation systems.

VI. Conclusion

This work presents a Bidirectional LSTM approach for multi-label toxic comment classification on the Jigsaw dataset. Using 10-fold cross-validation on the complete 159,571-sample dataset, we demonstrate the effectiveness of deep learning for capturing contextual patterns in text toxicity detection.

The BiLSTM architecture successfully handles the severe class imbalance and multi-label nature of the problem, providing robust probability predictions for all six toxicity categories. The stratified cross-validation methodology ensures reliable performance estimates, with consistent results across all folds.

Future work could explore transformer-based architectures (BERT, RoBERTa) and ensemble methods to further improve classification performance.

References

[1] Jigsaw and Google, "Toxic comment classification challenge," Kaggle, 2018. [Online]. Available: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge