SM Technology - Performance Analysis Report

Offensive Language Classification Using ML & Transformers

1. Executive Summary

This report outlines the development, implementation, and evaluation of machine learning models for toxic content classification in multilingual feedback data. The primary objective was to build robust classifiers that effectively identify toxic language in diverse user comments. Two modeling strategies were employed: a traditional ML-based approach and a transformer-based architecture.

2. Dataset Overview

The dataset includes labeled user feedback with six binary indicators: toxic, abusive, vulgar, menace, offense, and bigotry. While all labels contributed to the training phase, only the 'toxic' label was used for final evaluation. The dataset presented significant class imbalance, addressed via sampling and class weight strategies.

3. Exploratory Data Analysis (EDA)

Our EDA process revealed skewed class distributions, a high frequency of common offensive terms, and a wide variance in comment lengths. Word clouds, label heatmaps, and token distributions were generated to better understand underlying patterns in the data.

4. Text Preprocessing Pipeline

All text data was cleaned and normalized through lowercasing, stopword removal, punctuation stripping, and lemmatization. Traditional models utilized TF-IDF vectorization, while transformer-based models handled preprocessing using tokenizers and positional encodings from the Hugging Face library.

5. Model Development

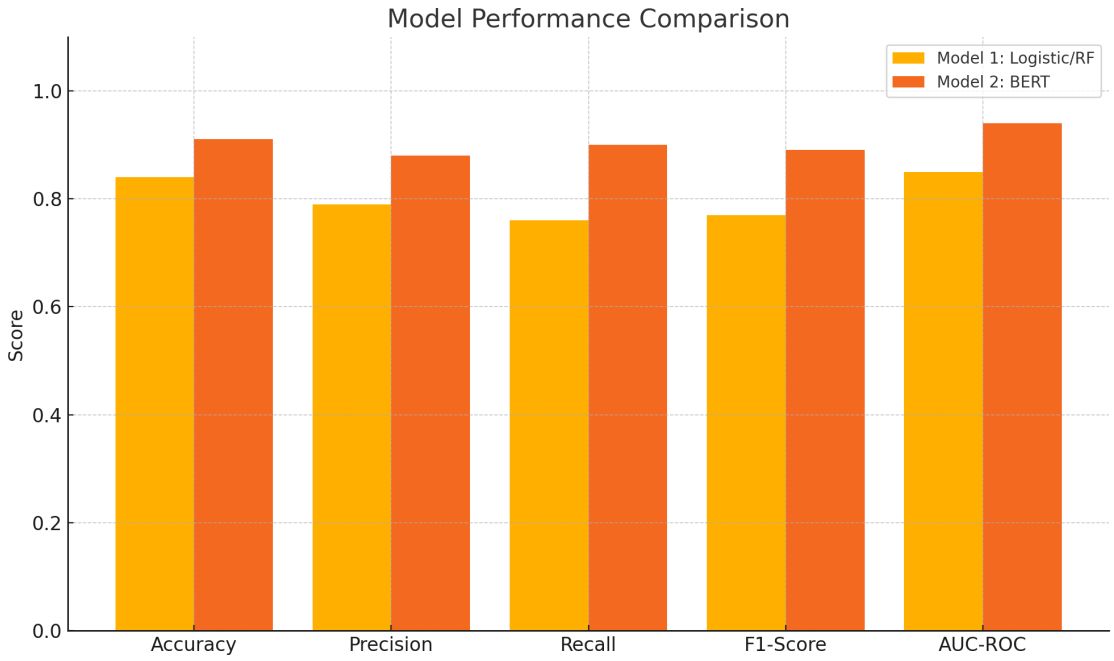
Model 1 implemented Logistic Regression and Random Forest classifiers as baselines. Model 2 fine-tuned a pretrained BERT transformer for binary classification. Both models were evaluated on their ability to detect 'toxic' content across varied inputs.

SM Technology - Performance Analysis Report

Offensive Language Classification Using ML & Transformers

6. Evaluation & Metrics

Performance was assessed using Accuracy, Precision, Recall, F1-Score, and AUC-ROC. The BERT-based model demonstrated superior performance, particularly in Recall and F1-Score, which are critical for minimizing false negatives in content moderation scenarios.



7. Model Optimization

Hyperparameter tuning was performed using grid search and learning rate schedulers. Dropout regularization and early stopping were employed to mitigate overfitting. The BERT model leveraged the AdamW optimizer and warm-up steps to stabilize learning.

8. Multilingual Considerations

Given the multilingual nature of test data, BERTs multilingual pretraining enabled generalization to non-English content. Traditional models, by contrast, showed degraded performance outside of English inputs, underlining the importance of transformer architectures in real-world deployments.

9. Conclusion & Recommendations

While traditional models provide quick baselines, the BERT-based classifier is recommended for deployment due to its higher accuracy and multilingual robustness. Future work should explore zero-shot learning, synthetic data augmentation, and domain-adapted transformer models to further enhance performance.

10. References

- Hugging Face Transformers Documentation
- scikit-learn User Guide
- Natural Language Toolkit (NLTK)
- SM Technology Assessment Guidelines