Detecting Toxic Comments: How Far Can AI Go?

Sunesh Praveen Raja Sundarasami, Aaron Cuthinho

Abstract

With the explosion of user-generated content online, ensuring safe and respectful digital spaces has become more important than ever. In this project, we address the challenge of multi-label toxic comment classification using a modern NLP pipeline. Our approach begins with thorough text cleaning and tokenization using Ekphrasis, tailored for the quirks of social media language. We extract features through a range of techniques, including TF-IDF, Word2Vec, GloVe, and Sentence-Transformer MiniLM embeddings. For modeling, we explore both classical algorithms (Logistic Regression, LightGBM, XGBoost), deep learning architectures (BiGRU, BiLSTM, Attention), and fine-tuned transformer models (BERT, RoBERTa, GPT-1, FLAN-T5). To tackle class imbalance and boost performance, we incorporate adaptive focal loss, oversampling, and threshold tuning. Our results highlight that transformer-based models, especially with advanced fine-tuning, excel at detecting nuanced toxic language. This demonstrates the real-world impact of recent NLP advances for content moderation.

Methodology

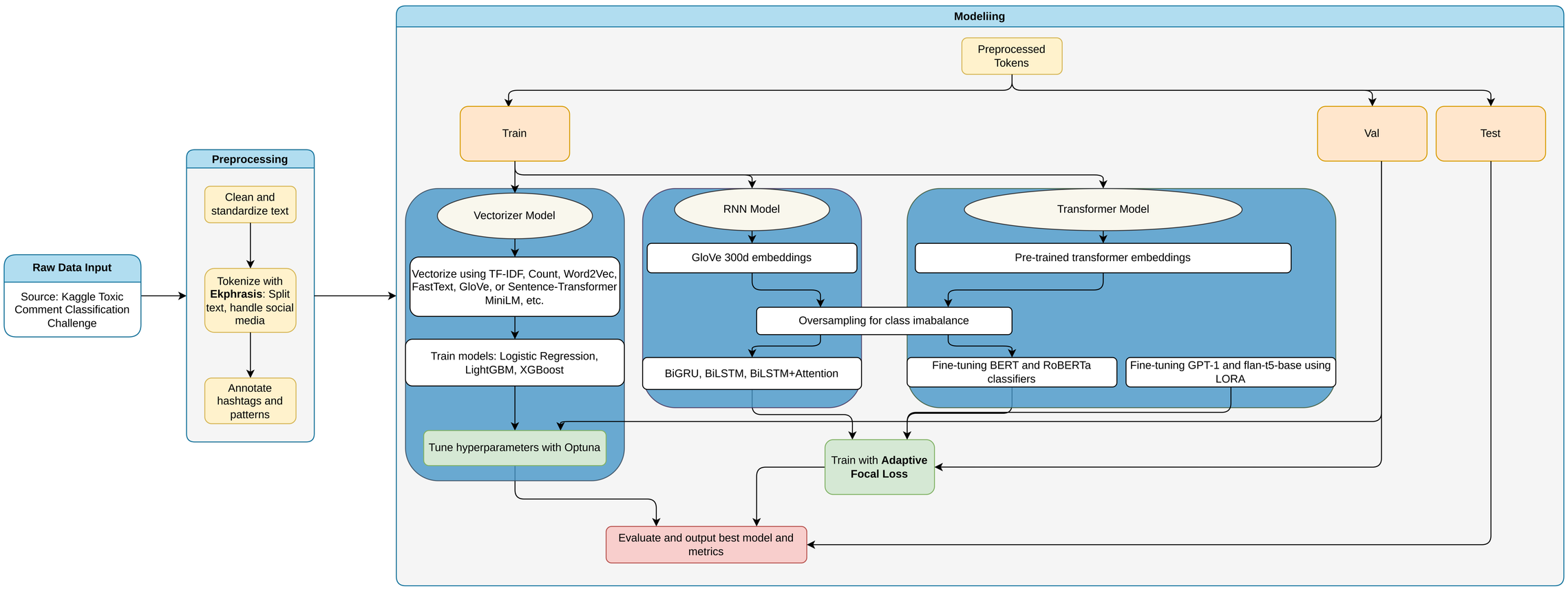
Pipeline overview:

1. **Preprocessing:** Clean and tokenize text (Ekphrasis).
2. **Features:** TF-IDF, Word2Vec, GloVe, Sentence-Transformer MiniLM, or transformer embeddings.
3. **Modeling:** Train classical models (Logistic Regression, LightGBM, XGBoost), RNNs (BiGRU, BiLSTM, Attention), and fine-tune transformers (BERT, RoBERTa, GPT-1, FLAN-T5).
4. **Optimization:** Adaptive focal loss, oversampling, threshold tuning.
5. **Evaluation:** F1-score, ROC-AUC, per-class analysis.

toxic	obscene	insult	severe_toxic	identity_hate	threat
15,294	8,449	7,877	1,595	1,405	478

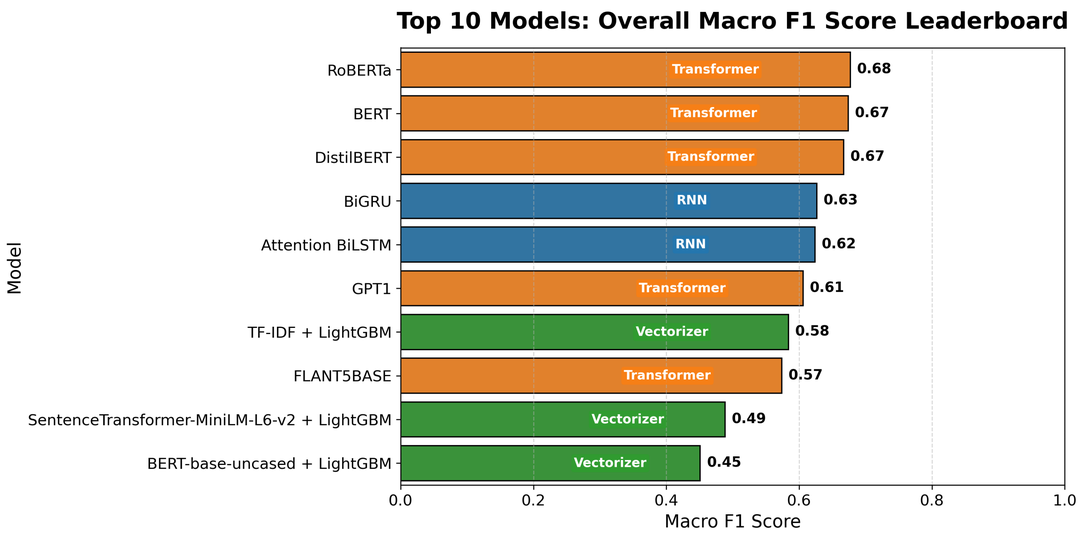
Table 1: Distribution of toxic comment labels in the dataset.

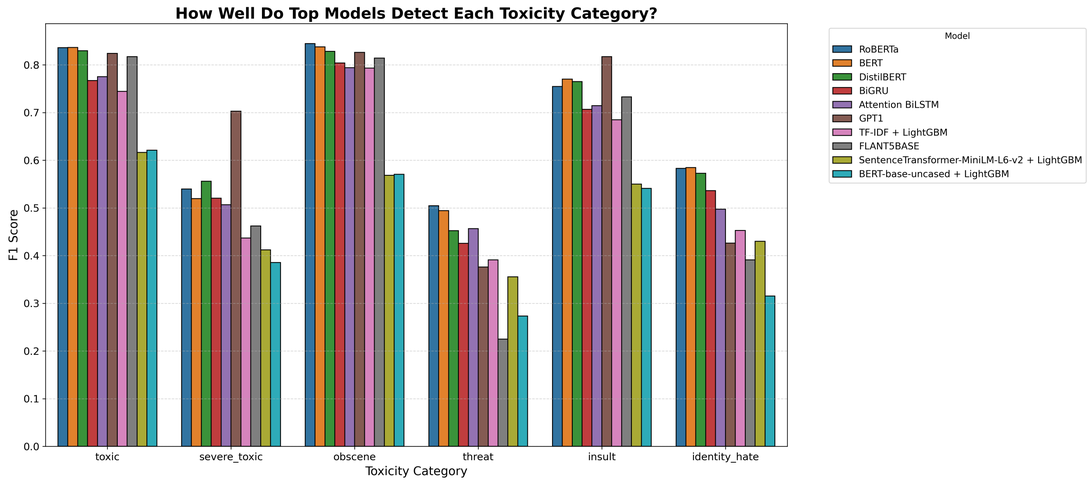
Pipeline Overview



Overview of the model pipeline for toxic comment classification

Results





Summary: Our results show that transformer models (RoBERTa, BERT, DistilBERT, GPT1) consistently outperform RNNs and vectorizer-based models both overall and across toxicity categories. However, rare classes like *threat* and *identity hate* remain difficult for all models. I found that even high-performing models tend to incorrectly flag neutral identity statements (e.g., "I'm Muslim", "I'm gay") as toxic, indicating bias learned from the dataset. In contrast, zero-shot large language models (LLMs) classify such statements as non-toxic, as expected. This highlights the importance of careful evaluation for unintended bias in toxic comment classifiers.

Link to Code and Usage Guidelines

Please scan the QR code to the side or visit the link mentioned below to access the code and usage guidelines.
`https://github.com/SuneshSundarasami/Multi_Label_Toxic_Comment_Classifier/`

References

[1] cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukierski. Toxic comment classification challenge. `https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge`, 2017. Kaggle, last accessed June 23, 2025.

[2] Jigsaw Wikipedia and Google. Wikipedia comments dataset. `https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data`, 2017. Kaggle, last accessed June 23, 2025.

Acknowledgement

We would like to thank our supervisors, Prof. Dr. Jörn Hees (joern.hees@h-brs.de) and Tim Metzler, M.Sc. (tim.metzler@h-brs.de) for providing valuable guidance and support during this project.

b-it Bonn-Aachen International Center for Information Technology

Contact

Sunesh Praveen Raja Sundarasami
Email: sunesh.sundarasami@smail.inf.h-brs.de
Aaron Cuthinho
Email: aaron.cuthinho@smail.inf.h-brs.de
Hochschule Bonn-Rhein-Sieg

Hochschule Bonn-Rhein-Sieg
University of Applied Sciences