Advancing Toxic Comment Classification: A Comparative Study of TF-IDF, SBERT, and Deep Learning Models

These authors contributed equally: Lucas Zhuang, Jiaxi Zhou, Kevin Li, Mark Yu

INTRODUCTION

Problem Statement. Online platforms face increasing challenges in moderating toxic comments, which can include **offensive, threatening, or hateful language**. Traditional keyword-based approaches often fail to capture contextual nuances and semantic relationships, leading to **misclassification** or **high false-positive rates**. Additionally, imbalanced toxicity labels pose difficulties for machine learning models, requiring strategies to improve **classification accuracy** while maintaining computational efficiency.

Objectives. This study explores various **machine learning and deep learning approaches** for multi-label toxic comment classification. We aim to:

1. Establishing a **baseline** with **TF-IDF + Naive Bayes** to assess its handling of imbalanced labels.
2. **Improve performance** using **TF-IDF + Logistic Regression**, with hyperparameter tuning.
3. **Enhance feature representation** with **SBERT embeddings**, evaluating impact on classification.
4. **Explore deep learning** by fine-tuning **RoBERTa**, a transformer-based language model.
5. **Compare traditional and deep learning methods** to determine the most effective approach.

Through this **comparative analysis**, we aim to identify the **optimal balance between performance, interpretability, and computational efficiency**, offering insights for **better content moderation** on online platforms.

DATASET ANALYSIS

Dataset Overview.

The dataset contains **159,571 comments**, each labeled with one or more of six toxicity categories: **toxic, severe toxic, obscene, threat, insult, and identity hate**. Labels are binary, where 1 indicates the presence of toxicity and 0 indicates its absence.

Exploratory Data Analysis (EDA).

Label Distribution. The dataset is highly imbalanced, with certain categories being significantly underrepresented:

Toxic: 9.58%

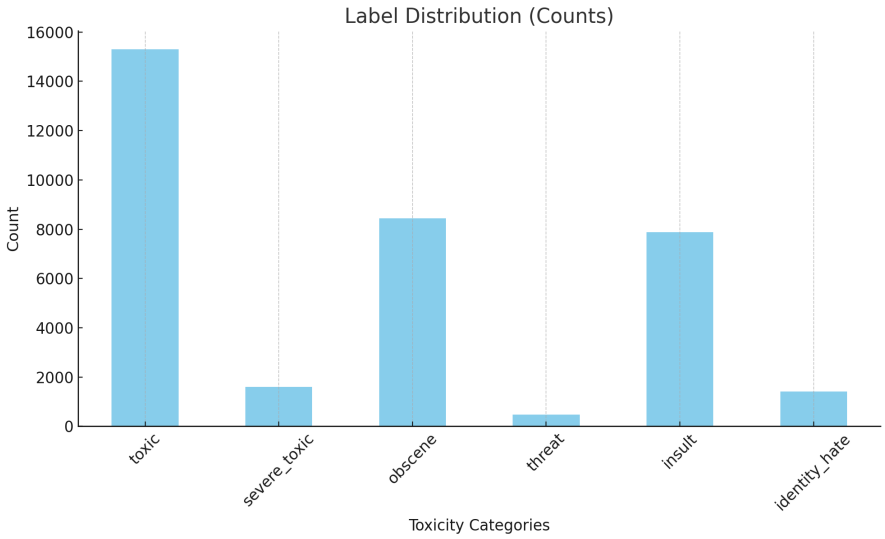
Severe Toxic: 0.99%

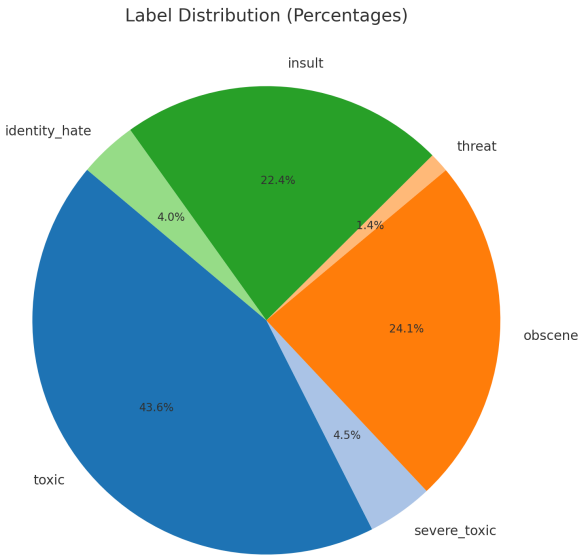
Obscene: 5.29%

Threat: 0.30%

Insult: 4.94%

Identity Hate: 0.88%



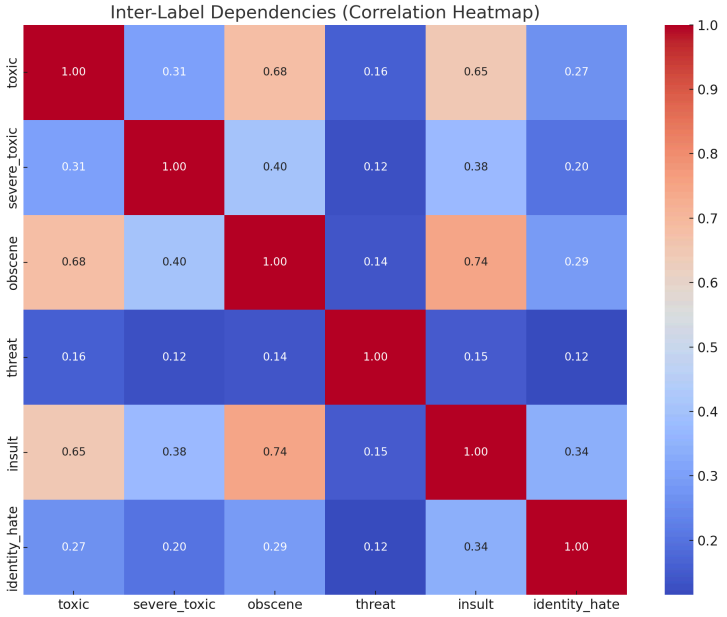


Label Counts:

- Toxic: 15,294 (9.58%)
- Severe Toxic: 1,595 (0.99%)
- Obscene: 8,449 (5.29%)
- Threat: 478 (0.30%)
- Insult: 7,877 (4.94%)
- Identity Hate: 1,405 (0.88%)

Inter-Label Dependencies. A correlation analysis revealed the following key insights:

- **Obscene & Insult (0.74)**: Strongest correlation, indicating significant overlap.
- **Obscene & Toxic (0.68)** and Insult & Toxic (0.65): High correlation, meaning these labels frequently co-occur.
- **Identity Hate & Insult (0.34)**: Moderate correlation, suggesting identity-based insults occur but are less frequent.



- **Threat has low correlation with other categories**, indicating it appears more independently.

Preprocessing Pipeline. To improve data quality and model performance, the following preprocessing steps were applied:

- **Lowercasing & Normalization**: Standardized text, removed excessive punctuation and elongated words.
- **Selective Punctuation Preservation**: Retained symbols like "@, #, !, ?, *" for context.
- **Character & Symbol Cleaning**: Removed non-essential characters for a cleaner dataset.

These steps ensured a clean, standardized, and structured dataset for improved model training and classification accuracy.

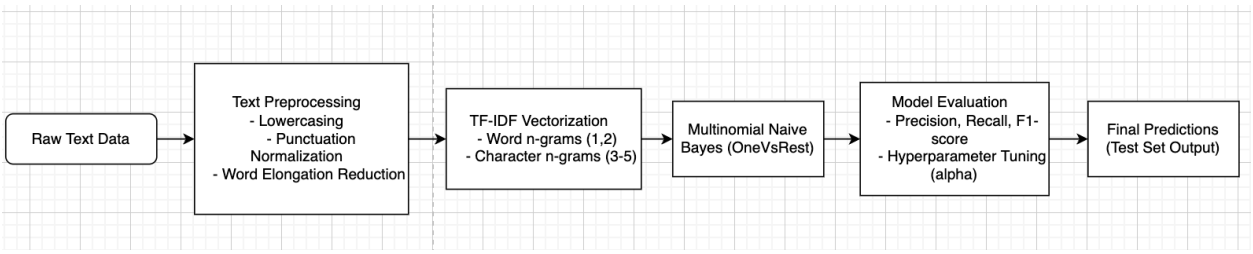
EXPERIMENTS

Experiment 1: TF-IDF + Naive Bayes *(Jiaxi Zhou)*

Method Details. This experiment employs **Term Frequency-Inverse Document Frequency (TF-IDF)** for feature extraction and **Multinomial Naive Bayes (MNB)** for classification within a **OneVsRest framework**. TF-IDF converts text into numerical features using:

- **Word-level TF-IDF:** Capturing unigram and bigram features to represent word patterns.
- **Character-level TF-IDF:** Extracting 3-5 character n-grams to capture subwords.

Purpose. **MNB**, a probabilistic classifier, is chosen for its efficiency in handling high-dimensional, sparse data. The **OneVsRest structure** enables independent classification for each toxicity label. This experiment establishes a **baseline model** to evaluate how well a simple probabilistic approach handles **imbalanced toxicity labels**, serving as a reference for more advanced techniques.



Experiment 2: TF-IDF + Logistic Regression *(Mark Yu)*

Method Details. To improve upon the Naive Bayes baseline, this experiment replaces MNB with **Logistic Regression**, a linear model better suited for **large, sparse textual data**. Key enhancements include:

- **Hyperparameter Tuning**: Randomized search optimization for regularization strength ($C \approx 21.37$) to balance bias-variance tradeoff.
- **TF-IDF Feature Refinements**: **Word-level TF-IDF**: Unigrams and bigrams, max features = 20,000. **Character-level TF-IDF**: 3-5 character n-grams, max features = 30,000.

Purpose. This experiment aims to **assess improvements over the Naive Bayes baseline** by using a more expressive linear model. Additionally, it evaluates the impact of hyperparameter tuning and class balancing on classification performance.

Experiment 3: SBERT Encoded Features + Logistic Regression *(Lucas Zhuang)*

Method Details. This experiment introduces **Sentence-BERT (SBERT) embeddings** as an alternative to TF-IDF for feature representation, combined with **Logistic Regression** for classification.

- **SBERT (Sentence-BERT)**: A variant of BERT optimized for sentence embeddings using a **Siamese network structure with mean-pooling**. Unlike TF-IDF, SBERT **preserves contextual meaning** by encoding semantic relationships within the text.
- **Feature Extraction**: SBERT embeddings replace sparse TF-IDF features with dense, high-dimensional representations, enabling improved text understanding.
- **Classification:** SBERT embeddings serve as input for a **OneVsRest Logistic Regression**, enabling independent multi-label classification.

Purpose. This approach seeks to **leverage deep semantic representations** for improved text classification. By capturing richer context, SBERT embeddings provide a **more advanced feature representation** than TF-IDF, potentially enhancing classification accuracy while maintaining computational efficiency with Logistic Regression.

Experiment 4: SBERT Encoded Features + XGBoost *(Kevin Li)*

Method Details. This experiment replaces **Logistic Regression** with **XGBoost**, a gradient boosting decision tree model, while retaining **SBERT embeddings** for feature representation.

- **SBERT Feature Extraction**: Provides dense, contextualized embeddings, capturing semantic relationships missing in TF-IDF.
- **Baseline Decision Tree**: Tested a standard decision tree to assess tree-based model performance on SBERT embeddings.

- **Gradient Boosting with XGBoost**: Addressed overfitting by optimizing tree learning with gradient boosting.

Purpose. This experiment evaluates whether **XGBoost can better utilize SBERT embeddings** compared to Logistic Regression, leveraging **non-linearity and boosting techniques** for improved classification.

Experiment 5: RoBERTa End-to-End Fine-Tuning *(Kevin Li)*

Method Details. This experiment fine-tunes **RoBERTa-base** for multi-label toxic comment classification, using context-aware feature extraction instead of static embeddings like **TF-IDF or SBERT**.

- Dynamic Feature Learning: Unlike TF-IDF and SBERT, RoBERTa learns task-specific representations during training.
- End-to-End Fine-Tuning: The model is trained directly on the classification task without modifying hyperparameters.

Purpose. This experiment evaluates whether fine-tuning RoBERTa improves classification performance compared to precomputed embeddings. Despite higher computational costs, RoBERTa outperformed all previous models, demonstrating its superior contextual understanding with minimal optimization.

DISCUSSION AND INSIGHTS

Model/Feature Transformation	Validation ROC-AUC	Kaggle Public ROC-AUC
Naive Bayes + TF-IDF	0.9449	0.94468
Logistic Regression + TF-IDF	0.9762	0.97422
Logistic Regression + SBERT	0.96844	0.96966
XGBoost + SBERT	0.9709	0.95832
RoBERTa	-	0.97828

Comparative Analysis: TF-IDF + Logistic Regression vs. SBERT + Logistic Regression

Feature Representation. **TF-IDF** captures term frequency but lacks semantic understanding, relying on word and character n-grams. **SBERT** generates dense, semantic-rich embeddings, preserving contextual relationships between words and phrases.

Model Complexity & Efficiency. **TF-IDF + Logistic Regression** is computationally efficient and interpretable, performing well on sparse data but lacking semantic depth. **SBERT + Logistic Regression** requires more resources due to embedding extraction but improves feature quality.

Performance Comparison.

- **TF-IDF + Logistic Regression (Untuned)** outperformed **SBERT + Logistic Regression** on the leaderboard, indicating that TF-IDF remains a strong baseline for toxic comment classification.
- **SBERT embeddings** improved contextual representation but did not provide a significant performance boost when paired with a simple linear model like Logistic Regression.

Both models perform well, but **TF-IDF + Logistic Regression (Untuned) achieves a slightly higher leaderboard score**. While SBERT embeddings offer richer representations, **Logistic Regression may not fully leverage their potential**.

Comparative Analysis: TF-IDF + Logistic Regression vs. Naive Bayes *(Jiaxi Zhou)*

Feature Representation. **Both models** use TF-IDF with word and character n-grams for feature extraction.

Model Complexity & Efficiency. **Naive Bayes** assumes feature independence, making it less flexible. **Logistic Regression** captures more complex relationships, improving classification accuracy.

Performance Comparison.

Logistic Regression **outperforms Naive Bayes**, especially after tuning. It effectively handles **class imbalance** with class_weight='balanced' and optimizes efficiently using **RandomizedSearchCV**, unlike Naive Bayes, which lacks direct balancing and uses less flexible tuning. **Overall, Logistic Regression is the superior choice** for imbalanced toxicity classification.

CONCLUSION

Summary of Findings. This study explored **traditional and deep learning approaches** for *toxic comment classification*, comparing TF-IDF with Naive Bayes, TF-IDF with Logistic Regression, SBERT embeddings, and deep learning models. **Logistic Regression with TF-IDF proved to be a strong baseline**, outperforming Naive Bayes. **SBERT embeddings** improved feature representation but did not significantly enhance performance with a linear model. These findings suggest that while traditional models remain competitive, deep learning methods may require more advanced architectures to fully utilize contextual embeddings.

Future Work. Future research can explore (1) **End-to-end transformer models** (e.g., **fine-tuned RoBERTa**) to fully leverage deep contextual representations. (2) **Hybrid models** combining TF-IDF and embeddings to balance efficiency and semantic understanding. (3) **Addressing dataset imbalance** using advanced resampling techniques or loss function adjustments. (4) Optimizing deep learning architectures will be key to **further improving accuracy and scalability** in toxic comment detection.