Performance Analysis Report

Project Title: Offensive Language Classification
Prepared By: *Md. Sajid Alam Chowdhury*
Email: sajid.chowdhury009@gmail.com

Objective:

The goal of this project was to develop a robust machine learning pipeline capable of identifying various forms of offensive content in user feedback. Each piece of feedback could be associated with none, one, or multiple offensive labels, making this a **multi-label classification** problem.

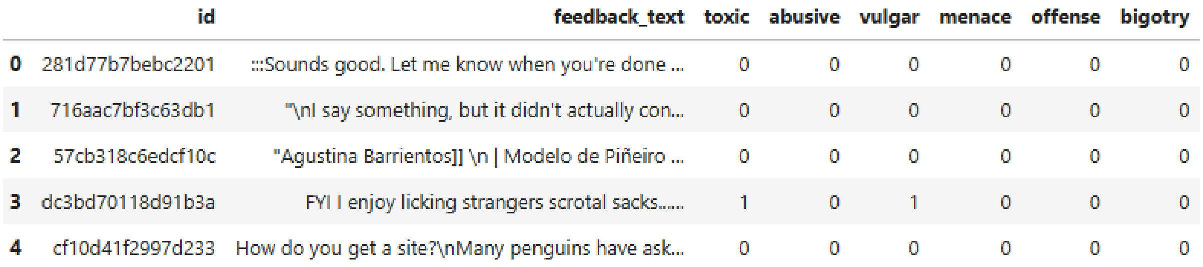
Dataset Overview:

Training Set:

- Contains English-language comments annotated with six binary labels:

  toxic, abusive, vulgar, menace, offense, bigotry

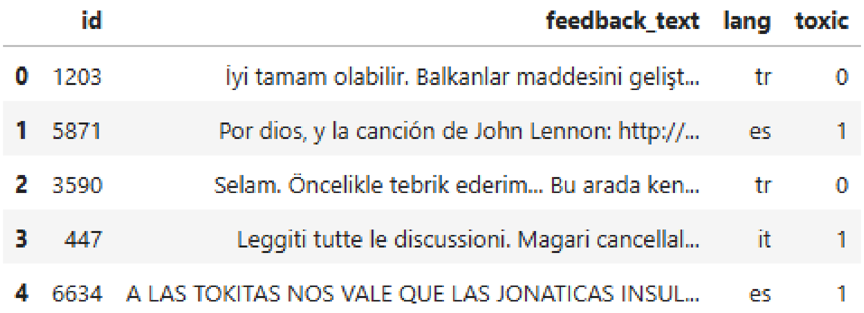
  Each label is independent and non-exclusive.



	id	feedback_text	toxic	abusive	vulgar	menace	offense	bigotry
0	281d77b7bebc2201	:::Sounds good. Let me know when you're done ...	0	0	0	0	0	0
1	716aac7bf3c63db1	"\nI say something, but it didn't actually con...	0	0	0	0	0	0
2	57cb318c6edcf10c	"Agustina Barrientos]] \n | Modelo de Piñeiro ...	0	0	0	0	0	0
3	dc3bd70118d91b3a	FYI I enjoy licking strangers scrotal sacks......	1	0	1	0	0	0
4	cf10d41f2997d233	How do you get a site?\nMany penguins have ask...	0	0	0	0	0	0

Validation and Test Set:

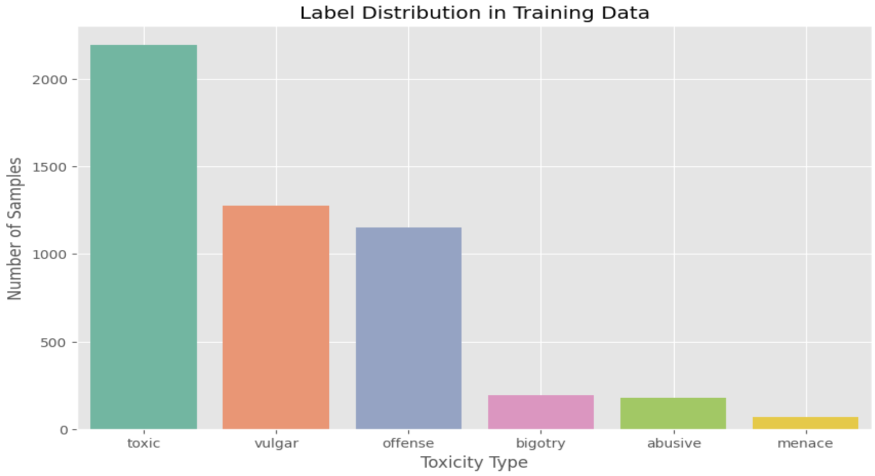
- Contains unlabeled feedback written in various languages with a specific column mentioning the corresponding language.
- Only the toxic label was provided for evaluation.
- To ensure compatibility with the English-only training data, the test data was translated to English using the Deep Translator library prior to preprocessing and prediction.



	id	feedback_text	lang	toxic
0	1203	İyi tamam olabilir. Balkanlar maddesini gelişt...	tr	0
1	5871	Por dios, y la canción de John Lennon: http://...	es	1
2	3590	Selam. Öncelikle tebrik ederim... Bu arada ken...	tr	0
3	447	Leggiti tutte le discussioni. Magari cancellal...	it	1
4	6634	A LAS TOKITAS NOS VALE QUE LAS JONATICAS INSUL...	es	1

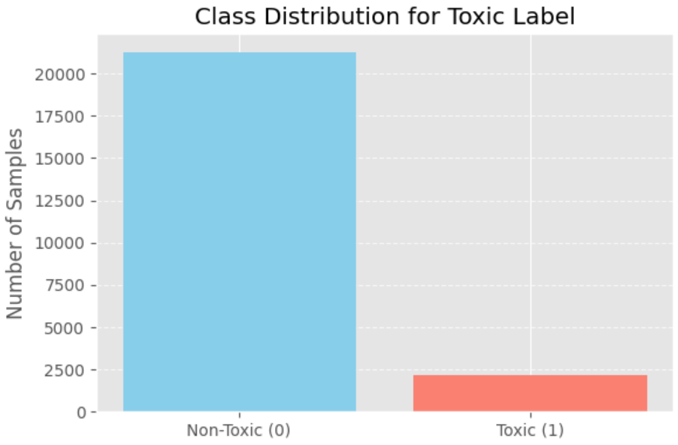
Problem Framing and Challenges:

- **Multi-Label Classification:**
  The Binary Relevance method was used, transforming the multi-label problem into independent binary classification tasks per label.



- **Language Difference:**
  A major challenge was the linguistic difference between the English-only training set and the multilingual test set. Even after translation, meanings, semantics and contexts might not be perfectly preserved, impacting model performance.

- **Class Imbalance:**
  The dataset exhibited significant label imbalance, particularly for labels such as menace and bigotry, which had far fewer positive samples than toxic or offense. This led to high accuracy (dominated by the majority class) but lower F1-scores, which are more sensitive to imbalanced data.



Preprocessing Steps:

- Translated non-English test data to English.
- Lowercased Convert text to lowercase
- Removed stop words, special characters, and punctuation

- Applied Tokenization to split sentences into words
- Stemming/Lemmatization to mormalize words to their root form
- Feature extraction methods:

  ML Models: Used TF-IDF vectorization and dense embeddings for textual features.

  DL Models (LSTM, GRU): Applied tokenization with learned or pre-trained embedding layers.

  Transformer Models (BERT, XLM-R): Used model-specific tokenization and contextual embeddings generated during fine-tuning.

Models Implemented:

Traditional Machine Learning Models

1. Logistic Regression (LR)
2. Random Forest (RF)

Both were used with Binary Relevance and TF-IDF features. Also, multilingual sentence embeddings was used to handle multilingual texts. Hyperparameter tuning was done using:

- GridSearchCV for LR
- RandomizedSearchCV for RF

Deep Learning Models

1. LSTM (Long Short-Term Memory)
2. GRU (Gated Recurrent Unit)

- The input texts were tokenized and transformed into padded sequences.

- An embedding layer was used to convert the integer sequences into dense vector representations.

- Both models were trained using the Binary Cross Entropy loss function, suitable for multi-label classification, combined with sigmoid activation at the output layer to independently handle each label.

Transformer-Based Models

1. BERT (base-uncased)
2. XLM-Roberta (XLM-R)

- The input texts were tokenized using the respective pre-trained tokenizers from Hugging Face Transformers.

- Texts were converted into input IDs, attention masks, and token type IDs, then passed through the pre-trained encoder (BERT/XLM-R).

- The pooled output from the encoder was passed through multiple dense layers with dropout for regularization. The final output layer used sigmoid activation to support multi-label prediction.

- Models were compiled with Binary Cross Entropy loss and trained using the AdamW optimizer.

Justification:

BERT: Strong contextual understanding for English texts.

XLM-RoBERTa: Ideal for multilingual data and cross-lingual generalization.

Evaluation Metrics:

Metrics used: (for ‘toxic’ label)

- Accuracy
- Precision, Recall, F1-Score
- ROC-AUC Curve
- Confusion Matrix

Evaluation was based on the translated test data, with known labels only for the toxic class.

Performance Summary (For Toxic Label):

Type	Model	Accuracy	Precision	Recall	F1-Score	AUC
ML	**Logistic Regression**	0.8065	0.6010	0.4409	0.5087	0.7850
	Random Forest	0.4480	0.2552	0.7454	0.3802	0.5722
DL	**LSTM**	0.7988	0.5651	0.4967	0.5287	0.7839
	GRU	0.8005	0.5570	0.5950	0.5754	0.8192
TF	**BERT**	**0.8337**	**0.6535**	0.5701	0.6089	**0.8723**
	XLM-RoBERTa	0.8268	0.6205	**0.6119**	**0.6162**	0.8666

Key Observations:

- XLM-Roberta outperformed all other models, particularly due to its multilingual training capabilities, which helped mitigate translation loss.
- Although accuracy was high across models, F1-scores were relatively lower due to:
  - **Class imbalance**, leading to poor recall for underrepresented classes.
  - **Translation discrepancies** introducing semantic drift between test and train data.

Conclusion:

- Traditional models like RF and LR provided solid baselines, but transformer-based models, especially XLM-Roberta, demonstrated superior performance.
- Binary Relevance was an effective strategy for managing label independence.
- Future improvements could include:
  - Applying class reweighting or resampling to handle imbalance
  - Fine-tuning multilingual models with translated training data
  - Experimenting with multi-label transformer architectures (e.g., BCE + sigmoid on top of BERT)