Multilingual Toxic Comment Classification

INFO 539: Statistical NLP Term Project
Student: Satwik Gudapati

---

1. Introduction

Online platforms face increasing challenges around toxic user comments. Automatic toxic comment classification, especially across multiple languages, is crucial to maintaining safe online spaces. In this project, we fine-tune a multilingual transformer model on a sampled subset of toxic comments to create an efficient multilingual toxicity classifier.

---

2. Dataset

We use the **Kaggle Jigsaw Multilingual Toxic Comment Classification** dataset.
Due to hardware (Mac MPS) memory limitations, we randomly sampled **1000 comments** from the full dataset to create a smaller working subset.

The dataset contains multilingual user comments labeled for multiple types of toxicity:

- Toxic
- Severe toxic
- Obscene
- Threat
- Insult
- Identity hate

---

3. Preprocessing

- Sampled 1000 rows randomly.
- Corrected multi-label format (each comment has six float labels).
- Tokenized using the `xlm-roberta-base` tokenizer with a maximum token length of 128.

---

4. Model and Training

- **Model:** Pretrained `xlm-roberta-base`
- **Fine-tuning task:** Multilabel classification
- **Batch size:** 2
- **Number of epochs:** 1
- **Learning rate:** 2e-5
- **Loss function:** Binary Cross Entropy with Logits Loss (BCEWithLogitsLoss)

Training was completed successfully.
Final Training Loss achieved: ~0.1869

---

5. Challenges

- **Kaggle API authentication error:** Resolved by manually downloading the dataset.
- **Memory Limitations:** On Mac M3 GPU (MPS backend), training large models can cause memory overflow. This was addressed by:
    - Sampling a small dataset (1000 rows)
    - Reducing batch size to 2
    - Limiting token length to 128

---

6. Results and Future Work

- **Results:** Successfully fine-tuned a transformer model with a final loss of ~0.1869 on the sampled dataset.
- **Future Work:**
    - Fine-tune on full dataset for better generalization
    - Evaluate zero-shot transfer across non-English languages
    - Experiment with larger models (XLM-Roberta-Large) and language-specific fine-tuning

---

7. Conclusion

This project demonstrated that transformer models like XLM-Roberta can effectively adapt to multilingual toxic comment detection, even with limited data and compute. With further scaling, this system can significantly enhance multilingual content moderation efforts.