LULEÅ UNIVERSITY OF TECHNOLOGY

Hate speech detection

Group 11

21th May 2025

Introduction

Hate speech detection has emerged as a critical task in the field of natural language processing (NLP), especially in social media moderation.

Goal: Detect and classify hate speech using NLP techniques.

This work presents a comparative analysis of transformer-based language models including

BERT, and RoBERTa, applied to the HASOC 2019 dataset.

Methodology - Dataset/Preprocessing

HASOC 2024
Hate Speech and offensive Content Identification in Multiple Languages

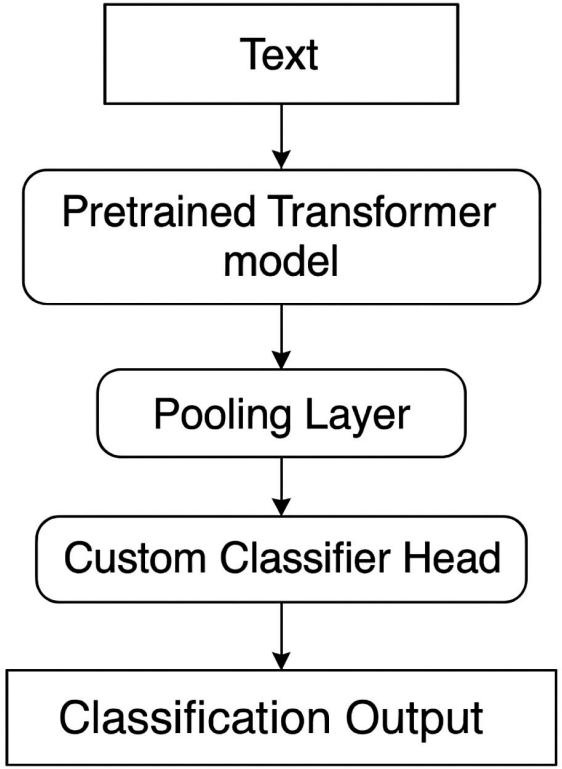
- Use HASOC 2019 dataset [1]
  - Forum and data challenge for multilingual research
  - Provide dataset for Hate Speech and Offensive Content Identification
  - Can be used for various hate speech detection tasks

- Preprocessing
  - Remove twitter usernames, URLs, hashtags, special characters
  - Convert sentences to lowercase

[1] https://hasocfire.github.io

Methodology - Model Architecture

1. Pretrained Transformer model
   a. BERT (base-uncased)
   b. RoBERTa (base, large, cardiffnlp/twitter-roberta-base-sentiment-latest)
2. Pooling Layer
   a. CLS
   b. Mean
   c. Max
   d. Attention-pooling
3. Custom Classifier Head
   a. Shallow ANN
   b. Deeper ANN
   c. CNN



Methodology - Training

- We used AdamW optimizer. And Set batch size = 16 and CrossEntropyLoss for binary classification.
- We varied hyperparameters such as epochs and learning rates.

- **For RoBERTa:**
  - Train one model with frozen layers.
  - Train another with all layers fine-tuned.

Training–cont

Experiment Tracking

- We used Weights & Biases (WandB) to:
  - Log training/validation accuracy, F1 score, loss.
  - Sweep over hyperparameters to find optimal settings.
  - Compare runs visually.

Model Evaluation

- We evaluated on the test set.
- We Computed:
  - Accuracy
  - Macro F1 Score
  - Confusion matrix (TP, FP, FN, TN)

Results

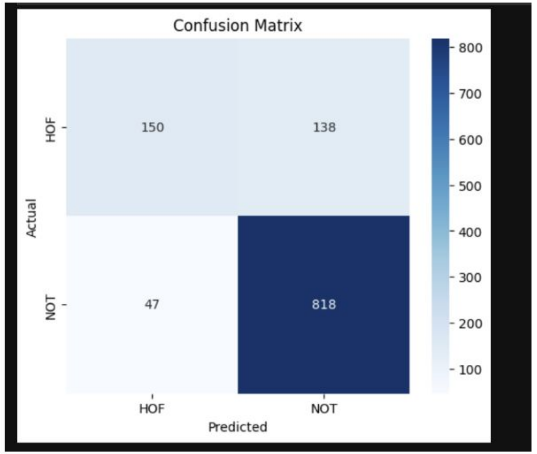


Fig. 4. Validation accuracy and loss for the Roberta transformer.

TABLE I
ROBERTA TEST RESULTS

F1	accuracy
0.89	0.83

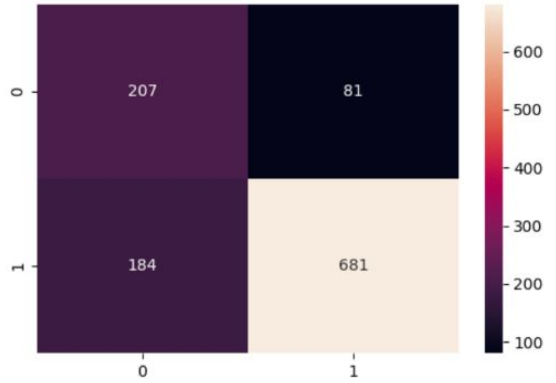


Fig. 6. Validation accuracy and loss for the BERT transformer.

TABLE II
BERT TEST RESULTS

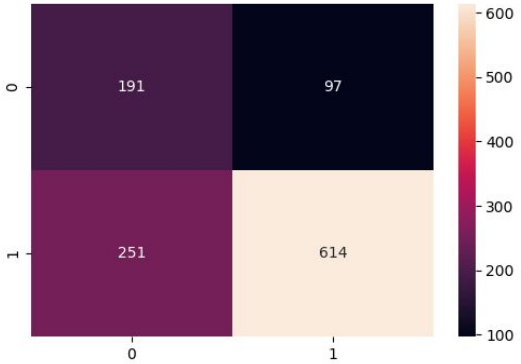
F1	accuracy
0.72	0.77

Additional results. XLNet

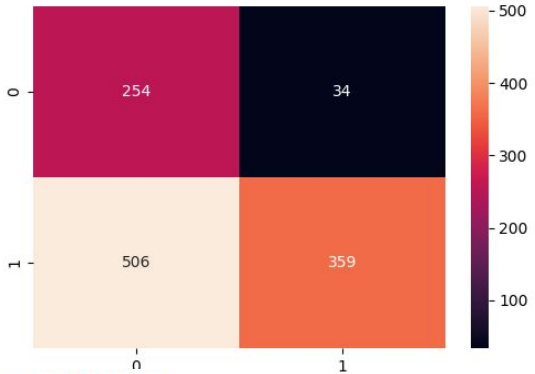
Aspect	Baseline paper setup (BERT / RoBERTa)	XLNet v1 – "2-layer / CNN / 128"	XLNet v2 – "full FT / CNN / 256"
Encoder checkpoint	bert-base, roberta-base, roberta-large, twitter-roberta	xlnet-base-cased	same xlnet-base-cased
Layers trainable	0 or top 1–2	2 / 12 (top layers)	12 / 12 (all layers)
Max-sequence length	128	128	256
Pooling in use	CLS / Mean / Max / Attention (ablation)	Max pooling inside CNN head	Max pooling inside CNN head
Classifier head	CNN, shallow FFN, deep FFN (grid)	CNN head (BaseCNNClassifier)	same CNN head (dropout 0.3 added internally)
Class-imbalance handlin	class weights / sampling / focal	class weights	class weights
Optimised params	head (+1–2 enc. layers if unfrozen)	head + 2 layers	all parameters
Learning rate	2 e-4 (head) or 2 e-5 (full FT)	2 e-5, no scheduler	2 e-6, linear warm-up 10 % (LR→0 by epoch 8)
Epochs run	5–10	20	20
Test F1 (HASOC 19)	RoBERTa ≈ 0.86	0.65 (Acc 0.70)	0.53 (Acc 0.53)
Main bottleneck	CLS may miss context	Too few layers, but solid generalisation	LR decays to 0 → weights stop learning

Additional results. XLNet

Conf.1



Config.2



XLNet Test Results for Partial vs. Full Fine–Tuning

Configuration	F1	Accuracy
Partial FT (top 2/12 layers, max_len=128)	0.6512	0.6982
Full FT (all 12/12 layers, max_len=256)	0.5277	0.5317

Discussion

- We compared fine-tuned vs. frozen Roberta model in accuracy and robustness
- We compared Roberta ,Bert models , and XLNET.
- Best Performance: Fine-tuned RoBERTa achieved F1 score of 0.89
- Frozen Models: Performed decently but less adaptable.
- BERT has lower performance than BERT, XLNet has lower performance than BERT and RoBERTa for our dataset

Fine-tuning Impact:

- Boosted model adaptability to hate speech detection.
- Helped reduce overfitting when partial layers were unfrozen.

Conclusion:

- XLNet’s architectural advantages did not translate on this small, short-text task: partial freezing underfits, and full fine-tuning overfits under the chosen schedule.
- Model choice and hyperparameter tuning must align with dataset characteristics
- For HASOC Task 1, RoBERTa proved the most robust and efficient.

Github Link

https://github.com/juliusatgit/HateSpeechDetection