Abusive Language Detection Using HateBERT LLM

Created By :

Siddhesh Bangar
Aditya Satheesan
Shivam Bhor
Saloni Deshmukh

Table Of Contents

Introduction

The rapid growth of online communication has led to an increase in abusive and harmful content across digital platforms. Identifying and moderating such content is a critical challenge for maintaining healthy online interactions.

This project aims to build a robust **Abusive Language Detection System** using **HateBERT Large Language Models (LLMs)**, specifically **hateBERT**. By fine-tuning the model on a custom dataset, it enhances the accuracy of abusive content classification.

The system is integrated into a **Streamlit** web interface for real-time detection and moderation of harmful language, providing a practical solution for safer online interactions.

Literature Survey

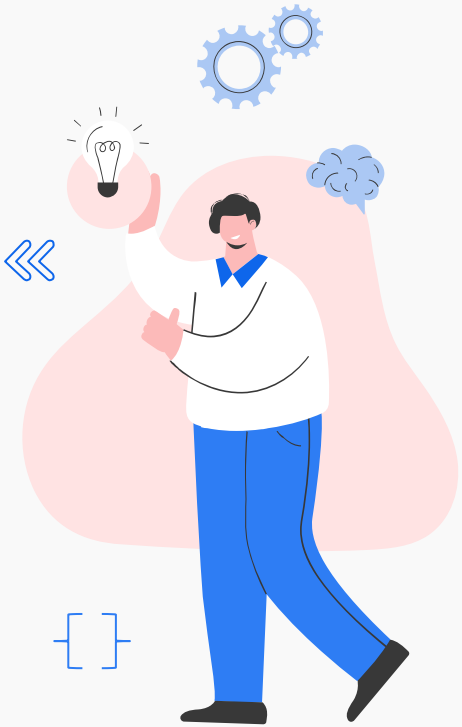
SN	Paper Name	Techniques	Author & Year of Publication	Advantages and Disadvantages
1.	Abusive Language Detection from Social Media Comments Using ML & DL Approaches	Machine Learning: NB, SVM, IBK, Logistic, JRip Deep Learning: CNN, LSTM, BLSTM, CLSTM	Muhammad Pervez Akhter et al., 2021	Advantages:- CNN achieved the best accuracy (96.2% Urdu, 91.4% Roman Urdu) One-layer DL models outperformed two-layer Disadvantages:- Roman Urdu has challenges due to lack of standard grammar and dictionary Processing Urdu script is complex
2.	Abusive and Threatening Language Detection in Urdu Using Boosting and BERT Models	Machine Learning: XGBoost, LightGBM Transformer-based: mBERT,	Mithun Das et al., 2021	Advantages: - dehatebert-mono-arabic achieved the best F1 score (0.88 for abusive, 0.54 for threatening content) Disadvantages- Significant imbalance in dataset for threatening language detection
3.	Abusive Language Detection using NLP	NLP techniques: N-gram features, linguistic features, syntactic features, and distributional semantics features	Kandarpa Venkata Abhiram, Panigrahi Srikanth (2022)	Advantages: - dehatebert-mono-arabic achieved the best F1 score (0.88 for abusive, 0.54 for threatening content) Disadvantages:- Significant imbalance in dataset for threatening language detection - Threatening language detection had low F1 scores

Literature Survey

SN	Paper Name	Techniques	Author & Year of Publication	Advantages and Disadvantages
4.	Deep Learning-based Approaches for Abusive Content Detection [7]	Various deep learning models: BiLSTM (Bidirectional Long Short-Term Memory), GRU (Gated Recurrent Unit), LSTM (Long Short-Term Memory) Feature selection via TF-IDF (Term Frequency-Inverse Document Frequency)	Simrat Kaur, Sarbjeet Singh, Sakshi Kaushal (2024)	Advantages: - dehatebert-mono-arabic achieved the best F1 score (0.88 for abusive, 0.54 for threatening content). Disadvantage:- Significant imbalance in dataset for threatening language detection - Threatening language detection had low F1 scores
5.	Abusive Words Detection on Reddit Using Machine Learning Algorithms	Machine Learning: SVM, Random Forest, XGBoost, CNN, Gradient Boosting Machine (GBM)	Madhurima Suseelan et al., 2024	Advantages: - Random Forest performed best with 99% accuracy, especially in recognizing abusive content. It does not have performance problems. Disadvantage: - Limited to sentiment analysis (upbeat, neutral, downbeat)

Objectives

- Detect abusive language using LLMs
- Fine-tune hateBERT on a custom dataset
- Real-time detection via Streamlit
- Improve accuracy with custom data



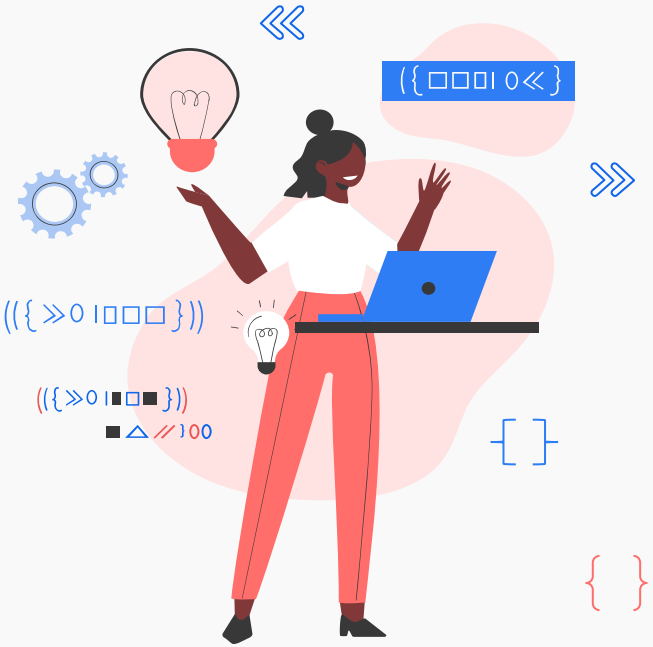
Abusive And Non-Abusive

Abusive Language	Non-Abusive Language
Contains insults, threats, or derogatory terms	Respectful and polite communication
Aims to harm or belittle individuals/groups	Supports constructive dialogue
Often includes hate speech or discrimination	Promotes understanding and inclusivity
Can incite violence or provoke strong reactions	Encourages positive interactions and feedback
May lead to emotional distress or psychological harm	Fosters a safe and supportive environment
Examples include slurs, harassment, and bullying	Examples include compliments, support, and advice

Proposed System

The Abusive Language Detection system gathers and preprocesses a diverse dataset of abusive and non-abusive language. The **hateBERT** model is fine-tuned for improved detection accuracy.

It features a user-friendly **Streamlit** frontend for real-time analysis and feedback on text classification, with performance evaluated through metrics like accuracy and F1-score.



Dataset Used

text	label
Why would you say something like that at a graduation? you are there to influence. Mistakes happen witl	0
Whatever!!! The pizza guy should have given their change back and let them give him a tip. He just	1
Everybody giving hate in the comments to someone who dedicated their life to improve womenâ€™s righ	0
1:07 when ur girl sees your dick	1
Well fuck Religion	1
Yeah maybe donâ€™t do that to your judge	0
Dawkins mentioned the marvelous gift of life, the gift of understanding..... My question is where did t	0
Lol suck one black guy and lady hahaha	1
Cry.. Cry.. Cry some more.. and then grow the fuck up one day	1
lol she's like one of them rats in new york	0

0 = Non-Abusive Text

1 = Abusive Text

System Requirements

SOFTWARE :

Windows 10 OS
Python 3.10.8
VS Code

HARDWARE :

8GB RAM
4GB GPU RTX
256GB SSD
i5 10th Gen Processor

Technologies Used

Transformers

Streamlit

Pandas

Scikit-learn

Torch

Result

Abusive Language Detection using HateBERT LLM

Enter a sentence or paragraph to analyze:

Detect Abusive Content

This app uses a fine-tuned HateBERT model for detecting abusive language in text.

Abusive Text

Abusive Language Detection using HateBERT LLM

Enter a sentence or paragraph to analyze:

Why don't you go and die, you ugly freak!

Detect Abusive Content

Abusive content detected!

This app uses a fine-tuned HateBERT model for detecting abusive language in text.

Non-Abusive Text

Abusive Language Detection using HateBERT LLM

Enter a sentence or paragraph to analyze:

All blacks are criminals.

Detect Abusive Content

Non-abusive content.

This app uses a fine-tuned HateBERT model for detecting abusive language in text.

Accuracy of Model

[348/348 02:22, Epoch 3/3]

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall
1	No log	0.246574	0.913420	0.954338	0.874477
2	No log	0.248980	0.924242	0.963636	0.887029
3	No log	0.253921	0.928571	0.932773	0.928870

```
TrainOutput(global_step=348, training_loss=0.23477205736883755, metrics=
{'train_runtime': 147.301, 'train_samples_per_second': 37.576,
'train_steps_per_second': 2.363, 'total_flos': 364079922854400.0, 'train_loss':
0.23477205736883755, 'epoch': 3.0})
```

Future Scope

Multi-language Support: Expand the system to handle abusive language in multiple languages.

Integration with Social Platforms: Deploy as a plugin for real-time moderation on social media and forums.

Conclusion

The Abusive Language Detection system using fine-tuned hateBERT effectively identifies harmful content, providing a practical solution for real-time moderation. It promotes safer online interactions and serves as a strong foundation for future enhancements like multi-language support and context-aware detection.

References

1. Akhter, M. P., Jiangbin, Z., Naqvi, S. I. R., AbdelMajeed, M., & Zia, T. (2021). Abusive language detection from social media comments using conventional machine learning and deep learning approaches. Multimedia Systems.

2. Das, M., Banerjee, S., & Saha, P. (2021). Abusive and Threatening Language Detection in Urdu using Boosting based and BERT based models: A Comparative Approach. FIRE 2021: Forum for Information Retrieval Evaluation

3. Kandarpa Venkata Abhiram, Panigrahi Srikanth (2022). "Abusive Language Detection Using NLP." *International Journal of Creative Research Thoughts (IJCRT)*, Volume 10, Issue 11. https://doi.org/IJCRT2211063

4. Suseelan, M., Boppuru, P. R., Ajith, K. A., & Swathy, V. S. (2024). Abusive Words Detection on Reddit Comments Using Machine Learning Algorithms. 2nd International Conference on Device Intelligence, Computing and Communication Technologies .

Thank You!