CS747 Final Project

Multi-Class Hate Speech Classification using NoahBERT

Name:	Pratish Mashankar	Sreeja Puthumana
G#:	G01354094	G01360679

ABSTRACT

In this project, we address the challenge of hate speech detection, specifically focusing on the differentiation between hate speech, offensive language, and neutral content, an underexplored area in existing research. Distinguishing between hate speech and offensive language is vital for online platforms to effectively filter harmful content, safeguard users, and foster a more respectful online community. Combining datasets from Ousidhoum et al. and Davidson et al., we curated a comprehensive dataset with 4000 offensive, 3000 neutral, and 2000 hate speech instances. Fine-tuning BERT resulted in a baseline accuracy of 74.9%. In response, we introduced NoahBERT, a modified model incorporating an additional layer and dropout, achieving an enhanced accuracy of 78.9%. Our work contributes to the critical task of improving online toxicity detection and fostering a safer digital environment.

1. INTRODUCTION

In the contemporary age of unrestricted social media freedom, it becomes imperative to establish stringent guidelines for the identification and restriction of harmful language, both at the individual and societal levels. BERT, with its contextualized word representations, excels in capturing intricate relationships between words in a sentence, making it suitable for sequence classification tasks. BertForSequenceClassification is built upon the BERT architecture, utilizing a bidirectional transformer encoder. The attention mechanism in BERT enables the model to consider the context from both directions, providing rich contextual embeddings for each word in a sentence.

Hate speech refers to any communication, conduct, or expression that offends, threatens, or insults a person or group based on attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender. Offensive speech generally encompasses expressions or content that is likely to cause discomfort, resentment, or displeasure. Normal speech typically aligns with common standards of politeness and respect. Identifying and classifying hate speech helps create safer online environments by mitigating the spread of harmful content and fostering a more respectful discourse. A binary classification model might oversimplify the complexities of language, making it challenging to capture the nuances between various forms of negativity. Multi-class models provide a more comprehensive representation of the training data. This report delves into the methodologies and advancements in text sentiment classification, exploring the application of cutting-edge techniques to accurately categorize sentiments expressed in diverse textual sources.

BERT, is a groundbreaking NLP model developed by Google in 2018, unlike traditional approaches, this model reads text bi-directionally. This enables a more thorough understanding of context and linguistic nuances. The BERT Tokenizer is a specialized tool that breaks down input text into smaller units, known as tokens, to facilitate the processing and understanding of hate speech tweets. Noah BERT which is a modified BERT version.

2. DATASET PREPARATION

The dataset we used in this study is a combined data from the MLMA English Hate Speech dataset by Ousidhoum et al. referred to as the MLMA dataset and the Automated Hate Speech Dataset by Davidson et al. referred to as the Davidson dataset. The amalgamation was necessitated by certain limitations in the individual datasets. The Davidson dataset exhibits a significant imbalance in class distribution, with an overwhelming majority of non-hate speech instances, posing challenges for machine learning models to learn and generalize patterns due to the skewed nature of the data. In the MLMA dataset, a limitation arises from the relatively small amount of data available for exploration

2.1 Davidson Dataset

This dataset consisted of 24,783 tweets that were divided into three classes: offensive, harmful, and normal. Offensive tweets include those tweets that may be perceived as inappropriate, objectionable, or disrespectful but may not necessarily include explicit hate speech. These tweets can range from being mildly inappropriate to more severe, causing discomfort or unease among readers. Hateful tweets are tweets that contain explicit hate speech or discriminatory language targeting individuals or groups based on attributes such as race, ethnicity, religion, gender, sexual orientation, etc. Hateful tweets go beyond mere offensiveness, expressing prejudice, discrimination, or hostility towards a particular individual or group. Normal tweets are those that include everyday messages shared on social media platforms that are neutral or positive in tone, normal tweets may include a range of content, such as updates, observations, or general thoughts. Figure 1a. elucidates the distribution of tweets within the dataset, providing a comprehensive overview of sentiment counts for each distinct class. We observe an extreme class imbalance between offensive and hateful speech. We hence also experiment with the MLMA dataset.

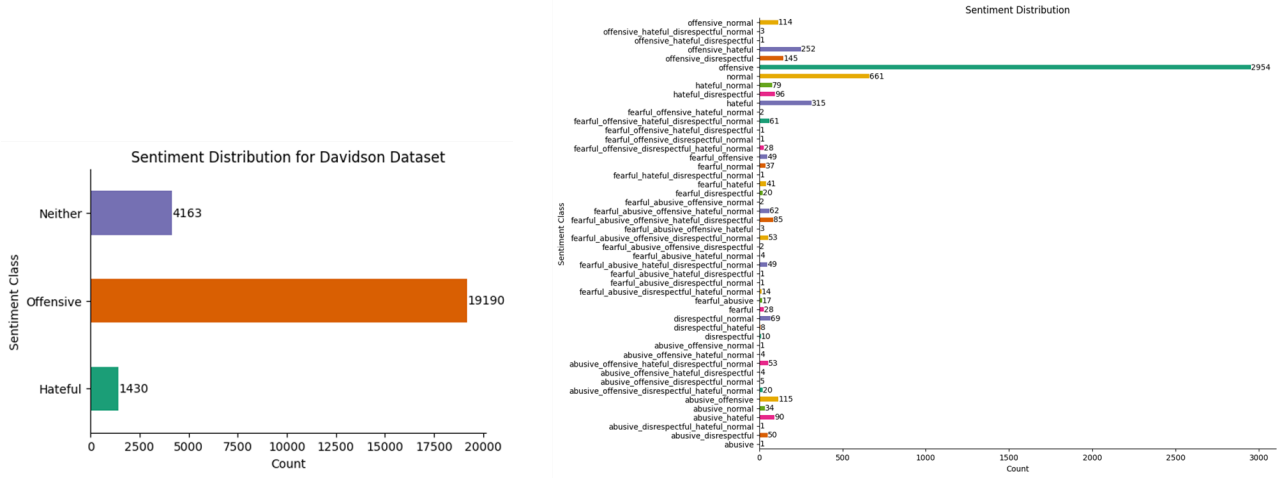

Figure 1*:* Sentiment distribution of a) Davidson dataset (left) and b) MLMA dataset (right)

2.2 MLMA Hate Speech Dataset:

The MLMA English hate speech dataset consisted of 5647 tweets along with the sentiment class. We have performed a feature engineering to remove trivial columns from the dataset. The dataset sentiments consist of combinations of five classes, offensive, normal, harmful, fearful, and abusive as observed in Figure 1b. In refining the

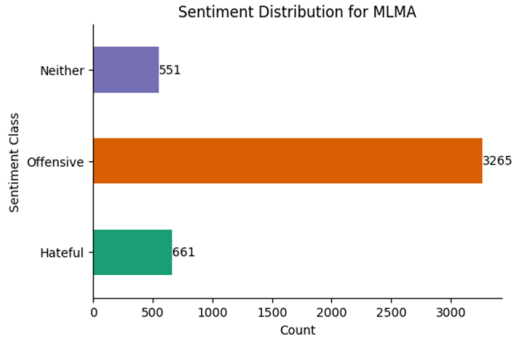

Figure 2*:* Sentiment distribution reorganized MLMA dataset

MLMA dataset for our research purposes, a strategic decision was made to streamline its classification schema from multiple classes to three, aligning it with the structure present in the first dataset. Initially, we excluded tweets containing elements of fearfulness and abuse. Subsequently, we included tweets classified as normal, ensuring the removal of any tweets exhibiting a combination of normalcy with other attributes. Following this, we further refined the dataset by excluding tweets that demonstrated a blend of both our existing classes, namely harmful and offensive. This adjustment facilitates a more cohesive comparative analysis between the two datasets The final distribution of the MLMA dataset is explained in Figure 2.

2.3 Final Distribution

As previously mentioned, the final dataset is a composite of both datasets. Upon combining these datasets, we observed data imbalance within the set. To address this, we formed a dataframe named 'neither_df' by concatenating rows from both datasets where the 'class' is labeled as 'neither.' From this dataframe, we sampled 3000 records. Subsequently, a 'hateful_df' dataframe was created by concatenating rows labeled as 'hateful' from both datasets, resulting in 1981 records. Additionally, we sampled 4000 records from the combined datasets belonging to the 'offensive' class. After successful sampling, all the sampled records were consolidated into a unified dataframe named 'combined_df.' The labels were then vectorized as 0, 1, 2, representing 'Neither,' 'Offensive,' and 'Hateful,' respectively. To introduce randomness, the rows were shuffled using a random seed of 42, resulting in the 'combined_df' dataset (Figure 2), suitable for training and testing purposes.

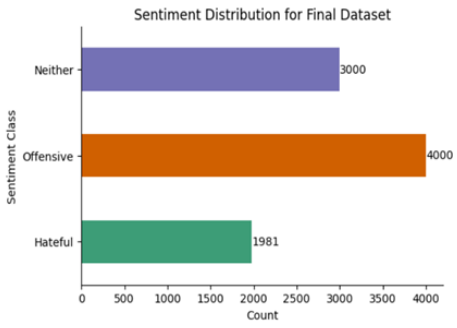

Figure 3*:* Sentiment distribution Final Dataset

3. METHODOLOGY

We have used Hugging Face's Transformers, an open-source library that offers versatile architectures and pre-trained models for natural language understanding including BERT. We have used the BertForSequenceClassification model from the Hugging Face model hub. We finetuned this model to perform the classification of the text into three classes- neither (0), offensive (1), and hateful (2). After tokenizing the data using

BertTokenizer, we prepared our training, validation, and test sets, reserving 80% of the data for training and 10% each for validation and evaluation. We then performed hyperparameter tuning with varying values for the AdamW optimizer with learning rates of 0.00001, 0.0001, and 0.001 for batch sizes of 35, 45, and 55. We observed the model's performance over 50 epochs.

To fine-tune the model on Hate Speech data for enhanced classification, we propose the NoahBERT (**No Ha**te Bert) model, a modification of the BERT model by introducing an additional feedforward layer and a dropout layer. The dropout layer's regularization hyperparameter, the dropout constant, was tuned to 0.1 after experimenting with values between 0.1 to 0.5. This augmentation was implemented through a custom model, BERTWithNoahClassifier, which integrated the pre-trained BERT transformer with newly added layers, and a linear classifier for the final prediction. The added feedforward layer, consisting of 256 neurons with ReLU activation, was strategically designed to capture and emphasize more nuanced features within the BERT embeddings. The model was trained using a cross-entropy loss function, and we observed promising results in terms of enhanced sentiment

classification. Additionally, to ensure the model's adaptability, hyperparameters such as the dimensions of the feedforward layer and the dropout rate were made configurable during model initialization. We again experimented with varying the learning rates and the batch size just as before. We again observed the model's performance over 50 epochs and determined the best number of epochs. With the best hyperparameters, namely: regularization constant, learning rate, batch size, and number of epochs, we determined the optimal evaluation accuracy for BERT and NoahBERT.

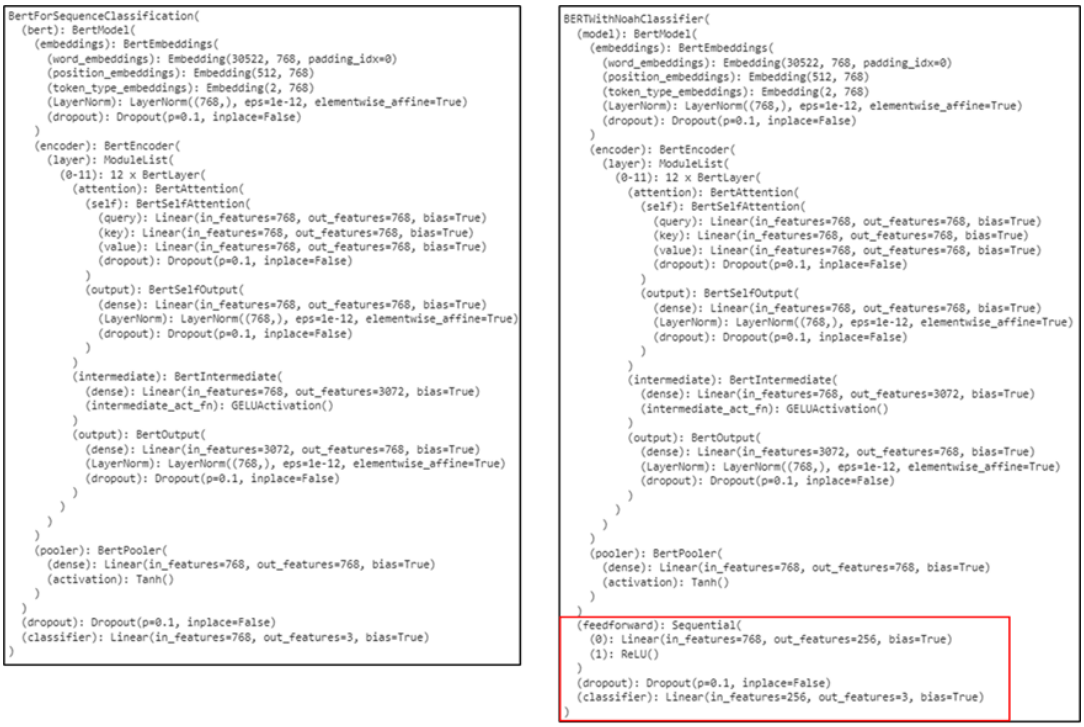

Figure 4*:* BERT architecture (Left) and NoahBERT architecture (Right)

We then performed an error analysis on both models, where we closely examined instances where the model misclassified test set examples by randomly sampling five misclassifications. Through a qualitative examination of these instances, we aimed to discern patterns or common linguistic properties that might be contributing to misclassifications. This in-depth analysis provided valuable insights into potential shortcomings or challenges faced by the model.

We finally performed a Robustness test on our NoahBERT model to determine if the NoahBERT is robust to grammatical and data anomalies. We concluded our research with possible future scopes for this project.

4. RESULTS

4.1 Bidirectional Encoder Representations from Transformers (BERT) Results

Figure 5 plots the accuracy and loss of the BERT model during the training and validation over 50 epochs. We observe that while the model's training accuracy climbed and plateaued above 95%, the validation accuracy lingered around 75%, exhibiting minimal improvement post-initial epochs. This disparity suggested a case of overfitting, where our model excelled with the training data but failed to generalize its learning to the validation data effectively. Similarly, the loss plots mirrored this trend. We observed that the BERT model gave the best validation accuracy of 74.7% after 3 epochs when

fine-tuned using the AdamW optimizer with a learning rate of 0.00001, and batch size of 45. We observed an evaluation or test accuracy of 74.9%, with 674 correct predictions out of 899.

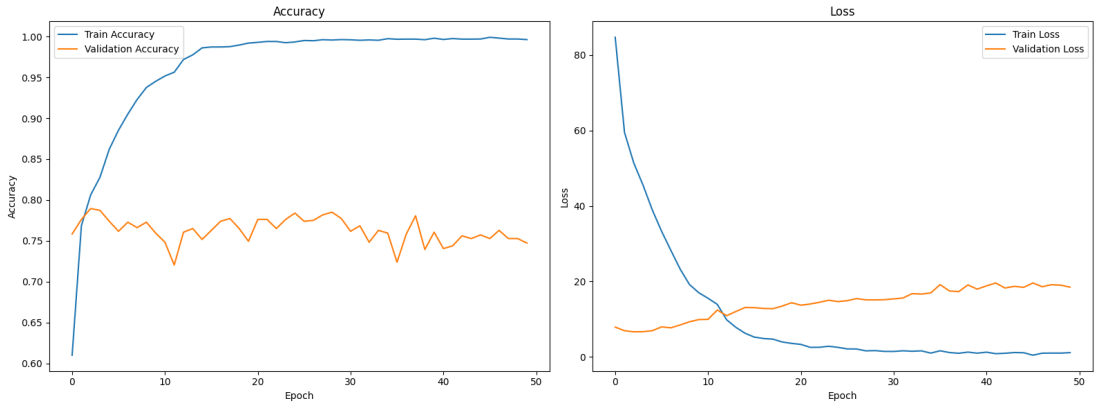

Figure 5: Accuracy and loss plot for BERT model.

4.2 No Hate BERT (NoahBERT) Results

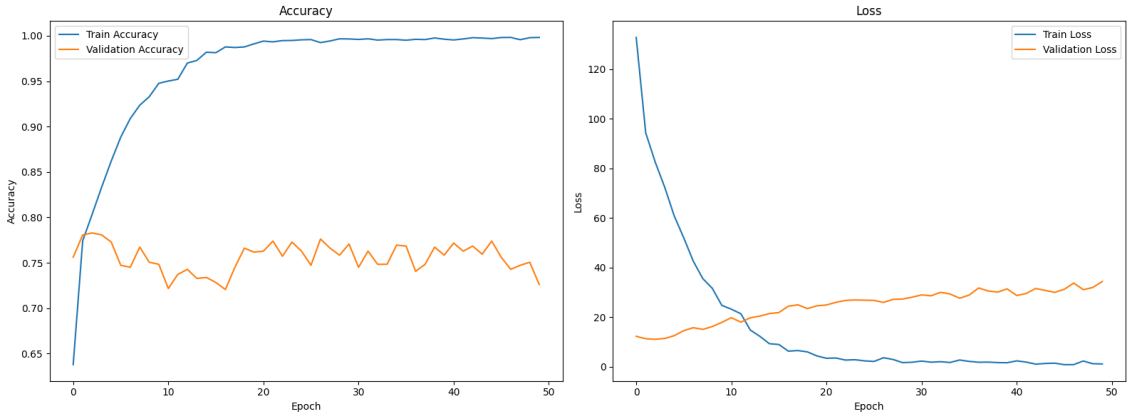

Figure 6: Accuracy and loss plot for NoahBERT model.

Figure 6 plots the accuracy and loss of the BERT model during the training and validation over 50 epochs. We again observe that the training accuracy consistently remained high, exceeding 95%, and the validation accuracy showed a stable pattern, maintaining around 75%. While still indicative of some overfitting, the model's performance is notably better than previous iterations. The loss plots further reflect this improvement, with the training loss decreasing to a low plateau and the validation loss, despite its higher baseline, displaying a downward trend. This suggests that the NoahBERT model has a strong learning capacity, although we feel further tuning might be necessary to bridge the gap between training and validation outcomes. We observed that the NoahBERT model gave the best validation accuracy of 78.06% after 3 epochs when fine-tuned using the AdamW optimizer with a learning rate of 0.00001, and batch size of 45. Our modifications to the BERT model resulted in a notable improvement in accuracy compared to the baseline BERT model. However, it is essential to note that the introduction of the extra feedforward layer led to increased computational complexity and execution time, emphasizing the trade-off between model performance and efficiency. We observed an evaluation or test accuracy of 76.53%, with 688 correct predictions out of 899.

5. ERROR ANALYSIS

In our pursuit of model improvement, a comprehensive error analysis was conducted to scrutinize instances where the hate speech classification models exhibited inaccuracies. The five randomly selected misclassified examples shed light on specific scenarios where the model faced challenges.

5.1 Error Analysis on BERT

```
Number of incorrect predictions: 207
Below are 5 random incorrect predictions:
Text: @ hovaa _ ok wat ever u say whitey. u prolly never even seen shaft.
Predicted label: 1
Correct label: 2
------------------------------
Text: rt @ alexishope _ 2014 : you got niggas & amp ; i got bitches but i want you... & # 128149 ; & # 127926 ;
Predicted label: 0
Correct label: 2
------------------------------
Text: mike pence out here following the girls? the faggot jumped out!!
Predicted label: 2
Correct label: 0
------------------------------
Text: whenever i hit 1000 followers i feel so self - destructive. like a nigger near a cop car.
Predicted label: 2
Correct label: 1
------------------------------
Text: @ user you can \ u2019t honestly be such an asshole can you? the shithole countries of the past were italy and \ u2026 @ url
Predicted label: 0
Correct label: 2
------------------------------
```

Figure 7: Five random misclassifications by BERT. 0: Neither, 1: Offensive, 2: Hate Speech

Figure 7 presents five random misclassification examples by the BERT model. We observe that in examples 3 and 4, The model likely misclassified these as hateful due to the presence of racially sensitive terms thus misjudging the context. Moreover, in examples 1, 2, and 5 the model failed to capture the offensive nature of the language, possibly because it focused on the mention of trivial words rather than the overall tone.

5.2 Error Analysis on NoahBERT

```
Number of incorrect predictions: 211
Below are 5 random incorrect predictions:
Text: @ stephdamartian they made up endearing names for us like nigger spook & amp ; monkey
Predicted label: 2
Correct label: 0
------------------------------
Text: @ user nah b i dont fuck with those ching chong ling long phones
Predicted label: 0
Correct label: 2
------------------------------
Text: the family of ron o'neal needs to sue wu - tang for that trash song they released in his name.
Predicted label: 2
Correct label: 1
------------------------------
Text: @ joebiden is such a retard! & # 8220 ; @ irritatedwoman : biden calls himself & # 8216 ; white boy & # 8217 ; in meeting with black pastors
Predicted label: 2
Correct label: 0
------------------------------
Text: we should just cancel the shithole countries and just give it to trump
Predicted label: 1
Correct label: 0
------------------------------
```

Figure 8: Five random misclassifications by NoahBERT. 0: Neither, 1: Offensive, 2: Hate Speech

Figure 8 presents five random misclassification examples by the NoahBERT model. We observe that in examples 1 and 4, the model was likely misled by the use of derogatory terms and misinterpreted their usage thus incorrectly classifying a neutral statement as hateful. In other examples, the model

may have misjudged the context and considered the statements more as hateful than offensive. We observe that both models struggle with sensitivity to specific terms and face difficulties in accurately interpreting the context of statements. However, our Noah BERT seems to have a better grasp of distinguishing between offensive and hateful language in some instances. It shows a potential improvement in differentiating between these two classes compared to BERT.

6. ROBUSTNESS TEST ON NoahBERT

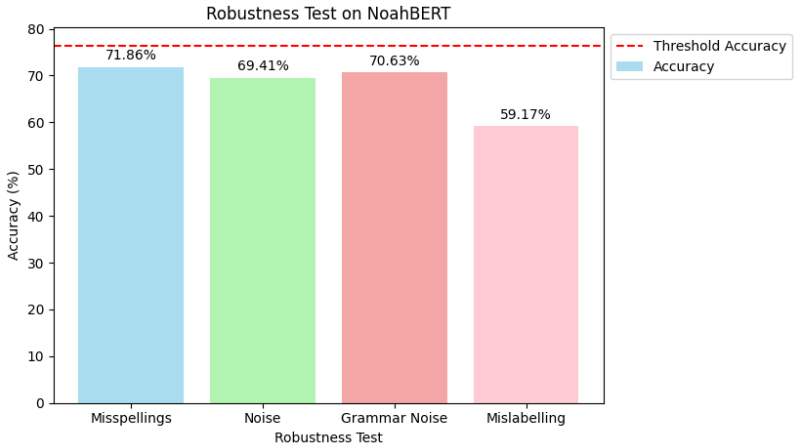

Figure 9: Robustness Performance of NoahBERT

To assess the NoahBERT model's robustness, we executed four distinct tests by modifying the Test set.

1. **Misspellings:** For every sentence in the tweet, we modified each token by randomly swapping each letter in the word with another letter in the same word. We observed a 71.86% accuracy.

2. **Noise:** We have introduced random letters to each tweet such that the effect generated would be the same as typo errors in the data. Noah BERT achieved 69.41% accuracy after introducing the noise.

3. **Grammar Noise:** We have performed this by using an instance of PorterStemmer, where we changed the words from the dataset into their root form by stemming them. NoahBERT was able to achieve 70.63%.

4. **Mislabelling:** We modified the Test set by changing the labels for 40% of the records in the test data. Here, NoahBERT showed 59.17% accuracy.

Upon observing a decreasing trend in the accuracies for the Tests, we can say that even though NoahBERT outperforms BERT in classifying Hate Speech, it is not robust to anomalies.

7. Conclusion

In conclusion, our exploration into multi-label Hate Speech classification using NoahBERT, a BERT-based model, has yielded promising results, showcasing the power of leveraging pre-trained transformer architectures for nuanced language understanding. The introduction of a custom model, NoahBERT, incorporating an additional feedforward layer, though not robust still contributed to a notable enhancement in accuracy, underscoring the effectiveness of model customization for specific tasks.

8. Future Work

In the future, we can focus on enhancing model scalability and improving the performance by introducing NoahBERT-Large, which will be trained on a larger corpus obtained using data augmentation. We can develop NoahRoBERTa, an optimized version of RoBERTa for robust hate speech classification. We can also delve into m-NoahBERT, a model that will augment the capabilities of enhanced BERT to address hate speech across various languages and linguistic nuances. Finally, we can explore Regularized-NoahBERT which will implement regularization techniques to address overfitting challenges. These endeavors collectively aim to advance the capabilities of hate speech classification models, catering to different linguistic contexts, addressing overfitting concerns, and enhancing the overall robustness of the models.

References:

1. MLMA: https://huggingface.co/datasets/nedjmaou/MLMA_hate_speech/tree/main

   Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and Multi-Aspect Hate Speech Analysis. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4675–4684, Hong Kong, China. Association for Computational Linguistics

2. https://github.com/t-davidson/hate-speech-and-offensive-language/tree/master

   Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. Proceedings of the International AAAI Conference on Web and Social Media, 11(1), 512-515. https://doi.org/10.1609/icwsm.v11i1.14955

3. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1810.04805

4. Chhabra, A., Vishwakarma, D.K. A literature survey on multimodal and multilingual automatic hate speech identification. Multimedia Systems 29, 1203–1230 (2023). *https://doi.org/10.1007/s00530-023-01051-8*

5. Mozafari, M., Farahbakhsh, R., & Crespi, N. (2019). A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media. *ArXiv*. /abs/1910.12574

6. HateBERT: Retraining BERT for Abusive Language Detection in English (https://aclanthology.org/2021.woah-1.3) (Caselli et al., WOAH 2021)

7. Mazari, A.C., Boudoukhani, N. & Djeffal, A. BERT-based ensemble learning for multi-aspect hate speech detection. Cluster Comput (2023). https://doi.org/10.1007/s10586-022-03956-x