Contextual Hate Speech Detection Leveraging RoBERTa: Overcoming Challenges in Online Communication

MSc Research Project

Master's in Data Analytics

Varun Bhalerao
Student ID: 22206884

School of Computing
National College of Ireland

Supervisor: Abubakr Siddig

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Varun Atul Bhalerao		
Student ID:	22206884		
Programme:	Masters in Data Analytics	**Year:**	2024
Module:	MSc Research Project		
Supervisor:	Abubakr Siddig		
Submission Due Date:	12/08/2024		
Project Title:	Contextual Hate Speech Detection Leveraging RoBERTa: Overcoming Challenges in Online Communication		
Word Count:	8142	**Page Count:**	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Varun Atul Bhalerao
Date:	12/08/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Contextual Hate Speech Detection Leveraging RoBERTa: Overcoming Challenges in Online Communication

Varun Bhalerao
x22206884

Abstract

The growth in online communication has increased hate speech in an exponential way, presenting an urgent need and serious challenges for any company that deals with the safe and inclusive maintenance of digital environments. Traditional methods of detection, one that uses simple keyword-based techniques often fail to capture the original, context-dependent nature of hate speech that manifests through slang, code words, and euphemisms. This paper presents the application of RoBERTa, a strongly optimized variant of BERT, toward increasing the accuracy and resiliency of hate speech detection. Training on larger datasets and longer training periods, and dynamic masking make RoBERTa much more powerful in understanding and processing human languages in their diversified and subtle contexts. Especially, it investigates whether RoBERTa can overcome the inefficiencies of early models in detecting hate speech efficiently across languages and cultural contexts, including traditional machine learning approaches and early deep learning models like CNNs and RNNs. Comparative analysis has shown that it outperforms traditional approaches in finding contextually nuanced hate speech, especially with other techniques like ensembling models or emotion recognition. The research aims to come up with a very accurate and versatile hate speech detection system that could work in different languages and across changing linguistic patterns. Hence, the result concludes the potential of transformer-based models in raising online safety and inclusivity, as shown by RoBERTa.

1. INTRODUCTION

There has been this exponential growth within online communication which has paralleled by large number of hate speech on those platforms, which quite improperly presents a big challenge to try and maintain a safe and friendly online community. Further, hate speech could also be considered as involving abusive language, threats, and discriminatory remarks against particular groups or individuals; it does project others toward violence, discrimination, and social divisiveness. This requires urgency in developing very effective detection and mitigation mechanisms due to the volume and overwhelmingly high speed with such harmful content can spread at. First, a system shall detect harmful words or phrases that trigger conflicts. However, traditional detection mechanisms have been proved outdated by the dynamic nature of online communication.

The language of hate speech is subtle, making detection a complex task. A word or even phrase may be normal in one context but offensive in another. Moreover, hate speech can also occur in hidden forms under the disguise of slang, code words, and euphemisms, which can easily avoid simplistic keyword-based techniques of detection. This task is also rendered difficult because Internet language keeps changing, where new terms and expressions pick up every day. Any effective hate speech detection system must understand such subtleties and adjust to them.

Recent advances in machine learning and natural language processing make very encouraging solutions to these challenges. Deep learning models have distinctly been impressing in knowing and processing human languages. Such models learn complex ways for which patterns and relationships can be represented within data, making them very effective at capturing subtle and context-dependent varieties of hate speech. Modern NLP models are based on advanced architectures and training methods that make them very proficient at tracking speech patterns across a variety of contexts. The research in this paper is focused on the application of a variant of the BERT model called RoBERTa in hate speech detection.

Understanding RoBERTa:

RoBERTa stands for Robustly Optimized BERT Pretraining Approach. This is a Facebook AI-developed Transformer-based model, extending the architecture of BERT by optimizing its pretraining with larger datasets (*RoBERTa: An optimized method for pretraining self-supervised NLP systems*, no date). The model has longer training periods along with better training techniques. It results in RoBERTa being able to achieve higher accuracy and better performance on NLP tasks such as hate speech detection.

Key Features of RoBERTa:

1. Training on Larger Datasets: BERT was trained with data from BooksCorpus and English Wikipedia (*What is the BERT language model? | Definition from TechTarget*, no date), while RoBERTa was further trained with an enhanced dataset that added Common Crawl News, OpenWebText, and Stories from Common Crawl (*FacebookAI/roberta-base · Hugging Face*, 2024). The large training set permits RoBERTa to learn more about the general statistical patterns of the language as well as the contextual information.
2. Long Training: RoBERTa is trained for a longer time compared to BERT; hence, it is able to understand and represent human languages in a more adaptive manner (*RoBERTa*, no date).
3. Dynamic Masking: RoBERTa dynamically masks input tokens at each training epoch, evolving the masking pattern so that the learning of the context is more robust (*RoBERTa*, no date).
4. Removed NSP: The Next Sentence Prediction (NSP) task of BERT, which carried few benefits, was removed by RoBERTa, and it was plugged into a solely Masked Language Modeling-based setup. In this way, a guarantee is provided that many more resources will be spent toward learning contextual information under RoBERTa.

Research Question:

How effective can the implementation of the RoBERTa model detect hate speech compared to other models presented in previous research?

2. RELATED WORK

The rapid rise of social media platforms has turned the detection of hate speech into a very important area of research. The possibility of very fast and far-reaching flow, with harmful content particularly at the lead, opens the requirement for efficient and accurate detection methodologies. In the last couple of years, especially after deep learning methods, obviously with BERT, NLP has improved a lot in optimizing the power of hate speech detection systems. This literature review fuses findings from major studies to picture a comprehensive overview of the state of hate speech detection research assisted by BERT and other advanced models of NLP.

2.2 BERT-Based Approaches for Hate Speech Detection

This way, BERT becomes a revolutionary model in NLP, thanks to its bidirectional training that understands the word in a sentence much more precisely than earlier models. Several studies have used the contextual understanding of BERT to enhance the task of hate speech detection on social media.

One of the studies used a fine-tuned version of BERT to detect hate speech on Twitter (Nayla, Setianingsih and Dirgantoro, 2023). In the paper, they decided to pick the BERT-base, uncased model since it is VERY good at understanding context for words, therefore very relevant here. Very key to this study were preprocessing steps: tokenization, normalization, and special character removal are some of the processes that cleaned the data and ready it for the model. The model was trained, and efficient with respect to accuracy, precision, recall, F1-score, among others, in measuring the model established its efficiency in hate speech detection on Twitter.

Another research on multilingual hate speech detection focused on how such challenges in non-English languages could be dealt with using BERT with Convolutional Neural Networks (Shukla, Nagpal and Sabharwal, 2022). More specifically, it was a study about Hindi-based hate speech; hence, using datasets like HASOC 2020 and HASOC 2021, the model can use BERT for contextual embeddings and CNNs to extract features that would capture the details in the Hindi language. The class imbalance issues are also dealt with in the study through techniques such as oversampling, making sure that the model does not favor the majority class at the expense of others while carrying out the hate speech detection.

The paper transfer learning approaches using BERT have been explored not only to enhance accuracy in hate speech detection but also to cover for the scarcity of labeled data (Mozafari, Farahbakhsh and Crespi, 2019). Upon fine-tuning a pre-trained BERT model over certain datasets pertaining to hate speech, researchers have been able to improve its accuracy of detection. Transfer learning enables knowledge the model has learned from a huge general corpus to be transferred into the more specific task of hate speech detection. In addition to reducing the need for extensive labeled data, this greatly improves generalization across different datasets.

2.3 Comparative Studies Involving BERT and Other Models

It has already been confirmed in the literature that BERT includes traditional machine learning models and early deep approaches. For instance, a comparative study that evaluated BERT-based models against traditional 1D CNNs using GloVe embeddings found that with BERT, it significantly outperformed the traditional CNN models (S *et al.*, 2024). The results from the study clearly showed that while 1D CNNs with the GloVe embeddings are good in some respects, in most cases they profoundly fail to capture the complex contextual relationships within text, which BERT does. This comparative analysis was achieved using a publicly available dataset downloaded from Kaggle. Some preprocessing steps included tokenization and normalization, and class imbalance was handled through oversampling techniques. Though requiring more computational resources, the BERT-based model consistently performed better in terms of accuracy, precision, recall, and the F1-score.

Another comparative study went a step further to include BERT, RoBERTa, DistilBERT, and XLNet in the evaluation for tasks related to emotion recognition, very close to hate speech detection (Acheampong, Nunoo-Mensah and Chen, 2020). In this research, it used the ISEAR dataset and really proved the case that transformer-based models like BERT and their variants give quite robust performance on most of the NLP tasks. The study has thus brought to the limelight that BERT, RoBERTa, and XLNet models perform very well in tasks that require nuanced understanding of context and emotion, which is quite critical in correctly detecting hate speech. This underlines their potential not only for emotion recognition but also in wider applications like hate speech detection, wherein understanding the underlying sentiment is important.

2.4 Ensemble and Hybrid Models

Other researchers have also attempted to come up with hybrid models for integrating BERT and other architectures, like CNN, to improve the performance of hate speech detection systems. This can be hybridized so that the model takes advantage of both BERT's contextual embeddings and CNN's feature extraction. For instance, in the case of a study that dealt with multilingual hate speech detection in Hindi languages, integration of BERT with CNN layers helped in better understanding of the language and improved model accuracy (Shukla, Nagpal and Sabharwal, 2022). These hybrid models can therefore go beyond the deep contextual understanding that BERT embeds to include spatial feature extraction capabilities of CNNs, therefore capturing the complexities of hate speech better across different languages and cultural contexts.

Ensemble models have also been suggested as a very promising way ahead in overcoming the challenges of hate speech detection (Ali, 2024). These ensemble techniques combine several models to counteract the weaknesses of each individual model and hence improve general detection accuracy. For instance, ensemble methods combining BERT with other NLP models have shown a little promise in adapting to new varieties of hate speech, which is key given the dynamic and constantly evolving nature of language on social media (Keshari, Malladi and Mittal, no date). Along this line, continuous updating of the model and use of ensemble techniques are highly recommended for maintaining its effect over time.

2.5 Traditional Machine Learning and Early Deep Learning Approaches

Before the transformer models like BERT in every task, traditional machine learning methods and early deep learning approaches had an important role in hate speech detection. This will include support vector machines, Naive Bayes, random forests, and earlier neural networks, including CNNs and RNNs (MacAvaney *et al.*, 2019). All of these models worked pretty well to some extent, especially scenarios on which hate speech language was plain and straightforward. However, in the case of complex cases, understanding exactly which context and subtleties of the language were necessary often eluded them.

Notable improvements accompanied the shift to deep learning, where models such as CNNs and RNNs offered better accuracy on handling sequential data. However, these models still could not capture the context of a sentence, a very basic requirement when it comes to the accurate detection of hate speech. It is through this limitation that transformer-based models, such as BERT, came into use, understanding context much more proficiently than their predecessors due to bidirectional training (MacAvaney *et al.*, 2019).

All of these studies, which compared earlier approaches to transformer models, concluded that BERT and variants were ahead of traditional models on most of the important metrics used to measure a model's performance, including accuracy, precision, recall, and F1 score. For instance, research comparing baseline CNNs and Bi-LSTMs with an added attention mechanism to pre-trained BERT and fine-tuned RoBERTa discovered that BERT-based models are much better in handling the complexities of hate speech detection. This clearly points to the failures of the old ways and how far modern NLP has progressed using these new models (Keshari, Malladi and Mittal, no date).

2.6 Addressing Multilingual and Cultural Challenges

The challenges of multilingual and culturally diverse environments in the hate speech detection task are unique in nature when compared to monolingual settings. These challenges range from confusion of offensive words and context dependency to biases naturally built into datasets. This has compelled research to insist on consideration of the foregoing challenges in the development of more effective and inclusive detection systems (Kovács, Alonso and Saini, 2021).

For example, hate speech will be expressed in different languages, dialects, or even code-switched sentences in multilingual contexts; all of these aspects make it quite a challenge for models trained on a single language. Fusion of external linguistic resources along with robust training methodologies is already known to improve the quality of models in such settings by a great amount. For instance, studies have turned to techniques of oversampling in order to treat class imbalances in multilingual datasets and ensure that the model does not disproportionately favor the majority class. Another approach is using BERT with CNNs in setting contextual embeddings and extracting features; this has proved quite effective in handling multilingual hate speech detection (Shukla, Nagpal and Sabharwal, 2022).

More importantly, considering the cultural context associated with hate speech is quite crucial in developing accurate detection systems. The use of language stretches so largely across cultures, and at times what is perceived as hate speech in one culture might not be so in another. For this reason, scholars have been calling for reviews on the models, incorporation of external cultural and linguistic resources to enhance adaptability and accuracy for hate speech detection systems across different cultural contexts (Lee, Jung and Oh, 2023).

2.7 Application of Emotion Recognition Techniques

Emotion recognition techniques are at a development stage but have been very promising in improving the accuracy of hate speech detection systems. Emotion recognition from text enables the understanding of sentiments and intentions tied to certain statements, hence forming the backbone of accurately identifying hate speech (Rezapour, no date).

Such studies on BERT, RoBERTa, DistilBERT, and XLNet have already established their capabilities of capturing very tiny emotional cues within the text. For example, research applying these transformer models to the ISEAR dataset for emotion recognition has presented how they can identify emotions such as anger, fear, and disgust with increased efficacy, typically associated with hate speech. These models can understand the emotional undertones of a statement, allowing them to classify more accurately whether a piece of text is hate speech or not (Acheampong, Nunoo-Mensah and Chen, 2020).

These results suggest that methods developed for emotion recognition could also be applied to hate speech detection and, in doing so, bring accuracy to the system. For example, a model in hate speech detection that uses emotion recognition may be better positioned to recognize hate speech expressed in a subtler or more indirect way, wherein the emotional context holds more of a key to grasping the intended meaning behind the words.

2.8 Benchmarking and Competitive Evaluation

Benchmarking studies and competitive evaluations are, therefore, of pivotal importance in making sense of the performance of models built for detecting hate speech. For example, SemEval-2020 was an Offensive Language Identification and Classification Competition describes a useful baseline against which to compare a wide array of different models while participants used different transformer-based models like BERT, the problem is solved by ensuring quality input that is fed with detailed preprocessing steps like tokenization and normalization (Alonso, Saini and Kovacs, 2020).

Results from SemEval-2020 have shown that proper metrics to assess models' performance, linked with accuracy, precision, recall, and F1-score, were very important in order to estimate the effectiveness of a model in hate speech detection. Certainly, this competition has revealed rather high potential regarding efficiency for BERT-based models, but it draws attention to areas where there is huge room for improvement. These benchmarking efforts provide a good platform for future research and help in choosing relevant models/techniques most impactful for hate speech detection.

The literature reviewed in the chapter on this topic has shown great improvement in the task of hate speech detection, specifically with a boost from models such as BERT. It is the case that transformer models, especially those presented with BERT, are proficient at handling context and capturing subtle variations in verbal semantics, which is influential in hate speech detection. However, further improvement could be done respect to multilingual data, language use evolution, and continuous model update issues.

Such challenges can be overcome by future research dealing with ensemble models, transfer learning, and event emotion recognition techniques. More robust and accurate hate speech detection systems will be realized operating across languages, cultures, and contexts by fine-tuning and developing these streams of approaches.

Research Gap:
Based on transformer-based models, with more specificity to RoBERTa, this paper focuses on boosting contextual and multilingual hate speech detection. The aim for this research is to develop a strong RoBERTa model fine-tuned for contextual understanding and efficiency in hate speech detection for both major and low-resourced languages. Specifically, it aims to achieve a much higher level of accuracy and resiliency in the detection of hate speech by the inclusion of more contextual characteristics, working with subtle and confusing language, and through hybrid and ensemble models—rather quite resilient to the two most frequent challenges modern systems representatively face: linguistic diversity and strong context dependency.

3. METHODOLOGY

In this research, which focuses on detecting hate speech, offensive language using a deep learning model known as RoBERTa which is an optimized version of BERT a FacebookAI model. The above section gives us an overview of the previous work done in the same field by various authors who used different models and methodologies which proved helpful for this research. Furthermore, the main goal of this research is to classify hate speech into 3 categories, such as hate speech, offensive language or neither. Hence, this section will focus on the method and on how it was done. The methodology which was followed by this research is none other than Knowledge Discovery in Database (KDD) (Alonso, Saini and Kovacs, 2020).

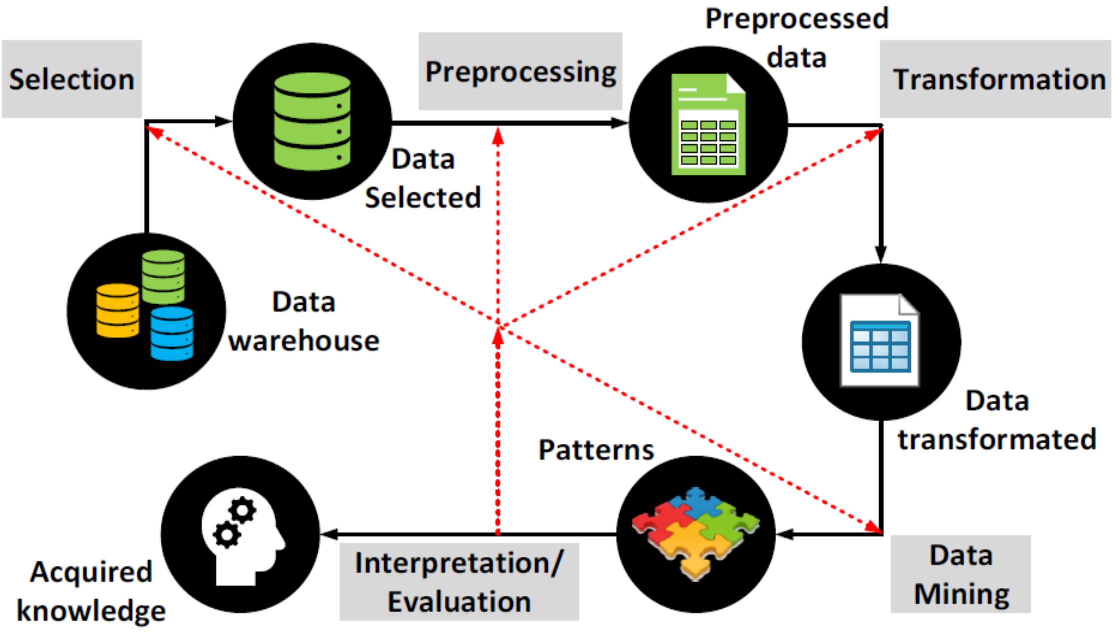

Figure 1: KDD methodology steps (Valencia et al., 2020)

3.1 Data Collection

Selection is the first stage of Knowledge Discovery in Databases, which includes the process of identifying and collecting the most relevant data needed for a particular task. In this work, what was mainly intended was to construct a model that could identify harmful language, including hate speech and offensive language on social media platforms. For accomplishing this, a dataset compiled with the mentioned objective has been opted (*Hate Speech and Offensive Language Dataset*, no date).

Dataset Overview

This paper uses a publicly available, human-annotated dataset of tweets, each labelled under one of the following three categories: Hate Speech, Offensive Language, and Neither. This multi-class labelling provides an explicit framework for the classification task where the model is supposed to learn from examples and correctly identify the type of language used in new, unseen tweets.

Table 1: Dataset overview

Column Name	Data type	Details
tweet_id	Integer	A unique identifier for each tweet.
tweet	String	The actual tweet text. Examples include offensive language, casual or slang expressions, or comments on various topics.
class	Integer	Classification label for the tweet (0, 1, 2, or 3).
count	Integer	Counts of occurrences, always 3 in this dataset.
hate speech	Integer	Indicates if the tweet contains hate speech (0 or 1).
offensive_language	Integer	Indicates if the tweet contains offensive language (0 or 1).
neither	Integer	Indicates if the tweet contains neither hate speech nor offensive language (0 or 1).

Data Selection Criteria

This dataset was chosen based on various technical requirements that were relevant for the training of robust machine learning models:

1. Relevance: The dataset is very relevant to the task of harmful content detection, making it more suitable for training a model that filters out or flags such content over social media.
2. Label Diversity: The multi-class nature of the dataset with distinct labels for hate speech, offensive language, and neutral content effectively exposes the models to a variety of linguistic patterns concerning harmful and nonthreatening classes.
3. Data Size: The dataset is big enough, running into thousands of tweets. A sizable dataset is needed for the training process. Deep learning models, specifically the transformer-based architectures like DistillBERT and RoBERTa, are known to require huge amounts of data to generalize well across highly diverse inputs.
4. Annotator Agreement: Columns on votes from annotators, such as hate speech, offensive language, neither bring further context into the consent of the annotators; this can be used to judge the reliability of the labels and could be utilized in model training by weighting examples using their level of agreement.
5. Real-World Applicability: The tweets represent real-world uses of the language, covering most of the informal, often fragmented, and contextual nature of social media text. This therefore ensures that models trained on this data are going to be more robust and applicable in real-world scenarios.

Dataset Preparation for Supervised Learning

This has made it particularly well suited for supervised learning tasks that aim to predict the class of a label according to tweet text. The tweet column included tweet text and a corresponding label in this structured format of the dataset, which makes the problem space well defined. In this structured format of the dataset, tweet_column represents the input feature, and class_column represents the target label for the models.

As for the data selection, a first inspection of the dataset was undertaken to verify quality and appropriate:

• Data Integrity Check: This will ensure that all entries in the dataset are complete, without any missing values in the critical tweet and class columns.
• Class distribution analysis: This is checking the spread of classes to find an imbalance that can affect performance and change the model on the training process.
• Duplicate Removal: A process of detecting and eliminating identical tweets prior to model training to prevent bias and overfitting.
The very strong points of choosing the dataset and its well-varied set of tweets hugged a platform for the following steps in the KDD process. This would help develop effective models of harmful language in the content of social media.

3.2 Data Preprocessing

Data preprocessing in the KDD process is a step for preparing and transforming raw data into forms suitable for analysis and modeling. In this case, it is acutely important for handling textual data, which is unstructured in nature and comes with certain challenges in NLP. For this study, where tweets are labeled to fall under three categories—Hate Speech, Offensive Language, neither—the data preprocessing was performed as follows:

3.2.1. Text Cleaning

Removing Special Characters and URLs: Many times, tweets contain special characters, URLs, and other irrelevant elements related to the semantic meaning of the text. These were removed in order to reduce noise and improve quality. For example, regular expressions have been utilised toward stripping URLs from tweets, while special characters were removed except when they are considered in valid fashion for the analysis.
Lowercasing: This was to ensure that everything in the dataset text was lowercase. All words shall be treated uniformly, regardless of their case—for instance, "Hate" and "hate." That will help avoid training the model to recognize one word in different cases as different entities.

3.2.2. Handling Class Imbalance

Resampling Technique: The dataset classes were quite imbalanced, with a few categories having far fewer examples compared to others. For that purpose, the following resampling techniques are used:
Upsampling: The minority classes, Offensive Language and Neither, were upsampled to further balance out the dataset. Examples from these classes were duplicated in order to match the number of samples present within the majority class: Hate Speech. Upsampling avoids biases rising in model training by learning from a more balanced representation of each class.

3.2.3. Tokenization

DistilBERT Tokenization: The DistilBERT model utilizes what is called the so-called DistilBertTokenizerFast, which provides tokenization specifically oriented to be applied for the DistilBERT model architecture. Tokenization means breaking down this text into smaller basic units, tokens, and encoding them then numerically. This involves the following operations:
Padding: This is a requirement to add some special tokens to make all sequences uniform in length.
Truncation: This method entails cutting off sequences beyond a maximum length to create uniformity in size for inputting.

Tokenization: The RoBERTa model employs RobertaTokenizerFast to handle tokenization for RoBERTa. As in DistilBERT, this involves the following steps:

Special Tokens and Segment IDs: These are specific tokens used by RoBERTa to represent different parts of the text, like segment IDs to distinguish between different input segments.

Padding and Truncation: Bringing all input sequences to a standard length by adding padding where necessary or truncating longer sequences.

3.2.4. Data Augmentation

Augmentation: Data augmentation was done on the RoBERTa model via word substitution with synonyms. It helped increase the diversity in the training dataset. The process involves replacing the words in the tweets using thesaurus or predefined list of synonyms. The model looks up thesaurus or the predefined list for synonyms and replaces them with it. It creates new variants of the original tweets helping the model generalize better to different wordings and context.

3.2.5 Feature Engineering

The feature engineering step involves two things. They are given below:

1. Tweet Length calculates the total number of words or tokens of each tweet. The feature helps the model to understand tweets which have too many words and can add more context to this classification.
2. The sentiment analysis was done using Text Blob, from which sentiment polarity scores had to be drawn for each tweet. Those scores in the form of their negative or positive content were lastly added as features to capture the emotional tone of the tweets.

3.2.6 Feature Scaling

Application of StandardScaler: To make sure that numeric features like tweet length and counts related to various classes contribute equally to the model's training process, feature scaling was done. It was done using StandardScaler. This method standardizes features by removing the mean and scaling to unit variance, which leads to much more stable and effective model trainings.

3.3 Data Transformation

Data transformation constitutes one of the most important stages in the KDD process. The process is primarily aimed at fine-tuning raw data so it can become much more suitable for effective analysis and model training. In this project, this will involve both feature engineering and advanced tokenization techniques to ensure the dataset is prepared for the training of the RoBERTa model.

Advanced Tokenization

Tokenization is the process for text, which consists of changing it into tokens that are basically the smallest units that any machine learning model can process. Now, this process becomes pretty complex in deep models like RoBERTa and involves a lot of sophisticated techniques that ensure the text is properly formatted.

1. **Using RobertaTokenizerFast**: In doing so, the raw text data would first be tokenized using the RobertaTokenizerFast, which would then feed this into the RoBERTa model for processing. This tokenizer is specifically developed to work with requirements from the RoBERTa architecture using special tokens such as [CLS] and [SEP] and segment IDs, key rows in differentiating parts of a given input text. The tokenizer also implements BPE—a

method for processing out-of-vocabulary words effectively by segmenting them into subword units. In this way, even unfamiliar words could be processed by the model without huge losses of information.

2. **Padding and Truncating**: Some padding and truncation work was done on the sequences of tokens he produced with a tokenizer to have them all of a persistent length. Padding is a method wherein more tokens are added to shorter sequences so that all the input sequences become if predefined, which often turns out to be instrumental in training the model in batches. Truncating is applied when such sequences are too long, ensuring that they don't exceed the size constraints of what can be fed into the model. In this project, though, the maximum sequence length was chosen by the distributions of tweet lengths in this dataset, which creates a balance between two conflicting goals: to preserve as much information as possible and computational limitations of the model.

Impact of Data Transformation

This data transformation process contributed meaningfully to making this dataset much more informative and compatible for deep learning models, specifically RoBERTa. Additional features engineered from this dataset included valuable context relating to tweet length and sentiment—two very critical factors in ascertaining the nature of content. Advanced tokenization made sure that text data was accurately and efficiently processed into a format suitable for the model while being mindful of technical limitations to deep learning.

Through these transformations, the dataset, therefore, became more aligned to the task of classifying tweets, hence improving the model performance for detecting hate speech and offensive language and no hurtful or neutral content.

3.4 Data Mining

The data mining phase is the most important part of a KDD process, and here advanced machine learning algorithms will be applied for preprocessing and transformations in search of patterns for building predictive models. In this study, two transformer-based models were used for the task of classification of tweets either as hate speeches, offensive languages, or neutral content. Those were DistillBERT and RoBERTa. Each model was fine-tuned for maximum possible performance.

For efficiency reasons, the use of DistillBERT as the baseline model was necessitated since it is faster and more lightweight compared to its larger predecessor, BERT. Essentially, DistillBERT is a distilled version of BERT, which not only makes it faster and lighter but also protects much of BERT's performance. In this study, an architecture of the DistilBertForSequenceClassification class was used, which is particularly designed for sequence classification problems. This model is pre-trained on large corpora and thus has plausible capabilities in understanding the nuances of a language that would better accomplish this task.

Fine-tuning of DistillBERT optimizer uses AdamW, a variant of Adam that carries weight decay to prevent overfitting. It basically puts regularization on the model weights in an effort to reduce overfitting. Here, the learning rate is 5e-5, which is really common when fine-tuning transformer models since it allows the model to learn new patterns while not diverging too much from those pre-trained weights. Training was done over three epochs, so the model had a chance to pass through the whole dataset several times. This was done on a GPU, which is also one of the requirements for handling such computational demands when training a transformer model; this significantly brings down the training time and improves performance. For instance, the performance of DistillBERT would be used as a benchmark to compare with that of the more complex RoBERTa model. The model was evaluated using similar key metrics as current works: accuracy, precision, recall, and the F1-score, on both the training and validation datasets.

For the base model, RoBERTa was selected since it has very excellent results on many NLP benchmarks. RoBERTa can be thought of as being an optimized variant of BERT, trained on a much larger corpus with more advanced training techniques making the model in capturing all the moods of language more powerful. This research study's architecture was based on the model TFRobertaForSequenceClassification, which is specifically designed for sequence classification tasks. This model is empowered by rich, pre-trained representations of language from RoBERTa in understanding the context and semantics of text.

To ensure high performance of the RoBERTa model, an optimization process for hyperparameters was performed using Optuna—one of the leading and flexible frameworks available for automatic hyperparameter tuning. Optimization in Optuna consist of defining an objective function that this framework aims to maximize; in the present case, this would be the accuracy of the model in the validation set. During the process, several key hyperparameters were tuned, such as the learning rate, batch size, number of epochs, and optimizer type. The learning rate is what will define how fast the model will update its weights during training, and Optuna will run a few learning rates to find the best value. Another hyperparameter that was varied is the batch size: it's the number of samples which shall be processed before updating the model's weights. This will find the best batch size for this dataset and this model architecture. The number of epochs was changed to identify at which point the model's performance began to plateau or degrade. Finally, different optimizers—most notably, Adam and AdamW—have been tested against one another to see which algorithm provides the best convergence behavior for the model.

Apart from hyperparameter tuning, data augmentation was done in terms of replacing some words in the tweets with their synonyms. It increased the diversity of training data without actually changing the real meaning of it, which itself would improve the generalization by the model. However, one major merit of this technique is its enhancement of generalization: making the model ready for deployment in a real-world scenario. It is fine-tuned on both the original and augmented datasets using the best set of hyperparameters identified by Optuna. During this stage, advanced tokenization techniques implemented by the RobertaTokenizerFast took a front seat in ensuring the text data was efficiently processed and fed into the model. It took several epochs for training, and with every successive epoch, the model gradually learned the patterns hidden in the data.

The final evaluation for the RoBERTa model was done using exactly the same metrics as those for the DistillBERT, thus making the results from these two models directly comparable. To be more specific, it was trying to answer the question of how this model could generalize to unseen data and classify tweets into correct categories. It is expected that the performance of this RoBERTa model will improve with augmentation and optimization, outperforming the baseline DistilBERT model. This will prove gains associated with a more complex model architecture coupled with a comprehensive way of tuning hyperparameters. This step shows that rigorous training and optimization are essential in bringing high-performance models to NLP that can be very helpful in performing real-world classification tasks.

3.5 Evaluation

The fine-tuned RoBERTa model for tweet classification was evaluated out-of-the-box with a few optimization trials. In the first optimization trial, with a learning rate of 1.21×10−61.21×10−6 Trained with only a 1e-6 learning rate for one epoch with Adam, the resulting model showed a validation accuracy of 89.69% that turned out to be the final test accuracy. This result, while still strong, really indicated that the very low learning rate and only one epoch did not allow the model to learn fully and generalize this data, since the performances dropped a lot lower compared to the following trials.

In the second trial, with a learning rate of 2.76×10−52.76×10−5 With these two epochs set, using the AdamW optimizer with the fine-tuning scheme, the model has consequently reached 91.99% of accuracy on the validation and test set. That was the best it could get—meaning that with that number of layers and heads, probably RoBERTa is expected to behave properly at its best in the number of accuracies. High accuracy speaks to its resiliency and capability to predict correctly between text pertaining to categories of hate speech, offense, and neutral, semantic, or lexical origin, showing that very likely the deep pre-training and fine-tuning of RoBERTa captured subtle patterns in language that specify category labels.

A third trial with a learning rate of 4.20×10−64.20×10-6 One epoch combined with the previously used optimizer Adam produced a classification rate of 91.38% for both validation and test sets. The performance did not degrade much as compared to the best trial, and the performance was still on a high value, exhibiting that RoBERTa generally generalized well despite a lower learning rate and reduced epochs. It probably showed even more the need for finding the right balance among these notions when attempting to hit the best performance possible with a pre-trained model.

The fine-tuned RoBERTa yielded test accuracy rated at 91.99% with 91.14% F1, which clearly showed its potent power for the classification task, and the performance of the DistilBERT model slightly goes away. DistilBERT obtained the average evaluation scores as follows: accuracy, 91.88%; precision, 91.12%; recall, 91.88%. These metrics suggest that while DistilBERT performs quite well, it might not achieve state-of-the-art results similar to RoBERTa. A slightly lower F1 score and accuracy suggest that although DistilBERT is lighter and more efficacious as a model than its deeper parent, RoBERTa, it fails to keep up on the deeper kind of contextual comprehension that the parent model brings in and that correlates with overall model power of representation. It does so, however, in a much more computational- and model-efficient manner, suggesting a compromise between those kinds of efficiency and model accuracy. This is a clear indication of how RoBERTa outperforms DistilBERT on the tasks, reflecting the model's great robustness in classifying hate speech, offensive language, and neutral content.

As can be seen, trial results underline the fact that RoBERTa was designed with an improved architecture and fine-tuned with optimal hyperparameters, making it more accurate and robust for performing this task. Though DistilBERT does serve efficiency and effectiveness in capacity, it does so by making gains in computational efficiency at the expense of the accuracy of the models. Model selection should be therefore driven by application needs, depending on whether performance or resource constraint is of interest.

4. Results

In this section, lets discuss about all the results which were obtained in this research. There were two models which were implemented the primary model for the research being RoBERTa model and for comparison of the results DistillBERT was also implemented. The future scope of the base paper, which had the same research done using RoBERTa model in a competition (Alonso, Saini and Kovacs, 2020), was to tweak hyperparameters according to the dataset as well as using a bigger dataset for training the model. It mentioned that the model then might give us better results. Hence, bigger dataset was used as well as hyperparameter optimisation was done using optuna to check if we get better results. For the evaluation we used a few metrics such as F1 score, accuracy, val_accuracy, and graphs such as ROC, precision recall curve, correlation matix. The results are shown below:

1. Classification Report:

Table 2: Classification report for DistillBERT

	Precision	**Recall**	**F1-score**	**support**
0	0.68	0.18	0.28	1144
1	0.95	0.96	0.95	15352
2	0.83	0.98	0.90	3330
Accuracy			0.92	19826
Macro Avg	0.82	0.71	0.71	19826
Weighted Avg	0.91	0.92	0.91	19826

Table 3: Classification report for RoBERTa

	Precision	**Recall**	**F1-score**	**support**
0	0.55	0.27	0.36	569
1	0.94	0.97	0.95	7672
2	0.90	0.91	0.90	1673
Accuracy			0.92	9914
Macro Avg	0.79	0.72	0.74	9914
Weighted Avg	0.91	0.92	0.91	9914

2. Confusion matrix:

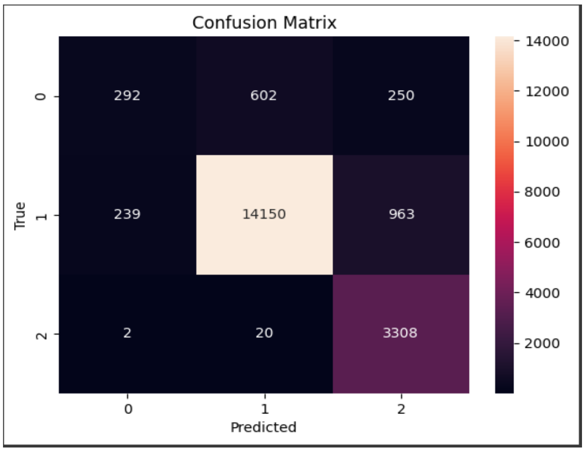

Figure 2: Confusion matrix for DistilBERT.

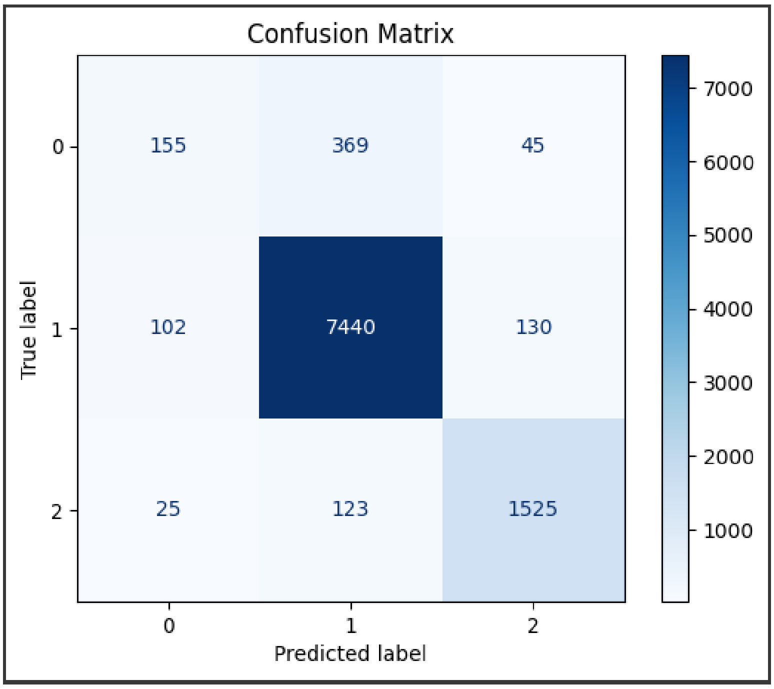

Figure 3: Confusion matrix for RoBERTa.

3. Precision Recall Curve:

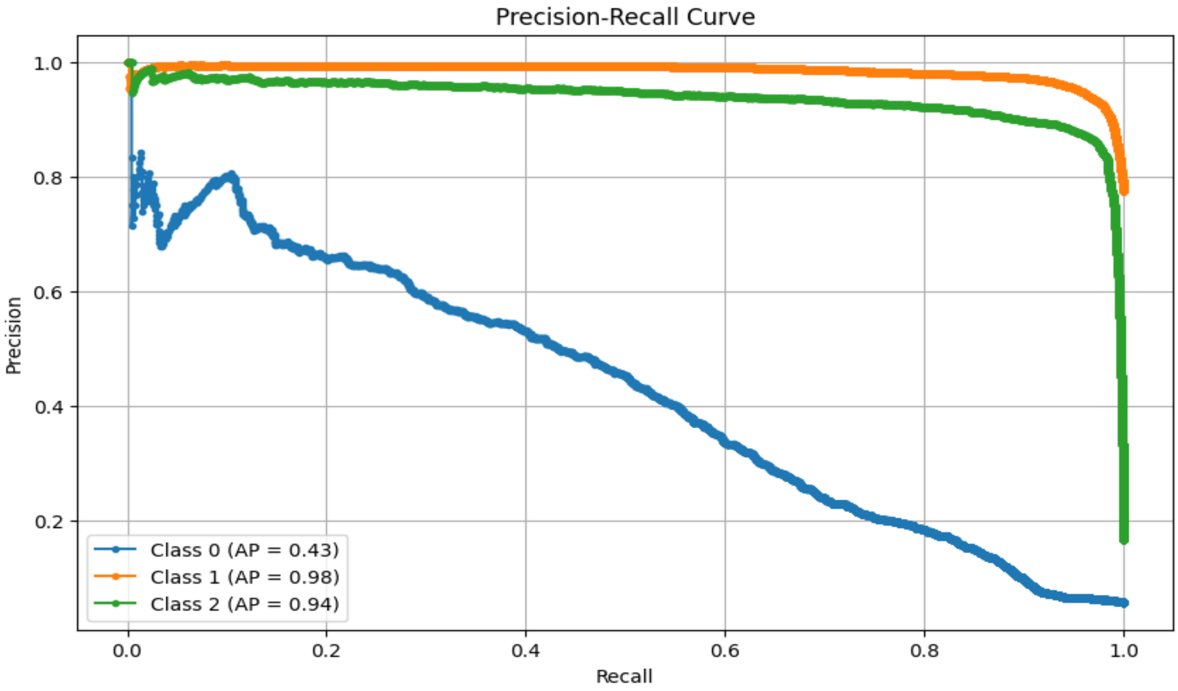

Figure 4: Precision recall for DistilBERT.

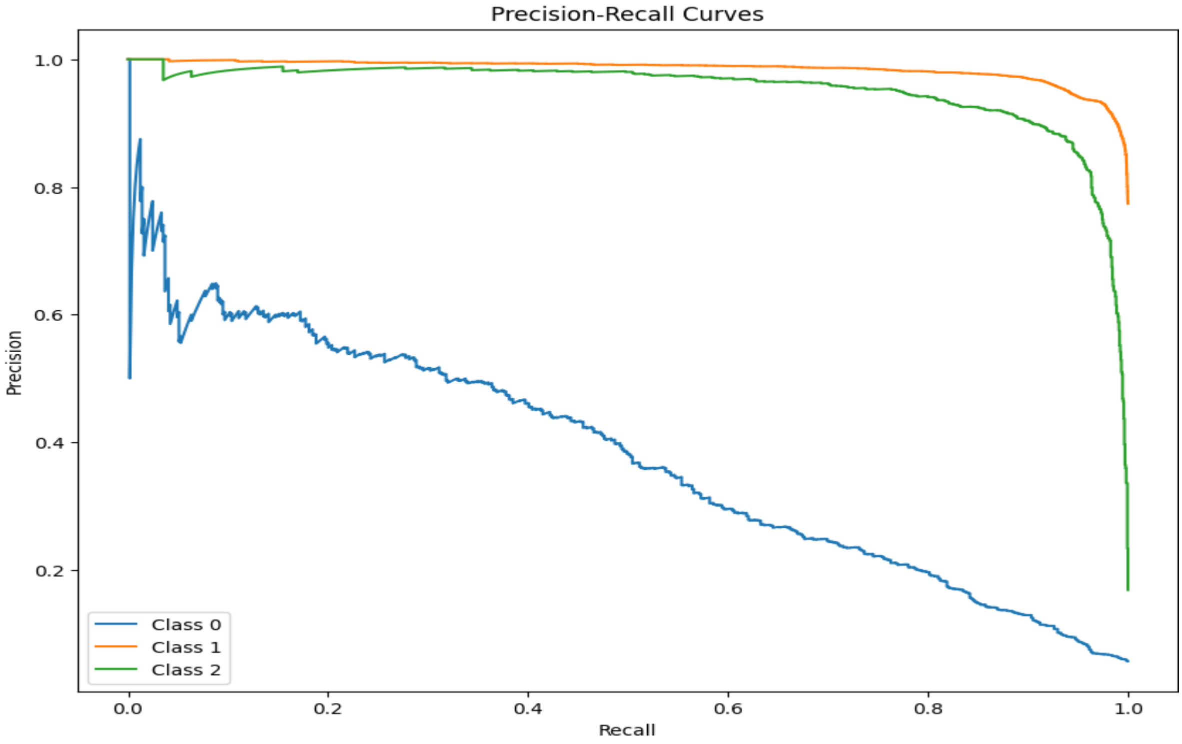

Figure 5: Precision recall for RoBERTa.

4. ROC Curve:

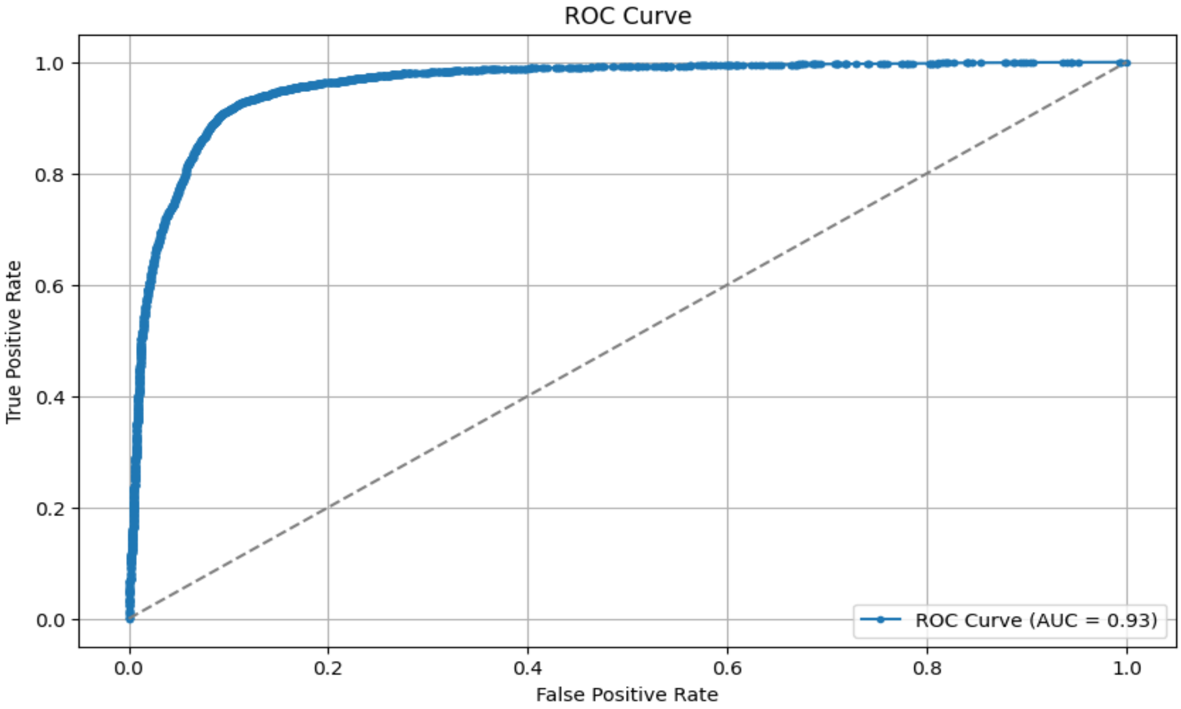

Figure 6: ROC Curve for DistilBERT.

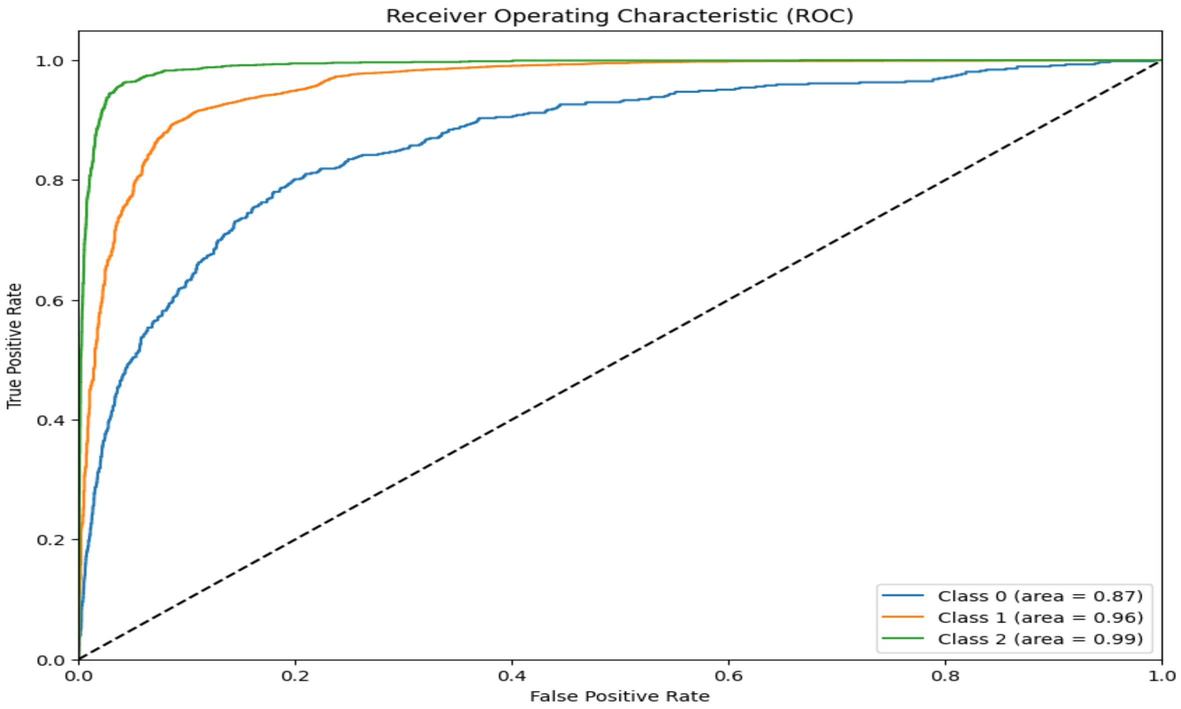

Figure 7: ROC curve for RoBERTa.

DistilBERT Model: In the class "Offensive," it contains the highest number of correct predictions in the confusion matrix, with 14,150 being correctly classified. However, it badly misclassified several cases belonging to the class "Neutral" and predicted them as "Offensive," about 602 of them. It also misclassifies, to a considerable extent, the class labeled as "Hate Speech," where 250 "Neutral" cases were predicted as "Hate Speech."

RoBERTa Model: The general performance of the RoBERTa model is better, reflected in the lower number of misclassifications. Take the "Neutral" instances where the number of misclassifications is 369, classified as "Offensive." More often than those readability metrics produced by the DistilBERT model. Also, the number of correct predictions for the class "Hate Speech" is high, that is, 1,525 correctly classified, demonstrating the strength of RoBERTa in distinguishing hate speech from the others.

The figures 4 and 5 both are for precision recall curve for DistilBERT and RoBERTa respectively. The precision recall curve shows the trade-off between the three different classes given in the dataset. Precision measures the true positives out of all positive predictions and recall measures the proportions of true positives out of all actual positives. The graph has a very slight difference, the class 1 label starts very close to 1.0 and then drops for DistilBERT, but for RoBERTa, the class 1 label starts very close to 1.0 and then stays straight indicating that RoBERTa effectively identifies positive instances with fewer false positives compared to DistilBERT, which is crucial for hate speech detection. The shape of the curve is near to the top right corner which is very ideal. It means that the model can maintain a high accuracy in distinguishing between classes.

The figures 6 and 7 both are for ROC curves for DistilBERT and RoBERTa respectively. The results show that DistilBERT with an AUC of 0.93 does a very good job at classifying between hate speech and non-hate speech. If the AUC (Area Under the Curve) is close to 1 that means the DistilBERT is quite correctly identifying hate speech while avoiding false positives.

However, the RoBERTa model shows more accurate results such as AUC for each class 0,1,2 being 0.87, 0.96, 0.99 respectively. This means that RoBERTa is much better than DistilBERT in classifying between classes.

Overall, these figures underline that RoBERTa is a more powerful model than DistilBERT for detecting different levels of hate speech, making it a more reliable choice for this kind of task.

Below is the distribution of classes after augmentation. Augmentation is one of the most important techniques in machine learning for the artificial extension and diversification of a dataset by the introduction of modified variations of existing data, which provides better generalization because a model would have been through enough variations of scenarios, reducing overfitting and improving general performance on unseen data. It is especially useful when, in certain problems, there may be too little data to help rebalance the classes and increase dataset size without added labeling cost, which ensures more flexibility in a model.

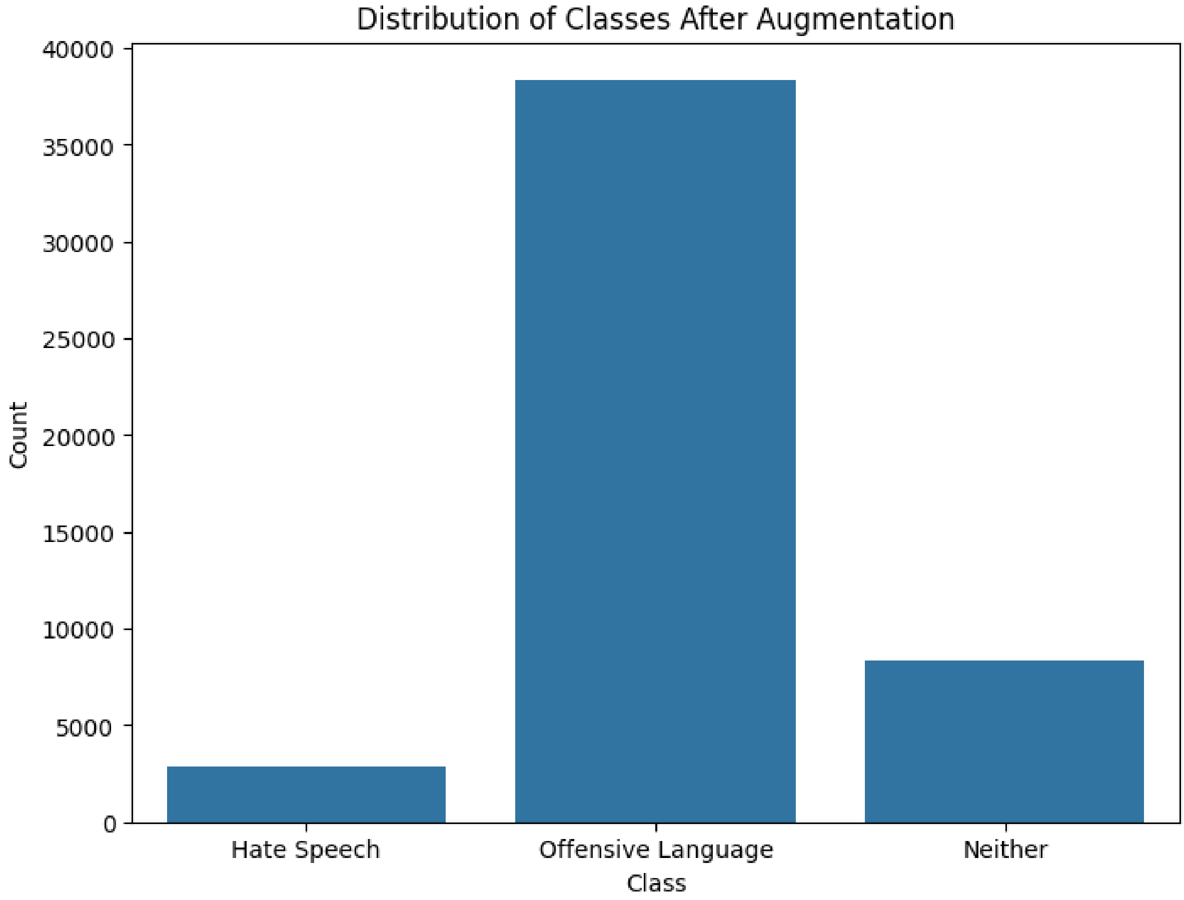

Figure 8: Distribution of classes after augmentation

And finally, below is the optimization path taken by optuna.

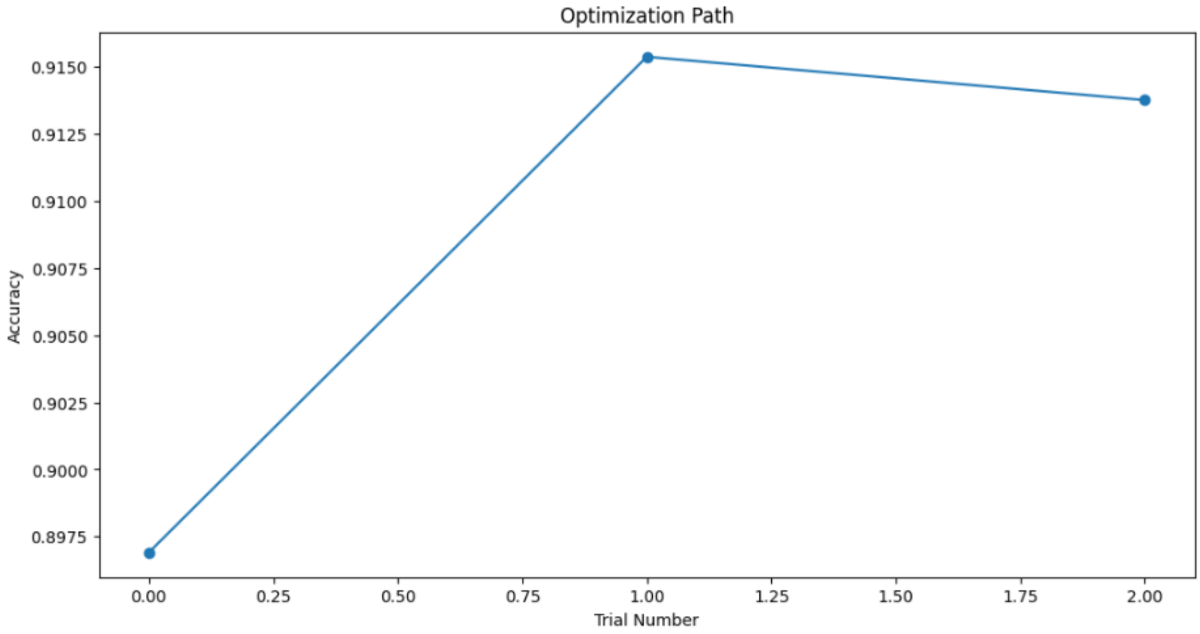

Figure 9: Optimization path by optuna.

5. Conclusion

The research work has been focused on the implementation and fine-tuning of two transformer-based models: RoBERTa and DistilBERT for hate speech and offensive language detection in online content. We are trying to solve some of the problems that make inherent complexities and subtleties of a human language difficult to comprehend within toxic communications by using state-of-the-art knowledge in natural language processing. After extensive experimentation, including hyperparameter optimization using Optuna, these models were fine-tuned to their best performances on our dataset. In this case, among all others, RoBERTa really showed quite outstanding performance with an accuracy of 91.99%. This outperforms the result of DistilBERT at an accuracy of 91.88%. One can easily make out the superior performance from RoBERTa, simply because of a greater model size and more extensive corpus during pretraining, hence better equipped to capture more subtle patterns of language use that associate with hate speech and offensive content.

It is evident from all our different experiments that a transformer-based model holds enormous potential when it comes to handling the challenges thrown toward text classification in the domain of hate speech detection. Although the fine-tuning process increased the accuracy of the models overall, it was mainly reflected in their performance with regard to generalization across a large number of offensive types, as was shown by the confusion matrix analysis. Off these, it was shown that RoBERTa misclassified between hate speech, offensive language, and neutral content at a lower percentage compared to DistillBERT, underlining its strength in telling fine lines of distinction between language. Furthermore, more stable and efficient training was engendered by advanced optimizers, such as AdamW in RoBERTa, further making a case for its position as a more suitable model for the task. This study frames attention for all on careful model selection, fine-tuning, and hyperparameter optimization toward the realization of high-performance text classification systems regarding the detection of hate speech and offensive language.

6. Future Scope

These findings of this paper open up avenues for promising future work in hate speech and offensive language detection. One important direction would be whether there is space for larger and more varied datasets to improve generalization across many languages and cultural contexts. Increasing the data to multilingual content and across different social media sites could bring about more inclusive, resilient models. This would involve the collection and labeling of data from underrepresented languages, covering a broader scope of online discourse and enhancing a model's capability to deal with a wide array of linguistic and cultural variations.

Another important line of future work, VML focuses on enhancing model interpretability and mitigating ethical issues. Transformer-based models, like RoBERTa and DistillBERT, have normally been criticized as fairly "black box" models, having deployment that enforces little transparency in relation to the deciding processes behind those models. Attention visualization or SHAP values could be added to explain what the model has predicted. This would increase the understandability and reduce biases in the model. Third, adding multimodal data such as text, images, videos, audio might improve detection as a result of a deeper overview of online content. In particular, strong emphasis on continuous learning and adaptation through active learning or federated learning could also better maintain the effectiveness of models against evolving language use and rising trends in online communications.

References

Acheampong, F., Nunoo-Mensah, H. and Chen, W. (2020) *Comparative Analyses of BERT, RoBERTa, DistilBERT, and XLNet for Text-based Emotion Recognition*.

Ali, J. (2024) ‘Addressing Challenges in Hate Speech Detection using BERT-based Models: A Review’, *Iraqi Journal for Computer Science and Mathematics,* 5, p. 1. Available at: https://doi.org/10.52866/IJCSM.2024.05.02.001.

Alonso, P., Saini, R. and Kovacs, G. (2020) ‘TheNorth at SemEval-2020 Task 12: Hate Speech Detection Using RoBERTa’, in A. Herbelot et al. (eds) *Proceedings of the Fourteenth Workshop on Semantic Evaluation. SemEval 2020,* Barcelona (online): International Committee for Computational Linguistics, pp. 2197–2202. Available at: https://doi.org/10.18653/v1/2020.semeval-1.292.

FacebookAI/roberta-base · Hugging Face (2024). Available at: https://huggingface.co/FacebookAI/roberta-base (Accessed: 10 August 2024).

Hate Speech and Offensive Language Dataset (no date). Available at: https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset (Accessed: 10 August 2024).

Keshari, N., Malladi, D. and Mittal, U. (no date) ‘Hate Speech Detection Using Natural Language Processing’.

Kovács, G., Alonso, P. and Saini, R. (2021) ‘Challenges of Hate Speech Detection in Social Media’, *SN Computer Science,* 2(2), p. 95. Available at: https://doi.org/10.1007/s42979-021-00457-3.

Lee, N., Jung, C. and Oh, A. (2023) ‘Hate Speech Classifiers are Culturally Insensitive’, in *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP). Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, Dubrovnik, Croatia: Association for Computational Linguistics, pp. 35–46. Available at: https://doi.org/10.18653/v1/2023.c3nlp-1.5.

MacAvaney, S. *et al.* (2019) ‘Hate speech detection: Challenges and solutions’, *PLOS ONE*. Edited by M. Huang, 14(8), p. e0221152. Available at: https://doi.org/10.1371/journal.pone.0221152.

Mozafari, M., Farahbakhsh, R. and Crespi, N. (2019) ‘A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media’. arXiv. Available at: https://doi.org/10.48550/arXiv.1910.12574.

Nayla, A., Setianingsih, C. and Dirgantoro, B. (2023) ‘Hate Speech Detection on Twitter Using BERT Algorithm’, in *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE). 2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, pp. 644–649. Available at: https://doi.org/10.1109/ICCoSITE57641.2023.10127831.

Rezapour, M. (no date) ‘Emotion Detection with Transformers: A Comparative Study’.

RoBERTa (no date) *Deepgram*. Available at: https://deepgram.com/ai-glossary/roberta (Accessed: 10 August 2024).

RoBERTa: An optimized method for pretraining self-supervised NLP systems (no date). Available at: https://ai.meta.com/blog/roberta-an-optimized-method-for-pretraining-self-supervised-nlp-systems/ (Accessed: 10 August 2024).

S, S. *et al.* (2024) ‘A Comparative Exploration in Text Classification for Hate Speech and Offensive Language Detection Using BERT-Based and GloVe Embeddings’, in *2024 2nd International Conference on Disruptive Technologies (ICDT). 2024 2nd International Conference on Disruptive Technologies (ICDT)*, pp. 1506–1509. Available at: https://doi.org/10.1109/ICDT61202.2024.10489019.

Shukla, S., Nagpal, S. and Sabharwal, S. (2022) ‘Hate Speech Detection in Hindi language using BERT and Convolution Neural Network’, in *2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS). 2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pp. 642–647. Available at: https://doi.org/10.1109/ICCCIS56430.2022.10037649.

Valencia, A.M. *et al.* (2020) ‘Proposal for a KDD-based procedure to obtain a set of intelligent systems training applied to the identification of failures in hydroelectric power plants’, *Journal of applied research and technology*, 18(6), pp. 376–389. Available at: https://doi.org/10.22201/icat.24486736e.2020.18.6.1364.

What is the BERT language model? | Definition from TechTarget (no date) *Enterprise AI*. Available at: https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model (Accessed: 10 August 2024).