International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Online Textual Hate Content Recognition using Fine-tuned Transformer Models

Sneha Chinivar [1]* Roopa M S [2] Arunalatha J S [1] Venugopal K R [1]

[1]University Visvesvaraya College of Engineering, India
[2]Dayanand Sagar College of Engineering, India
* Corresponding author's Email: schinivar@gmail.com

Abstract— The popularity, anonymity, and easy accessibility of social media have enabled it as a convenient platform to outspread hate speech. Hate speech can take many forms, viz., racial, political, LGBTQ+, religious, gender-based, nationality-based, etc., overlapping and intersecting with numerous forms of persecution and discrimination, leading to severe harmful impacts on society. It has become crucial to address the problem of online hate speech and create an inclusive and safe online environment. Several techniques have already been investigated to address the issue of online hate speech and have obtained reasonable results. But, their contextual understanding needs to be stronger, and it is quite a complex task as they need larger datasets to take complete advantage of the model's architecture. In this work, we explored the usage of transformer-based pre-trained models, particularly Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized BERT (RoBERTa), to fine-tune them further to detect online hate speech efficiently. Our approach performed well and improved Accuracy and F1-score metrics results by 9.65 percent, Precision and Recall by 10.28 and 8.96 percent, respectively, compared to state-of-art methods with a subsampled dataset, limited resources and time.

Keywords- BERT Embedding-based Ensemble Technique, Online Hate Speech, Social Media, Transformer-based Models.

I. INTRODUCTION

The advent of social media platforms has entirely changed how individuals interact, learn, think, and go about their everyday lives. Social media has come a long way from being a means to be in touch with distant family and friends, a medium for entertainment and marketing, to today's appealing platforms where anybody can voice their opinion. It has turned into a platform where it can bring actual impact on society. But sadly, few people are squandering these platforms to spread hate [1].

Social media is being used by a few to instigate people against issues that can go beyond ideology, culture, and political boundaries and to fuel and maintain the issue's intensity. It is being used as a launch pad to instigate violence in the physical society.

Online hate is not just a sociological problem anymore. It has now turned into a technological problem, too; the New Zeland incident of 2019 at Christchurch demonstrated that when the killer posted a document online expressing his intentions and details of the attack before the episode and live-streamed the attack for about 17 minutes on a social media platform [3]. Fig. 2 represents the results of the study conducted in Sweden in 2016 between March and September, where we see how online posts against refugees directly correlate to an actual physical attack on them [4]. A study conducted by [2] states that between 2019 and mid-2021, on average, for every 1.7 seconds, race or ethnicity-based hate speech content gets posted online. The report further states that discussion over transphobic hate speech has increased by 10%, homophobic hate speech by 2%, acephobia by 14%, and biphobia and queerphobia by 9% and 5%, respectively. In addition, the report cites a 242% rise in discussion over cancel culture, causing a real-life impact on an individual's relationships, careers, and mental well-being.

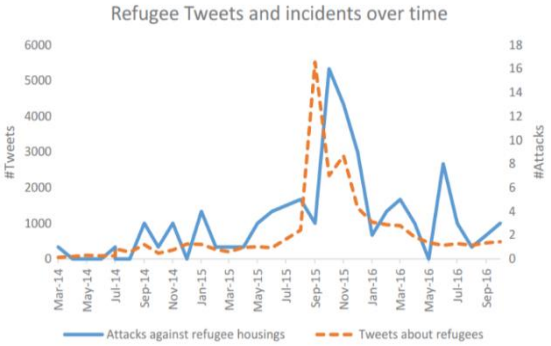

Figure. 1 Results of the Study Conducted in Sweden

Therefore, combating this offensive online behaviour has become inevitable now. The main challenge in addressing this online issue is to make the model automatically understand the text's context. In this work, we have fine-tuned two well-known

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

transformer-based models, namely Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized BERT (RoBERTa) to understand the context of the text better and recognize online hate speech efficiently.

The transformer-based models encode the textual content and their context into numerical representations called word embeddings. These models were pre-trained on the large corpus for them to have a notion of the language beforehand. So that when they are fine-tuned for any downstream task, they become capable of performing a specific task with minimal resources, a smaller dataset, and nominal training time. Thus, in this work, we have fine-tuned BERT and RoBERTa transformer models to perform online hate recognition efficiently.

The main contributions of this work are as follows:

(i) Fine-tuned BERT and RoBERTa models on various hyperparameters to detect online textual hate content efficiently and became the first to evaluate the Binary Hate Speech (BHS) dataset [5] by applying Transformer-based Models.

(ii) Compared the performances of the most popular transformer models and showed their efficiency against the state-of-art concerning Precision, Recall, F1-Score, and Accuracy metrics for a specific, and reduced BHS dataset.

The following section outlines the literature survey; in Section 3, we describe the methodology and brief the experimental setup and results in Section 4. The results are discussed in Section 5, and finally, Section 6 contains the conclusions.

II. LITERATURE SURVEY

Researchers have put considerable effort into tackling the problem of online hate speech in recent years. In this section, we briefly discuss a few of the contemporary works done to completely buckle down online offensive behaviour.

Elzayady et.al., [6] implemented a personality trait feature-based framework to recognize hate speech on the Twitter dataset. Masari et.al., [7] proposed an ensemble technique to identify multiple aspects of hate speech. This approach combined the Bidirectional Encoder Representations from Transformers (BERT) base model with varied deep learning techniques and tried to reduce the misclassification of online posts and enhance the precision of hate speech detectors.

Valle-Cano et al., [8] presented a multi-modal named SocialHaterBERT that combined HaterBERT, a hate speech classifier, and SocialGraph, a representation of users' social context, to identify Twitter hate speech automatically. The model utilizes the characteristics of users within the social networks and the posted content's context to detect online hate. This approach showed its efficiency over the models that solely depend on the textual content to identify online hate content.

Solovev et.al., [9] presented a study that explored the relationship between hate speech and moralized language on online platforms. This study analyzed whether the moralized language's presence in the source post can be a robust hate speech predictor of its commensurable replies. The study offered a unique insight into the underlying mechanisms for the escalation of hate on social media platforms and helped automate hate speech recognition.

Basak et.al., [10] proposed a solution to tackle the problem of online shaming on the Twitter platform by performing multi-class classification of tweets. Based on the obtained categorization, a web application named BlockShame was designed to mute Twitter social media offenders. Ayo et.al., [11] put forward a neural network model that uses hybrid embeddings of Term Frequency – Inverse Document Frequency (TF-IDF) and Long Short-Term Memory (LSTM) to extract meaningful features. These features are given to a better version of the Cucko Search (CS) neural network model to automate detecting Twitter hate speech.

To deal with online hate that comes with the happening of any calamitous event in the real world, Rudra et.al., [12] presented a classifier to accurately identify whether the tweet contains any derogatory content against a community or not during any disastrous event. In addition, this work proposed a real-time system named DisCom developed by collecting tweets posted during any unfortunate hazardous events to recognize communal tweets in any upcoming disaster episode. This system would help the concerned authorities make necessary decisions such as filtering, promoting distinctive contents etc.

Ayo et.al., [13] developed a clustering model based on probability to address the problem of classifying Twitter's hate speech. The proposed model is robust to imbalanced data, fragmentation issues, setting up of a threshold, and imprecision. Kapil et.al., [14] proposed a framework of deep learning-based multi-task learning that utilizes multiple source's information to detect online hate speech. This approach performed consistently well in terms of F1-Score and accuracy metrics compared to models designed for single-task learning.

Mathew et.al., [15] introduced a benchmarked dataset, HateXplain, to identify social media hate speech. Varied state-of-art models viz Convolution Neural Network (CNN) – Gated Recurrent Unit (GRU), Bi-Directional Recurrent Neural Network (BiRNN), BiRNN-Attention, BERT are tested on the data and evaluated it from varied aspects of recognizing online hate. While Prasad et.al., [16] presented a multi-modal classifier called Character Text Image Classifier (CTIC) to perform online hate speech classification. CTIC was developed based on BERT, EfficientNet, and Capsule Network models and got

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

trained with varied sampling techniques and selective training on the dataset consisting of text and its associated images.

Ghosal et.al., [17] presented a language-independent framework that utilizes the emotional and semantic context of the text to detect online hate. The proposed approach showed the significance of the text's parts-of-speech, analysis of the term's co-occurrences, and distance between context terms and terms in recognizing online textual hate.

Besides those approaches mentioned above to address online offensive behaviour, the advancement of research in the field of Natural Language Processing (NLP) made researchers explore varied embedding techniques from the elementary frequencybased methods like Bag of Words (BoW), TF-IDF to the latest transformer-based models like BERT, RoBERTa to understand the context of the text precisely and identify online hate efficiently using various machine learning and deep learning based classifiers [18],[19],[20],[21]. Designing an efficient model that can understand the textual content and its context accurately and proficiently recognize offensive online content is yet to be achieved. In this paper, we have endeavoured to address the problem of online hate by fine-tuning the transformer-based models, which internally consist of an embedding layer and a classifier to classify the online text into its appropriate hate or non-hate category efficiently.

III. PROBLEM STATEMENT

In this section, we define the problem statement for recognizing online hate speech. Given a set of textual tweets T = {t1. . . . . tN}, where N represents the total number of tweets in the input data, our fine-tuned transformer model aims

1) To efficiently categorize tweets into their appropriate class, i.e., hate or not-hate.
2) To enhance the Accuracy, Precision, Recall and F1-Score metrics of the state-of-art method's with a randomly selected subsampled dataset.

In particular, tweets are given individually as input to two transformer-based models, i.e., BERT and RoBERTa. These models internally consist of an embedding layer to convert the textual content to vectors and a classifier, a single layer fully connected neural network, which categorizes the tweet text given as input to its appropriate binary class.

IV. METHODOLOGY

In this section, we describe the transformer-based pre-trained architecture to detect hate speech from social media texts. We have deployed two well-known heavily pre-trained transformer-based models, i.e., BERT and RoBERTa, and fine-tuned their end classifier layer for hate speech identification task by varying a few hyperparameters such as the number of layers, batch size, epoch, learning rate etc. Since BERT and RoBERTa transformer models are pre-trained on a large corpus, they already know the language. With some tuning, they performed exceptionally well with limited data in the hate speech identification task. They saved much time, effort, and resources to train the model from scratch. The architecture diagram of our proposed approach is depicted in Figure. 2.

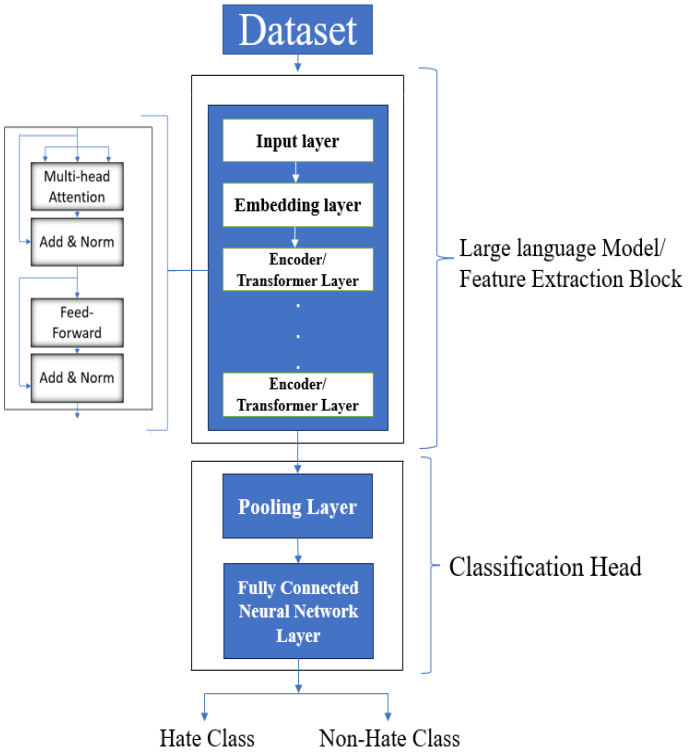

Figure. 2 Architecture to Detect Online Hate Speech

A. Dataset

In this work, we have used the Binary Hate Speech (BHS) dataset [5], which is developed using a Python package called Twython [22] to fetch tweets from the standard search Application Program Interface (API) of Twitter social media. Domain-specific keywords and their varied combinations are also used to extract tweets between January 2018 and May 2020. The tweets collected are filtered out utilizing multiple lexicons to improve the dataset's task specificity. In total, the BHS dataset consists of 10,242 tweets. Of these, 5121 tweets belong to the hate class, and the rest to the non-hate class. In this work, we have used randomly sampled 50% of this dataset to pull down the myth that deep learning architectures require an enormous amount of data to perform well. Out of 5121 tweets, 70%, i.e., 3592 tweets used for training, 20%, i.e., 1014 for testing and the rest 10%, i.e., 514 for validation and the same is depicted in Figure. 3.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

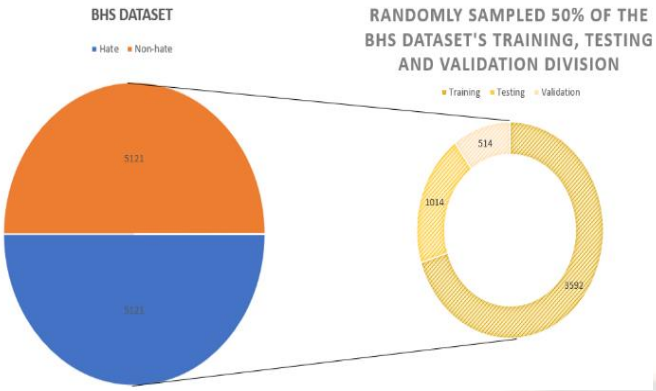

Figure. 3 BHS Dataset Division

B. Fine-tune Transformer-based Models

Transformer-based models are deep learning models built on transformer architecture, a type of neural network. Fine-tuning a pre-trained transformer model helps to achieve better performance on a particular downstream task, i.e., hate speech identification with limited data, as it allows us to take advantage of representations of the language got from its pre-training process and circumvent the requirement to train the model from scratch by saving time and resources.

In this work, to perform efficient hate speech detection, two transformer models have been used,

(i) Fine-tuned BERT model
(ii) Fine-tuned RoBERTa model

The utilized fine-tuned Transformer models consist of essential components, specifically Input Layer, Embedding Layer, Transformer or Encoder Layer, Pooling Layer and Classifier Layer.

The input layer of the models takes the raw textual data as input and uses SentencePiece tokenizer, a variant of WordPiece tokenizer, to convert the given text to tokens. Even though both BERT and RoBERTa model uses the same tokenizer, the way they handle the tokenization process with special characters and the casing of words varies. Since RoBERTa is trained comparatively on a larger corpus, the tokens it generates are unique from BERT. In addition, this layer adds two unique tokens, i.e., [CLS] and [SEP], to provide the model with additional information regarding the text structure given as input. [CLS] token is added to the start of each input text sequence to represent the task to be performed in classification and concedes the model to consider the entire sequence's content instead of individual tokens to perform the classification task. [SEP] tokens are used in BERT to represent the sequence with more than one input sequence, whereas it is used to separate segments of a single input sequence in RoBERTa.

The output of the input layer is given to the Embedding Layer to map the tokens to their corresponding vector representations from the pre-trained BERT in the case of fine-tuned BERT model and pre-trained RoBERTa in the case of fine-tuned RoBERTa model as both the models are pre-trained on different corpora.

The output of the Embedding Layer is given to the BERT and RoBERTa Transformer Layer of fine-tuned BERT and RoBERTa models, respectively. The Transformer Layer of both the models consists of a stack of twelve Encoders, internally consisting of a Multi-head Self Attention Layer and a Feedforward Layer. Multi-head Self Attention layer encodes the input sequence and captures the context of each token by calculating the attention scores between them. The output of the Self-Attention Layer is given to the Feed-forward network, which is a fully connected neural network that captures the complex and non-linear relation between words of the input sequence by applying a non-linear transformation to each token of the sequence independently. Normalization of each sublayer is performed before passing them to the next layer by adding Add & Norm Layer, which adds a residual connection between the input and output of both sublayers to mitigate vanishing and exploding gradient problems.

The Pooling Layer of both models aggregates the output from the Transformer Layer to a fixed-size vector representation of the entire input sequence.

Finally, the Classifier Layer takes the representations from the Pooling Layer and applies linear transformation followed by the Sigmoid activation function to perform binary classification of hate speech identification task.

The weights of pre-trained layers of BERT and RoBERTa transformers are freezed and trained only in the final Classification Layer for recognizing online hate on 70% of the dataset, i.e., training data during the fine-tuning process. Thus, the transformer-based models learned more reasonable representations and performed online hate detection tasks better with a smaller dataset. The optimal hyperparameters of both models, such as the Number of Layers, Epoch, Learning Rate, Batch size etc., for the hate speech recognition task, are selected using the validation dataset. Finally, the model is evaluated on the test dataset to determine its performance.

V. EXPERIMENTAL SETUP AND HYPERPARAMETER TUNING

The experiments conducted in this study is carried out on a 32 GB RAM system having a GPU of type Quadro M4000. A Python version of 3.9.13 has been used to write the code, and a Hugging Face version 4.20.1 library is used to fine-tune transformer models. A PyTorch framework and a Scikit learn library has been used for the implementation.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

A. Training Details of Fine-tuned Transformer-based Models

The selected hyperparameters for fine-tuned transformer models to perform hate speech identification downstream task on the dataset given by Ghosh et.al., [5] are presented in Table I.

Table 1. TRANSFORMER MODELS HYPERPARAMETER FOR HATE SPEECH DETECTION TASK

Hyperparameter	Chosen Value
Batch Size	16
Epoch	10
Neural Network Depth	01
Neurons Count per Each Layer	128
Learning Rate	0.00002 (2e-5)
Dropout	0.5
Optimizer	AdamW
Adam epsilon	0.00000001 (1e-8)
lr-AdamW Momentum	0.7

With this setup, fine-tuned BERT and RoBERTa took 44 minutes & 6 seconds, and 1 hour, 7 minutes & 59 seconds, respectively, to get trained.

B. Fine-tuning of BERT for Hate Speech Identification

BERT, developed by Google, is a transformer-based language model pre-trained on 800 million words of BooksCorpus and 2,500 million words of English Wikipedia corpus. It uses a multi-layer bi-directional transformer encoder to generate deep bi-directional vector representations of the textual data [23].

Fine-tuning is the process of making a pre-trained model adapt itself to perform any particular downstream task, viz., classification, language inference, question answering etc., that may have different data distribution and set of labels. Finetuning allows utilizing the knowledge of pre-trained models to perform any particular task and helps to achieve better performance and faster training. The generalization ability of pre-trained models helps to enhance the performance metrics of a specific downstream task with a smaller amount of labelled data.

While fine-tuning the pre-trained BERT, we kept the BERT model as the starting point and added a single neural network followed by a sigmoid function. Usually, while fine-tuning the model, embedding and classifier layers are trained simultaneously to improve the performance of the downstream task. In this work, we have trained only the classifier layer using the cross-entropy loss function as our dataset is relatively small and achieved better performance by just training the final layer. Thus fine-tuned BERT captured the specificity (fine feature) required to detect hate speech and enhanced the model's overall performance.

C. Fine-tuning of RoBERTa for Hate Speech Identification

RoBERTa, a variant of BERT developed by Facebook AI Research (FAIR) in 2019, addresses some of the limitations and issues of the BERT model.

The significant variations done to BERT to create RoBERTa are as follows.

(i) Training Data: RoBERTa model is pre-trained on a more extensive and diverse dataset than BERT. It is pre-trained on a combination of BookCorpus, which consists of 800 million words, and English Wikipedia, which consists of 2.5 billion words and varied web pages. In addition, it is pre-trained for a longer duration than BERT, with larger batch sizes and more training steps allowing the model to learn more complex patterns and relationships between data.

(ii) Byte-Pair Encoding: RoBERTa utilizes the variant of the Byte Pair Encoding algorithm to handle rare and out-of-vocabulary words efficiently. Since RoBERTa's vocabulary size and training data are more, the resulting tokens are better than BERT.

(iii) Dynamic Masking: RoBERTa uses a dynamic masking technique during pre-training instead of masking the same tokens as BERT.

(iv) No Next Sentence Prediction (NSP) Objective: During pre-training, Roberta is trained only for the Masked Language Modelling (MLM) task, whereas BERT is pre-trained for both MLM and NSP objectives.

These modifications allowed the model to explore data thoroughly and learn more conceptualized text representations and made the model more robust and efficient. Thus, when it is fine-tuned by adding a single neural network layer called Classifier Layer to the end of the pre-trained RoBERTa model and trained it for hate speech detection using cross-entropy loss, we were able to get enhanced performance metrics in comparison to the state-of-art methods for all four metrics viz., Precision, Recall, F1-Score and Accuracy.

D. BERT Embedding-based Ensemble Technique

As a baseline, we experimented with a combination of the BERT embedding technique and ensembled deep learning models to detect online hate. We used BERT transformer model only to generate the embedding vectors of the text, unlike fine-tuned BERT model. The obtained vector representations are given independently to

(i) CNN, a deep learning model that utilizes convolutional layers to extract useful features for the classification task, pooling layer to reduce the dimensionality of the resultant vector, followed by

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

fully connected neural networks to perform binary classification of hate speech detection task [24]. For our experiment, 2 convolutional layers are used in sequence, each layer having 150 filters (size 2-4). The output from both the layers is merged to get a result of the same shape as that of the individual layer. Max pooling of size 2 is applied after each convolutional layer to extract the most important features and it is passed through 2 dense layers, each having 100 neurons and ReLu activation function, and finally through the output layer containing softmax activation function.

(ii) Long Short-Term Memory (LSTM), a type of Recurrent Neural Network (RNN) capable of processing sequential data and learning long-term dependencies between time steps [25]. For the task of hate speech detection, LSTM with 3 gates i.e., forget, input, and output and 128 neurons are used after Bidirectional LSTM (BiLSTM) containing 256 neurons. The output is passed through word-attention layer and then through 2 dense layers containing 100 neurons and ReLu activation function each, and eventually through the softmax activation function for final classification.

(iii) Gated Recurrent Unit (GRU), a type of RNN similar to LSTM with fewer parameters [26]. Bidirectional GRU (Bi-GRU) with 256 neurons are used for our task, followed by a GRU layer containing 128 neurons. The output from GRU is passed through the word attention layer and then through 2 dense layers consisting of 100 neurons and ReLu activation, and then finally through the output layer consisting of softmax activation.

(iv) Stacked CNN over LSTM model, where CNN block is used to extract features of input data and LSTM for sequence prediction.

(v) Stacked CNN over GRU model, where CNN will be stacked on top of GRU model and performed hate speech identification.

Finally, the result obtained from all these five models is ensembled and analyzed in the final ensembled model i.e., BEET's performance.

VI. RESULTS AND DISCUSSIONS

In this section, we discuss the experimental results of the proposed approach. Fine-tuning the transformer models to perform binary classification of hate speech recognition acutely cut down the resources, time, and data needed as transformers have certain weights fixed during the pre-training process resulting in models having some knowledge of the language beforehand.

The proposed approach is evaluated using four varied metrics; they are

1) Accuracy, a fundamental metric that measures the model's overall performance by dividing the number of correct classifications by the total number of classifications performed.

2) Precision, a metric that measures the proportion of instances correctly classified as a positive class, i.e., hate speech in our work, out of all instances classified as the positive class.

3) Recall, a metric that measures a model's effectiveness in finding out all its positive classes. It does this by dividing the number of correctly classified positive classes by the total number of correctly identified positive and incorrectly identified positive instances as negative classes.

4) F1-Score, a metric that combines the Precision and Recall scores into a single score by calculating their harmonic mean to provide a balanced measure of the model's performance.

The evaluation results of both the fine-tuned BERT and RoBERTa models are presented in Table 2. Figure. 4(a)-(d) and Figure. 5(a)-(d) represent the graphs of the fine-tuned transformer model's all four metrics against the Trainer or Global Step of validation dataset, a separate unseen data the model is witnessing for the first time after being trained. Trainer/Global Step is a measuring unit that gives each batch count and is calculated by dividing the dataset size by the product of batch size and epochs. Batch is a training dataset's subset utilized to update model parameters during training. Batch size is a hyperparameter that sets out the number of training examples utilized in one epoch. Epoch specifies the number of times the model has seen the complete training data.

Table 2. EVALUATION RESULTS

Fine-tuned BERT Model				
	Accuracy	Precision	Recall	F1-Score
Validation	0.9610	0.9759	0.9455	0.9604
Test	0.9763	0.9820	0.9704	0.9761
Fine-tuned RoBERTa Model				
Validation	0.9785	0.9959	0.9610	0.9782
Test	0.9852	0.992	0.9783	0.9851

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

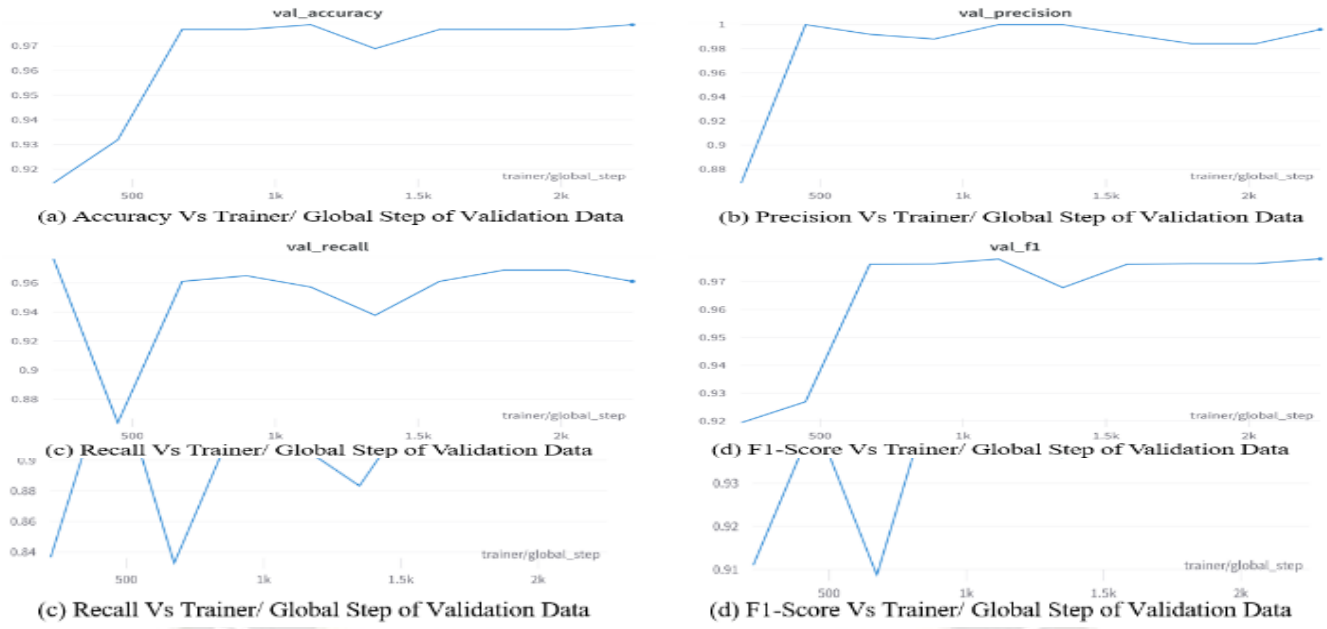

Figure. 4 Fine-tuned BERT's Evaluation Metric Graphs of Validation Data

Figure. 5 Fine-tuned RoBERT's Evaluation Metric Graphs of Validation Data

Figure. 5(a)-(d) clearly shows the efficiency of fine-tuned RoBERTa model on the unseen dataset and its capability to effectively get generalized to new data and not to get overfit to the training dataset. The efficiency of the fine-tuned BERT model is depicted in Figure. 4(a)-(d), and the obtained result could be more satisfying when compared to the fine-tuned RoBERTa. Fine-tuned BERT model's validation data with respect to all four metrics show a lot of steep and sharp curves until it passes around 2000 Global Steps. Comparatively fine-tuned RoBERTa model's validation data concerning all four metrics shows less variation of curves, proposing that the model quickly becomes generalized to the new data.

Table 3. compares all four metrics of our proposed approach with the state-of-art methods that have used the same dataset. In addition, the Table includes the baseline BEET model, and our previous work of recognizing online hate using fine-tuned A Light BERT (ALBERT) is included. The results clearly show transformer-based models' efficiency in identifying online hate speech. In particular, the fine-tuned RoBERTa model exceeded all other approaches' performance and proved its efficiency

Table 3. COMPARISON OF THE PROPOSED APPROACH PERFORMANCE WITH THE STATE-OF-ART METHODS

Approach	Accuracy	Precision	Recall	F1-Score
SEHC [5]	0.8887	0.8892	0.8887	0.8886
BEET (Baseline)	0.8474	0.8477	0.8474	0.8473
Fine-tuned ALBERT	**0.9977**	**0.9508**	**0.9546**	**0.9527**
Fine-tuned BERT	**0.9763**	**0.9820**	**0.9704**	**0.9761**
Fine-tuned RoBERTa	**0.9852**	**0.9920**	**0.9783**	**0.9851**

concerning Precision, Recall, and F1-Score metrics. The accuracy metric of fine-tuned ALBERT beats the performance of the fine-tuned RoBERTa model by a negligible number. Thus demonstrated, by fine-tuning pre-trained transformer models, identification of online hate speech can be made efficiently with less effort, time, and resources.

A. Ablation Study

We experimented with a baseline model named BEET and three varied transformer models to study the role of different hyperparameters of the varied transformer models. We started our experiment with BERT embedding-based ensembled deep learning models (BEET). Then, instead of just using the BERT generated embeddings, we experimented with the BERT transformer model (Embeddings + Classifier) and fine-tuned it with randomly chosen hyperparameters of 1 neural network

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

Table 4. ABLATION STUDY EVALUATION METRICS

Fine-tuned RoBERTa with	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Neural Network Depth = 01	0.9881	0.9881	0.9881	0.9881
Number of Neurons Count per Each Layer = 128				
Number of Instances per Batch = 02				
Neural Network Depth = 01	0.9822	0.9803	0.9842	0.9822
Number of Neurons Count per Each Layer = 128				
Number of Instances per Batch = 04				
Neural Network Depth = 01	0.9861	0.9881	0.9842	0.9861
Number of Neurons Count per Each Layer = 128				
Number of Instances per Batch = 08				
Neural Network Depth = 02	0.9822	0.988	0.9763	0.9821
Number of Neurons Count per Each Layer = 128 and 32				
Number of Instances per Batch = 08				
Neural Network Depth = 03	0.9871	0.9881	0.9861	0.9871
Number of Neurons Count per Each Layer = 256,128 and 32				
Number of Instances per Batch = 08				
Neural Network Depth = 04	0.9832	0.9919	0.9743	0.983
Number of Neurons Count per Each Layer = 512,256, 128 and 32 Number of Instances per Batch = 08				

layer, 128 neurons in that one layer, batch size 16, 10 epochs, and learning rate 0.00002. Then, with the same hyperparameter as BERT's, we experimented with ALBERT, a variant of BERT, and found the Accuracy metric improved by 2.14%, but the Precision metric gets reduced by 3.12%, Recall by 1.58% and F1-Score by 2.34% in comparison to fine-tuned BERT model. Further, we experimented with another variant of BERT named RoBERTa with the same hyperparameters as BERT. We found that the Accuracy metric improved by 0.89%, Precision by 1%, Recall by 0.79%, and F1-Score by 0.9% compared to the fined-tuned BERT model, proving the efficiency of fine-tuned RoBERTa model in comparison to all other approaches adopted in identifying online hate speech.

Since we randomly chose the hyperparameter initially, we started experimenting with varied hyperparameters of fine-tuned RoBERTa model. We began by experimenting with different numbers of instances per batch and then with the neural network depth and neuron count per layer. The results obtained from these experiments are listed in Table 4.

We even experimented with the number of epochs but found optimal results with ten epochs, so we fixed the model's number of epochs to 10. Thus, we finally realized the efficiency of fine-tuned RoBERTa model with initially chosen random hyperparameters viz., one neural network layer, 128 neurons in that one layer, batch size 16, 10 epochs and 0.00002 learning rate in recognising online hate speech with limited resources.

VII. CONCLUSIONS

Online hate speech is spreading like wildfire in recent years with an increase in the use of online social networks. Thus, in this study, we attempted to address the problem of online hate speech detection by fine-tuning the transformer-based models. This work notably used BERT and RoBERTa transformer models to fine-tune them to identify online hate speech efficiently. Since these transformer models are pre-trained with larger datasets beforehand, their contextual understanding capability of the text is high. Usage of these models enabled us to recognize online textual hate content with a smaller dataset,

limited training time, and resources compared to state-of-the-art methods.

The proposed work put forward a step to combat online hate efficiently in social media texts. It can be further extended with varied other advanced transformer models like T5, XLNET etc., and ensembling the fine-tuned transformer model is also an interesting experiment that can be carried out in the future and analyze whether they can further improve the efficiency of the online offensive content identification task. In addition to the textual data, recently, there has been an increase in the spreading of online hate messages via memes, gifs, images, and videos. Combating these non-textual hate content has become vital and an interesting future direction to pursue.

REFERENCES

[1] S. Chinivar, M. S. Roopa, J. S. Arunalatha, and K.R.Venugopal, "Online offensive behaviour in socialmedia: Detection approaches, comprehensive review and future directions," Entertainment Computing, p. 100544, 2022.

[2] D. T. Label, "Uncovered: Online Hate Speech in the Covid Era,"https://www.brandwatch.com/reports/online-hate-speech/ view/#

[3] BBC, "Christchurch shootings: 49 dead in New Zealand mosque attacks," https://www.bbc.com/news/world-asia-47578798, 2019.

[4] M. Khamaiseh, "The Problem with Hate Speech: How the Media has Fuelled its Rise," https://institute.aljazeera.net/en/ajr/ article/1697, 2021.

[5] S. Ghosh, A. Ekbal, P. Bhattacharyya, T. Saha, A. Kumar, and S. Srivastava, "SEHC: A Benchmark Setup to Identify Online Hate Speech in English," IEEE Transactions on Computational Social Systems, 2022.

[6] H. Elzayady, M. S. Mohamed, K. M. Badran, and G. I. Salama, "A Hybrid Approach based on Personality Traits for Hate Speech Detection in Arabic Social Media," International Journal of Electrical and Computer Engineering, vol. 13, no. 2, p. 1979, 2023.

[7] A. C. Mazari, N. Boudoukhani, and A. Djeffal, "BERT-based Ensemble Learning for Multi-aspect Hate Speech Detection," Cluster Computing, pp. 1–15, 2023.

[8] G. del Valle-Cano, L. Quijano-Sanchez, F. Liberatore, and ´ J. Gomez, "SocialHaterBERT: A Dichotomous Approach for ´ Automatically Detecting Hate Speech on Twitter through Textual Analysis and User Profiles," Expert Systems with Applications, vol. 216, p. 119446, 2023.

[9] K. Solovev and N. Prollochs, "Moralized Language Predicts ¨ Hate Speech on Social Media," PNAS Nexus, vol. 2, no. 1, p. pgac281, 2023.

[10] R. Basak, S. Sural, N. Ganguly, and S. K. Ghosh, "Online Public Shaming on Twitter: Detection, Analysis, and Mitigation," IEEE Transactions on computational social systems, vol. 6, no. 2, pp. 208–220, 2019.

[11] F. E. Ayo, O. Folorunso, F. T. Ibharalu, and I. A. Osinuga, "Hate Speech Detection in Twitter using Hybrid Embeddings 14 and Improved Cuckoo Search-based Neural Networks," International Journal of Intelligent Computing and Cybernetics, vol. 13, no. 4, pp. 485–525, 2020.

[12] K. Rudra, A. Sharma, N. Ganguly, and S. Ghosh, "Characterizing and Countering Communal Microblogs during Disaster Events," IEEE Transactions on Computational Social Systems , vol. 5, no. 2, pp. 403–417, 2018.

[13] F. E. Ayo, O. Folorunso, F. T. Ibharalu, I. A. Osinuga, and A. Abayomi-Alli, "A Probabilistic Clustering Model for Hate Speech Classification in Twitter," Expert Systems with Applications, vol. 173, p. 114762, 2021.

[14] P. Kapil and A. Ekbal, "Leveraging Multi-domain, Heterogeneous Data using Deep Multitask Learning for Hate Speech Detection," arXiv preprint arXiv:2103.12412, 2021.

[15] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "Hatexplain: A Benchmark Dataset for Explainable Hate Speech Detection," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 17, 2021, pp. 14 867–14 875.

[16] N. Prasad, S. Saha, and P. Bhattacharyya, "A Multimodal Classification of Noisy Hate Speech using Character Level Embedding and Attention," in 2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021, pp. 1–8.

[17] S. Ghosal and A. Jain, "HateCircle and Unsupervised Hate Speech Detection incorporating Emotion and Contextual Semantic," ACM Transactions on Asian and Low-Resource Language Information Processing, 2022.

[18] N. H. Cahyana, S. Saifullah, Y. Fauziah, A. S. Aribowo, and R. Drezewski, "Semi-supervised Text Annotation for Hate Speech Detection using K-Nearest Neighbors and Term Frequency-Inverse Document Frequency," Int. J. Adv. Comput. Sci. Appl, vol. 13, no. 10, pp. 147–151, 2022.

[19] S. Akuma, T. Lubem, and I. T. Adom, "Comparing Bag of Words and TF-IDF with Different Models for Hate Speech Detection from Live Tweets," International Journal of Information Technology, pp. 1–7, 2022.

[20] M. Almaliki, A. M. Almars, I. Gad, and E.-S. Atlam, "ABMM: Arabic BERT-Mini Model for Hate-Speech Detection on Social Media," Electronics, vol. 12, no. 4, p. 1048, 2023.

[21] T. Hettikankanamge and A. Pinidiyaarachchi, "Multi-label Emotion Classification of Tweets with Transformer Models," 2023.

[22] R. McGrath, "twython 3.9.1," https://pypi.org/project/twython/, 2021.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2018.

[24] Zhang and B. Wallace, "A Sensitivity Analysis of (and practitioners' guide to) Convolutional Neural Networks for Sentence Classification," arXiv preprint arXiv:1510.03820, 2015.

[25] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] K. Cho, B. Van Merrienboer, D. Bahdanau, and Y. Bengio, "On ¨ the Properties of Neural Machine Translation: Encoder-Decoder Approaches," arXiv preprint arXiv:1409.1259, 2014.