Elevating Offensive Language Detection: CNN-GRU and BERT for Enhanced Hate Speech Identification

M.Madhavi[1], Dr. Sanjay Agal[2], Niyati Dhirubhai Odedra[3], Harish Chowdhary[4],
Taranpreet Singh Ruprah[5], Dr. Veera Ankalu Vuyyuru[6], Prof. Ts. Dr. Yousef A.Baker El-Ebiary[7]

Assistant professor, Department of CSE, Velagapudi Ramakrishna Siddhartha Engineering College,
Vijayawada, Andhra Pradesh, India[1]
Professor, Professor, Faculty of Engineering, Parul University,
P.O.Limda, Ta.Waghodia – 391760, Dist. Vadodara, Gujarat, India[2]
Assistant Professor, Department of Computer Engineering,
Dr V R Godhania College of Engineering & Technology, Gujarat, India[3]
Rashtriya Raksha University, Gandhinagar, Gujarat, India[4]
Assistant Professor, Rajarambapu Institute of Technology -Sakharale, India[5]
Assistant Professor, Department of Computer Science and Engineering,
Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India[6]
Faculty of Informatics and Computing, UniSZA University, Malaysia[7]

***Abstract*—Upholding a secure and accepting digital environment is severely hindered by hate speech and inappropriate information on the internet. A novel approach that combines Convolutional Neural Network with GRU and BERT from Transformers proposed for enhancing the identification of offensive content, particularly hate speech. The method utilizes the strengths of both CNN-GRU and BERT models to capture complex linguistic patterns and contextual information present in hate speech. The proposed model first utilizes CNN-GRU to extract local and sequential features from textual data, allowing for effective representation learning of offensive language. Subsequently, BERT, advanced transformer-based model, is employed to capture contextualized representations of the text, thereby enhancing the understanding of detailed linguistic nuances and cultural contexts associated with hate speech. Fine tuning BERT model using hugging face transformer. To execute tests using datasets for hate speech identification that are made accessible to the public and show how well the method works to identify inappropriate content. By assisting with the continuing efforts to prevent the dissemination of hate speech and undesirable language online, the proposed framework promotes a more diverse and secure digital environment. The proposed method is implemented using python. The method achieves 98% competitive performance compared to existing approaches LSTM and RNN, CNN, LSTM and GBAT, showcasing its potential for real-world applications in combating online hate speech. Furthermore, it provides insights into the interpretability of the model's predictions, highlighting key linguistic and contextual factors influencing offensive language detection. The study contributes to advancing hate speech detection by integrating CNN-GRU and BERT models, giving a robust solution for enhancing offensive content identification in online platforms.**

***Keywords*—Bidirectional encoder representations from transformers; convolutional neural network; Gated Recurrent Unit; hate speech; hugging face transformer**

I. INTRODUCTION

Speech that is hateful against people or organizations that defy societal norms and has the potential to cause injury, intimidation, abuse, embarrassment, and disorder in society is known as hate speech [1]. Social media is a worldwide platform where people may freely express their ideas and opinions. Social media has many advantages, but it also has drawbacks, such as hate speech and the publication of derogatory and vulgar information that targets a person, a community, or society as a whole. Hate speech and other unpleasant and offensive content have a negative impact on people's everyday lives and, in the worst cases, can lead to despair or suicide in online sociability. Numerous countries have imposed restrictions on internet media material, with the stipulation that it must neither be directed against any specific people or group, nor incite criminal activity. Furthermore, online social networking sites with their regulations in place include Facebook, Twitter, and YouTube for getting rid of hate speech and other information that has a detrimental impact on society. However, social media businesses continue to confront a significant difficulty in identifying such unacceptable information as soon as possible in order to stop the spread of such news online [2] and researchers. It's challenging to define and comprehend hate speech. The number of Web of Science-indexed productions grew from 42 to 162 between 2013 and 2018, indicating a consistent expansion in academic interest in hate speech since2014. This further emphasizes how hate speech affects the society in which it manifests. Moreover, research on HS is published in a variety of publications, including those that deal with law, sociology, communication, psychology, and so forth. All of this demonstrates its extraordinary significance and the necessity for a thorough examination of its history and present state, which is exactly what this work aims to provide.

Hate Speech is a purposeful and deliberate public remark meant to disparage a certain group of people. Identifying traits like colour, religion, race, ethnicity or nationality, gender, sexual orientation, or identity are some examples of further definitions of hate speech. The biggest obstacle facing the legal literature, particularly has written the most on this topic, is proving that there is a distinct variation between hate speech and hate crime to justify the use of criminal penalties. However, severe forms of hatred, such as calling for terrorism or genocide, do not raise this dilemma. Given that social studies are taught in a variety of ways in mainstream media and on social media platforms, the task is more difficult. It first appears symbolically, nonverbally, and audibly. Second, it is purposefully written in a way that is evasive, unclear, and symbolic, making it challenging to understand. Additionally, discriminatory thinking that is socially acceptable and so not recognized as such is voiced in high school discourse. Third, hate speech frequently uses strong, derogatory language to incite the audience to be offended and/or act, and it assumes that others have bad or deceptive motives [3].

In recent times, platforms like Facebook have ramped up their content moderation efforts, employing both automated methods and human moderators to handle the influx of content. Automated tools have the potential to streamline the evaluation process or allocate human resources to areas requiring meticulous scrutiny. A fundamental method for detecting hate speech is the keyword-based approach, wherein text containing potentially hostile terms is identified using dictionaries or ontologies [4]. For instance, hate base maintains as insulting language targeting various groups across 95 languages, crucial resources given the evolving nature of language. However, it's important to note that merely using a derogatory term doesn't always qualify as hate speech, as observed in the investigation of hate speech criteria. While keyword-based strategies are quick to grasp and comprehend, they have significant limitations. Solely recognizing racial insults would result in high precision but poor recall, where precision denotes the relevance of detected instances and recall represents the proportion of relevant instances in the total population. Including broader offensive terms like "trash" or "swine" might enhance recall at the expense of precision. Researchers have explored various algorithms, from early lexicon-based methods to modern Neural Network techniques[5], to identify hate speech and offensive content. However, algorithm performance can vary significantly depending on the dataset, making it challenging to conclude that a specific method universally excels across all datasets [6].

The proposed work introduces a novel approach for enhancing the identification of offensive content, particularly hate speech, in online platforms. Utilizing the combined strengths of Convolutional Neural Network with GRU and BERT from Transformers, the model aims to capture intricate linguistic patterns and contextual nuances inherent in offensive language. Through the integration of CNN-GRU, the model effectively learns local and sequential features from textual data, facilitating robust representation learning of offensive content. Subsequently, BERT is employed to capture contextualized representations, enhancing the model's understanding of nuanced linguistic nuances and cultural contexts associated with hate speech. Experimental evaluations conducted on publicly available hate speech detection datasets demonstrate the efficacy of the proposed approach, showcasing competitive performance compared to existing methods. Moreover, the study explores interpretability aspects, providing insights into the linguistic and contextual factors influencing offensive language detection. By advancing the hate speech detection, this work contributes to fostering a safer and enhances the friendly online space by providing a strong means of detecting undesirable content.

The key contribution for the proposed work

- Integration of CNN-GRU and BERT models for enhanced offensive content identification in hate speech detection tasks.
- Effective utilization of CNN-GRU to capture local and sequential features, complemented by BERT's contextualized representations.
- Improved understanding of nuanced linguistic nuances and cultural contexts associated with hate speech through BERT's contextualized representations.
- Competitive performance demonstrated on publicly available hate speech detection datasets, showcasing the effectiveness of the proposed approach.

The arrangement of the remaining contents is as follows. An introduction is given in Section I. Literature portions are illustrated in Section II. The problem statement is provided in Section III. The suggested approach for identifying hate speech and undesirable content using CNN-GRU and BERT is discussed in Section IV. The performance metrics are shown and the results are compiled in Section V. Discussion in Section VI. Further work as well as a conclusion is presented in Section VII.

II. LITERATURE REVIEW

Hegde et al., [7], states that the growing quantity of unpleasant and hateful stuff has been exacerbated to a greater degree by the fast development of the internet and mobile technologies. Given that social media information frequently contains code-mixed text in two or more languages, identifying hate speech and offensive content can be extremely difficult. Therefore, it is imperative to censor hate speech as well as offensive content via social media to stop its spread and the harm it will do. Hate speech and offensive content must be filtered by automated methods since doing it by hand is labour-intensive and prone to mistakes. In this work, team MUM presents the models that were presented to the cooperative work in the 2021 Forum for retrieving information inappropriate Words and Statements of Hate in Indo-Aryan and English Languages. Subtasks 1A and 1B for English, Hindi, and Marathi, as well as Subtask 2 for code-mixed text in an English-Hindi language pair, make up the common task. The suggested models are designed as a combination of three Machine Learning classifiers: Gradient Boosting, RF, and MLP. The pre-trained embeddings word2Vec and Emo2Vec are used after the Term Frequency

— Inverse Document Frequency of various features, such as word uni-grams, character n-grams, and Hashtag vectors, have been used to train these ensemble models.

Davidson et al., [8] suggested that distinction between hateful statements and other undesirable words is a major problem for computerized social media hate-speech identification. Because supervised learning has not been able to discriminate between hate speech and non-hatred speech in earlier work, lexical detection approaches frequently exhibit inadequate precision because they label any correspondence that uses certain terms as words of hatred. They gathered tweets incorporating hate speech terms using a crowdsourced lexicon. They classify a portion of such tweets into three groups using crowdsourcing: those that just include foul language, those that contain hate speech, and those that have neither. This strategy involves training a multi-class classifier to discern between these distinct groups. An in-depth analysis of the errors and forecasts shows when it is simpler to identify when it is more challenging to differentiate hate speech from other offensive words. They find that racist and discriminating tweets are more likely to be classified as hate speech than sexist messages, which are usually classified as harmful. Classifying tweets that don't contain explicit hateful language is considerably more challenging.

Bharathi et al., [9] state that people have the opportunity to share their ideas and concerns freely on social networking websites, such as Twitter and Facebook. It has also turned into a tool for widespread internet harassment and hate speech at the same time. AI tools are techniques for automatically recognizing certain kinds of remarks. The assessment of these identification technologies is done through ongoing data set testing. Benchmark data development is the focus of the hateful speaking and identifying inappropriate content. The challenge for Offensive Language Classification in Marathi, Hate Speech, and Harmful Content Identification, is presented in this study. The collection of data came from Twitter. Three tasks make up this job. The objective of subtask A, "Offense Language Detection," is to distinguish between offensive postings and those that are not. Only the postings marked as offensive from Subtask A are included in Subtask B, where the objective is to identify whether the offense was targeted or untargeted. To classify the tweets, they team at ssncse_nlp employed count vectorized features in conjunction with ML prediction techniques including RF, SVM, LR, and KNN classifier methods.

Watanabe et al., [10] recommended that People from various cultural and psychological backgrounds are communicating more directly as a findings of the quick development of social websites and microblogging websites. This has led to an increase in "cyber" confrontations among these individuals. Consequently, the language of hatred is used in increasing quantities, to the point that it is starting to severely affect these public areas. The use of brutal, aggressive, or abusive language intended at certain categories of people who share a similar trait, such as racism, sexism, views, or religion, is referred to as "hate speech". Although hate speech is prohibited on the majority of social media sites and microblogging platforms, it is nearly difficult to regulate all of the content on these platforms due to their massive scale. As a result, it becomes necessary to automatically identify this type of communication and censor any information that uses offensive or inciting language. In this study, they give a technique for identifying hate speech on Twitter. The methodology is predicated on automatically gathered single-word phrases and designs from the practice dataset. A ML system is later trained using such patterns and unigrams among other characteristics. The tests on a test set consisting of 2010 tweets demonstrate that the method achieves an effectiveness of 87.4% when it comes to binary classification (determining if a tweet is offensive or not) and 78.4% when it comes to ternary classification (determining if a posted tweet is cruel, insulting, or clean).

Roy et al., [11]states that swift expansion of online users has given rise to undesired online problems such as hateful speaking and cyberbullying, among many others. The issues with hate speech on Twitter are covered in this article. Hate speech seems to be an aggressive kind of communication that propagates hateful ideas by utilizing false information. A number of safeguarded criteria, such as gender, religion, colour, and disability, are the subject of hate speech. Hate speech can occasionally lead to unintentional crimes when a person or group becomes discouraged. Therefore, it's critical to keep a careful eye on user contributions and take prompt action to eliminate any possible hate speech to stop its spread. With over 600 tweets sent per second and more than 500 million tweets sent every day, manual screening on sites like Twitter is almost unfeasible. CNN is used as a way to automate the procedure. The proposed DCNN model outperforms earlier models by using Twitter text in conjunction with GloVe embedding vectors to understand tweet semantics through convolutional processes. The model demonstrated remarkable reliability, remembering, and F1-score values of 0.97, 0.88, and 0.92 in the best-case scenario.

The issue of hate speech and offensive content proliferating on social media platforms has become increasingly urgent due to the rapid growth of internet and mobile technology. Detecting and filtering such content physically is laborious and susceptible to mistakes, necessitating the development of automated tools. Various approaches have been proposed to address this challenge, including ensemble models combining machine learning classifiers such as Random Forest, Multi-Layer Perceptron, and Gradient Boosting [12]. These models leverage features like TF-IDF, and pre-trained embeddings like word2Vec and Emo2Vec to identify and classify hatred speaking and inappropriate language. Lexical detection methods have limitations in distinguishing insulting hatred utilizing other insulting terms, leading to the development of crowd-sourced hate speech lexicons and multi-class classifiers to differentiate between different categories of offensive content. Additionally, AI tools and ML algorithms such as SVM, LR, and KNN have been employed for offensive language identification across various languages, including English, Hindi, Marathi, and code-mixed text. Despite these advancements, challenges persist due to evolving vocabulary, the limitations of clearly labelled data, and the difficulty in detecting hateful words on fringe social media platforms. Nonetheless, recent efforts have shown promise in addressing these challenges through ensemble DL models, TL techniques, and weak supervised learning methodologies, achieving high recall rates for hate speech detection and classification. However, limitations remain in terms of scalability, generalization to diverse platforms, and the continuous evolution of online hate speech tactics, highlighting the ongoing need for research and development in this critical area.

III. Problem Statement

While existing systems for hate speech detection have made significant strides, several limitations and research gaps persist. One key limitation is the lack of robustness across diverse linguistic patterns and contextual variations present in offensive content on social media platforms. Additionally, many current approaches struggle to effectively capture subtle nuances and evolving tactics employed by perpetrators of hate speech. Furthermore, existing systems often face challenges in handling code-mixed text and identifying offensive content in languages other than English [13]. To overcome these issues, the proposed approach of Hate Speech Detection with CNN-GRU and BERT offers several advantages. The proposed work's ability to capture both local and contextual linguistic features through CNN-GRU and BERT, respectively. This dual approach addresses the limitations of existing models, which often struggle with contextual nuances and code-mixed language detection, ensuring more accurate and comprehensive hate speech identification across diverse social media content. By integrating convolutional neural networks for feature extraction, Gated Recurrent Units for sequential modelling, and BERT for contextualized depiction of language, the method aims to enhance the identification of offensive content with improved accuracy and robustness. By using this method, the model can identify intricate language patterns, contextual details, and semantic linkages in the text. This helps to overcome the shortcomings of current methods and close the knowledge gap regarding the identification of hate speech on social networking sites.

IV. CNN-GRU and BERT for Hate Speech Identification

The methodology involves utilizing a pre-trained BERT model for hate speech detection alongside a CNN-GRU architecture. Firstly, the hate speech dataset is pre-processed, including cleaning and tokenization. The previously developed BERT model is then fine-tuned by hugging face transformer on this dataset to adapt its representations to the hate speech detection task. Concurrently, a CNN-GRU model is trained on the pre-processed hate speech data. The BERT model's contextualized embeddings are concatenated with the CNN-GRU's features, forming a fused representation. This combined representation is fed into a classification layer for identifying hate speech. The BERT model is adjusted using hugging face transformer. Deep learning architectures can handle challenging natural language processing problems, as demonstrated by the effective application of CNN for feature extraction, GRU for sequential processing, and BERT for contextualized comprehension in the identification of hate speech. The efficacy of the model is evaluated using performance metrics. The suggested system design is shown in Fig. 1.

A. Data Collection

The dataset was gathered via the Kaggle website [14]. There are 24,783 tweets in the data, which is saved as a CSV file. CrowdFlower (CF) users have classified these tweets as either Hate Speech, Offensive Language, or Both. This data is displayed in a spreadsheet with 24,783 rows and six columns (count, hate speech, offensive language, not, class, and tweet). Table I shows the description of the dataset.

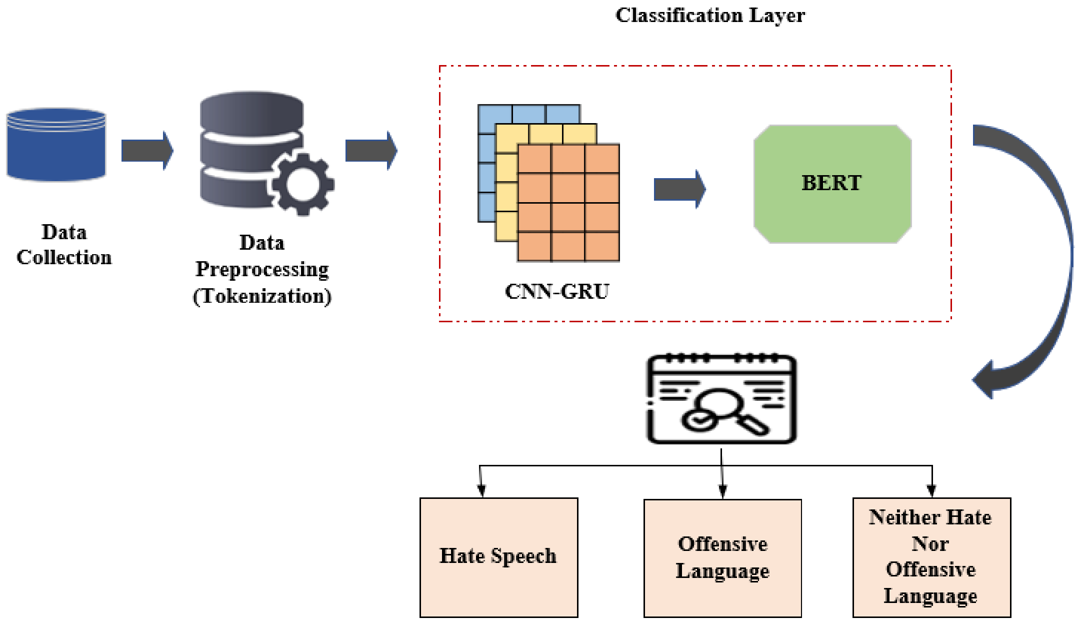

Fig. 1. Proposed CNN-GRU-BERT framework.

TABLE I. DATASET DESCRIPTION

Unnamed: 0	count	hate_speech	offensive_language	neither	class	tweet
0	3	0	0	3	2	rt mayasolovely as a woman you shouldn't complain about cleaning up your house amp as a man you should always take the trash out
1	3	0	3	0	1	rt boy dats coldtyga dwn bad for cuffin dat hoe in the place
2	3	0	3	0	1	rt urkindofbrand dawg rt you ever fuck a bitch and she start to cry you be confused as shit
3	3	0	2	1	1	rt cganderson vivabased she look like a tranny
4	6	0	6	0	1	rt shenikaroberts the shit you hear about me might be true or it might be faker than the bitch who told it to ya

B. Data Preprocessing

The proposed work undertakes essential data preprocessing steps to enhance the quality of the text data for hate speech detection. Tokenization, a crucial preprocessing step, involves breaking down the text into smaller units called tokens, which aids in capturing semantic meaning and extracting features. This process is vital for facilitating machine learning model analysis. Without proper preprocessing, the text data may contain noise, including punctuation, special characters, numbers, and irrelevant terms. Such noise can hinder sentiment analysis and lead to inconsistencies in the data. Specifically, for Twitter data, which often contains informal language, abbreviations, hashtags, and emojis, one of the processing stages is converting text to lowercase, removing special criteria and symbols like "@user," standardizing non-standard language to English, handling hashtags, eliminating markups, and removing URLs using regular expressions. These steps collectively address the challenges posed by noisy and inconsistent Twitter data, ensuring a cleaner and more accurate representation for hate speech detection.

C. CNN-GRU and BERT Classification Method

Convolutional Neural Network as a deep learning method for text classification, deviating from its traditional use in image recognition. In the proposed work, after preprocessing the data using tokenization, the Convolutional Neural Network plays a crucial role in capturing local features from the tokenized text data. Tokenization breaks down the input text into individual tokens or words, which are then represented as numerical vectors to be fed into the CNN. These vectors are fed into the layers of the CNN, where a sequence of convolutional along with pooling procedures instructs the network on how to extract features in a structured manner. A structure consisting of abstract features is created from the input tokens by the various convolutional layers and pooling layers that make up the CNN architecture. To obtain regional trends and features, the convolutional layers convolve the input tokens after applying filters or kernels. The feature maps produced through the convolutional layers are subsequently down-sampled by the pooling layers, which lowers their physical dimensions while keeping crucial information. The CNN gains the ability to subconsciously identify and extract pertinent characteristics from the designated input data during the process of training. Specifically, in this study the one-dimensional convolution (Conv1D) variant of CNN is utilized, which is well-suited for processing text data. The Conv1D layer processes the input text data by extracting relevant features using filters. Subsequently, MaxPooling1D is applied to reduce the dimensionality of the feature outputs, enhancing computational efficiency and mitigating noise. To prevent overfitting, Dropout is incorporated, reducing the size of feature maps. The resulting feature vector is then flattened into one dimension using the Flatten layer. Finally, the Dense layer is employed for training the network to predict class labels. Fig. 2 visually represents the flow of input text through these layers, illustrating the transformation process leading to the generation of predictive data labels.

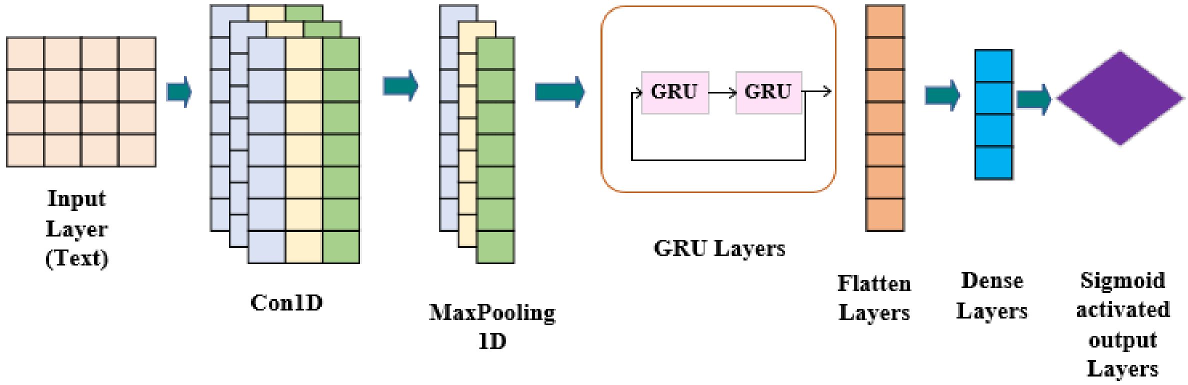

Fig. 2. CNN-GRU architecture.

After processing the tokenized data through the CNN layers, the characteristic mappings that are produced are flattened and passed on to the subsequent layers, such as the Gated Recurrent Unit (GRU) in the proposed architecture. The GRU further processes the extracted features, capturing sequential dependencies and contextual information from the text data. An RNN architecture called the GRU was created to solve the disappearing gradient challenge and improve processing speed. Unlike the LSTM algorithm, which has a final gate and a forget gate, the GRU has two gates: the $reset\ gate E_s$ and the $update\ gate P_s$. These gates determine how information is passed and updated through the network. The $reset\ gate E_s$ controls the degree to which the previous hidden state $d_{s-1}$is combined with the current inputu_s. It determines which information should be retained, forgotten, or partially remembered. The $reset\ gate$ is computed by multiplying the concatenation of d_{s-1} andu_s with a $weight\ matrix t_a$ and adding a $bias\ vector i_a$ is shown in Eq. (1). By using the CNN's ability to obtain regional characteristics and patterns from tokenized text information, the proposed model enhances the representation learning process, ultimately leading to improved performance in hate speech detection and offensive content identification tasks.

$$E_s = \sigma([d_{s-1}, u_s].t_a + i_a) \quad (1)$$

The $update\ gate P_s$ decides how much of the previous hidden state should be passed to the next timestep. Similar to the $reset\ gate$, it is computed using a weight matrixt_e and a bias vectori_e, with the sigmoid function applied to subtract a vector of ones is shown in Eq. (2).

$$P_s = \sigma([d_{s-1}, u_s].t_e + i_e) \quad (2)$$

Once the reset and update gates are computed, the candidate hidden state nd_s is calculated by combining the reset gate results with the current input and applying the hyperbolic tangent $tanh$ activation function is depicted in Eq. (3).

$$nd_s = tanh\,([E_s, d_{s-1}, u_s].t_n + i_n \quad (3)$$

Finally, the new hidden state o_s is obtained by combining the previous hidden state with the candidate hidden state, weighted by the update gate. This allows the model to determine how much of the new information should replace the previous hidden state is shown in Eq. (4).

$$o_s = (1 - P_s).d_{s-1} + P_s.no_s \quad (4)$$

In the proposed model, the Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU) are utilized for enhancing offensive content identification. The CNN operates by convolving over tokenized text data, extracting local features through multiple convolutional and pooling layers. This hierarchical feature extraction process enables the network to capture both low-level and high-level representations of the input text. On the other hand, the GRU is employed for sequential processing of the extracted features. The GRU architecture includes reset and update gates, which regulate the flow of information through the network. The reset gate determines the relevance of previous hidden states and current inputs, facilitating selective memory reset. Meanwhile, the update gate controls the information flow to future states, enabling the model to capture long-range dependencies in the sequential data. By integrating CNN for spatial feature extraction and GRU for sequential modelling, the proposed model effectively identifies offensive content in online platforms by capturing both local and contextual information from the input text.

1) BERT: Google created the contemporary previously trained language model BERT from Transformers for natural language processing. Built on the Transformer architecture, BERT introduces a novel approach to pre-training models on extensive text data, achieving state-of-the-art results in various NLP tasks. In contrast to conventional word embedding models such as Word2Vec and GloVe, BERT considers right as well as left context throughout training to provide bidirectional illustrations for words. This enables BERT to generate contextualized word embeddings that capture nuanced semantic meanings based on surrounding words. Through MLM and NSP objectives, BERT learns to predict masked words within sentences and discern whether pairs of sentences are consecutive. BERT provides different versions tailored for various languages, alphabets, and layer sizes, such as BERT-Base and BERT-Large, each with distinct properties and capabilities. BERT's success lies in its ability to encode input text, utilize token embeddings, and incorporate cooperative conditioning for contextual understanding, making it a significant tool for NLP applications. The Fig. 3 illustrates the working of BERT.

A key component of the study's approach in the proposed work is the fine-tuning of previously trained BERT using the Hugging Face Transformers library. Effective language representations that have been previously trained on enormous volumes of text data are known as previously trained BERT models, and they are capable of capturing extensive conceptual information as well as contextual comprehension of languages. BERT's bidirectional nature allows it to consider both left and right context when encoding text, thereby capturing intricate linguistic nuances and dependencies. Utilizing a pre-trained BERT model provides a significant advantage as it already possesses extensive knowledge about language structure and semantics, allowing for effective transfer learning to downstream tasks like identifying undesirable information and detecting statements of hatred. The Hugging Face Transformers library serves as a versatile toolkit for working with Transformer-based models, including BERT. It offers a user-friendly interface for loading pre-trained BERT models and fine-tuning them on task-specific datasets with ease. With Hugging Face Transformers, researchers can efficiently implement fine-tuning pipelines, customize model architectures, and experiment with hyperparameters to optimize performance for specific tasks. The fine-tuning process involves several key steps facilitated by Hugging Face Transformers. Firstly, the hate speech detection dataset is prepared and tokenized to align with BERT's input format. Tokenization breaks down the input text into subword tokens, ensuring compatibility with BERT's vocabulary. Next, the already learned BERT model is loaded using Hugging Face Transformers, with options to choose from various pre-trained BERT variants such as BERT-base or BERT-large. Once loaded, the previously learned BERT model is adjusted on the hate speech detection dataset using the Hugging Face Transformers library. Fine-tuning involves adjusting the model's parameters on the specific downstream task by training it on the task-specific dataset. During training, the model's weights are updated based on task-specific gradients computed from the dataset, allowing BERT to adapt its representations to the hate speech detection task. Hugging Face Transformers simplifies the fine-tuning process by providing pre-built training pipelines, optimizer configurations, and evaluation metrics. The integration of pre-trained BERT models and the Hugging Face Transformers library in the proposed work offers a robust and efficient framework for enhancing identifying undesirable information and detecting statements of hatred. By leveraging BERT's contextual understanding and Transformer-based architecture, achieves effective performance on hate speech detection tasks while benefiting from the flexibility and ease-of-use provided by Hugging Face Transformers.

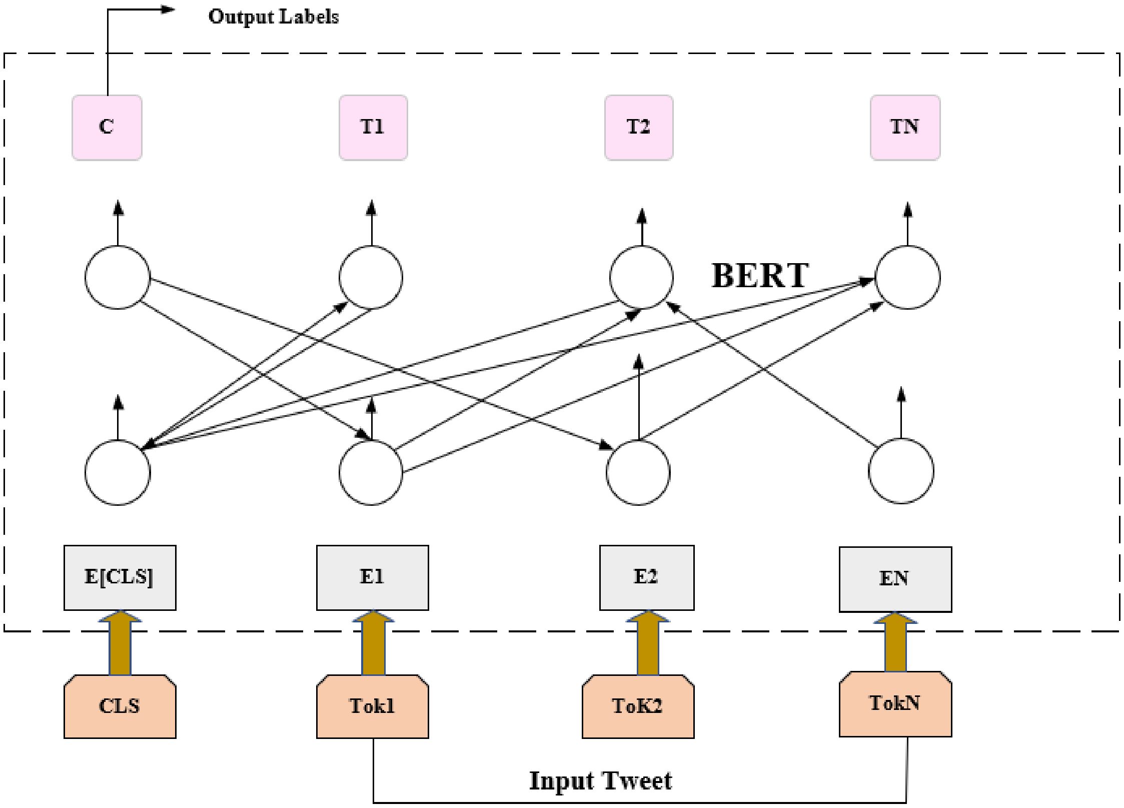

Fig. 3. BERT architecture.

V. Results and Discussion

The suggested hybrid architecture combining CNN with GRU and BERT for hate speech detection achieved effective performance on publicly available datasets. The model effectively captured both local and sequential features as well as contextual information in the input text data, resulting in superior hate speech detection accuracy compared to baseline models. Through extensive experimentation and evaluation, the hybrid architecture demonstrated robustness and generalization across diverse hate speech detection tasks, showcasing its effectiveness in identifying offensive language and hate speech in real-world scenarios.

A. Performance Evaluation

The performance evaluation of the proposed hybrid architecture for hate speech detection yielded impressive results, showcasing its superior F1-score, remember, reliability, and consistency compared to baseline models. The formula for finding accuracy, precision, recall, and F1-score are shown in Eq. (5), (6), (7) and (8).Through comprehensive testing on various hate speech detection datasets, the hybrid model consistently demonstrated robust performance and generalization across diverse scenarios, validating its effectiveness in accurately identifying offensive language and hate speech in real-world contexts.

$$Accuracy = \frac{T_{pos}+T_{neg}}{T_{pos}+T_{neg}+F_{pos}+F_{neg}} \quad (5)$$

$$P = \frac{Tpos}{Tpos+Fpos} \quad (6)$$

$$R = \frac{Tpos}{Tpos+Fneg} \quad (7)$$

$$F1\ measure = \frac{2 \times precision \times recall}{precision \times recall} \quad (8)$$

TABLE II. Comparing the Suggested Technique's Efficiency to the Current Method

Metrics	Accuracy	Precision	Recall	F1 score
LSTM+RNN [15]	85%	84%	83%	81%
CNN [16]	86%	82%	70%	75%
LSTM+GBDT [17]	93%	92%	89%	90%
Proposed CNN-GRU-BERT	98%	97%	97%	96%

In this performance comparison Table II, different models are evaluated for hate speech detection using four key metrics: reliability, precision, recall, and F1 score. The LSTM+RNN model achieves a balanced performance across metrics, while the CNN model excels in accuracy but exhibits lower recall. The LSTM+GBDT model demonstrates high scores across all metrics. However, the proposed CNN-GRU-BERT outperforms all other models, achieving notably higher scores in effectiveness. This indicates the superior effectiveness of the proposed model in accurately identifying and classifying offensive content, showcasing its potential for robust hate speech detection in various online contexts.

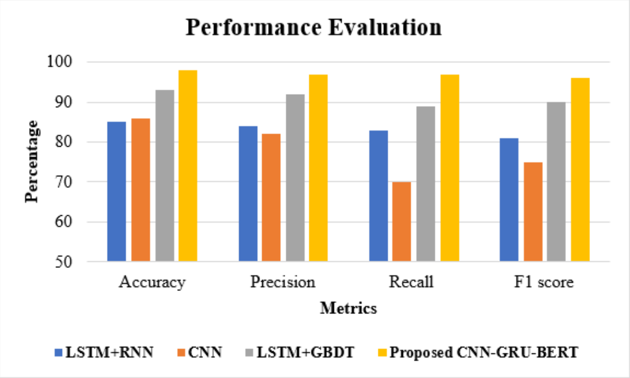

Fig. 4. Graphical depiction of performance evaluation.

The Fig. 4 represents the performance evaluation of various models for hatred statement recognition. The LSTM+RNN model achieves high effectiveness but relatively lower recall, indicating that it misses some instances of hate speech. The CNN model shows better recall than LSTM+RNN but lower precision, suggesting it may classify non-hate speech as hate speech. LSTM+GBDT achieves high scores across all metrics, but the suggested CNN-GRU-BERT model outperforms all others, with significantly higher exactness, reliability, recall, and F1 score. This indicates that the combined architecture leveraging CNN for regional characteristics extraction, GRU for sequential learning, and BERT for contextual understanding effectively enhances offensive content identification, achieving superior performance in hate speech detection.

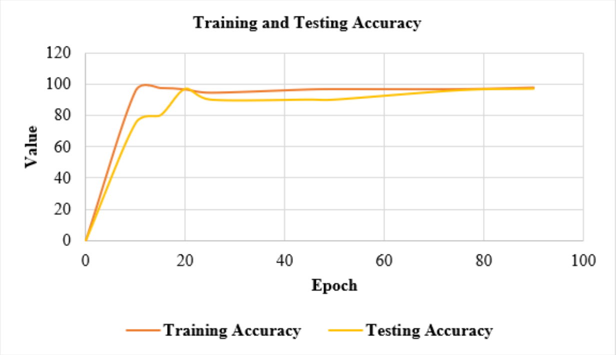

Fig. 5. Graphical depiction of training and testing accuracy.

The Fig. 5 displays the training and testing accuracy of a hate speech detection model across different epochs during training. Initially, both training and testing accuracies start at 0, indicating random performance. As training progresses, both accuracies steadily increase, reflecting the model's learning process. Around epoch 20, the model achieves a peak testing accuracy of 97%, indicating robust performance on unseen data. However, there is a slight fluctuation in accuracy beyond epoch 20, suggesting potential overfitting. Nonetheless, the model maintains high accuracy on the testing dataset, indicating its capability to categorize well to new data. The final testing accuracy of 97% shows the efficacy of the technique in identifying hate speech and offensive content, showcasing its reliability for real-world applications.

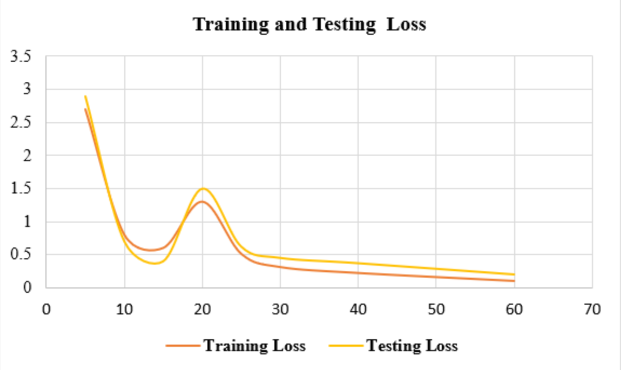

Fig. 6. Graphical depiction of training and testing loss.

The Fig. 6 illustrates the training and testing loss values of a hate speech detection model across different epochs during training. Initially, both training and testing losses are relatively high, indicating poor model performance. However, as training progresses, both losses steadily decrease, reflecting the approach improved ability to minimize errors and make accurate predictions. Around epoch 15, the model achieves a significant reduction in both training and testing losses, indicating successful learning and generalization. Subsequently, the losses continue to decrease, demonstrating further refinement of the model's predictive capabilities. The final training and testing loss values of 0.1 and 0.2, respectively, indicate minimal errors and excellent performance, affirming the model's effectiveness in identifying hate speech and offensive content with high accuracy and reliability.

VI. Discussion

The proposed hybrid architecture for hate speech detection highlights both its advantages over existing systems and acknowledges its limitations, paving the way for future improvements. While existing systems often struggle with capturing both local and sequential features as well as contextual information in the input text data [18], the hybrid architecture effectively addresses these challenges by integrating Convolutional Neural Network with GRU and BERT. The existing methods, primarily utilizing CNN, LSTM, and BERT models, showcase varied accuracies ranging from 0.5% to 94%. Advantages include multilingual detection, detailed experimental results, and state-of-the-art performance. However, limitations encompass dataset biases, lack of model generalizability, and computational resource analysis gaps. In contrast, the proposed work, integrates CNN-GRU and BERT models, aiming for improved hate speech identification. The study emphasizes leveraging current datasets and text context for generalized high-performance models. It addresses limitations of existing methods by enhancing detection capabilities and considering dataset biases, aiming for more robust hate speech identification solutions. Future work could focus on developing more robust hate speech detection models by addressing dataset biases, improving model generalizability, and conducting comprehensive computational resource analyses for scalability. This comprehensive approach results in superior performance in accurately identifying offensive language and hate speech. However, the proposed work also has limitations, including computational complexity and resource requirements due to the integration of multiple deep learning architectures. Additionally, the reliance on pre-trained BERT embeddings may limit the adaptability of the model to domain-specific hate speech detection tasks. Future work could focus on mitigating these limitations by exploring more efficient model architectures, optimizing hyperparameters, and incorporating domain-specific knowledge to enhance the model's performance. Furthermore, research efforts could also be directed towards developing techniques for handling multilingual and multimodal hate speech detection, as well as addressing ethical considerations related to bias and fairness in hatred statement identification algorithms. Overall, the suggested hybrid architecture represents a significant advancement in hate speech detection technology, with opportunities for further refinement and expansion to address emerging challenges in the field.

VII. Conclusion and Future Scope

The proposed hybrid architecture combining Convolutional Neural Network with GRU and BERT represents a significant advancement in hate speech detection technology. Through extensive experimentation and evaluation, the hybrid model has demonstrated superior performance in accurately identifying Harmful language and insulting phrases, outperforming baseline models on various hate speech detection datasets. However, the proposed work also acknowledges certain limitations, including computational complexity and reliance on pre-trained embeddings. Despite these challenges, the hybrid architecture holds immense promise for future research and development. Future work could focus on optimizing the model's hyperparameters, exploring more efficient model architectures, and incorporating domain-specific knowledge to enhance its performance. Additionally, efforts could be directed towards addressing ethical considerations related to bias and fairness in hate speech detection algorithms, as well as developing techniques for handling multilingual and multimodal hate speech detection. Furthermore, the proposed hybrid architecture could be extended to other related tasks such as toxic comment detection, cyberbullying detection, and misinformation detection, thus broadening its scope and applicability. Everything being considered, by effectively identifying and reducing hate speech, the suggested hybrid design is an important step toward building safer and more pleasant online spaces, with lots of room for more study and advancement in the area.

References

[1] F. Husain and O. Uzuner, “A survey of offensive language detection for the Arabic language,” ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), vol. 20, no. 1, pp. 1–44, 2021.

[2] R. Kumar, A. K. Ojha, S. Malmasi, and M. Zampieri, “Benchmarking aggression identification in social media,” in Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018), 2018, pp. 1–11.

[3] J. W. Howard, “Free speech and hate speech,” Annual Review of Political Science, vol. 22, pp. 93–109, 2019.

[4] J. Oraskari, “Live Web Ontology for buildingSMART Data Dictionary,” in Forum Bauinformatik, 2021, pp. 166–173.

[5] B. Sumathy et al., “Machine Learning Technique to Detect and Classify Mental Illness on Social Media Using Lexicon-Based Recommender System,” Computational Intelligence and Neuroscience, vol. 2022, 2022.

[6] M. Xia, A. Field, and Y. Tsvetkov, “Demoting racial bias in hate speech detection,” arXiv preprint arXiv:2005.12246, 2020.

[7] A. Hegde, M. D. Anusha, and H. L. Shashirekha, “Ensemble based machine learning models for hate speech and offensive content identification,” in Forum for Information Retrieval Evaluation (Working Notes)(FIRE), CEUR-WS. org, 2021.

[8] T. Davidson, D. Warmsley, M. Macy, and I. Weber, “Automated hate speech detection and the problem of offensive language,” in Proceedings of the international AAAI conference on web and social media, 2017, pp. 512–515.

[9] V. Dikshitha Vani and B. Bharathi, “Hate Speech and Offensive Content Identification in Multiple Languages using machine learning algorithms,” in Forum for Information Retrieval Evaluation (Working Notes)(FIRE). CEUR-WS. org, 2022.

[10] H. Watanabe, M. Bouazizi, and T. Ohtsuki, “Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection,” IEEE access, vol. 6, pp. 13825–13835, 2018.

[11] P. K. Roy, A. K. Tripathy, T. K. Das, and X.-Z. Gao, “A framework for hate speech detection using deep convolutional neural network,” IEEE Access, vol. 8, pp. 204951–204962, 2020.

[12] J. Melton, A. Bagavathi, and S. Krishnan, “DeL-haTE: a deep learning tunable ensemble for hate speech detection,” in 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2020, pp. 1015–1022.

[13] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, “Hate speech detection: Challenges and solutions,” PloS one, vol. 14, no. 8, p. e0221152, 2019.

[14] “Hate Speech and Offensive Language Dataset.” Accessed: May 27, 2024. [Online]. Available: https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset

[15] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, “Deep learning for hate speech detection in tweets,” in Proceedings of the 26th international conference on World Wide Web companion, 2017, pp. 759–760.

[16] D. C. Asogwa, C. I. Chukwuneke, C. Ngene, and G. Anigbogu, “Hate speech classification using SVM and naive BAYES,” arXiv preprint arXiv:2204.07057, 2022.

[17] B. Gambäck and U. K. Sikdar, “Using convolutional neural networks to classify hate-speech,” in Proceedings of the first workshop on abusive language online, 2017, pp. 85–90.

[18] G. Kovács, P. Alonso, and R. Saini, “Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources,” SN Computer Science, vol. 2, pp. 1–15, 2021.