The Effectiveness of Different Deep Learning Models in Detecting Hate Speech on Social Media

***Abstract*—Recent changes in social media have made it more difficult to control the propagation of hate speech. Deep learning models for automated hate speech identification have been proposed as a possible answer to this problem. In this study, we evaluate how well several deep learning models can identify hate speech on social media. On a dataset of social media posts that include both hate speech and non-hate speech content, we run experiments. We contrast the results of various models, including attention-based models, long short-term memory (LSTM), and convolutional neural networks (CNNs). We also look at how other factors, such as the amount of training data used and the use of pre-trained word embeddings, affect how well these models perform. Our findings show that in terms of detecting hate speech, attention-based models outperform CNN and LSTM models. In conclusion, this work offers insights into the effectiveness of deep learning models for detecting hate speech and can help with the creation of more precise and effective models for controlling hate speech on social media.**

***Index Terms*—Convolutional neural networks (CNNs), Long short-term memory (LSTM), embeddings, BERT, CONV1D, GRU, Tokenization, Hate speech**

I. Introduction

With the rise of social media came new difficulties in controlling the spread of hate speech, which is a language that targets people or groups based on their particular characteristics. Serious repercussions like inciting aggression and promoting discrimination may result from this. Therefore, it is now essential for preserving a respectful and secure online environment to identify hate speech on social media.

Deep learning models have been suggested as an answer for the automatic detection of hate speech in order to handle this problem. These models are effective at identifying hate speech because they can extract significant features from text and learn from huge amounts of data.

Through experiments on a dataset of social media posts with both hate speech and non-hate speech material, this study assesses how well different deep learning models work at identifying hate speech on social media. Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and attention-based models are among the models that are compared while exploring the effects of various variables on the performance of these models, including the size of the training data and the use of pre-trained word embeddings.

This study's primary goal is to shed light on the effectiveness of deep learning models for identifying hate speech and serve as a roadmap for the creation of more precise and effective models for policing hate speech on social media. The findings of this research can assist social media platforms and policymakers in creating efficient plans for controlling hate speech and promoting an inclusive and safe online community.

II. Literature Review

The development of extensively annotated datasets and advanced classification techniques has facilitated the monitoring of users' sentiments and viewpoints on social media platforms, thereby simplifying the process. However, detecting hate speech on social media is difficult, especially when comparing prediction models over time and taking subject shifts in online discourse into consideration. This essay presents a case study on the "Contro l'odio" platform, which monitors instances of hate speech directed towards foreign individuals on the Italian Twitter platform. The research assesses the AlBERTo transformer-based neural network classifier's temporal robustness and evaluates various training approaches to enhance the classifier's performance over time. In this study by Florio et al., the authors investigate the durability over time of algorithms developed for the purpose of forecasting hate speech. Additionally, it evaluates the influence of the quantity of training data and the temporal scope on the precision of these predictions. [1]. The research contends that Alberto performs better with a sufficient time window while needing less annotated data than conventional classifiers and that supervised classification models have limitations on data affected by events. As topics and linguistic patterns in social media language change considerably over time, the study emphasized the need to track the robustness of hate speech detection systems.

The article by Zhou et al. addresses the problem of hate speech on social media and the demand for automated hate speech detection [2]. The authors of this study concentrate on the implementation of machine learning methodologies, specifically deep learning techniques such as ELMo, BERT, and CNN, for the purpose of categorising hate speech. The authors suggest using fusion methods to combine the outputs of various classifiers in order to improve overall classification performance, but each method has benefits and drawbacks of its own. The classifier fusion framework and the variables influencing the fusion outcomes are described in the paper. The authors' experiments indicate that the performance of hate speech detection is significantly improved through fusion processing. However, the writers advise that in future work, early cooperation before classification should receive more focus.

The paper by Das et al. implemented numerous baseline models, including m-BERT, XLM-Roberta, IndicBERT, and MuRIL, for the automatic identification of hate speech [3]. They continue to research interlingual transfer mechanisms, such as language transfer, joint training with language transfer, ELFI, etc., to improve classification performance. The XLM-Roberta model operates best when there are enough training cases. Models like MuRIL, however, succeed when the Romanized and Original Bengali are combined for joint training because they can take advantage of the semantic relationship between the two. The study has some limitations. The first drawback is that it ignores any external context that might be relevant for the task of detecting hate speech, such as the user's gender, posting history, or profile bio. Furthermore, although being extremely effective, the transformer-based models utilized in the study haven't been put to the test against adversarial scenarios.

Conventional and rule-based methods of machine learning limit the ability to identify hate speech. Deep learning techniques have produced effective results, but frequently only pay attention to semantic characteristics and ignore sentiment features. The study by Zhou et al. suggests a model for detecting hate speech based on sentiment knowledge sharing (SKS), and its performance is evaluated against a number of baselines, including SVM, LSTM, GRU, CNN-GRU, BiGRU-Capsule, Universal Encoder, BERT, and GPT [4]. The researchers discovered that hybrid neural networks outperform simple RNNs and that models based on neural networks outperform feature-based models. SKS surpasses all other models on both datasets, reaching an F1 value of nearly 90% on the DV dataset and a performance improvement of over 10% on the SE dataset when compared to the next best model. BERT performs better on the DV dataset. The authors come to the conclusion that SKS is superior to other models in terms of functionality, usability, and parameter count.

The study conducted by Shubhang et al. highlights the necessity of implementing automated detection mechanisms for hateful speech on social media platforms, owing to its escalated occurrence in recent times [5]. The statement underscores the challenge of detecting instances of hate speech in English language written materials and the deleterious effects it can have on both individuals and the wider community. It is suggested that a hybrid NLP (natural language processing) model, integrating convolutional and recurrent layers, be utilized for the identification of hate speech on social media platforms. The accuracy of the model, which utilizes the Bi-GRU-LSTM-CNN classifier, is 77.16%.

The article by Boishakhi et al. discusses on the detection of hate speech using a combination of textual and visual information [6]. The study used a dataset of tweets collected during the 2017 French presidential election, containing both textual and visual data. The authors extracted textual features using natural language processing techniques and visual features using pre-trained convolutional neural networks. Subsequently, the tweets were categorised into two distinct groups, namely hate speech and non-hate speech, through the utilisation of machine learning algorithms such as logistic regression, decision tree, and support vector machine. The findings demonstrated that utilizing a multi-modal strategy led to greater accuracy than relying just on textual or visual elements. According to the authors, the multi-modal technique may be helpful in identifying hate speech on websites that have both textual and visual content, like social media. The research emphasizes how crucial it is to take into account both textual and visual information when attempting to identify hate speech, and it contends that using several modalities might increase the efficacy of such efforts on online platforms.

This article by Ziqi et al. presents a comprehensive survey of the field of deep learning, encompassing its historical development and fundamental principles [7]. It cites a particular paper as a source of information. As per the authors' assertion, deep learning is a machine learning paradigm that employs deep neural networks. The authors deliberate on the primary implementations of deep learning, encompassing computer vision, natural language processing, and speech recognition. The authors proceed to expound upon the fundamental principles of deep learning, encompassing neural networks, Backpropagation, activation functions, Convolutional Neural Networks, recurrent neural networks, and generative models. The authors elucidate the operational mechanisms of each of these theoretical constructs and provide illustrative instances of their practical application. Additionally, the authors furnish a compilation of resources intended for individuals seeking to expand their knowledge on the subject of deep learning. In summary, the present study offers a succinct and enlightening exposition of the domain of deep learning and its fundamental principles, rendering it a valuable reference for individuals with an interest in this swiftly progressing realm of inquiry.

The paper by Kumar et al. proposes a deep learning-based model for Twitter hate speech detection [8]. The paper uses a benchmark dataset of approximately 25,000 annotated tweets and contrasts the F1 score and accuracy of the proposed model to those of a conventional machine learning classifier. The paper reports that the proposed model outperforms the baseline classifier and demonstrates the effectiveness of deep learning techniques for detecting hate speech.

Naidu et al. performed a study in which they compared various deep-learning models for the purpose of identifying hate speech on Twitter. The models that were evaluated included CNN, RNN, LSTM, and BiLSTM [9]. The assessment of the models is predicated on their precision, recall, precision, and F1 score, utilising a dataset comprising 16,000 annotated tweets. The study showcases the superior performance of BiLSTM in comparison to other models and posits that models based on deep learning can effectively identify instances of hate speech on social networking sites.

The research by Rottger et al. employs HATECHECK, a comprehensive suite of functional evaluations, to assess mod-

els designed for detecting hate speech [10]. The present study puts forth a series of metrics aimed at evaluating the efficacy of models in identifying and emphasising instances of hate speech. The significance of these models lies in their relevance to the task of online content moderation. The present study aims to draw attention to the inadequacies of assessing hate speech models through the utilisation of performance indicators derived from pre-existing datasets. The present body of research posits that biases and limitations inherent in datasets result in models that exhibit an over-reliance on keywords and a lack of generalizability. A collection of 29 functional assessments made up of the HATECHECK tool are used to assess the presence of both hateful and non-hateful contrasts. The present investigation aims to establish and validate the functionalities of a model through a series of tests. The implementation of a comprehensive annotation process is crucial in ensuring the reliability and precision of assessments. The present study evaluates the efficacy of HATECHECK, a tool designed to assess the performance of natural language processing models in detecting hate speech. The present study conducted an evaluation of performance biases and keyword sensitivity. The present study aims to introduce the HATECHECK tool as a means of advancing hate speech detection models.

The purpose of the study by Ali et al. is to provide an overview of the still-emerging field of research on the computerized recognition of racial microaggressions [11]. One of the paper's major contributions is a comprehensive review of the research on this developing topic from an emotional and scientific standpoint, which helps us comprehend the variables that affect the automatic identification of small-scale abuse. The study offers a solution for this problem by showcasing an automated system for identifying racial microaggressions. The suggested approach uses hand-crafted lexicons of somewhat racist terms and phrases that could be construed as small-scale assaults in addition to machine learning techniques to identify microaggressions. The findings of the research demonstrate that the suggested technique beats other options in regard to the F-measure and accuracy.

In order to detect insults in quick text conversations, the research study by Ombui et al. used a hierarchical framework of fundamental characteristics mapped into low-level characteristics [12]. In order to autonomously recognise hate speech produced during Kenya's 2012 and 2017 presidential campaigns, the research employed trained machine learning to build a system for classification using 48k tweets in particular. According to the initial results, the precision is 0.74, that's better than the reference value for the identical data set annotated by humans.

This study by Gao et al. detects online hate speech due to its prevalence. The authors say context is key for recognising subtle and innovative hate speech [13]. They present a Fox News user comment dataset with extensive context information, including screen names, thread comments, and the news story. Feature-based logistic regression and neural network hate speech identification are studied. Neural network models use text compositional interpretations, while logistic regression uses target comment and context features. Ensemble models are recommended by the authors. Context-aware logistic regression and neural networks outperform context-free models. Ensemble models beat strong baseline systems by 10% in F1-score. This research provides context-aware online hate speech detection algorithms.

III. Data Collection and Preprocessing

We collected the data from Kaggle. Except for ctweet dataset, datasets use binary notation. (it has 3 labels). The process of preprocessing for the textual data categories of "Food," "Sarcasm," and "Tweet" entails the elimination of punctuation marks, tags that function as names, and HTML components. Subsequently, the text is converted to a lowercase format. The datasets of Reddit, ctweet, and Hazardous necessitate preprocessing, thereby leaving the decision of how to proceed in your hands. The emotional perspective and data structure exhibit notable dissimilarities.

A. Data Collection

TABLE I
Dataset table

text	score
you will love these planters' quality that you would	1
these gummi bears caused my wife and I terrible gas even	0
these almonds are wonderful I eat a handful with an Atkins	1
I try to steer clear of HFCS and trans fats i bought	0
these drinks are literally the healthy non-soda	1

The present study employs a dataset that has been partitioned into multiple subsets, with each subset fulfilling a specific role in the analytical process. In order to accurately represent their binary nature, the binary datasets within the overall dataset are assigned numerical values of 0 and 1. The ctweet subset is noteworthy in that it introduces a fourth value to the existing values of 0, 1, and 2 in the 'Y' column, thereby expanding the range of emotional expression.

The dataset comprises three primary subsets, namely ctweet, stweet, and food. To classify sentiments into negative (0) or positive (1), these subsets undergo sentiment analysis that determines whether they are positive or negative. The category known as ctweet is inclusive of a classification for sentiments that are unbiased in nature, which is represented by the numerical designation of 2.

B. Data Preprocessing

The present section breaks down the preprocessing methodologies utilized to preprocess the data with the aim of identifying hate speech through a range of deep learning models. The preprocessing procedures involve several steps, including handling missing values, randomizing the dataset, dividing it into training and validation subsets, conducting tokenization, and applying embedding methodologies.

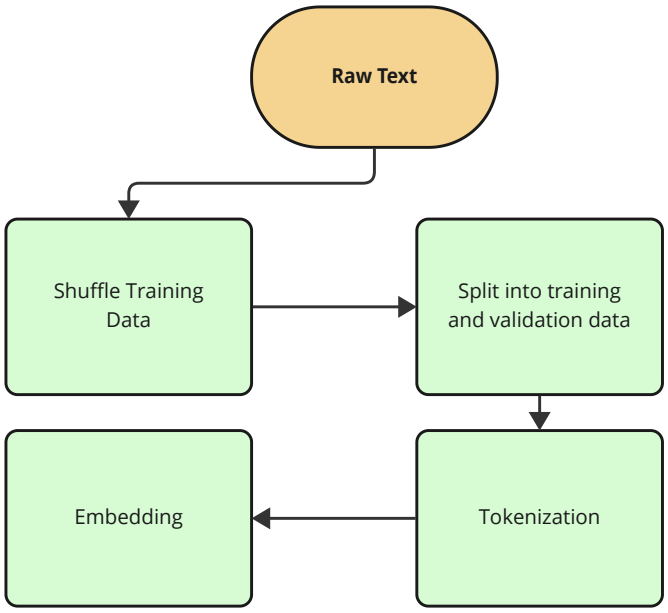

Fig. 1. Preprocessing techniques used

The collected data were first examined for any occurrences of missing values or null entries. By employing the Pandas library, null values were identified and removed, resulting in data completeness and integrity. Additionally, the dataset underwent a process of eliminating any inconsistencies or anomalies, such as erroneous assessments or events without corresponding textual evaluations.

To address the possibility of bias and promote the randomness of the data, a shuffling procedure was executed. Randomising the set of data was done with the aim of mitigating any potential influence of consecutive patterns on the success of the model throughout the training and evaluation phases.

Following this, the dataset was subjected to partitioning into discrete training and validation sets. The act of segregating data into discrete training and testing subsets is an essential phase in providing an accurate evaluation of the efficacy of deep learning models. The training set serves the purpose of enabling the training of models, while the validation set is employed to assess their generalization abilities and identify any possible occurrences of overfitting. After performing preprocessing on the textual data, the process of tokenization was carried out. This involved dividing the reviews into distinct tokens or lexical units. This methodology facilitates the conversion of textual information into numerical representations that are appropriate for integration into deep learning algorithms. Furthermore, the utilization of embedding techniques was employed to capture both the contextual information and connotative meaning of the text.

IV. Methodology

We tested four deep learning models Convolutional Neural Network (CONV1D), Bidirectional Encoder Representations from Transformers (BERT), Gated Recurrent Unit (GRU), and Long Short-Term Memory (LSTM) —to identify hate speech in text data. These models were selected because they performed well on Natural Language Processing (NLP) tests and were good at identifying the context of text input.

The dataset utilized in this research consisted of tweets that were categorized as either containing offensive language or not containing hate speech. Preprocessing of the dataset involved deleting all URLs, mentions, and special characters. The text was then changed to lowercase for similarity once the data had been tokenized.

With a ratio of 70:15:15, we divided the dataset into training, authentication, and test data. We used the validation set to tune hyperparameters and avoid overfitting while the models were trained using the training set. The efficiency of each model was assessed using the test dataset.

Using a global max-pooling layer, a layer that is dense with a sigmoid activation function, and a one-dimensional convolutional layer, the CONV1D model was built. In the study, the Adam optimizer was used to train the models, and the binary cross-entropy loss function was the loss function used. A 64-person batch size was used throughout the model's training process, which lasted for 10 iterations.

The BERT model utilized in our research was derived from the pre-existing BERT base uncased model of the Hugging Face Transformers library. The optimization of the model was performed through the utilization of an AdamW optimizer and a binary cross-entropy loss function. The training process of the model involved three iterations, each utilizing a batch size of 16. The models of Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) were endowed with an embedding layer, a recurrent layer, and a dense layer that employed a sigmoid activation function. The study utilized binary cross-entropy as the loss function and trained the models with the Adam optimizer. The training process involved 10 epochs and 64 batches.

Using measures for precision, recall, F1-score, and accuracy, we assessed the models' performance. In order to establish which model performed best for hate speech identification, we also compared the models.

Overall, using this technique, we were able to assess how well various deep learning models performed when detecting hate speech in text data.

A brief description of the models used is given below:

A. LSTM

Long Short-Term Memory (LSTM) is a type of artificial recurrent neural network that is used in deep learning, which is able to learn order dependency in problems including sequence prediction. The discipline of deep learning uses this kind of artificial neural network. This is because they have the ability to store information for extended periods of time, which

explains the phenomenon. The fundamental idea of an LSTM consists of the cell state and all of its multiple gates. The state of the cell acts as a channel via which information about relative relationships can go from the beginning to the end of the sequence chain. It is similar to the memory of the network. A variety of neural networks make up the gates, which are responsible for deciding whether or not a piece of information may be stored in the cell state. During their training, the gates are likely to acquire what information is critical to recall and what details may be ignored. The memory gate, input gate, and output gate are examples of additional gates. The Forget gate is responsible for deciding whether or not information from prior procedures should be kept for later evaluation. The input gate evaluates the data from the currently active phase to decide which pieces of information from that phase should be added. The output gate is the one that decides which hidden state comes next.

B. GRU

The GRU is a component of more recent generations of recurrent neural networks. It is analogous to an LSTM in a variety of respects and functions similarly. In the same way, as LSTM can, it can address the problem of the gradient disappearing in ordinary RNNs. This is because the GRU network cells make use of an adjustable gate method, which enables them to efficiently handle both historical data and the data that is now being received. The GRUs chose to transmit information using the concealed state as opposed to the cell state because it was more secure. It just has access to a resetting gate and an updating gate; these are the only two gates it has. The update gate performs operations that are analogous to those of the forget and input gates of an LSTM. It determines which data should be deleted and which new data should be included. The reset gate is another type of gate that is utilized in the process of determining how much-outdated data should be erased. GRUs are able to be taught more quickly than LSTMs due to the fact that they contain fewer tensor processes. Researchers often make use of both options in order to determine which of them is superior for the given use scenario.

C. BERT

BERT was provided with particular training on both Wikipedia and Google's Books Corpus. Because of this knowledge, BERT was able to acquire a comprehensive comprehension of our world as well as the English language. MLM makes bidirectional learning from text easier to do or more mandatory by covering one word in a phrase and requires BERT to use the words that come before and after the covered word in order to guess the word that is being covered up. Because it was trained on such a large text corpus, the architecture of BERT is able to better grasp the language, learn diversity in data patterns, and perform well across a variety of applications that require natural language processing. Due to the fact that it is bidirectional, BERT is able to collect information from both sides of the context of a token as it is being trained. Although it just adds a very thin layer to the core notion, BERT may be used to a broad variety of language issues. The procedure of classifying emotional analysis exhibits similarities to that of sentence classification, in which a classification layer is overlaid onto the transformer output for the CLS token.

D. CONV1D

Convolutional Neural Networks, often known as CNNs, have demonstrated impressive performance in a variety of computer vision applications and have demonstrated promise in natural language processing (NLP) tasks such as text classification. CONV1D is a kind of CNN that was developed specifically to manage one-dimensional sequences, such as text data. Its primary focus is on speed and accuracy.

The input text is first processed by a CONV1D model, where it is transformed into a sequence of word embeddings, and then it is passed through a number of convolutional layers. The process of extracting local features from the input sequence is accomplished by the convolutional layers, which utilize a sliding window mechanism. Subsequently, a nonlinear activation function, such as ReLU, is applied to generate a new set of features. The convolutional layers are credited with the aforementioned process. The conclusive classification outcome is produced by feeding the produced feature maps into a densely connected layer that is equipped with a sigmoid activation function. The current stage culminates the process of categorization. The feature maps generated serve as the basis for producing a feature vector of a predetermined length by means of the global pooling layer.

The capability of the CONV1D model to extract both local and global information from the input sequence is one of the model's many advantages. The global pooling layer may be able to grasp the overarching significance of the input sequence, while the convolutional layers may be able to pick up localised information such as grammatical structures and n-grams.

In the research study that we conducted on the identification of hate speech, we utilised a CONV1D model that consisted of one-dimensional convolutional layers, a global max-pooling layer, and a dense layer with a sigmoid activation function. All of these layers were interconnected. During the training process of the model, the binary cross-entropy was utilised as the loss function in conjunction with the Adam optimizer. A 64-person training batch was used throughout 10 training epochs to train the model. The performance of the model was evaluated based on a number of different metrics, including accuracy, recall, F1 score, and precision.

V. Results and Analysis

These various models: LSTM, GRU, CONV1D, and BERT were used for the sentiment analysis of the reviews of hate speech detection that had been gathered from social media. The analysis of 12 reviews yielded accuracy figures of 99.34%, 99.32%, 99.32%, and 99.04% respectively.

Table II measures up the sentiment analysis results produced by these models. A higher f1 score reflects a better balance among both recall and precision. Recall considers how well the model can identify all instances of positive sentiment, whereas precision measures whether the model can categorize positive sentiment.

The models had been found to be extremely accurate, most likely as a result of the investigation's perfect use of the data. The models were able to accurately identify the review sites' sentiments thanks to the appropriate and flawless use of data. The complexity of natural language and the subjective data of sentiment analysis, where different people may perceive the same sentiment differently, may also play a role in the high accuracy attained.

In total, 4 machine-learning models were utilized in this study. When it comes to reviews with poor grammar or reviewers who occasionally assign incorrect review ratings, we would want to determine just how impactful these models are. The models were effective in sentiment analysis.

TABLE II
COMPARISON OF 3 DIFFERENT MODELS

Metric	**LSTM**	**GRU**	**CONV1D**	**BERT**
Accuracy	99.34%	99.32%	99.32%	99.04%
F1 Score	0.99	0.99	0.99	0.98
Precision	0.99	0.99	0.99	0.98
Recall	0.99	0.99	0.99	0.98

LSTM exhibited the highest level of performance among the models. The models (LSTM, GRU, CONV1D, and BERT) were not even specifically trained to detect hate speech from social media sentiment analysis. Rather, these models had been initially trained on sizable datasets before becoming adjusted or tailored to suit this particular situation. Hugging face squad and Meta generated the pre-trained model LSTM, GRU, CONV1D, and BERT. To add, a high rate of accuracy in sentiment analysis could well lead to flawless use of data and reviews to the comparatively tiny dataset size and model boundaries. Let's assume the data has different stages of sentiment that are demanding for the model to understand (including sarcasm and irony). In certain cases, it could turn out tricky for such a model to define the sentiment, leading to less accuracy.

The findings of this study offer an overview of the sentiment of feedback on hate speech detection that was gathered from social media the results indicate higher accuracy. Achieving a bigger and more varied dataset may well be obliged to further increase the precision of sentiment analysis models for hate speech detection which were taken from social media. This might be done with the additional aid of different types of social media. Extra added preprocessing or extraction of features of the data, as well as the use of different models or techniques for sentiment analysis, may very well be assumed in further studies in addition to a larger dataset.

REFERENCES

[1] K. Florio, V. Basile, M. Polignano, P. Basile, and V. Patti, "Time of your hate: The challenge of time in hate speech detection on social media," *Applied Sciences*, vol. 10, no. 12, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/12/4180

[2] Y. Zhou, Y. Yang, H. Liu, X. Liu, and N. Savage, "Deep learning based fusion approach for hate speech detection," *IEEE Access*, vol. 8, pp. 128 923–128 929, 2020.

[3] M. Das, S. Banerjee, P. Saha, and A. Mukherjee, "Hate speech and offensive language detection in Bengali," in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online only: Association for Computational Linguistics, Nov. 2022, pp. 286–296. [Online]. Available: https://aclanthology.org/2022.aacl-main.23

[4] X. Zhou, Y. Yong, X. Fan, G. Ren, Y. Song, Y. Diao, L. Yang, and H. Lin, "Hate speech detection based on sentiment knowledge sharing," 01 2021, pp. 7158–7166.

[5] S. Shubhang, S. Kumar, U. Jindal, A. Kumar, and N. R. Roy, "Identification of hate speech and offensive content using bi-gru-lstm-cnn model," in *2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, 2023, pp. 536–541.

[6] F. T. Boishakhi, P. C. Shill, and M. G. R. Alam, "Multi-modal hate speech detection using machine learning," in *2021 IEEE International Conference on Big Data (Big Data)*, 2021, pp. 4496–4499.

[7] Z. Zhang and L. Luo, "Hate speech detection: A solved problem? the challenging case of long tail on twitter," *CoRR*, vol. abs/1803.03662, 2018. [Online]. Available: http://arxiv.org/abs/1803.03662

[8] A. Kumar, V. Tyagi, and S. Das, "Deep learning for hate speech detection in social media," in *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*, 2021, pp. 1–4.

[9] T. A. Naidu and S. Kumar, "Impact of deep learning models on hate speech detection," in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2021, pp. 1–5.

[10] P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, and J. Pierrehumbert, "HateCheck: Functional tests for hate speech detection models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 41–58. [Online]. Available: https://aclanthology.org/2021.acl-long.4

[11] O. Ali, N. Scheidt, A. Gegov, E. Haig, M. Adda, and B. Aziz, "Automated detection of racial microaggressions using machine learning," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2020, pp. 2477–2484.

[12] E. Ombui, L. Muchemi, P. Wagacha, A. Gichamba, and M. Karani, "Leveraging hierarchical features for hatespeech identification in short message texts," in *2019 IEEE AFRICON*. IEEE, 2019, pp. 1–5.

[13] L. Gao and R. Huang, "Detecting online hate speech using context aware models," in *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. Varna, Bulgaria: INCOMA Ltd., Sep. 2017, pp. 260–266. [Online]. Available: https://doi.org/10.26615/978-954-452-049-6$_0$36