Healthcraft Frontiers
https://www.acadlore.com/journals/HF

Racism and Hate Speech Detection Using QAHA Based Hybrid Deep Learning Model: LSTM-CNN

Praveen Kumar Jayapal[1], Kumar Raja Depa Ramachandraiah[2], Kranthi Kumar Lella[3*]

[1] DiSTAP, Singapore-MIT Alliance for Research and Technology, 138602 Singapore, Singapore
[2] Faculty of Information and Communications Technology, Universiti Teknikal Malaysia Melaka, 76100 Melaka, Malaysia
[3] School of Computer Science and Engineering, VIT-AP University, 522237 Vijayawada, India

[*] Correspondence: Kranthi Kumar Lella (kranthi.kl@vitap.ac.in)

Received: 10-12-2023 **Revised:** 11-17-2023 **Accepted:** 11-23-2023

Citation: Jayapal, P. K., Ramachandraiah, K. R. D., & Lella, K. K. (2023). Racism and hate speech detection using QAHA based hybrid deep learning model: LSTM-CNN. *Healthcraft Front.*, *1*(1), 1-14. https://doi.org/10.56578/hf010101

Abstract: Twitter, a predominant platform for instantaneous communication and idea dissemination, is often exploited by cybercriminals for victim harassment through sexism, racism, hate speech, and trolling using pseudonymous accounts. The propagation of racially charged online discourse poses significant threats to the social, political, and cultural fabric of many societies. Monitoring and prompt eradication of such content from social media, a breeding ground for racist ideologies, are imperative. This study introduces an advanced hybrid forecasting model, utilizing convolutional neural networks (CNNs) and long-short-term memory (LSTM) neural networks, for the efficient and accurate detection of racist and hate speech in English on Twitter. Unlabelled tweets, collated via the Twitter API, formed the basis of the initial investigation. Feature vectors were extracted from these tweets using the TF-IDF (Term Frequency-Inverse Document Frequency) feature extraction technique. This research contrasts the proposed model with existing intelligent classification algorithms in supervised learning. The HateMotiv corpus, a publicly available dataset annotated with types of hate crimes and ideological motivations, was employed, emphasizing Twitter as the primary social media context. A novel aspect of this study is the introduction of a revised artificial hummingbird algorithm (AHA), supplemented by quantum-based optimization (QBO). This quantum-based artificial hummingbird algorithm (QAHA) aims to augment exploration capabilities and reveal potential solution spaces. Employing QAHA resulted in a detection accuracy of approximately 98%, compared to 95.97% without its application. The study's principal contribution lies in the significant advancements achieved in the field of racism and hate speech detection in English through the application of hybrid deep learning methodologies.

Keywords: Cyberstalkers; Artificial hummingbird algorithm; Quantum-based optimization; Long short-term memory; Racism detection

1. Introduction

In the digital age, the prevalence of social media has revolutionized the way individuals communicate and express themselves. A notable trend observed is the uninhibited expression of thoughts and opinions by users, often leading to the oversharing of personal information (Benítez-Andrades et al., 2022). The anonymity provided by social networks emboldens many users to post their emotions and thoughts without filters, sometimes disregarding the potential harm to others (Sadiq et al., 2021). Particularly, individuals with racist ideologies exploit social media to disseminate their beliefs, asserting their right to free expression (Benitez-Andrades et al., 2021). This unfettered expression is not confined to fanatical religious, racial, or political views; it extends to extreme bigotry, including sexist behavior that transgresses the bounds of hate speech (Lee et al., 2022). The voluminous nature of interactions on social media platforms renders manual monitoring and response to the myriad of comments, messages, and data virtually unfeasible (Macherla et al., 2023). Furthermore, the scarcity of official data on hate crimes underscores the prevailing issues in accurately addressing such content on these platforms

(Alnazzawi, 2022). Despite these challenges, the rich data available on social media are pivotal for user data processing. Data mining plays a critical role in this context, uncovering hitherto unknown patterns within datasets and enabling rapid, informed research and decision-making processes (De Souza & Da Costa-Abreu, 2020). In the realm of Natural Language Processing (NLP), LSTM and CNN are prominent neural network architectures. LSTM excels in handling sequential data, while CNN is adept at detecting patterns and features in text data. The amalgamation of these two architectures suggests a hybrid model that capitalizes on the strengths of both LSTM and CNN for effective hate speech detection.

Social media platforms, notably Facebook and Twitter, have raised significant concerns regarding the prevalence of inappropriate language in user posts (Vanetik & Mimoun, 2022). The manifestation of racism on these platforms is multifaceted, encompassing both overt and covert forms (Arcila-Calderón et al., 2021). Instances include the utilization of counterfeit profiles for disseminating racist remarks. Historically linked to ethnicity, racism now proliferates based on skin tone, country of origin, language, cultural background, and predominantly religious beliefs. Online remarks and posts inciting racial tensions have threatened the social, political, and cultural equilibrium of various nations (Reddy et al., 2023). The rapid dissemination of racist ideologies via social media underscores the urgency of identifying and eliminating such content (Jia et al., 2021).

Exposure to racist comments and tweets on social media has been associated with various mental and physical health conditions, leading to adverse health outcomes (Bisht et al., 2020). Three distinct forms of racism identified in online interactions include institutional racism, personally mediated racism, and internalized racism (Toliyat et al., 2022). Personally mediated racism occurs when an individual experiences or witnesses prejudice based on race (Herodotou et al., 2020). Consequently, racism in society inflicts psychological stress on individuals, heightening the risk of chronic diseases (Istaiteh et al., 2020). Racist groups and individuals are increasingly employing sophisticated methods to promote cyber racism (Kokatnoor & Krishnan, 2020). Sentiment analysis has gained prominence for its application in analyzing social media content for purposes like hate speech detection and racism identification (Kaya & Alatas, 2022). Recent advancements in automatic detection methods aim to address the issue of abusive content (Rodríguez-Sá nchezet al., 2020). Machine and deep learning approaches have proven their efficacy in various domains, including sentiment analysis (Mozafari et al., 2020; Pitropakis et al., 2020).

This study, therefore, employs a hybrid deep learning model to analyze racist tweets, with the following contributions: first, implementation of pre-processing techniques, such as TF-IDF and Bag of Words, for enhanced classification accuracy; second, utilization of the LSTM model within CNN-LSTM for effective detection of racist and offensive language; third, application of the QAHA to refine AHA performance, thereby optimizing the hyper-parameters of CNN-LSTM; finally, deployment of multiple methods for the detection of racist and offensive speech on a publicly available dataset. The subsequent sections of this study are structured as follows: Section 2 reviews related works on Twitter data analysis. Section 3 elucidates the proposed model, while Section 4 presents the experimental analysis. Section 5 concludes the study and outlines future research directions.

2. Related Works

The pervasive nature of hate speech on social media has catalyzed significant research efforts. Lee et al. (2022) explored sentiment analysis to detect tweets laden with racist content, employing Gated Convolutional Recurrent Neural Networks (GCR-NNs). This stacked ensemble model, integrating Gated Recurrent Units (GRUs), CNNs, and Recurrent Neural Networks (RNNs), demonstrates a synergistic improvement over individual components. In GCR-NNs, GRUs serve to extract salient features from raw text, which are then processed by CNNs to assist RNNs in making accurate predictions. Comparative analyses with existing models underscored the GCR-NN's efficacy, achieving an accuracy of 98% and a 97% detection rate for racist tweets.

Peng et al. (2023) employed a sophisticated Bidirectional Encoder Representations from Transformers (BERT) model, specifically optimized for Twitter sentiment classification, to analyze the tone of approximately one million tweets related to the Black Lives Matter (BLM) movement from July 2013 to March 2021. This model, tested on the Sentiment 140 dataset, achieved unparalleled results among machine learning models, registering an accuracy of 0.94 in the testing phase. The study utilized metrics such as retweet counts and word counts in tweets to visualize key concepts and milestones of the BLM movement. Public opinion analysis revealed varied degrees of support for issues like social justice and police brutality. The implications of this research extend to the promotion and analysis of social and political movements.

Ghosal & Jain (2023) introduced the first unsupervised detection system, which encompasses the HateCircle algorithm, hate tweet classification, and code-switch data preparation techniques. The HateCircle method, employing word co-occurrence analysis, determines the hate orientation of each phrase. A multi-class system for hate tweet categorization was developed using part-of-speech tagging, Euclidean distance, and Geometric median methods. The system proved more effective in identifying hate content in local scripts compared to Roman script, advocating its use in code-switch data preparation. Utilizing an enhanced hate lexicon in conjunction with various dictionaries, the system achieved a maximum F1-score of 0.74 in the Hindi dataset and 0.88 in the Bengali dataset. Comparative evaluation of the proposed parts of speech tagging and Geometric detection strategies with the

HateCircle method and hate tweet identification framework showed that HateCircle attained maximum accuracies of 0.73 and 0.78 on the Hindi and Bengali datasets, respectively. This study demonstrates the efficacy of contextual detection research incorporating a language-independent component in combating the spread of subtly harmful content on social media.

Ali et al. (2023) proposed novel graph-based algorithms for identifying hate material on social media. Utilizing Twitter, a dataset was created for testing and validation purposes, involving the extraction and annotation of tweets by language experts. The authors introduced a custom LSTM-GRU model to categorize hate speech into distinct classes. Applied to the compiled dataset, the model achieved an accuracy of 98.14 percent. The Girvan-Newman method was employed to identify key individuals and intraclass communities on Twitter. This approach is significant for monitoring social media to detect potential disruptions, including the identification of hate tweets and groups.

In a separate study, Agarwal et al. (2023) explored enhancing the efficiency of automatic hate speech detection on Social Media Platforms (SMPs) by parallelizing traditional ensemble learning techniques. The research involved parallelizing three popular hate speech detection algorithms—bagging, A-stacking, and random subspace—and their performance was compared with their serial counterparts across various high-dimensional datasets. These datasets encompassed diverse topics such as the COVID-19 pandemic and the 2020 US farmers' agitation in India (2021). The parallel models demonstrated a considerable increase in speed and efficiency, validating their suitability for the intended applications. This study underscored the importance of generalization by testing the models in a cross-dataset environment, finding that the parallelized algorithms maintained accuracy comparable to their serial versions.

Joloudari et al. (2023) proposed future research directions for the development of a BERT-based model tailored for sentiment analysis. In this approach, a deep CNN architecture captures the hierarchical structure of tweet embeddings, while the BERT model accumulates contextual representations of words, efficiently delineating the intricate semantics of tweets related to COVID-19. Comparative analysis with existing sentiment analysis techniques demonstrated that the BERT-deep CNN models excel in real-time classification of sentiments in COVID-19-related tweets. The study's findings contribute significantly to understanding public sentiment, offering insights that are crucial for policymakers in discerning public opinion, identifying misinformation, and formulating emergency response strategies. This research sets a new benchmark for future studies in sentiment analysis of social media data in crisis contexts and furthers the development of sentiment analysis methodologies.

Saleh et al. (2023) explored the effectiveness of using domain-specific word embedding for the automatic identification of hate speech. This method assigns negative connotations to specific terms to detect coded words effectively. Additionally, the application of the transfer learning language model (BERT), known for its proficiency in various NLP tasks, was examined for hate speech classification. Experimental results revealed that a bidirectional LSTM-based model, employing domain-specific word embedding, achieved an F1-score of 93% on a balanced dataset comprising existing hate speech datasets. In contrast, the BERT model attained a 96% F1-score. The performance of pre-trained models was found to be influenced by the volume of training data. Despite the disparity in corpus size, the first method, focusing on domain-specific data during training, outperformed the BERT model, trained on a larger corpus. The study highlighted the advantage of creating large pre-trained models from rich domain-specific content for contemporary social media platforms.

Nagar et al. (2023) introduced a novel methodology for the detection of hate dialogue on Twitter. This method integrates the author's content, social context, and linguistic characteristics to enhance the accuracy of hate speech detection. Incorporating textual content and the surrounding social environment, the approach employs an encoder to assimilate the unified features of the authors. The adaptability of this framework allows the use of various text encoders to capture the textual properties of the material, rendering it suitable for a broad spectrum of existing and future language models. The efficacy of this method was validated on two distinct Twitter datasets, demonstrating significant improvements over current state-of-the-art approaches. The results highlighted the importance of considering social context in enhancing the identification of hate speech on Twitter.

Liu et al. (2023) proposed BotMoE, a system designed to detect fraudulent bots on Twitter by using multiple modalities of user information, including metadata, textual content, and network structure. The system incorporates a Mixture-of-Experts (MoE) layer, which considers the Twitter communities to augment domain generalization and adaptability. BotMoE constructs modal-specific encoders for metadata attributes and textual structure, subsequently employing a MoE layer that categorizes users into appropriate groups based on community expertise. The final stage involves an expert fusion layer that amalgamates user representations from metadata, text, and graph perspectives, ensuring consistency across all modalities. Extensive trials indicated that BotMoE significantly surpassed existing methods in identifying sophisticated and stealthy bots, demonstrating reduced reliance on training data and enhanced generalization capabilities for new and unknown user populations.

Mnassri et al. (2023) addressed the challenge of unbalanced and sparsely labeled datasets by introducing a learning strategy that incorporates external emotional variables from diverse corpora. This study utilized BERT and mBERT, the latter focusing on cross-lingual identification of abusive material. Leveraging the shared encoder of transformers, the model concurrently recognizes abusive content and incorporates emotional aspects. This

approach facilitates rapid learning through the use of auxiliary information and enhances data efficiency by minimizing overfitting through shared representations. The research indicated that incorporating emotional intelligence significantly improved the accuracy of databases in recognizing hate speech and abusive language. Notably, multi-task models exhibited fewer errors than single-task models in both hate speech identification and aggressive language detection tasks, presenting an intriguing development in this field.

Almaliki et al. (2023) introduced a method for the precise identification of Arabic anti-Semitism on Twitter. This study implemented the Arabic BERT-Mini Model (ABMM) for detecting online bigotry. Twitter data were analyzed using bidirectional encoder representations from the model, categorizing the findings into typical, abusive, and hateful classes. Comparative tests were conducted against current state-of-the-art methods, and the results demonstrated that the ABMM model excelled in detecting Arabic hate speech, achieving a highly encouraging score of 0.986.

Gite et al. (2023) explored the application of Ant Colony Optimization (ACO) as an optimization strategy, integrating it with four machine learning models utilizing various feature selection and extraction methods on K-Nearest Neighbour (KNN) and Logistic Regression (LR). The objective was to demonstrate the differences between findings using comparative analysis. The proposed feature selection and extraction methods facilitated the improvement of the machine learning models' efficiency. This study considered both numerical datasets for stroke prediction and textual datasets for hate speech detection. The text dataset, compiled from Twitter API data encompassing tweets with positive, negative, and neutral emotions, utilized the TF-IDF method in conjunction with ACO. The application of ACO to the Random Forest model resulted in a significant accuracy enhancement, reaching up to 10.07 percent.

Fazil et al. (2023) presented a Bidirectional LSTM (BiLSTM) network for identifying xenophobic content. The model employed a multi-channel setup using contemporary word representation techniques to capture semantic relationships across various time frames using multiple filters of different kernel widths. The network processed encoded representations from several channels, with the output of a stacked 2-layer BiLSTM being combined and transmitted through a dense layer, subsequently weighted by an attention layer. The classification was conducted using a sigmoid function in the output layer. The performance of this model was evaluated on three Twitter datasets using four assessment metrics. Comparative analysis with five state-of-the-art models and an aggregate of baseline models indicated superior performance of the proposed model. The ablation study revealed that the removal of channels and the attention mechanism significantly impacted the model's performance. An empirical study was conducted to determine the optimal settings for the model's word representation methods, optimization algorithms, activation functions, and batch size.

3. Material and Methods

For the automatic detection of online hate crimes, access to annotated corpora is indispensable. In the absence of a standardized benchmark, researchers have been compelled to collect and categorize data on hate crimes independently. This research aims to fill the gap in literature focusing specifically on the identification and motivations of hate crimes, which has been previously unexplored, thus hampering the understanding of prevalent hate crime causes and their mitigation.

3.1 Corpus Construction

The HateMotiv corpus was developed through the collection of Twitter posts spanning nine years (1 January 2010–30 December 2019), using the TweetScraper too (Alnazzawi, 2022). It is important to note that the presence of terms such as "hate crime" or "hate crimes" in a tweet does not necessarily imply the endorsement or incitement of violence against a specific group. Twitter users commonly employ hashtags to associate their posts with specific events or topics (Burnap & Williams, 2016). Consequently, prevalent hashtags related to hate crimes were identified using the "Hashtagify" application (https://hashtagify.me/) (Hashtagify). Hashtags including "hate crime," "racist," "racism," and "Islamophobia" were among those selected for compiling relevant tweets. These hashtags, found to align closely with the FBI’s categorization of hate crimes, were employed as keywords to extract suitable tweets. An English instructor with extensive annotation experience selected these keywords, resulting in a query that returned 23,179 tweets containing the specified hashtags. To optimize the resources for manual annotation, a subset of 5,000 tweets was randomly selected for further analysis.

3.2 Annotation Process

Each tweet was annotated by two native English-speaking annotators (Cogit). Uniform standards were applied in the annotation process, focusing on identifying the type of hate crime and the motivation behind it. The corpus was annotated for four categories and causes of hate crimes, as outlined in Table 1. Regular discussions were held between the annotators and the judge overseeing the annotation process to address any arising inconsistencies or challenges.

Table 1. Glossed entity classes for HateMotiv corpus

Class Category	Explanation
Hate crime type	Hate crime type refers to categories identified by the FBI, including physical assault, verbal abuse, and incitement to hatred
Motivation	Motivation refers to the underlying motive for hate crimes, such as bias related to racism, religion, disability, and unknown

The HateMotiv corpus analysis revealed that physical assault constitutes the most common type of hate crime, while verbal abuse is the least frequently recorded category on Twitter. A notable observation from the data is the primary role played by the inability to accept diversity in terms of skin color and nationality in the perpetration of various forms of hate crimes. Conversely, hate crimes attributed to disability and negative attitudes towards disabled individuals constituted a minor percentage of the overall causes.

Furthermore, sexism or gender-based discrimination emerged as the second most prevalent justification for hate crimes, following racial prejudice. Notably, a proportion of hate crimes were committed with no apparent motive, reflecting the perpetrators' inherent biases or mental health issues. This is captured in the corpus under the term "unknown motive." The data indicated that assaults were the most frequent type of hate crime committed for reasons classified as unknown. However, the incidence of crimes committed for indeterminate reasons remains relatively low, accounting for approximately 0.011% of all categorized hate crimes.

3.3 Cleaning and Visualizing Data

The analysis of emojis within tweets was employed as a preliminary method to gauge the tone of the message. However, the primary focus was on textual data, which required extensive cleaning and preprocessing. This process involved the following four steps:

Step 1: Filtering. The first step involved the removal of hypertext links (e.g., http://google.com) and user handles typically starting with the "@" symbol on Twitter. This was crucial to eliminate irrelevant data and focus on the content of the tweets.

Step 2: Tokenization. In the second step, a Bag of Words representation was created. This involved the exclusion of punctuation and question marks to facilitate the appropriate representation of large datasets.

Step 3: Stop-word removal. Commonly occurring words such as "a," "an," and "the," which do not contribute to the analysis, were eliminated in the third step.

Step 4: N-gram creation. This step involved generating n-grams, defined as sequences of 'n' words or characters extracted from the text. While unigrams and bigrams have distinct utilities, this study utilized unigram tokens for tweet preparation. The decision to focus on unigrams was based on their comprehensive data coverage.

The selection between constructing unigrams or bigrams should be guided by the specific objectives of the study. Bigrams, such as "not good," are effective in conveying emotions succinctly, making them particularly suitable for sentiment analysis and product reviews. Conversely, unigrams offer comprehensive data coverage. In this research, the focus was on unigram tokens for tweet preparation, to evaluate the efficacy of various stemmers and lemmatizers. It was found that while lemmatizers could deconstruct compound words into their elements, this process did not significantly enhance accuracy compared to the categorization models applied. Post-cleanup of text documents, tokenization was employed for more detailed analysis, necessitating the transformation of these tokens into feature vectors. Feature vectors serve as a crucial representation in the training of classification algorithms.

Two transformation techniques were compared: the Bag of Words method and the TF-IDF method. The Bag of Words approach, a straightforward transformation strategy, utilizes the corpus's diverse words as features, with each column indicating the frequency of a specific term's occurrence. Despite its computational simplicity, this method provides limited insights beyond word frequency. The TF-IDF method, on the other hand, combines the frequency of a term's occurrence in the text with its distribution across different document types to assign a weight to each word. This implies that commonly occurring words across various text types are assigned a lower weight. The feature vectors generated using these methods were successfully prepared for use in training classification models.

3.4 Classification Using the Proposed Architectures

This study introduced three distinct deep learning models for hate speech classification: a LSTM network, a CNN, and a hybrid model combining both. The performance of each model in executing classification tasks was evaluated comprehensively. The CNN model, characterized by higher efficiency and a manageable number of trainable parameters, was ultimately selected for deployment on System-on-Chip Field Programmable Gate Arrays (SoC-FPGAs). The choice was influenced by hardware limitations, as SoC-FPGAs, known for their high performance and low power consumption, are suitable for edge computing applications, but only the CNN model

was compatible with this hardware.

3.4.1 LSTM architecture

As depicted in Figure 1, the employed LSTM architecture incorporates two LSTM layers, each consisting of 100 LSTM cells. These layers are designed to accurately represent the sequential relationships between the features and the labels of hate speech data. The LSTM layers process the incoming data, discerning complex connections between characteristics and labels. Subsequently, two fully connected layers receive the output from the LSTM layers and generate a final prediction based on the processed data.

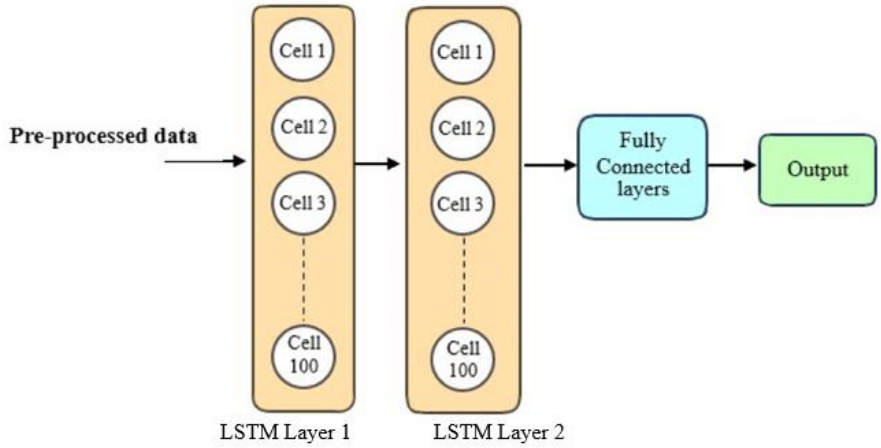

Figure 1. Proposed LSTM architecture

3.4.2 CNN architecture

Figure 2 presents the CNN model structure, comprising several CNN layers followed by pooling layers, originally developed for image classification tasks. For the purpose of hate speech classification, data arrays, represented as images, were input into the model. The CNN was then trained to identify pertinent features and make predictions about the input data concerning the classification of racist content.

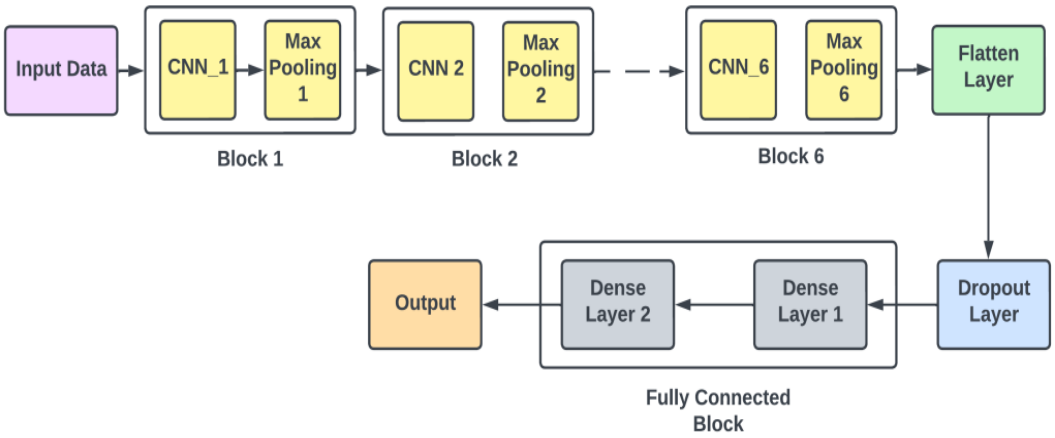

Figure 2. Proposed CNN architecture

3.4.3 CNN-LSTM architecture

The hybrid architecture, depicted in Figure 3, amalgamates the strengths of CNN and LSTM networks to enhance outcomes for complex deep learning tasks. This model harnesses the LSTM's ability to model temporal correlations among features in conjunction with the CNN's capacity to extract pivotal information from the data. The CNN simplifies the feature extraction process by identifying key elements, while the LSTM maintains the temporal relationships within the data.

The CNN model was ultimately chosen for hardware implementation due to its superior performance and efficiency compared to the LSTM and CNN-LSTM models. The selection was also influenced by the model's simplicity in terms of understanding and implementation on Field Programmable Gate Arrays (FPGAs). Hardware

limitations were a contributing factor, as the FPGA was compatible only with the CNN architecture.

In summary, the study presented three distinct deep learning models - CNN, LSTM, and their hybrid - for the classification of racist content. Extensive testing led to the selection of the CNN model for implementation on SoC-FPGAs due to its high performance and computational efficiency. The CNN model demonstrated superiority over LSTM models in terms of accuracy and processing efficiency, with all models capable of analyzing temporal aspects of data.

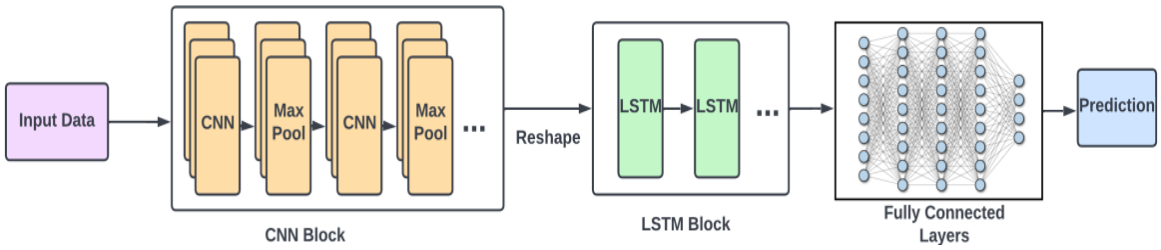

Figure 3. Proposed CNN-LSTM architecture

3.4.4 Model implementation

Three models were implemented for comparative analysis: an LSTM model, a CNN, and a CNN-LSTM hybrid. The LSTM model consists of two layers, each with 100 units, following an input layer that processes 1200 features sequentially. The initial LSTM layer receives data in the format of (1, 1200) and outputs in the form of (1, 100), which is then fed into the subsequent LSTM layer. A series of fully connected layers compute the probability distribution for each class after the network learns the association between features and labels.

The proposed CNN architecture comprises six CNN blocks, each containing a CNN layer followed by a max pooling layer. After the CNN blocks, a flatten layer converts the aggregated features into a one-dimensional array. A dropout layer with a 50% rate is introduced to prevent overfitting, randomly eliminating neurons to reduce dependency on training data. The output from the CNN layers is passed to fully connected layers, which produce the probability distribution for each class.

In the CNN-LSTM hybrid:

- Each of the four CNN processing units includes a CNN layer for feature extraction and a max pooling layer to capture the most salient features.
- A reshape layer then transforms the output from the CNN blocks, converting 3D CNN output into 2D LSTM input.
- The CNN block's output is fed into an LSTM layer to learn features that evolve over time.
- Finally, a flatten layer followed by two fully connected layers generates a probability distribution for the classes.

3.4.5 Parameter tuning

The performance of the implemented deep learning models can be significantly influenced by the tuning of hyperparameters. These models necessitate meticulous adjustment of several hyperparameters, which affect memory and compute complexity. This section outlines the additional hyperparameters that facilitate the selection of a specific approach for given scenarios. It is observed that superior outcomes often require extensive tuning of these hyperparameters.

Hyperparameter optimization can be mathematically defined as:

$$x^* = \arg\min_{x \in X} f(x) \quad (1)$$

where, $f(x)$ represents the objective function to minimize the error measured on the validation set, and x^* is the set of hyperparameters within the domain X that yields the lowest error. The goal is to identify the hyperparameter values that result in optimal performance on the validation set metric. The selection between manual and automatic hyperparameter tuning establishes a balance between the maximal computational cost of automated models and the in-depth knowledge required for manual selection. In this study, a QBO model was utilized to fine-tune the CNN-LSTM model's hyperparameters.

QBO

This section delineates the principles of QBO, an approach utilized for feature selection in this study. In QBO, binary representation is used, where '1s' indicate features to be retained and '0s' denote features to be discarded. The quantum bit (Q-bit) operations in QBO involve each feature being represented as a superposition, characterized by a complex integer. This is mathematically expressed as:

$$q = \alpha + i\beta = e^{i\theta}, |\alpha|^2 + |\beta|^2 = 1 \tag{2}$$

where, α and β correspond to the two potential states of the Q-bit, namely '0' and '1'. The angle of q is adjusted using the arctan function.

The primary objective of QBO is to determine the change in the value of q. This process is conducted using a calculation method $\Delta\theta$.

$$q(t+1) = q(t) \times R(\Delta\theta) = [a(t)\beta(t)] \times R(\Delta\theta) \tag{3}$$

where, $R(\Delta\theta)$ stands for the rotation matrix associated with a change of $\Delta\theta$ in the angle, defined as:

$$R(\Delta\theta) = \begin{bmatrix} \cos(\Delta\theta) & -\sin(\Delta\theta) \\ \sin(\Delta\theta) & -\cos(\Delta\theta) \end{bmatrix} \tag{4}$$

The optimal solution, denoted as X_b, is predetermined to set the value of the parameters influencing q. The binary representation of a solution X_i is represented by its j-th bit X_{ij}, while the j-th bit of X_b at time t is denoted as X_{bj}. As reported in (Srikanth et al., 2018), the angle vector in QBO is capable of assuming one of eight distinct values, allowing for varied adjustments in the Q-bit representation.

The primary objective of QBO is to balance the exploration and exploitation potential. Initially, the data is split into a 70% training set and a 30% testing set. Random numbers are then used to determine the fitness of each agent. The agent with the lowest fitness score is selected as the best agent. The AHA (Zhao et al., 2022) is employed for exploitation, and the solution is updated iteratively until the termination criteria are met. Following this, the implemented QAHA is evaluated on the reduced dimensionality of the test set based on the optimal solution. The QAHA will be expounded in the subsequent sections.

A. First stage

Initial agents representing the population are generated, with each solution comprising D Q-bits. Consequently, each solution X_i can be expressed as:

$$X_i = \left[q_{i1} | q_{i2} | \ldots | q_{iD} \right] = \left[\theta_{i1} | \theta_{i2} | \ldots | \theta_{iD} \right], i = 1, 2, \ldots N \tag{5}$$

where, X_i refers to the superposition of probabilities of selecting or not selecting features.

B. Second stage

The process of updating agents until a specified criterion is met constitutes a critical phase in the application of the QAHA. Initially, the binary representation of each solution, denoted as X_i, is determined through an equation. This representation involves a random charge $rand \in [0,1]$ and a parameter β as defined in Eq. (2).

$$BX_{ij} = \begin{cases} 1 & if \quad |\beta|^2 > rand \\ 0 & otherwise \end{cases} \tag{6}$$

Subsequently, the fitness value for each agent is computed. This computation is achieved by training a CNN-LSTM classifier, with the features derived from BX_{ij} serving as the model's hyperparameters. The fitness value is formulated as:

$$Fit_i = \rho \times \gamma + (1-\rho) \times \left(\frac{|BX_{ij}|}{D} \right) \tag{7}$$

where, $|BX_{ij}|$ represents the error rate in classifying features using the CNN-LSTM classifier, and denotes the total number of features employed. A normalization factor within the range [0, 1] ensures parity in fitness levels across different agents. The LSTM model is preferred due to its simplicity, efficiency, and reliance on a singular tuning parameter. Its ability to retain information from the training set contributes to its effectiveness, particularly when other classifiers might not yield desired results.

The subsequent stage involves identifying the most optimal agent X_b, characterized by the minimum fitness value Fit_b. This step is pivotal in the QAHA process as it determines the most suitable set of features for the

classification task at hand.

C. Third stage

The test set is narrowed down to features equivalent to those in the binary representation of X_b. The reduced-dimension test set is then used to apply the trained classifier for predictions. Subsequently, the output quality is thoroughly evaluated. The computational cost of QAHA is determined by the initial population size, population size N, fitness evaluation N_{Fit}, and maximum iteration count.

$$O(QAHA) = O(T \times C \times N + T \times N \times D + T \times D/2) + O(N \times D) \tag{8}$$

In summary, the complexity of QAHA is given by:

$$O(QAHA) = O\left(T \times N_{Fit} \times N + T \times N \times D + T \times D/2\right) \tag{9}$$

4. Results and Discussion

4.1 Hardware and Software Used for the Experiments

For preprocessing the tweets and applying deep learning methodologies, a Jupyter notebook, scripted in Python 3.6, was employed. The computational tasks were executed on a workstation equipped with the following hardware specifications: an Intel Core i7-9700K processor operating at 3.60GHz, 32.0GB of RAM, and a 6GB NVIDIA GeForce graphics card. For text preprocessing, the spaCy and nltk libraries were utilized, facilitating the efficient processing of the datasets.

4.2 Performances Metrics

The performance of the model was evaluated using the confusion matrix, a tool that provides four distinct outcomes: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The effectiveness of the model was determined through the calculation of various metrics derived from the confusion matrix.

$$Accuracy = (TN + TP)/(TN + TP + FN + FP) \tag{10}$$

$$Sensitivity = \left(\frac{TP}{TP + FN}\right) \tag{11}$$

$$Specificity = \left(\frac{TN}{TN + FP}\right) \tag{12}$$

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$F - Measure = \frac{2TP}{2TP + FP + FN} \tag{14}$$

4.3 Analysis of the Proposed Classifier

The efficacy of the proposed model was compared with that of generic deep learning models. The uniqueness of the dataset used in this study, which had not been previously employed for validation analysis, necessitated this comparison. The results of the analysis are presented in Table 2 and Table 3.

The analysis of the hybrid model's performance is detailed in Table 2. The Autoencoder (AE) model demonstrated an Area Under the Curve (AUC) of 0.758 and achieved an accuracy of 87.72%. Precision was recorded at 78.14%, recall at 89.92%, and the F-measure at 88.67%. Subsequently, the Deep Belief Network (DBN) model exhibited an AUC of 0.854, accuracy of 89.17%, precision of 70.91%, recall of 85.69%, and an F-measure of 82.33%. The RNN model registered an AUC of 0.687, with an accuracy of 90.28%. The precision was noted at 64.17%, recall at 86.66%, and F-measure at 80.24%. Following this, the CNN model showed an AUC of 0.947, an accuracy of 92.78%, precision of 91.94%, recall of 90.61%, and an F-measure of 86.86%. The LSTM model recorded an AUC of 0.957 and an accuracy of 94.34%. Its precision was 92.45%, recall 93.78%, and F-measure

91.36%. Finally, the combined CNN-LSTM model achieved the highest performance with an AUC of 0.967, accuracy of 95.97%, precision of 96.84%, recall of 97.24%, and an F-measure of 94.13%.

Table 2. Analysis of the proposed hybrid model without QAHA

Classification	AUC	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
AE	0.758	87.72	78.14	89.92	88.67
DBN	0.854	89.17	70.91	85.69	82.33
RNN	0.687	90.28	64.17	86.66	80.24
CNN	0.947	92.78	91.94	90.61	86.86
LSTM	0.957	94.34	92.45	93.78	91.36
CNN-LSTM	0.967	95.97	96.84	97.24	94.13

Table 3. Analysis of the proposed hybrid model with QAHA

Classification	AUC	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
AE	0.9343	91.56	84.56	90.66	90.67
DBN	0.8923	90.12	85.2	92.57	86.54
RNN	0.9082	93.22	78.7	90.67	88.67
CNN	0.9135	94.43	93.6	94.01	91.54
LSTM	0.9544	93.23	94.7	96.62	95.09
CNN-LSTM	0.9829	98.45	98.8	99.56	97.70

Table 3 presents the outcomes from the assessment of the hybrid model integrated with QAHA. The AE model exhibited an AUC of 0.9343, achieving an accuracy of 91.56%. Precision was recorded at 84.56%, recall at 90.66%, and the F-measure at 90.67%. The DBN model attained an AUC of 0.8923, accuracy of 90.12%, precision of 85.2%, recall of 92.57%, and an F-measure of 86.54%. The RNN model demonstrated an AUC of 0.9082 and achieved an accuracy of 93.22%. Its precision was noted at 78.7%, recall at 90.67%, and F-measure at 88.67%. The CNN model showed an AUC of 0.9135, an accuracy of 94.43%, precision of 93.6%, recall of 94.01%, and an F-measure of 91.54%. Further, the LSTM model registered an AUC of 0.9544, with an accuracy of 93.23%, precision of 94.7%, recall of 96.62%, and an F-measure of 95.09%. Lastly, the CNN-LSTM model, representing the pinnacle of this research, achieved an AUC of 0.9829, an exceptional accuracy of 98.45%, precision of 98.8%, recall of 99.56%, and an F-measure of 97.70%.

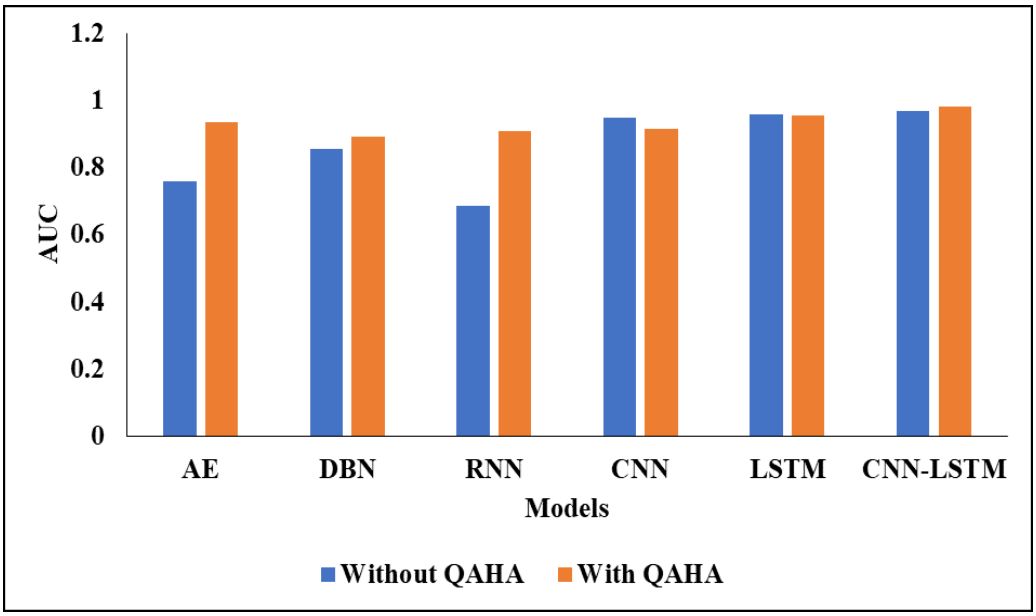

Figure 4. AUC comparison

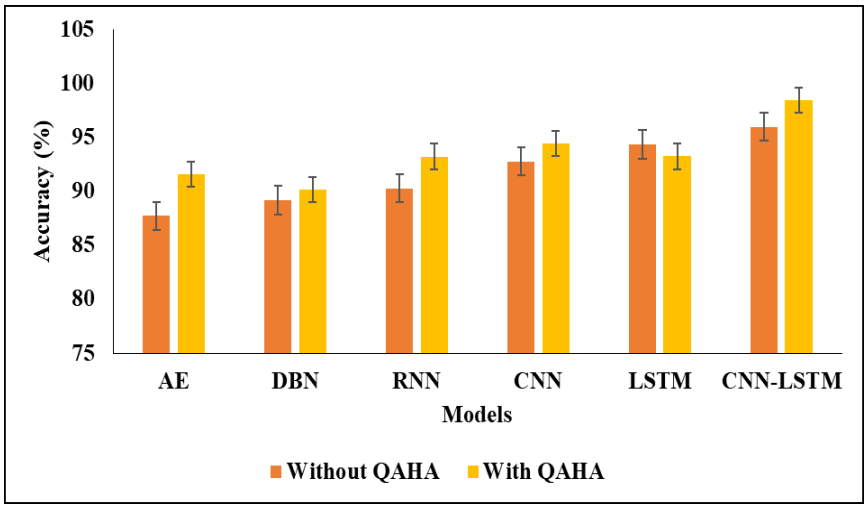

Figure 5. Accuracy analysis with and without the optimization model

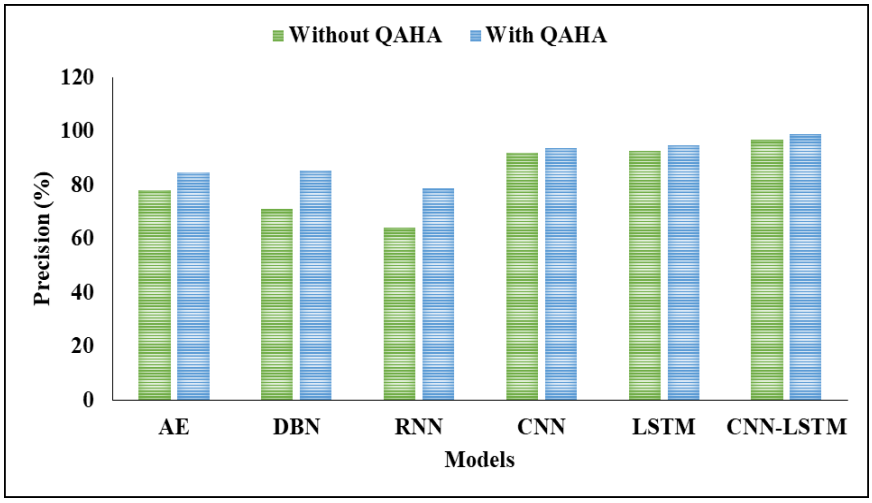

Figure 6. Graphical representation of various deep learning models

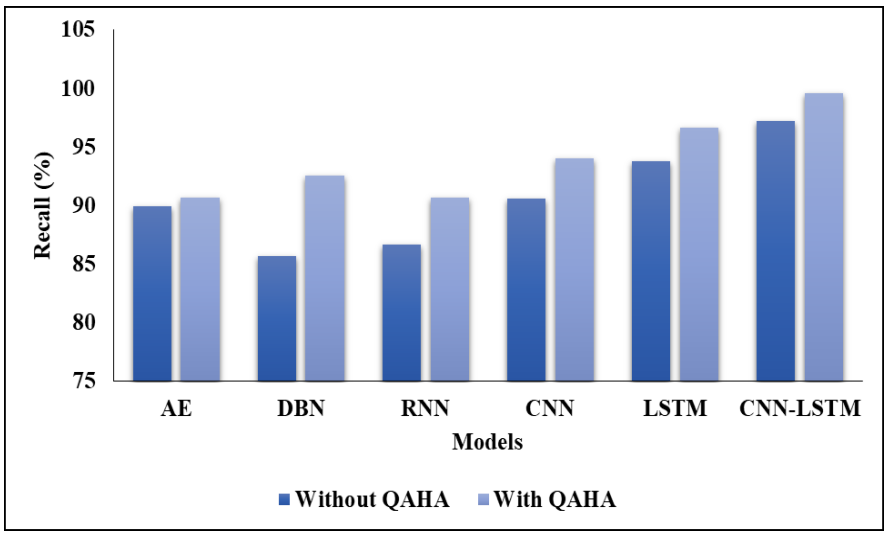

Figure 7. Recall analysis

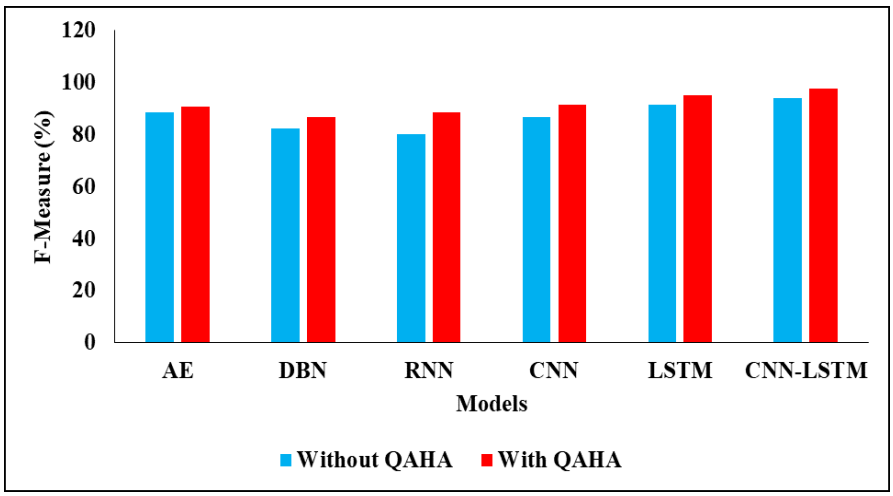

Figure 8. Validation analysis of QAHA

Figure 4, Figure 5, Figure 6, Figure 7 and Figure 8 provide graphical representations of these results, including AUC comparison (Figure 4), accuracy analysis with and without the optimization model (Figure 5), and recall analysis (Figure 7), among others. The validation analysis of QAHA is depicted in Figure 8. The analysis revealed that the implementation of QAHA significantly enhanced the performance of the deep learning models. The CNN-LSTM model with QAHA outperformed the other models, demonstrating the effectiveness of the proposed hybrid approach in the context of hate speech classification.

5. Conclusions and Future Work

The emergence of racist content on social media platforms, particularly Twitter, has necessitated the development of automated detection and removal mechanisms. This study has adopted a sentiment analysis approach to identify racist tweets, focusing on specific phrases and words. Following data preprocessing, neural network classification was conducted using LSTM, CNN, and a hybrid CNN-LSTM model. The experimental results demonstrated that the CNN and hybrid models significantly outperformed the LSTM model in both phases of the analysis. It was found that, despite its lower execution time, LSTM's complexity rendered it less suitable for SoC-FPGAs compared to the CNN model. The CNN's simpler architecture and high accuracy underscored its appropriateness for SoC-FPGA implementation. Furthermore, the QAHA was employed to optimize hyperparameters, enhancing the classification accuracy of the proposed model.

The dataset used in this study is publicly available, offering a valuable resource for future research into the automatic detection and prediction of hate crimes and their underlying motivations, including racism. This accessibility to the scientific community could spur further investigations into this domain. The experimental study revealed that the proposed model achieved superior performance compared to baseline models, with accuracy and recall rates exceeding 95% and 96%, respectively. Understanding the factors contributing to online hate crimes through advanced deep-learning techniques can be instrumental in curbing detrimental biases and reducing the incidence of crimes driven by such biases. Future work in this area aims to refine and employ sophisticated deep-learning methods to train models in recognizing the root causes of hate crimes shared online.

Data Availability

The data used to support the research findings are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflict of interest.

References

Agarwal, S., Sonawane, A., & Chowdary, C. R. (2023). Accelerating automatic hate speech detection using parallelized ensemble learning models. *Expert Syst. Appl., 230*, 120564. https://doi.org/10.1016/j.eswa.2023.120564.

Ali, M., Hassan, M., Kifayat, K., Kim, J. Y., Hakak, S., & Khan, M. K. (2023). Social media content classification

and community detection using deep learning and graph analytics. *Technol. Forecasting Social Change, 188*, 122252. https://doi.org/10.1016/j.techfore.2022.122252.

Almaliki, M., Almars, A. M., Gad, I., & Atlam, E. S. (2023). ABMM: Arabic BERT-Mini model for hate-speech detection on social media. *Electronics, 12*(4), 1048. https://doi.org/10.3390/electronics12041048.

Alnazzawi, N. (2022). Using Twitter to detect hate crimes and their motivations: The hatemotiv corpus. *Data, 7*(6), 69. https://doi.org/10.3390/data7060069.

Arcila-Calderón, C., Amores, J. J., Sánchez-Holgado, P., & Blanco-Herrero, D. (2021). Using shallow and deep learning to automatically detect hate motivated by gender and sexual orientation on Twitter in Spanish. *Multimodal Technol. Interact., 5*(10), 63. https://doi.org/10.3390/mti5100063.

Benítez-Andrades, J. A., González-Jiménez, Á., López-Brea, Á., Aveleira-Mata, J., Alija-Pérez, J. M., & García-Ordás, M. T. (2022). Detecting racism and xenophobia using deep learning models on Twitter data: CNN, LSTM and BERT. *PeerJ Comput. Sci., 8*, e906. https://doi.org/10.7717/peerj-cs.906.

Benitez-Andrades, J. A., González-Jiménez, Á., López-Brea, Á., Benavides, C., Aveleira-Mata, J., Alija-Pérez, J. M., & García-Ordás, M. T. (2021). BERT model-based approach for detecting racism and xenophobia on Twitter data. In Research Conference on Metadata and Semantics Research, pp. 148-158. https://doi.org/10.1007/978-3-030-98876-0_13.

Bisht, A., Singh, A., Bhadauria, H. S., Virmani, J., & Kriti. (2020). Detection of hate speech and offensive language in Twitter data using LSTM model. In *Recent Trends in Image and Signal Processing in Computer Vision*. pp. 243-264. https://doi.org/10.1007/978-981-15-2740-1_17.

Burnap, P. & Williams, M. L. (2016). Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Sci., 5*, 11.

De Souza, G. A. & Da Costa-Abreu, M. (2020). Automatic offensive language detection from Twitter data using machine learning and feature selection of metadata. In 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, pp. 1-6. https://doi.org/10.1109/IJCNN48605.2020.9207652.

Fazil, M., Khan, S., Albahlal, B. M., Alotaibi, R. M., Siddiqui, T., & Shah, M. A. (2023). Attentional multi-channel convolution with bidirectional LSTM cell toward hate speech prediction. *IEEE Access, 11*, 16801-16811. https://doi.org/10.1109/ACCESS.2023.3246388.

Ghosal, S. & Jain, A. (2023). HateCircle and unsupervised hate speech detection incorporating emotion and contextual semantics. *ACM Trans. Asian Low-Resour. Lang. Inf. Process., 22*(4), 1-28. https://doi.org/10.1145/3576913.

Gite, S., Patil, S., Dharrao, D., Yadav, M., Basak, S., Rajendran, A., & Kotecha, K. (2023). Textual feature extraction using ant colony optimization for hate speech classification. *Big Data Cognitive Comput., 7*(1), 45. https://doi.org/10.3390/bdcc7010045.

Herodotou, H., Chatzakou, D., & Kourtellis, N., (2020). A streaming machine learning framework for online aggression detection on Twitter. In 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, pp. 5056-5067. https://doi.org/10.1109/BigData50022.2020.9377980.

Istaiteh, O., Al-Omoush, R., & Tedmori, S. (2020). Racist and sexist hate speech detection: Literature review. In 2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA), Valencia, Spain, pp. 95-99. https://doi.org/10.1109/IDSTA50958.2020.9264052.

Jia, B., Dzitac, D., Shrestha, S., Turdaliev, K., & Seidaliev, N. (2021). An ensemble machine learning approach to understanding the effect of a global pandemic on Twitter users' attitudes. *Int. J. Comput. Commun. Control., 16*(2).

Joloudari, J. H., Hussain, S., Nematollahi, M. A., Bagheri, R., Fazl, F., Alizadehsani, R., Lashgari, R., & Talukder, A. (2023). BERT-deep CNN: State of the art for sentiment analysis of COVID-19 tweets. *Social Network Anal. Min., 13*(1), 99. https://doi.org/10.1007/s13278-023-01102-y.

Kaya, S., & Alatas, B. (2022). A new hybrid LSTM-RNN deep learning based racism, xenomy, and genderism detection model in online social network. *Int. J. Adv. Networking Appl., 14*(2), 5318-5328.

Kokatnoor, S. A., & Krishnan, B. (2020). Twitter hate speech detection using stacked weighted ensemble (SWE) model. In 2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), Bangalore, India, pp. 87-92. https://doi.org/10.1109/ICRCICN50933.2020.9296199.

Lee, E., Rustam, F., Washington, P. B., El Barakaz, F., Aljedaani, W., & Ashraf, I. (2022). Racism detection by analyzing differential opinions through sentiment analysis of tweets using stacked ensemble GCR-NN model. *IEEE Access, 10*, 9717-9728. https://doi.org/10.1109/ACCESS.2022.3144266.

Liu, Y., Tan, Z., Wang, H., Feng, S., Zheng, Q., & Luo, M. (2023). BotMoE: Twitter bot detection with community-aware mixtures of modal-specific experts. *arXiv Preprint*, arXiv:2304.06280. https://doi.org/10.48550/arXiv.2304.06280.

Macherla, H., Kotapati, G., Sunitha, M. T., Chittipireddy, K. R., Attuluri, B., & Vatambeti, R. (2023). Deep learning framework-based chaotic hunger games search optimization algorithm for prediction of air quality index. *Ing. Syst. Inf., 28*(2), 433-441. https://doi.org/10.18280/isi.280219.

Mnassri, K., Rajapaksha, P., Farahbakhsh, R. & Crespi, N. (2023). Hate speech and offensive language detection using an emotion-aware shared encoder. *arXiv Preprint,* arXiv:2302.08777. https://doi.org/10.48550/arXiv.2302.08777.

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). Hate speech detection and racial bias mitigation in social media based on BERT model. *PloS One, 15*(8), e0237861. https://doi.org/10.1371/journal.pone.0237861.

Nagar, S., Barbhuiya, F. A., & Dey, K. (2023). Towards more robust hate speech detection: Using social context and user data. *Soc. Netw. Anal. Min., 13*(1), 47. https://doi.org/10.1007/s13278-023-01051-6.

Peng, J., Fung, J. S., Murtaza, M., Rahman, A., Walia, P., Obande, D., & Verma, A. R. (2023). A sentiment analysis of the Black Lives Matter movement using Twitter. *STEM Fellowship J., 8*(1), 56-66. https://doi.org/10.17975/sfj-2022-015.

Pitropakis, N., Kokot, K., Gkatzia, D., Ludwiniak, R., Mylonas, A., & Kandias, M. (2020). Monitoring users' behavior: Anti-immigration speech detection on Twitter. *Mach. Learn. Knowl. Extr., 2*(3), 11. https://doi.org/10.3390/make2030011.

Reddy, N. V. R. S., Chitteti, C., Yesupadam, S., Subbaiah, V., Desanamukula, S. S. V., & Bommagani, N. J. (2023). Enhanced speckle noise reduction in breast cancer ultrasound imagery using a hybrid deep learning model. *Ing. Syst. Inf., 28*(4), 1063-1071. https://doi.org/10.18280/isi.280426. Rodríguez-Sá nchez,F., Carrillo-de-Albornoz, J., & Plaza, L. (2020). Automatic classification of sexism in social networks: An empirical study on Twitter data. *IEEE Access, 8*, 219563-219576. https://doi.org/10.1109/ACCESS.2020.3042604.

Sadiq, S., Mehmood, A., Ullah, S., Ahmad, M., Choi, G. S., & On, B. W. (2021). Aggression detection through deep neural model on Twitter. *Future Gener. Comput. Syst., 114*, 120-129. https://doi.org/10.1016/j.future.2020.07.050.

Saleh, H., Alhothali, A., & Moria, K. (2023). Detection of hate speech using BERT and hate speech word embedding with deep model. *Appl. Artif. Intell., 37*(1), 2166719. https://doi.org/10.1080/08839514.2023.2166719.

Search and Find the Best Twitter Hashtags. Hashtagify. https://hashtagify.me/

Srikanth, K., Panwar, L. K., Panigrahi, B. K., Herrera-Viedma, E., Sangaiah, A. K., & Wang, G. G. (2018). Meta-heuristic framework: Quantum inspired binary grey wolf optimizer for unit commitment problem. *Comput. Electr. Eng., 70*, 243-260. https://doi.org/10.1016/j.compeleceng.2017.07.023.

Toliyat, A., Levitan, S. I., Peng, Z., & Etemadpour, R. (2022). Asian hate speech detection on Twitter during COVID-19. *Front. Artif. Intell., 5*, 932381. https://doi.org/10.3389/frai.2022.932381.

Training Data for AI, ML with Human Empowered Automation. Cogit. https://www.cogitotech.com/about-us

Vanetik, N. & Mimoun, E. (2022). Detection of racist language in French tweets. *Information., 13*(7), 318. https://doi.org/10.3390/info13070318.

Zhao, W., Wang, L., & Mirjalili, S. (2022). Artificial hummingbird algorithm: A new bio-inspired optimizer with its engineering applications. *Comput. Meth. Appl. Mech. Eng., 388*, 114194. https://doi.org/10.1016/j.cma.2021.114194.