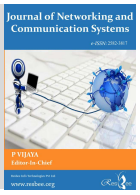
Vol.8 No.1 Jan 2025

Journal of Networking and Communication Systems

Received 11 December, Revised 20 December, Accepted 4 January



Identification and Blocking of Hate Speech in English-Hausa Code-mixed Language on Facebook: A Hybrid Deep Learning Framework

Hyellamada Simon

Department of Computer Science, Federal Polytechnic, Mubi, Nigeria
hyellasimon@gmail.com

Abstract: The rise of social media platforms, especially Facebook, has revolutionized communication and interaction worldwide. However, the unrestricted nature of these platforms has also given rise to the proliferation of hate speech, posing significant challenges to maintaining a safe and inclusive online environment. In this paper, a novel framework using hybrid deep learning models was proposed for the detection and blocking of hate speech on Facebook English-Hausa code-mixed language. The proposed hybrid deep learning model combines the strengths of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to form a Convolutional – Long Short-Term Memory (C-LSTM) with glove word embeddings to efficiently identify and block hate speech content. The datasets used in this work were collected from Facebook and were labeled manually as hate speech or non-hate speech. This paper also analyses the model's architecture, data preprocessing techniques, and the evaluation metrics presented, hence demonstrating its effectiveness in real-world scenarios where the experimental results show that the hybrid C-LSTM model generated 0.90 accuracy which outperformed the CNN, LSTM, and the baseline-models. The findings in this paper suggest that hate speech in Facebook English-Hausa code-mixed language can be detected and blocked instantly before posting.

Keywords: Hybrid Deep Learning Framework, English-Hausa Code-Mixed Language, Hate Speech Detection and Blocking, Facebook, Social Media, Word Embedding.

Nomenclature

Abbreviation	Expansion
NLP	Natural Language Processing
DL	Deep Learning
FCNN	Fuzzy-based Convolutional Neural Network
NB	Naive Bayes
CNN	Convolutional Neural Networks
LSTM	Long Short-Term Memory
ML	Machine Learning
BERT	Bidirectional Encoder Representations from Transformers
USE	Universal Sentence Encoder
SERC	Self-Annotated Reddit Corpus
GRU	Gated Recurrent Unit
RNN	Recurrent Neural Networks
DEA	Dolphin Echolocation Algorithm
RF	Random Forests
SVM	Support Vector Machines

1 Introduction

Social media platforms have become integral to the lives of millions of users globally. Among these platforms, Facebook stands out as one of the most popular and influential, allowing users to share their thoughts, ideas, and emotions with a vast audience. Social networking sites like Facebook and others made it necessary for users to create profiles and make lists of friends and peers with whom to connect and share fascinating content, including posting and reading other users' posts and comments [1]. However, the anonymity and lack of moderation on such platforms have led to the rampant spread of hate speech, discrimination, harassment, rumors, fake news, etc. that could trigger a catastrophic catastrophe.

Free speech is strongly related to political expression of ideas and serves as a necessary precondition for governmental institutions. Freedom of expression is a common right that promotes democratic involvement in society. The fundamental features of digital communications reduce the costs of information distribution and allow people to easily cross geographical boundaries, which have changed the conditions of speech in the digital age.

Generally, hate speech is predominately motivated by religious hatred and is characterized by the use of deadly words like kill, murder, and destroy as well as quantity words like many or million [2]. It is also more informal, and angrier, and frequently openly attacks the target through name-calling. Detecting and blocking hate speech is a complex challenge due to the varied and nuanced nature of the content involved. Traditional methods often struggle to cope with the intricacies of language and context, demanding more sophisticated techniques such as deep learning to tackle this problem effectively.

Numerous studies have been undertaken to address hate speech on social media sites. Initial methodologies predominantly depended on keyword matching and rule-based frameworks, which were inadequate in identifying nuanced expressions of hate speech. Recent breakthroughs in NLP and deep learning have facilitated the development of more robust models by researchers.

2. Literature Review

Anand et al. [3] employed NLP in conjunction with deep learning methodologies. The method proposed multilingual offensive language detection. The dataset was sourced from Facebook, Twitter, and YouTube, and subsequently underwent pre-processing to segment the data, remove noise, and filter out stop words. Features were chosen for segmented data via an FCNN. They employed an ensemble design incorporating a Bi-LSTM model and a hybrid NB architecture to extract certain properties from the dataset and classify them. An algorithmic classification method was utilized to evaluate the utilization of unpleasant terminology based on the emotions conveyed in the text. Through experimental analysis, unacceptable content in many languages was identified with 98% accuracy, an F-1 score of 92.5%, and a 45% RMSE, utilizing data from YouTube, Twitter, and Facebook.

A hybrid deep learning methodology was built in [4] for an abusive text recognition system capable of identifying English text deemed "not suitable for work" (NSFW). The hybrid model for the textual dataset is founded on CNN and LSTM networks. They further categorize data employing five ML methodologies to facilitate result comparison.

Subsequent, Kar & Debbarma [5] introduced an effective feature extraction and hybrid methodology for deriving diverse features from code-mixed texts. A quantum search optimization approach was subsequently utilized to enhance the recovered attributes. This method decreases data complexity during the detecting phase. They utilized a hybrid diagonal-gated recurrent neural network to identify hate speech and assess the sentiment conveyed in the phrase. The authors evaluated the feasibility of the FE-DGRNN approach utilizing the HASOC 2019 dataset. The dataset facilitates the multilingual assessment of the approach's efficacy, as it comprises postings in Hindi, English, and German. The simulation results indicated that the accuracy of FE-DGRNN for Task-1, Task-2, and Task-3 on the multilingual code-mixed texts dataset was 87.74%, 88.98%, and 84.74%, respectively. The FE-DGRNN method significantly enhances accuracy, precision, recall, and F-measure relative to existing classifiers, suggesting its potential as a reliable and efficient instrument for sentiment analysis and hate speech detection in multilingual and code-mixed texts.

Khan et al. [6] proposed BiCHAT to train Twitter representations to recognize hate speech. A specialized BiLSTM integrated with a deep CNN and a Hierarchical Attention-based deep learning model was employed. The model utilized tweets as input and implemented an attention-aware deep convolutional layer after the BERT layer. The convolutionally encoded data was subsequently delivered via an attentive bidirectional LSTM network. The researchers evaluated the tweet's rudeness with a Softmax layer. The suggested model surpassed the state-of-the-art (SOTA) and baseline models, with enhancements of 8%, 7%, and 8% in precision, recall, and f-score, respectively. Three benchmark datasets from Twitter were utilized to train and assess the model. BiCHAT has commendable performance in training and validation accuracy, exhibiting improvements of 5% and 9 % respectively. The authors also investigated the influence of various neural network components on the model. Upon examination, they determined that the exclusion of the deep convolutional layer significantly affected the performance of the recommended model. Their investigation encompassed the impact of various activation functions, batch sizes, optimization algorithms, and embedding strategies on the performance of the BiCHAT model.

Sharma et al. [7] indicate that the utilization of automated social media analysis tools is on the rise, prompting worries regarding the reliability of analytics and demonstrating that the mere presence of sarcastic comments can substantially diminish the precision of computerized sentiment analysis. The many NLP algorithms suggested to date are constrained by textual proximity and context, limiting their

applicability across diverse material kinds. The researchers combined three sentence-based models—Unsupervised Learning LSTM-based autoencoder, BERT, and USE—to create an innovative hybrid model. The transformer-based models BERT and USE analyze entire sentences rather than focusing solely on individual words, unlike word-based context. The model was assessed using three separate real-world social media datasets: Twitter, headlines, and the SARC. The accuracy ratings were 83.92%, 90.80%, and 92.80%, respectively.

Ghosh et al. [8] developed a hybrid deep learning approach for the detection of hate speech in Bangla social media. It integrated CNN, LSTM, FastText embedding, and GRU. The experiment utilized a publicly available dataset including 30,000 samples for the model's training and evaluation. The hybrid model developed to detect hate speech in Bangla attained an accuracy of 90%, demonstrating high sensitivity and specificity. Multiple pertinent DL techniques were assessed on the identical dataset; nevertheless, none surpassed the performance of the constructed hybrid model. The resilience of the hybrid model greatly augmented its suitability for hate speech identification.

Murshed et al. [9] developed the DEA-RNN model, a hybrid deep learning framework, to identify cyberbullying. The proposed DEA-RNN model integrates Elman-type RNN with the DEA to improve the parameters of the Elman RNN and decrease training duration. The hybrid model was evaluated using a dataset of 10,000 tweets, and its performance was juxtaposed with that of advanced algorithms including RF, RNN, SVM, Bi-directional LSTM, and Multinomial NB. The findings of the experiment indicate that, across all three testing conditions, the hybrid DEA-RNN model exhibited superior performance. In this case, it has attained an F1-score of 89.25%, specificity of 90.94%, precision of 89.52%, recall of 88.98%, and accuracy of 90.45%.

Aulia and Budi [10] introduced a technique for utilizing machine learning models to autonomously identify hate speech in extensive Indonesian texts. The researchers manually assessed 906 texts, resulting in a new dataset of hate speech from Facebook in Indonesia. The experiment revealed that the SVM classification algorithm, utilizing TF-IDF, word unigrams, character quad-grams, and lexical features, achieved an F1-score of 85%, the highest result obtained. A study used a paradigm for the efficient identification of contentious themes that precede hate speech on Facebook. The researchers employed graph, emotion, and sentiment analysis methodologies to examine messages on prominent Facebook sites. This approach successfully identified web pages that promote hate speech on sensitive topics within the comment sections [11].

Sreelakshmi et al. [12] implemented a model for the automatic detection of hateful content with ML techniques, considering code-mixed social media communications in Hindi and English. The Facebook dataset utilized a pre-trained word embedding library to represent 10,000 samples of hate and non-hate data through FastText. The performance of FastText features surpassed that of doc2vec and word2vec features in comparative analysis. Research on a code-mixed dataset indicates that character-level features yield the most favorable results.

Abuzayed and Elsayed [13] formulated a model for detecting hate speech in Arabic commentary. They investigated two separate word representations and 15 ML techniques, including SVM and CNN. In trials including 8000 labeled datasets, distributed term representation surpasses statistical bag-of-words representation, while neural learning models exceed classical learning approaches. The integrated CNN and RNN architecture surpassed the majority-class baseline, which achieved a macro F1-score of 0.49, by attaining a macro F1-score of 0.73.

Fortuna et al. [14] built a model for detecting hate speech with Multi-Layer Perceptron, Parallel Random Forest, Boosted Logistic Regression, and Extreme Gradient Boosting. The researchers integrated data from many classification systems, encompassing aggression and toxicity, to demonstrate that training on analogous data yielded marginally improved results on Facebook and other social media platforms. The researchers determined that, even for humans, the task of addressing hate speech is intricate and demanding, leading them to conclude that employing more generalized models may be beneficial.

Das et al. [15] used an encoder-decoder-based ML model to classify user comments written in Bengali on Facebook platforms. A dataset of 7,425 Bengali comments, categorized into seven distinct types of hate speech, was utilized for the algorithm's creation and evaluation. The researchers successfully encoded and extracted local information from the statements utilizing 1D convolutional layers. The authors utilized the attention mechanism, LSTM, and GRU-based decoders to predict types of hate speech. The trial yielded a 77% accuracy rate. The results indicate that the attention-based decoder outperformed the other two encoder-decoder methods.

A method for identifying online violence, including hate speech on social media, was introduced [16]. The researchers employed various classifiers, including DL models based on CNN, attention-based models, and Google AI's pre-trained language model BERT. Aggression was classified into three categories: overt aggression, covert aggression, and nonaggression. The authors created a user interface for a web browser plugin to scan inappropriate comments on users' timelines on social media platforms such as Facebook and Twitter. The authors achieved a weighted F1-score of 0.64 with the TRAC Facebook English dataset and

a weighted F1-score of 0.62 with the Hindi dataset. They achieved a weighted F1-score of 0.58 from the English Twitter dataset and an F1-score of 0.50 from the Hindi Twitter dataset.

Similarly, [17] introduced a strategy for identifying foul language online with a Danish dataset composed of user-generated content from Facebook and Reddit. The dataset was annotated to document diverse categories of hostile comments and their respective targets. The authors devised four automatic classification systems, each intended to function for both Danish and English. The classifier's output yielded an impressive macro-averaged F1 score for detecting offensive language in both Danish and English. The experiment demonstrated the efficacy of both English and Danish languages in identifying types and targets of hostile language through an algorithmic approach to detecting various forms of offensive language. The model was constructed using the Baseline, Learned-BiLSTM Classifier, Fast-BiLSTM Classifier, AUX-Fast-BiLSTM Classifier, and Logistic Regression Classifier.

Del-Vigna et al. [18] identified various kinds of hate on Facebook to distinguish among the numerous varieties. In accordance with the defined taxonomy, they annotated crawled comments using input from up to five distinct human annotators. The authors created and executed two text classifiers utilizing morpho-syntactical attributes, word embedding lexicons, and sentiment polarity. The classifiers utilize two separate learning algorithms: SVM for the first and LSTM for the second. They evaluated the two learning algorithms to determine their efficacy in classifying hate speech. The findings illustrate the efficacy of the two classification techniques evaluated on Facebook text that has been manually annotated for hate speech. The SVM achieved an accuracy of 72.95% and an F1-score of 0.79, whereas the LSTM attained an accuracy of 75.23% and an F1-score of 0.081.

CNNs and LSTMs were employed [19] to develop a model utilizing two claim datasets obtained from online user comments. They attained a performance markedly superior to the state of the art for one dataset (which classifies arguments as factual versus emotional) and a performance comparable to the state of the art on the other dataset (which categorizes propositions based on their verifiability) without employing any elaborate, costly feature set. Their approach utilizes a generalized, simple, and efficient methodology suitable for claim categorization across many datasets and tasks. The CNN model utilizing word2vec embeddings achieved a macro average F1 score of 70.47%, whereas the LSTM model with word2vec attained a score of 65.60%.

In [20], an approach utilizing NLP technology was created for detecting objectionable content on social media platforms, specifically Facebook and Twitter. The researchers considered the linguistic and metalinguistic characteristics utilized by the Italian language to differentiate between Twitter and Facebook posts, owing to the distinct usage patterns of the two platforms and the character constraints inherent to Twitter communications. The researchers considered two distinct datasets from two separate online social networks. The researchers employed SVM, one-layer, and two-layer BiLSTM to develop their models. The model was constructed and assessed using datasets from Twitter and Facebook. The F1 scores recorded were 0.8288 for Facebook and 0.7993 for Twitter.

Research indicates that CNNs have achieved remarkable outcomes in numerous NLP applications, including sentiment analysis and text classification. RNNs are particularly adept at capturing the sequential characteristics of language. Consequently, by utilizing the advantages of both CNN and LSTM architectures, including:

The efficacy of CNNs in capturing localized features and patterns in text is advantageous for detecting specific phrases or word combinations that may signify hate speech. LSTMs are proficient in capturing long-range dependencies in sequential data. In the realm of hate speech detection, LSTMs facilitate the comprehension of contextual links among words in sentences, thereby capturing the intricate nuances of hate speech.

CNNs can acquire word-level characteristics by employing filters over several segments of the input text, therefore catching significant local patterns. LSTMs can process word sequences, acquiring representations that encapsulate the sequential characteristics of phrases and paragraphs. This aids in comprehending the overarching context of a text. LSTMs can preserve context and memory across extended sequences, which is essential for comprehending the dynamic context in hate speech detection tasks. CNNs can encapsulate local context and patterns, offering supplementary information to enhance the sequential comprehension of LSTMs.

LSTMs can manage sequences of varying lengths, rendering them appropriate for processing messages of diverse lengths typically encountered on social networking platforms. Incorporating pooling layers in CNNs enables the network to consolidate information from diverse parts of the text, facilitating adaptability to changing text lengths. Both CNNs and LSTMs can utilize dropout and other regularization methods to mitigate overfitting, hence improving the model's generalization capacity in hate speech detection tasks. CNNs can produce feature maps that emphasize particular local patterns within the input text. This facilitates the interpretation of which words or phrases contribute to hate speech detection.

This work proposes a hybrid C-LSTM deep learning model that integrates the advantages of CNNs and RNNs to promptly identify and mitigate incidents of hate speech on Facebook. The suggested model

sought to enhance the accuracy and robustness of hate speech detection on the platform by leveraging local feature extraction through CNNs and sequential context modeling via RNNs. This approach is innovative, as the literature studied and researchers' expertise indicates that no prior study has been conducted on the automatic detection and censorship of hate speech in code-mixed English and Hausa on Facebook before its dissemination to the public.

The subsequent portions of this study are outlined as follows: Section 2 examines the research approach, detailing the data collecting and preprocessing methods, model design, and experimental setting. Section 3 delineates the results and discussion, whereas Section 4 comprises the conclusion of the work.

2. Methods

2.1 Data Collection and Preprocessing

To develop and train the hybrid DL model, a dataset containing 17375 posts and comments was collected from Facebook. The data collection covered posts and comments from numerous critically debatable areas such as religion, racism, sports, and politics. The dataset in Fig. 1 was carefully collected and labeled as either hate speech (1) or non-hate speech (0) by human annotators to ensure accuracy.

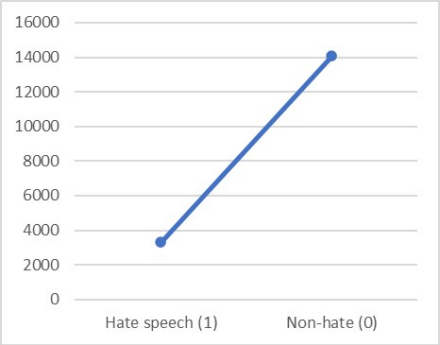


Fig. 1. *Labeled Facebook dataset*

The data depicted in Fig. 1 was labeled according to the definition of hate speech as an overt disparagement of an individual based on recognized protected characteristics, including race, sexual orientation, national origin, religion, caste, sex, gender, gender identity, and serious illness or disability [30]. The dataset comprised 17,375 postings and comments, with 3,294 samples classified as hate speech (1) and 14,081 samples categorized as non-hate speech (0). Example of hate speech and non-hate speech in Hausa-English code-mixed language is provided in Table 1.

Table 1. *Sample of hate speech and non-hate speech*

Label	Text
Hate speech (1)	…most of the people killed in those villages in northern Nigeria are arna whose lives are worthless…
Non-hate speech (0)	…nobody deserves to die in Nigeria regardless of the person's beliefs. Adini ba dole after all…

Preprocessing of the dataset involved the conversion of emojis and emoticons to the various word equivalence, lowercasing, removal of stop words and special characters, tokenization, and vectorization of data (Fig. 3).

2.2 Model Design

The hate speech detection structure in Fig. 2 depicts the actions that can take place in the model to automatically detect and block instances of hate speech and then allow users to edit the text devoid of hate speech.

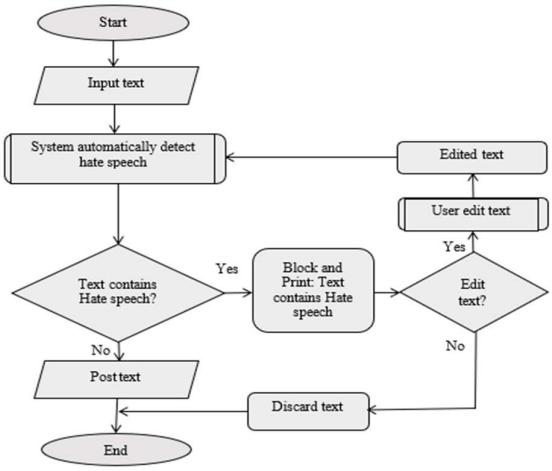


Fig. 2. *The Hate Speech Detection Structure*

2.3 Experimental Setup

The various processes required in creating a well-designed model to automatically identify and block hate speech (Fig. 3). The stages include collecting data and cleaning the datasets, converting text data into features that deep learning algorithms can understand, dividing datasets into train sets and test set, creating the predictive model using a hybrid C-LSTM, and finally evaluating the model using standard evaluation metrics (Accuracy, Precision, Recall, & F1-scores).

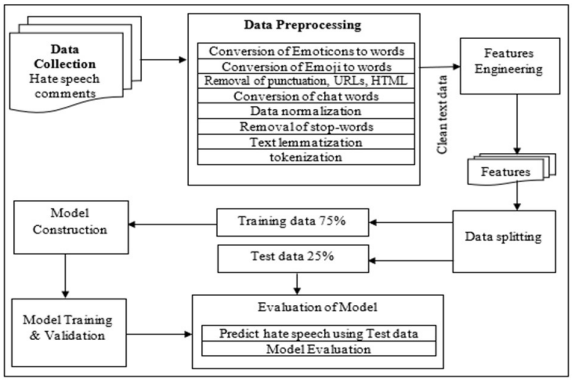


Fig. 3. *Framework for the Hybrid C-LSTM model.*

2.4 Word Representations

Numerous feature representation schemes have been published in the literature to represent text in a way that DL algorithms can comprehend. Bag-of-While using n-grams or specific patterns as features, word feature representation techniques struggle with data sparsity and are unable to fully capture the context of the data [21]. This paper employed word embeddings to address these challenges, encapsulating both syntactic and semantic information in a textual dataset.

Consequently, the hybrid C-LSTM model presented in this research was constructed utilizing pre-trained word embedding features from Global Vectors for Word Representation (GloVe). Word embeddings rely on a vector space model that captures the level of similarity between individual word vectors to provide information on the underlying meaning of words [22]. Numerous research [23] [24] indicate that word embeddings have been effectively employed to categorize offensive and abusive online text. The pre-trained word embedding models Word2Vec, FastText, and GloVe have been effectively employed for text classification [25] [17].

2.5 Hybrid C-LSTM Model Development

2.5.1 CNN Layer

The initial component of the hybrid model comprises a CNN. The CNN analyzes textual data via numerous convolutional and max-pooling layers, allowing the model to acquire hierarchical features from the input

text. In the realm of hate speech identification, CNN analyzes the tokenized textual input via numerous convolutional layers. Each convolutional layer utilizes filters to extract local characteristics from the input text, identifying patterns and signals characteristic of hate speech. Max-pooling layers succeed the convolutional layers, decreasing the dimensionality of the acquired features while preserving the most significant information. This approach enables CNN to discern the most critical elements of the input text for hate speech identification.

2.5.2 LSTM Layer

The subsequent component of our model is an LSTM network. The LSTM effectively captures sequential dependencies in text, enabling it to comprehend the context and intent of the language employed.

2.5.3 Fusion Layer

The outputs from both the CNN and LSTM layers are combined and transmitted through a fusion layer. This layer integrates the acquired features from both components, producing a cohesive representation of the input text that adeptly captures both local and contextual information.

2.5.4 Fully Connected Layer

The consolidated representation from the fusion layer is input into fully connected layers for classification. These layers execute advanced reasoning and decision-making to ascertain the presence of hate speech in the input. This article integrates the benefits of both designs into a unified framework for the detection and prevention of hate speech. CNNs are unable to fully grasp long-range dependencies in input phrases due to the localized nature of their convolutional and pooling layers. This limitation can, however, be effectively surmounted by a singular recurrent layer [26].

2.6 Hyper-parameter Tuning

Hyperparameters, including learning rate, batch size, the quantity of convolutional and recurrent layers, and the amount of hidden units in the LSTM, were optimized utilizing the validation set. Hyperparameter tuning is an essential process in deep learning, as the selection of suitable values can profoundly influence the model's efficacy. Consequently, this research employed four concealed layers, with the initial three triggered by the ReLU activation function. The output layer was activated using a sigmoid activation function. Although the dataset comprises merely two classes, the output layer was constructed using two neurons. The loss function utilized was binary_crossentropy, the activation function employed was sigmoid, the Adam optimizer was applied, the batch size was set to 128, the verbosity level was 1, and the dropout rate was 0.2. The model training was halted after ten epochs due to computational limitations. The Sklearn module in Python was utilized to partition the original datasets into training and testing sets. The datasets were randomly shuffled and divided into a training and validation set comprising 75%, and a test set comprising 25%. Due to computational limitations, the datasets' split method was chosen over the more complex cross-validation methods like k-fold validation. The data split approach was selected with the presumption that the train set's sample size was sufficiently large to prevent the overfitting issue.

The model was trained via supervised learning, which led to the creation of a more competent text classifier. A fully connected dense layer makes up the model classification module. Keras library was used for the implementation of the front end of the hybrid C-LSTM model. The back end was implemented using TensorFlow in Python programming on Jupyter Notebook [27]. The training and testing of the model was done on a laptop computer with the following specifications: hardware- Intel CPU Core i3-4005U at 1.70GHz and 1.70GHz speed, with a 4GB RAM, 500GB hard drive, software- Microsoft Windows 10 operating system, Python 3 on Jupyter notebook.

2.7 Evaluation Metrics

To evaluate the performance of the model using the test set, several evaluation metrics were employed. Commonly used metrics [28] [29] for binary classification tasks like hate speech detection are presented as follows:

Precision: The ratio of true positive predictions (accurately recognized instances of hate speech) to the total number of positive predictions generated by the model.

$$Precision = \frac{TP}{(TP + FP)} \quad (1)$$

Recall: The ratio of true positive predictions to the total number of real positive cases in the test set.

$$Recall = \frac{TP}{(TP + FN)} \quad (2)$$

The evaluations for specific classes are simply averaged to determine the macro average precision and recall.

$$MacroAvgPrecision = \frac{\sum_{C=1}^{C} Precision}{C} \quad (3)$$

$$MacroAvgRecall = \frac{\sum_{C=1}^{C} Recall}{C} \quad (4)$$

The macro F1-Score is the harmonic mean of macro-precision and macro-recall.

$$MacroAvg\ F1 - score = 2 * (\frac{MacroAvgPrecision * MacroAvgRecall}{MacroAvgPrecision^{-1} + MacroAvgRecall^{-1}}) \quad (5)$$

Accuracy: The ratio of accurate predictions (including true positives and true negatives) to the total instances in the test set.

$$Accuracy = \frac{TP+TN}{(TP + TN+FP + FN)} \quad (6)$$

Error Rate: The ratio of the wrong predictions to the total number of instances in the test set.

$$Error\ Rate = \frac{FP+FN}{(TP + TN+FP \quad FN)} \quad (7)$$

Where: True-Positive (TP), True-Negative (TN), False-Positive (FP), False-Negative (FN), and the total number of classes (C).

The demonstration of the hybrid C-LSTM model's evaluation performance is presented in Fig. 4 (the confusion matrix).

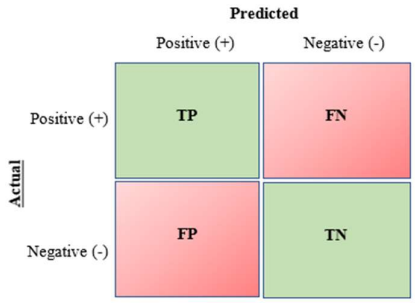


Fig. 4. *Confusion Matrix*

3. Results and Discussion

To evaluate the hybrid C-LSTM model, several metrics were used, including precision, recall, accuracy, weighted average, and macro average f1-scores. However, the macro average f1-score was the particular measure chosen to determine how well the hybrid model can detect hate speech considering the imbalanced nature of the dataset used. Also, extensive testing was conducted on a real-world dataset obtained from Facebook to assess its robustness. The model's ability to leverage the strengths of both CNNs and RNNs proved crucial in detecting hate speech, as it effectively captured both local patterns and contextual information.

The efficiency of the hybrid model was evaluated by applying the trained model to the test set's unseen samples after the models had been trained using samples from the training sets. The evaluation results from the experiments are presented in Fig. 5 showing the actual performance of the developed hybrid of C-LSTM model to detect hate speech on Facebook.

	precision	recall	f1-score	support
0	0.94	0.96	0.95	3702
1	0.63	0.54	0.58	517
accuracy			0.90	4219
macro avg	0.78	0.75	0.76	4219
weighted avg	0.90	0.90	0.90	4219

Fig. 5. *Evaluation Result*

The confusion matrix for evaluating the C-LSTM hybrid model's performance is presented in Fig. 6.

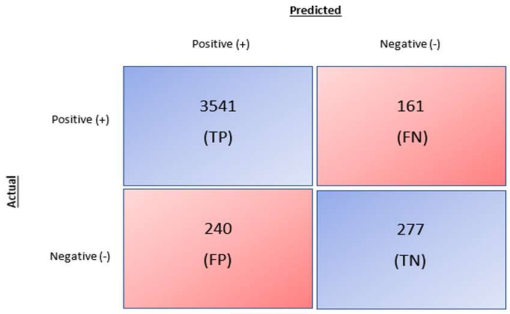


Fig. 6. *The Confusion Matrix*

The evaluation outcome of the hybrid C-LSTM model for hate speech detection, illustrated in Fig. 5, demonstrates impressive efficiency on the test set. The model attained an accuracy of 0.90, signifying its ability to correctly classify a substantial majority of cases, effectively differentiating between hate speech and non-hate speech content. The MacroAvgPrecision score of 0.78 indicates that the model accurately identifies 78% of occurrences classified as hate speech, hence reducing the likelihood of false positive identifications. A MacroAvgPrecision of 0.78 indicates potential for enhancement in the precise identification of hate speech occurrences.

The MacroAvgRecall score of 0.75 signifies that the model identifies 75% of the actual instances of hate speech in the test set. Although a MacroAvgRecall of 0.75 is comparatively high, it indicates that the model continues to overlook certain instances of hate speech, perhaps owing to the intricate and dynamic characteristics of hate speech content, particularly in English-Hausa code-mixed expressions. The macro average F1-score of 0.76 is a balanced metric that accounts for both precision and recall, offering a thorough assessment of the model's efficacy in both hate speech and non-hate speech categories. An F1-score of 0.76 indicates that the model achieves a satisfactory equilibrium between precision and recall, demonstrating competent performance in both areas.

The error rate assessed via (7) from the confusion matrix in Fig. 6, at 0.0950, signifies commendable performance, with roughly 9.5% of instances in the test set misclassified by the hybrid C-LSTM model, which is minimal given the imbalanced dataset context. An accuracy of 0.90 is promising, suggesting that the model may serve as a dependable instrument.

4.1. Comparison of the Hybrid C-LSTM Model with Baseline Models

The performance efficiency of the hybrid C-LSTM model was compared to the baseline models, with findings displayed in Table 2. The efficiency of the hybrid C-LSTM model in the experiment was assessed through accuracy, juxtaposing it with two other deep learning methodologies (CNN and LSTM) as well as baseline models [18] [15]. The DL models in this experiment were trained, assessed, and tested using supervised learning on the Facebook English-Hausa code-mixed dataset, utilizing features from the pre-trained GloVe word embedding model. Fig. 7 presents the findings based on the accuracies of the three models (CNN, LSTM, and C-LSTM) to detect instances of hate speech on Facebook English-Hausa code-mixed comments.

Table 2. *Comparison of the Hybrid C-LSTM Model with Baseline Models*

AUTHOR(S)	MODEL	ACCURACY
[8]	LSTM+GRU	0.77
[11]	SVM+WORD2VEC	0.73
	LSTM+WORD2VEC	0.75
THIS PAPER	CNN+GLOVE WORD EMBEDDINGS	0.89
	LSTM+GLOVE WORD EMBEDDINGS	0.86
	C-LSTM+GLOVE WORD EMBEDDINGS	0.90

The result comparison is presented graphically in Fig. 6 showing efficiencies based on performance accuracies.

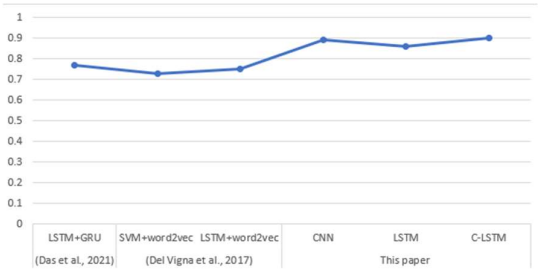


Fig. 7*. Graphical presentation of the model results based on accuracy*

Table 2 presents the results of the CNN, LSTM, and hybrid C-LSTM models using the Facebook English-Hausa code-mixed dataset for training as well as testing. Similarly, Fig.7 has presented the baseline results for comparison with the hybrid model's performance. The hybrid C-LSTM model achieved an accuracy of 0.90, outperforming CNN with 1% and LSTM with 4% in a task to detect hateful speech on Facebook English-Hausa code-mixed dataset. Similarly, considering the performance of the result generated by the hybrid C-LSTM model to identify hate speech, and comparing such performance with the baseline models shows that the hybrid model in this experiment outperformed the LSTM model [16] by 13%, SVM and LSTM [15] with 17% and 15% respectively.

An accuracy of 90% indicates that the hybrid C-LSTM model is an effective instrument for identifying and mitigating hate speech on Facebook, hence fostering a safer online environment. Continuous monitoring and model changes are crucial to mitigate potential difficulties and improve its efficiency.

4. Conclusion

The proposed hybrid C-LSTM deep learning model offers an efficient method for identifying and preventing hate speech on Facebook in code-switched English-Hausa languages. By combining the strength of convolutional, recurrent neural networks, with the Glove pre-trained word embedding characteristics. The hybrid C-LSTM model has demonstrated promising results in hate speech detection on Facebook, with a balanced F1-score with reasonably high accuracy and a minimal error rate, which proposes that the model maintains a suitable balance between precision and recall. The model attained equilibrium between local and sequential linguistic characteristics, markedly improving its capacity to identify hate speech in English-Hausa code-mixed language. The research illustrates the promise of advanced deep learning approaches in combating online hate speech transmission and building a safer and more inclusive online environment. The findings in this study suggest that the suggested model has demonstrated its potential to restrict the usage of hate speech on Facebook using English-Hausa code-mixed languages since the hate speech would be recognized and prevented before reaching the public. This will protect the poster and its potential target from the hazards that may result from such posts. While the model successfully identifies a significant portion of hate speech content, further optimizations, and improvements can be done to enhance precision and recall.

Future research could explore detecting and blocking hate speech in Facebook multimedia content since most toxicity comments these days are propagated via multimedia content.

Compliance with Ethical Standards

Conflicts of interest: Authors declared that they have no conflict of interest.

Human participants: The conducted research follows the ethical standards and the authors ensured that they have not conducted any studies with human participants or animals.

References

[1] H. Simon, B.Y. Baha, E.J. Garba, "Trends in Machine Learning on Automatic Detection of Hate Speech on Social Media Platforms: A Systematic Review", FUW Trends in Science & Technology Journal, vol. 7, no. 1, pp. 1 – 16, e-ISSN: 24085162, 2022a.

[2] M. ElSherie, V. Kulkarni, D. Nguyen, W. YangWang, E. Belding, "Hate Lingo: A Target-Based Linguistic Analysis of Hate Speech in Social Media", In Proceedings of the Twelfth International AAAI Conference on Web and Social Media, pp. 42-51, 2018.

[3] M. Anand, K.B. Sahay, M.A. Ahmed, D. Sultan, R.R. Chandan, B. Singh, "Processing in Computation for Offensive Language Detection in Online Social Networks by Feature Selection and Ensemble Classification Techniques", Elsevier: Theoretical Computer Science, vol. 943, no. 17, pp. 203-218, 2023.
[4] P.N. Fale, K.K. Goyal, K. Shivani, "A Hybrid Deep Learning Approach for Abusive Text Detection", AIP Conference Proceedings, vol. 2753, no. 1, 2023.
[5] P. Kar, S. Debbarma, "Multilingual Hate Speech Detection Sentimental Analysis on Social Media Platforms using Optimal Feature Extraction and Hybrid Diagonal Gated Recurrent Neural Network", Springer - Journal of Supercomputing, 2023.
[6] S. Khan, M. Fazil, V.K. Sejwal, M.A. Alshara, R.M. Alotaibi, A. Kamal, A.R. Baig, "BiCHAT: BiLSTM with Deep CNN and Hierarchical Attention for Hate Speech Detection" Elsevier Journal of King Saud University - Computer and Information Sciences, vol. 34, pp. 4335-4344, 2022.
[7] D.K. Sharma, B. Singh, S. Agarwal, H. Kim, R. Sharma, "Sarcasm Detection over Social Media Platforms using Hybrid Auto-Encoder-Based Model", MDPI Electronics, vol. 11, no. 2844, 2022.
[8] T. Ghosh, A.A.K. Chowdhury, M.H.A. Banna, M.J.A Nahian, M.S. Kaiser, M. Mahmud, "A Hybrid Deep Learning Approach to Detect Bangla Social Media Hate Speech", In Proceedings of International Conference on Fourth Industrial Revolution and Beyond, Springer, Singapore, vol. 437, 2022.
[9] B.A.H. Murshed, J. Abawajy, S. Mallappa, M.A.N. Saif, H.D.E. Al-Ariki, "DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform", IEEE Access, vol. 10, pp. 25857-25871, 2022.
[10] N. Aulia, I. Budi, "Hate Speech Detection on Indonesian Long Text Documents Using Machine Learning Approach", Association for Computing Machinery (ACM), pp. 164-169, 2019.
[11] A. Rodriguez, C. Argueta, Y. Chen, "Automatic Detection of Hate Speech on Facebook Using Sentiment and Emotion Analysis", IEEE, pp. 169-174, 2019.
[12] K. Sreelakshmi, B. Premjith, K.P. Soman, "Detection of Hate Speech Text in Hindi-English Code-mixed Data", In Procedia Computer Science, vol. 171, pp. 737-744, 2020.
[13] A. Abuzayed, T. Elsayed, "Quick and Simple Approach for Detecting Hate Speech in Arabic Tweets", In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a shared Task on Offensive Language Detection, pp. 109-114, 2020.
[14] P. Fortuna, J. Ferreira, L. Pires, G. Routar, S. Nunes, "Merging Datasets for Aggressive text Identification", In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, pp. 128–139, 2018.
[15] A.K. Das, A. Al-Asif, A. Paul, M.N. Hossain, "Bangla Hate Speech Detection on Social Media Using Attention-based Recurrent Neural Network", Journal of Intelligent Systems, vol. 30, pp. 578–591, 2021.
[16] S. Modha, P. Majumder, T. Mandl, C. Mandalia, "Detecting and Visualizing Hate Speech in Social Media: A Cyber Watchdog for Surveillance", Expert Systems with Applications, vo. 113725, 2020.
[17] G.I. Sigurbergsson, L. Derczynski, "Offensive language and Hate Speech Detection for Danish", , 2019.
[18] F. Del-Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, M. Tesconi, "Hate Me, Hate Me Not: Hate Speech Detection on Facebook", In Proceedings of the 1st Italian Conference on Cybersecurity", pp. 86–95, 2017.
[19] C. Guggilla, T. Miller, I. Gurevych, "CNN- and LSTM-based Claim Classification in Online User Comments", In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, pp. 2740–2751, 2016.
[20] C. Bosco, F. Dell'Orletta, F. Poletto, M. Sanguinetti, M. Tesconi, "Overview of the EVALITA 2018 Hate Speech Detection Task", 2018.
[21] F. Almeida, G. Xexeo, "Word Embeddings: A Survey. Computation and Language", Journal of Computation and Language, pp. 1–10, 2019.
[22] V.Q. Lee, T. Mikolov "Distributed Representations of Sentences and Documents", Computation and Language, 2014.
[23] L.S. Altin, A. Bravo, H. Saggion, "LaSTUS/TALN at SemEval-2019 Task 6: Identification and Categorization of Offensive Language in Social Media with Attention-based Bi-LSTM model, In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, Minnesota, USA, pp. 672–677, 2019.
[24] A. Umar, S. A. Bashir, L.C. Ochei, and I. A. Adeyanju, "Profiling Inappropriate Users' Tweets Using Deep Long Short-Term Memory (LSTM) Neural Network", i-manager's Journal on Pattern Recognition, vol. 5, no.4, pp. 27-43, 2019.
[25] S. Agrawal, A. Awekar, "Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms", Springer, BBC News Article, pp. 141–153, 2018.
[26] A. Hassan, A. Mahmood, "Deep learning approach for sentiment analysis of short texts", In Proceedings of the 3rd international conference on control, automation, and robotics (ICCAR-2017), pp. 705–710, 2017.
[27] H. Simon, B.Y. Baha, E.J. Garba, "A Multi-platform Approach Using Hybrid Deep Learning Models for Automatic Detection of Hate Speech on Social Media", Bima Journal of Science and Technology, vol. 6, no. 2, pp. 2536-6041, 2022b.
[28] P. Kapil, A. Ekbal, "A Deep Neural Network-Based Multi-Task Learning Approach to Hate Speech Detection", Knowledge-Based Systems. Pp. 106458, 2020.
[29] M. Grandini, E. Bagli, G. Visani, "Metrics for Multi-Class Classification: An Overview", Statistics Machine Learning, 2020.
[30] Facebook, "Community Standards", 2019. https://www.facebook.com/communitystandards/objectionable_content.