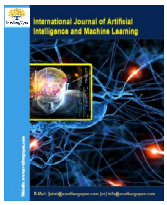
https://doi.org/10.51483/IJAIML.6.1.2026.1-17 ISSN: 2789-2557

SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/



Research Paper Open Access

Modeling Multi-Class Mal-Information Detection: A Comparative Analysis of Machine Learning and Deep Learning Approaches

Andualem Woldegiorgis[1*], Mohammed Abebe[2], Durga Prasad Sharma[1,3] and Worku Jimma[4]

[1]Faculty of Computing and Software Engineering, Arba Minch University, Ethiopia. E-mail: andualemwoldegiorgis@bongau.edu.et
[2]Faculty of Computing and Software Engineering, Arba Minch University, Ethiopia. E-mail: mohammed.abebe@amu.edu.et
[3]Expat Professor, Arba Minch University, Ethiopia. E-mail: dp.shiv08@gmail.com
[4]Faculty of Computing and Informatics, Jimma University, Ethiopia. E-mail: worku.jimma@ju.edu.et

Article Info

Volume 6, Issue 1, January 2026
Received : 27 September 2025
Accepted : 21 December 2025
Published : 20 January 2026
doi: 10.51483/IJAIML.6.1.2026.1-17

Abstract

The rapid growth of social media as a platform for communication and information sharing has raised concerns about its negative impact on social cohesion and peace. Harmful online content can fuel intergroup hatred, violence, mass killings, and deepen social and political polarization that promote prejudice and hostility. Therefore, detecting and mitigating harmful content like hate, offensive speech, and harassment is critically important. Amharic, Ethiopia's official working language, is widely spoken across a country rich in diverse religions, ethnicities, and cultures. However, research on harmful content detection remains limited due to scarce linguistic resources, small datasets, and limited adoption of advanced technologies. This study aims to improve Amharic mal-information detection through robust multi-class classification tasks. A dataset of 13,683 Amharic texts was used to train and evaluate on machine learning models such as (RF, SVM, LR, NB), as well as DL models such as (CNN, LSTM, BiLSTM, CNN-LSTM, RNN, and ensemble models). Feature engineering, i.e., Bag-of-Words, TF-IDF, GloVe, FastText, and Word2Vec, was applied. The results show that CNN, stacking ensemble, and RF achieved 94% accuracy, followed by BiLSTM, SVM, and LSTM. Future research should focus on scalable, well-annotated, multimodal, and explainable AI to address the dynamic nature of social media.

Keywords: *Mal-information, Hate speech, Social media, Offensive speech, Machine learning, Deep learning, Amharic*

1. Introduction

Nowadays, social media platforms have become an essential part of our daily lives, linking lots of people across the globe. They offer venues for communication, information exchange, and community involvement.

* *Corresponding author: Andualem Woldegiorgis, Faculty of Computing and Software Engineering, Arba Minch University, Ethiopia. E-mail: andualemwoldegiorgis@bongau.edu.et*

Despite social media's advantages, there is a rising worry about the spread of misinformation, disinformation, and mal-information. One of mal-information classes is hate speech, which refers to any form of communication (written, spoken, or symbolic) that encourages violence, prejudice, and opposition towards individuals or groups based on attributes including race, ethnicity, religion, gender, sexual orientation, and other protected characteristics (Rascão, 2020; Cohen-Almagor and Stamile, 2021). Additionally, hate speech is a powerful term that fuels intergroup hatred, mass killing, and even genocide, promotes opposition, and expands divisions (Megersa and Minaye, 2023). It also creates an environment where prejudice booms, destroying societal harmony and separating a polarized, dangerous world for future generations (Vasist *et al.*, 2024). Hate speech remains a concept that lacks a universally recognized definition, and the academic discourse reflects this ambiguity, as evidenced by a range of definitions provided by leading scholars in the field (Anderson and Barnes, 2023; Jahan and Oussalah, 2023a; Alatawi *et al.*, 2021; Jahan and Oussalah, 2023b; Alhejaili, 2025). Ethiopia released a new proclamation on the prevention and suppression of hate speech and disinformation in March 2020. According to the law, any statement that provokes hatred, discrimination, or violence against people or groups because of characteristics including race, gender, religion, ethnicity, or disability is considered hate speech under the law (The Federal Democratic Republic of Ethiopia, 2020). The general definition provided by the United Nations' policy defines hate speech as any communication spoken, written, or behavioral that attacks or discriminates against a person or group based on their identity, including factors like religion, ethnicity, nationality, race, color, descent, gender, or other characteristics (United Nations & Strategy and Plan of Action on Hate Speech, 2023).

The cultural and ethnic variety of Ethiopia is reflected in its multilingual landscape, comprising more than 80 languages. Amharic, a Semitic language, serves as the official working language across various Regional States of Ethiopia. Written using Geez script, it comprises 275 alphabets, primarily consisting of consonant-vowel pairs and a "Fidel" writing system with 34 base characters and six supplementary characters (Gebremichael *et al.*, 2022; Abate *et al.*, 2020; Bakala and Kekeba, 2021).

Mal-information is spreading harmful truths or falsehoods that aim to harm and threaten individuals, social groups, organizations, or countries, such as hate speech and harassment (Aïmeur *et al.*, 2023; Islam *et al.*, 2020; Bockting *et al.*, 2023; Hutson, 2023). The propagation of hate speech on the internet had a profound effect on people's lives, communities, and societies. In addition to undermining the core values of diversity, equality, and inclusion, hate speech also maintains prejudice, stereotypes, and societal divides. The enormous volume and rapidity of material creation on social media platforms are one of the main obstacles to combating hate speech. Since millions of posts, comments, and messages are exchanged daily, and therefore it is nearly challenging for human or manual moderators to evaluate and filter each instance of hate speech manually. To overcome such challenges, automatic techniques for identifying hate content and preventing hate speech are extensively needed for early warnings.

Numerous studies have been conducted on Amharic language text mining and hate speech detection (Abate *et al.*, 2020; Bakala and Kekeba, 2021; Mossie *et al.*, 2018; Tesfaye and Kakeba, 2020). These studies focus on detecting hate speech in social media texts employing various machine learning and deep learning models to improve accuracy and efficiency. However, Amharic remains under-resourced, lacking essential NLP tools, which limits the development of advanced language applications. Challenges such as character redundancy, similar-sounding words, spelling variations, inconsistent abbreviations, and borrowed terms further complicate Amharic text classification (Neshir *et al.*, 2021). Existing studies (Tesfaye and Kakeba, 2020; Abebaw *et al.*, 2022; Debele and Woldeyohannis, 2022) mainly rely on limited textual data and focus on binary classification, often neglecting the potential of multimodal approaches (e.g., audio, video, and images) for hate speech detection in Ethiopian languages. In addition, shortage of publicly available datasets, annotated stop words, and linguistic dimensions. Most research to date uses unimodal, text-based methods to classify offense, abuse, hate, or harassment, limiting broader application. This study addresses these gaps by introducing a comprehensive, annotated multiclass dataset for Amharic and evaluating a wide range of machine and deep learning models. It offers empirical insights into effective models for detecting hate and offensive speech in Ethiopia's multicultural, multi-ethnic context. By incorporating social media factors and applying unimodal multiclass classification, the study contributes significantly to the detection of mal-information (hate speech, offensive speech, harassment). Additionally, it emphasizes the importance of stop word development and covers the way for future

research on multilingual information disorder detection through advanced models. The main aim of the study is to investigate and develop a deep learning model for multiclass hate speech detection on social media, focusing on Amharic. It evaluates state-of-the-art solutions, comparing machine learning and deep learning models and ensemble learning to identify the most effective approach. A dataset of 13,683 text samples and 985 stop words was collected from Facebook, Telegram, and Twitter. Additionally, 364 hate speech keywords related to gender, race/ethnicity, and religion were identified and categorized into four classes such as Hate Speech and offensive, Normal (Non-Hate Speech), Racial/Ethnic Hate Speech), and Religious Hate Speech, to support the analysis and serve as word or sentence indicators. The study also reviews and evaluates up-to-date approaches to hate and offensive speech detection, comparing the performance of traditional machine learning, deep learning, and ensemble learning models. Therefore, this study attempts to address some of the following important questions:

RQ1: What are the existing state-of-the-art alternative solutions viable to adopt existing mal-information detection?

RQ2: What are the most effective multi-class machine learning and deep learning models for detecting mal-information?

RQ3: What contextual research gaps exist for future work in linguistic localization of scalable social media datasets and in developing context-aware deep learning models capable of handling multimodal, multilingual, and multi-class classification tasks?

This study contributes by providing an extensive analysis of state-of-the-art Machine Learning (ML) and Deep Learning (DL) models applied to mal-information and disruptive content on social media. It synthesizes evaluation metrics and benchmark datasets to identify the most effective approaches for fine-tuning in real-world scenarios. Proposed a scalable dataset, prepared Amharic stop words, and developed a novel multi-classification model that can accurately classify abusive information into respective classes. Furthermore, the study proposes a deep learning model that achieves superior accuracy in the textual multi-class classification of mal-information. It also presents localized research gaps, including the scarcity of linguistic, multimodal, multilingual, standardized, and well-annotated datasets, and recommends scalable data collection and annotation pipelines to inform and guide future research efforts.

2. Literature Review

One commonly utilized technique for hate speech identification is a word meaning disambiguation (Geetanjali and Kumar, 2025). Topic modeling (Araque and Iglesias, 2022; Chhabra and Vishwakarma, 2023), dictionary-based and part-of-speech approaches (Chiche and Yitagesu, 2022), rule-based and bag-of-words approaches (De Santis *et al.*, 2025), and n-grams (Mohd and Awang, 2021) and TF-IDF (Term Frequency-Inverse Document Frequency) (Mohd and Awang, 2021) are commonly employed techniques that consider the occurrence and arrangement of words or subsequences of words in a text. These approaches analyze the frequency and co-occurrence patterns of specific words or groups of words associated with hate speech, enabling the identification of potentially offensive or hateful content.

The study (Ertel, 2018) claims that the detection of hate speech has become a significant issue in recent years, driven by the availability of vast amounts of data and advancements in text processing techniques. Other studies (Sevani *et al.*, 2021; Ertel, 2018; Alpaydin, 2020) revealed that machine learning models for hate speech detection can be supervised, unsupervised, or semi-supervised. Supervised learning is most common, relying on annotated data to train prediction models. Common algorithms include SVM, Random Forest, Naïve Bayes, Logistic Regression, Decision Tree, LSTM, CNN, and K-Means clustering.

The authors in (Faris *et al.*, 2020) put efforts into hate speech detection for Arabic using word embedding and deep learning models. They recommended recurrent convolutional networks, which include LSTM and convolutional network layers. In this paper, word embedding was implemented using the Word2Vec and AraVec packages and performance evaluation metrics. The study concluded that AraVec achieved better outcomes, with a precision of 68.965%, an accuracy of 66.564%, a recall of 79.768%, and an F1 measure of 71.688%.

The authors in (Aljohani *et al.*, 2024) proposed a hybrid CNN, Random, and grey wolf optimizer to identify hate speech in social media interactions in the Arabic language. The four steps in the system's operation

include preprocessing, encoding, and model training by machine learning, and evaluating the trained model. The experiment collected and trained the models on separate datasets of 5846 and 6023 tweets classified into three classes, i.e., hateful, abusive, and normal content. The result shows almost the same result on both datasets and achieved a high accuracy result of 97.83%, F1-score of 97.83%, and precision of 97.84%.

The study (Al-Hassan and Al-Dossari, 2022) explored and analyzed hate speech detection in Arabic tweets, classifying them into five categories: non-hate, religious, racial, sexist, and general hate. Using 11,000 annotated samples, the study compared an SVM baseline with four deep learning models. The finding showed that a CNN-LSTM model achieved the best performance with 72% precision, 75% recall, and a 73% F1 score. Another study (Vo *et al.*, 2025; Huu *et al.*, 2019) addressed automated hate speech detection on Vietnamese social networks using a multi-class model trained on 25,431 samples, classifying content as hate, offensive, or clean. Various algorithms were tested, with Logistic Regression performing best, achieving an F1-score of 67.76% using k-fold cross-validation.

A very important study (Ruwandika and Weerasinghe, 2018) explored hate speech detection combined with machine learning and lexicon-based methods for English social media hate speech detection. Among five models tested, Naïve Bayes with TF-IDF achieved the best F-score (0.719). Supervised models outperformed unsupervised ones, though K-Means showed promise with limited features. Another study (Kocoñ *et al.*, 2021) conducted a comprehensive literature review on cyberbullying on social networking sites, examining its characteristics, research methods, and theoretical foundations. Using social cognitive theory, the authors developed an integrated framework highlighting the roles of perpetrators, victims, and bystanders and identified gaps for future research. In study (Arshad *et al.*, 2023), researchers explored feature engineering, machine learning, and deep learning for detecting offensive and hateful speech in Urdu. Using a dataset of 7,800 tweets, they applied transfer learning with FastText and multilingual BERT (RoBERTa). The RoBERTa-based multi-class model outperformed baseline methods, achieving a macro F1-score of 0.82. Other researchers in a study (Mehmood *et al.*, 2024) developed a hate speech detection model for Roman Urdu using a newly annotated dataset of 30,000 samples. The researchers applied deep learning models, including Bi-LSTM, LSTM, and CNN. The proposed Bi-LSTM model achieved the best performance with 87.5% accuracy and an F-score of 0.885.

Researchers in the study (Mossie *et al.*, 2018) focused on Amharic hate speech detection using 6,120 social media post comments. They employed Word2Vec and TF-IDF for feature selection, with Naïve Bayes and Random Forest classifiers achieving accuracies of 79.83% and 65.34% respectively. While the study contributed to the field, it was limited by a small dataset and a lack of stop word handling and dimensionality considerations. In another study (Tesfaye and Kakeba, 2020). Researchers developed an automated detection model using recurrent neural networks. They compiled a labeled dataset from Facebook posts and comments by activists, applied word n-grams and Word2Vec for feature extraction, and trained LSTM and GRU models. The RNN-LSTM model achieved a high accuracy of 97.9%.

Another study (Debele and Woldeyohannis, 2022) explored hate speech on social media as a significant issue, causing discrimination and violating human rights. This study aimed to identify Amharic hate speech using a multi-modal approach, combining audio and textual elements using a deep learning model. The researchers collected 1,459 extracted audio files from YouTube videos and used the Google Speech-to-Text API to transliterate them into text scripts. The experimental results showed that the multi-modal model with BILSTM outperformed in detecting Amharic hate speech with an accuracy of 88.15%. An alternative multi-channel CNN model was proposed in (Abebaw *et al.*, 2022), investigated to use multi-channel CNN models for under-resourced languages with an inadequate training dataset. The researchers used a limited dataset, having 2000 labeled social media comments. The results showed that the proposed multi-channel CNN model outperformed the single-channel CNN models but underperformed the baseline Support Vector Machine (SVM). Finally, the findings suggested that the MC-CNN model could be used as a substitute solution for hate speech detection in cases of dataset scarcity. A study (Ganfure, 2022) explored hate speech detection in the Ethiopian Afaan Oromo language. The researcher collected data from different posts and comments from the 35,200 Afaan Oromo dataset. The study was modeled using deep learning models, including CNN, LSTMs, BiLSTMs, LSTM, GRU, and CNN-LSTM. The results show that CNN and Bi-LSTM excel beyond all the other investigated models, with an F1-score of 87%. Another study (Ababu and Woldeyohannis, 2022) explored the

detection and classification of hate speech in Afaan Oromo language. The study applied classical, ensemble, and deep learning algorithms with feature engineering techniques, i.e., BoW, TF-IDF, Word2Vec, and Keras, on a dataset of 12,812 instances. BiLSTM with pre-trained Word2Vec outperformed other models, achieving 84% accuracy for eight-class classification and 88% for binary classification.

The review revealed that most existing studies on Mal-information in Ethiopian languages rely heavily on unimodal, text-based analysis and focus primarily on binary classification. There is a significant lack of integration of multiple data modalities, such as audio, video, images, and memes, which are increasingly used in online communication. Additionally, the studies often lack public datasets, annotated stop words, and support for multi-class classification, which limits their applicability and generalizability. This study addresses these limitations by supporting a comprehensive, multi-class classification approach tailored to the detection of various forms of harmful content, hate speech, harassment, and offensive language, within the context of Ethiopian languages, particularly Amharic. The goal is to develop a robust and accurate model that enhances linguistic localization and contributes meaningful advancements to the field of Mal-information detection.

3. Research Design and Methodology

This study followed a structured methodology that began with problem identification, a comprehensive literature review, and an evaluation of related work. This was followed by identifying research gaps, collecting and analyzing data, developing the model, and finally presenting and discussing the results. Sources included Scopus (18), IEEE Xplore (10), Web of Science (2), Springer (6), ACM Digital Library, and Google Scholar (12), and other covering studies published between 2018 and 2025. The main keywords used were "Mal-information", OR "Hate Speech" OR "Harassment" OR "Offensive Speech", OR "Deep Learning" OR "Machine Learning", "Social Media".

3.1. Data Collection and Preprocessing

Hate speech detection presents a major challenge in the fields of Natural Language Processing (NLP), machine learning, and deep learning, requiring researchers and developers to apply a range of advanced techniques to correctly classify and address occurrences of hate speech. There are different large number of publicly available datasets for foreign languages such as English (Ali *et al.*, 2022), European, and Asian (Aluru *et al.*, 2020). Amharic language, being a widely spoken and resourced language within Ethiopia, remains significantly underrepresented in publicly accessible datasets for hate speech detection (Ayele *et al.*, 2022). This gap hinders research and the development of effective classification models. To overcome this challenge, the study collected 13,683 samples of Amharic social media content from July 2023 to September 2025. Sources included prominent national and regional media outlets (e.g., EBC, FBC, OBN, OMN, VOA, EPA, AMECO, Abbay TV) as well as content from public figures, activist platforms, journalists, and accounts with over 10,000 followers to ensure a broad and balanced representation of perspectives. The dataset was curated to reflect various forms of hate speech in Amharic, covering offensive, hateful, and harassing language. To support more accurate text analysis, 985 commonly used Amharic stop word terms with minimal semantic value were identified and excluded from the processing pipeline. For model development, the dataset was preprocessed and divided into 80% training and 20% testing sets using stratified sampling, preserving the proportional distribution of hate speech classes. This methodological approach enabled the creation of balanced and representative data splits, improving model generalization and performance in detecting hate speech through different categories.

3.2. Description of Dataset

The dataset has four classes, speech types were numerically labeled as follows: Hate Speech and offensive = 1, Normal Non-Hate Speech = 2, Racial/Ethnic Hate Speech) = 3, and Religious Hate Speech = 4. In total, 13,483 Amharic texts were collected and categorized into four distinct classes, as shown in Table 1. Among these, 2,291 instances are general hate speech, featuring offensive or threatening language directed at individuals or groups. Non-hate speech includes 7,440 instances of neutral, positive, or unrelated content. Racial or ethnic hate speech accounts for 2,651 instances, involving discriminatory language against specific groups, while 1,301 instances represent religious hate speech targeting particular beliefs or faiths. To balance the dataset and to improve model fairness, an inverse class frequency weight-balancing technique was applied during model training, assigning higher weights to the minority class. This method helps prevent the model from

being biased toward the majority class and enhances its ability to detect patterns in underrepresented categories. During this stage, 985 frequently used Amharic stop words with minimal semantic value were removed to reduce noise and improve data quality. This detailed classification enables a deeper understanding of different hate speech types, supports more focused mitigation strategies, and fosters efforts toward building a more inclusive and respectful digital environment. Additionally, this approach enables stakeholders to effectively address specific issues related to gender, race, ethnicity, religion, and other characteristics.

Table 1: Distribution of the Amharic Hate and Offensive Speech Dataset

Language	Hate and Offensive		Non-Hate Speech (Neutral)	Total
Amharic	General Hate and Offensive, and Gender-Based	2,291	7,440	13,483
	Racial/Ethnic Based Hate Speech	2,651		
	Religious Hate Speech	1101		
Total		**6,043**	**7,440**	**13,483**

3.2.1. Performance Evaluation Metrics

We evaluated the proposed models using standard performance metrics, including accuracy, recall, precision, and F1-score, which were calculated based on the formulas provided in Equations (1) to (4). In addition, the task of hate speech identification is approached as a classification problem, aiming to determine whether a posted comment is hateful, offensive, and gender-based, Racial/Ethnic hate speech, or religious hate speech. To assess the performance of hate speech detection, a confusion matrix is utilized. A concise description of the associated metrics is provided below.

$$Accuracy = \frac{TP + TN}{TP + TF + FP + FN} \quad ...(1)$$

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

$$1\,Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad ...(3)$$

$$Precision = \frac{TP}{TP + FP} \quad ...(4)$$

Where, True Positive (TP) = Hate, offensive, and gender-based speech predicted as Hate; True Negative (TN) = Real posts/comments predicted as Non-Hate Speech; False Positive (FP) = Hate, offensive, and gender-based hated posts predicted as Non-Hate Speech (Normal); False Negative (FN) = Non-Hate Speech (Normal) posts predicted as Hate.

4. Results

This section presents a comprehensive analysis of hate speech detection in the Amharic language texts. This paper performed an extensive review and analysis of up-to-date research studies in the domain that address issues and challenges in the problem of hate speech detection. The study focused on evaluating and analyzing hate and offensive speech detection by selecting existing machine learning and deep learning algorithms with their respective performance evolution metrics and developing a robust detection model by using tools, i.e., Tensor Flow, Pandas, Scikit-learn, Anaconda, and Google Colab Cloud Service. In addition, the experiment was modeled using four machine learning models, i.e., Random Forest, SVM, Logistic Regression, Naïve Bayes, five deep learning models (CNN, LSTM, BiLSTM, CNN-LSTM, RNN), two ensemble learning models (Voting, Bagging, Stacking), and Feature Engineering techniques (GloVe, FastText, Bag-of-Words, TF-ID). The evaluation metrics used in this study, accuracy, precision, recall, and F1 score, provide valuable insights into the effectiveness of the models in classifying hate speech in multi-classes, i.e., General Hate, Non-Hate Speech, Racial/Ethnic, and Religious Hate. The applied hyper-parameters are detailed in Table 2, while Table 3

Table 2: Hyper-Parameter Selection

Parameter	Search Space	Selected Value
Embedding Dimension	128, 256, 512	128, 256
Activation Function	Relu, Softmax	Relu, Softmax
Loss Function	Binary_crossentropy, Categorical_crossentropy	Categorical_crossentropy
Dropout Rate	0.1, 0.2, 0.3	0.2, 0.3
Learning Rate	1e-3, 5e-4, 1e-4	1e-3 (0.001),
Epochs	5, 20, 25, 30	5
Optimizer	Adam, AdamW, SGD	Adam
Batch Size	16, 32, 64	32

summarizes the performance and evaluation results of the selected machine learning, deep learning, and ensemble learning models.

The experimental result, as depicted in Table 3 and Figure 1, shows that the CNN model achieved the highest overall performance with an accuracy, precision, recall, and an F1-score, showing that it efficiently captured key textual features in the dataset. The Stacking Ensemble and Random Forest models performed almost equally well, reflecting their strength in combining multiple classifiers and handling feature variability efficiently. The BiLSTM model also demonstrated strong performance, showing its ability to capture sequential and contextual dependencies in the data, though it slightly underperformed compared to CNN and ensemble approaches. Models such as the Voting Ensemble and SVM achieved moderate results, suggesting stable but less powerful generalization capabilities. Other machine learning models like Logistic Regression, Bagging Ensemble, and LSTM yielded fair results, achieving accuracies between 0.83 and 0.85, likely limited by their simplicity in handling complex linguistic relationships. In addition, Naïve Bayes, CNN-LSTM hybrid, and RNN performed poor accuracy, which indicates issues such as under-fitting, vanishing gradients, or ineffective architecture design. In general, the results show that deep convolutional

Table 3: Model Evaluation Results

Model	Accuracy	Precision	Recall	F1-Score
CNN	0.943905	0.944196	0.943905	0.943906
Stacking Ensemble	0.943602	0.94374	0.943602	0.94349
Random Forest	0.943299	0.943985	0.943299	0.943263
BiLSTM	0.934506	0.935407	0.934506	0.934565
Voting Ensemble	0.902668	0.904436	0.902668	0.902791
SVM	0.896301	0.897349	0.896301	0.896468
Logistic Regression	0.847483	0.851012	0.847483	0.847109
Bagging Ensemble	0.836264	0.840517	0.836264	0.835813
LSTM	0.834445	0.839674	0.834445	0.835576
Naïve Bayes	0.766828	0.790003	0.766828	0.765508
CNN-LSTM	0.51789	0.414898	0.51789	0.404475
RNN	0.338387	0.323862	0.338387	0.197331

and ensemble-based models are the most effective for this task, while simpler or unstable recurrent architectures struggle to achieve competitive.

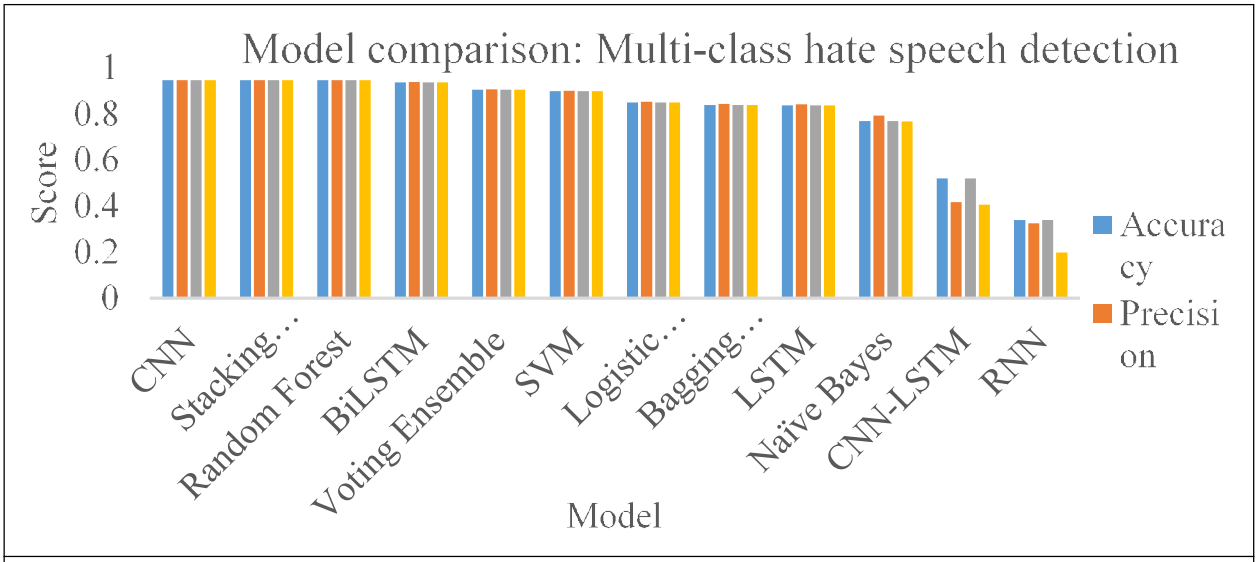


Figure 1: Summary of Model Evaluation Results

Table 4: Summary of the Existing State of the Art Related Works, Key Contribution and Key Gaps

Ref.	Year	Problem Addressed	Dataset Used	Data Types	Techniques Used	Evaluation Metrics		Key Contribution	Research Gaps
						ACC	F1		
Bakala Defersha and Kekeba Tune	2021	Hate Speech (Afaan Oromo)	Afaan Oromo Hate Speech Detection	Text	ML (SVM, NB, LR, DT, and RF)	67%	64%	The study demonstrated that ML models can effectively identify Afaan Oromo hate speech.	The study achieves lower accuracy and F1 score. Lacks analysis and detailed comparison with existing state-of-the-art studies, and fails to integrate multimodal and conduct bi-classification.
Mossie *et al.*	2018	Hate Speech (Amharic)	Custom Dataset	Text	Machine Learning (NB, and RF)	79.83%	65.34%	The study uses Apache Spark and machine learning algorithms (Random Forest and Naïve Bayes) to classify posts as 'hate' or 'not hate'.	Lacks analysis with other ML models and a detailed comparison with existing state-of-the-art studies. Lacks integrate multimodal and conduct bi-classification, and a Small dataset size.
Tesfaye and Kakeba	2020	Hate Speech (Amharic)	Custom Dataset	Text	Deep Learning (LSTM and GRU)	97.9%	-	The study developed an LSTM-based model for detecting hate speech in Amharic Facebook posts. The GRU model, although effective, performed at a lower rate of 88%.	The study fails to address the consideration of long sentences with and without comparison with BiLSTM. There is a lack of publicly available datasets and a compression of existing deep learning models.

Table 4 (Cont.)

Abebaw *et al.*	2022	Hate Speech (Amharic)	Custom Dataset	Text	Deep Learning (CNN)	An F1 score of 81.3%		The study proposes a multi-channel CNN (MC-CNN) that outperforms traditional single-channel CNNs, with the baseline single-channel CNN achieving 78.2% accuracy and the SVM model achieving 92.5%.	Small dataset sizes are used and fail to generalize; The study only considers the F1 score rather than other performance evaluation metrics; Lacks integration of multimodal and conducts bi-classification; Lacks a detailed analysis proposed model.
Debele and Woldeyohannis	2022	Hate Speech (Amharic)	Custom Dataset	Audio and Texts	DL (LSTM, BiLSTM, GRU, and BiGRU)	The accuracy of LSTM and BiLSTM is 84% and 88.15%, respectively		The study uses deep learning to detect hate speech in Amharic.	Uses a small dataset size and fails to generalize. Lacks to integrate multimodal and conduct bi-classification. The study did not address any other offensive speech.
Aljohani *et al.*	2024	Hateful and abusive	L-SHAB and T-HSAB	Text	CNN with an attention mechanism and (GWO)	97.83%	97.83%	Proposed a hybrid model combining CNN, attention, and optimized RF, achieving better performance than individual models and classifiers.	The study fails to address the consideration of long sentences with and without comparison with BiLSTM; it lacks detailed analysis with other deep learning models; Performed only on textual.
Ababu *et al.* and Ganfure	2022 and 2025	Malinformation Hate Speech (Afaan Oromo)	Custom Dataset	Text	CNN, BiLSTM, CNN-BiLSTM, and BiGRU	78.05%	78%	The proposed model effectively detects and classifies offensive speech.	The study lacks focus on bilingual hate speech detection in Ethiopian languages, particularly Amharic and Afaan Oromo, and fails to address the complexities of code-mixing in bilingual communication.
Aljohani *et al.*	2024	Malinformation (Hateful and Abusive)	L-SHAB and T-HSAB	Text	CNN with an attention mechanism and (GWO)	97.83%	97.83	The study proposed a hybrid approach combining a CNN with an attention mechanism and an optimized RF, demonstrating superior performance.	The study fails to address the consideration of long sentences with and without comparison with BiLSTM. Lacks detailed analysis with other deep learning models. Performed only on textual data.

Table 4 (Cont.)

Alatawi *et al.*	2021	Malinformation (Hateful)	Twitter and Stormfront	Text	BERT and BiLSTM	0.84	0.80 $ 0.75	A study addressed the detection of white supremacy in the English language, distinguishing between white supremacist and non-white supremacist.	Lacks the integration of other modalities, limited scope on one language, and does not discuss the long-term effectiveness of the proposed models.
Navya *et al.*	2025	Malinformation (Homophobic and Transphobic)	LT-EDI-EACL	Text	BERT and RoBERT	-	0.95%	The study demonstrates a transformer-based model that effectively classifies comments as homophobic, transphobic, or non-anti-LGBT+ in both English & Malayalam.	The study lacks detailed analysis results with a comparison between other existing models.

As depicted in Table 4 and Figure 2, research studies on mal-information, specifically hate speech, offensive speech, and homophobic and transphobic detection, reveal several common research gaps. Many analyses, such as those from (Bakala and Kekeba, 2021; Mossie *et al.*, 2018), demonstrate lower accuracy and lack detailed comparisons with state-of-the-art techniques. The studies have a limited research effort on the unification of multimodal and multi-class classification. Other studies in (Tesfaye and Kakeba, 2020; Abebaw *et al.*, 2022) did not consider long sentences, had limited small dataset sizes, and focused primarily on singular evaluation metrics like F1 score. Additionally, studies in (Navya *et al.*, 2025; Alatawi *et al.*, 2021) present limitations in language coverage and the lack of discussions on long-term effectiveness, emphasizing the need for more comprehensive evaluation methods in future research. In general, the meta-analysis results revealed that deep learning models achieved higher accuracy compared to traditional machine learning models.

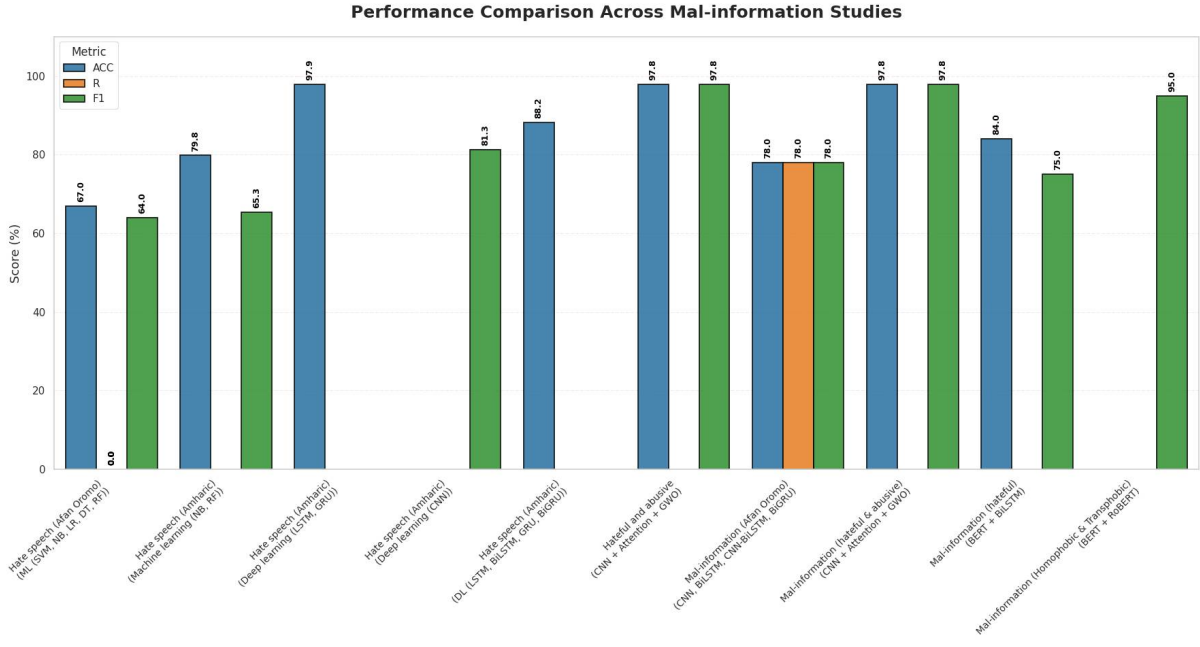


Figure 2: Summary of the Existing State of the Art Related Works and Techniques Used

Table 5: Summary of the Linguistic Localization and Techniques Used toward Malinformation

Ref.	Problem Addressed	Dataset Used	Data Types	Techniques Used	ACC	R	F1	Remarks
Debele and Woldeyohannis (2022)	Hate Speech Amharic	6,497 files	Audio & Text	DL (LSTM, BiLSTM, GRU, BiGRU)	84-88.15%	–	–	Small dataset; lacks generalization and multimodal expansion beyond audio-text; no coverage of other information disruption types.
Tesfaye and Kakeba (2020)	Hate Speech, Amharic	Custom dataset (30,000)	Textual	DL (LSTM, GRU)	97.9%	–	–	Lacks comparison with BiLSTM and other models; absence of public datasets; lacks model compression and multimodal integration.
Mossie *et al.* (2018)	Hate Speech, Amharic	Custom dataset (6,120)	Textual	ML (NB, RF)	79.83%	–	65.34%	Small dataset size; lacks comparison with other ML models and multimodal integration.
Ababu *et al.* (2025) and Ganfure (2022)	Hate Speech, Afaan Oromo	Custom dataset	Textual	CNN, BiLSTM, CNN-BiLSTM, BiGRU	78.05%	78%	78%	Lacks bilingual hate speech detection (Amharic–Afaan Oromo); no handling of code-mixed language.
Bakala and Kekeba (2021)	Hate Speech Afaan Oromo	Afaan Oromo Hate Speech Detection	Textual	SVM, NB, LR, DT, RF	67%	–	64%	Lower accuracy and F1 score; lacks analysis and comparison with state-of-the-art; lacks multimodal integration and bi-classification.
Abebaw *et al.* (2022)	Hate Speech, Amharic	2,000 annotated social media comments	Textual	DL (CNN)	–	–	81.3%	Small dataset; lacks full metric evaluation (only F1); lacks multimodal and bi-classification integration.

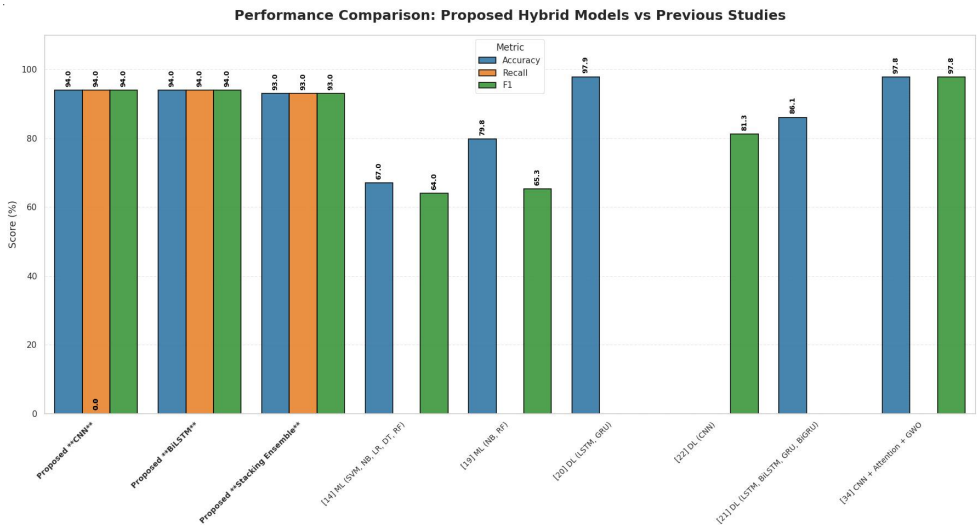


Figure 3: Summary of Proposed Model with Existing State-of-the-Art Malinformation Detection Models

The summary result of Linguistic Localization and Localization and Techniques used toward mal-information shows that Deep learning models, such as CNN, LSTM, and GRU, have achieved a comparatively

Table 6: Comparative Analysis of the Proposed Model and State-of-the-Art Techniques

Ref.	Techniques Used	Evaluation Metrics		
		Accuracy	Recall	F1-Score
Bakala and Kekeba (2021)	ML (SVM, NB, LR, DT, and RF)	67%		64%
Mossie *et al.* (2018)	Machine learning (NB, and RF)	79.83%		65.34%
Tesfaye and Kakeba (2020)	Deep learning (LSTM and GRU)	97.9%	-	-
Abebaw *et al.* (2022)	Deep learning (CNN)			81.3%
Debele and Woldeyohannis (2022)	DL (LSTM, BiLSTM, GRU, and BiGRU)	84% and 88.15%		
Aljohani *et al.* (2024)	CNN with an attention mechanism and Grey Wolf Optimizer (GWO)	97.83%		97.83%
Proposed Model	CNN	**94%**	**94%**	**94%**
	BiLSTM	**94%**	**94%**	**94%**
	Stacking ensemble	**93%**	**93%**	**93%**

optimum accuracy (Table 5). However, the lack of scalable social media datasets in the Ethiopian context, localized context-aware models capabilities towards handling multimodal, i.e., visual and textual, multilingual, multi-class classification with an optical optimization technique and its appropriate annotations tasks still persist as serious research gaps and can be explored.

As shown in Figure 3 and Table 6, we compared existing machine learning and deep learning models for mal-information, i.e., hate speech, offensive detection across various datasets. Traditional models such as SVM, NB, LR, DT, and RF achieved moderate results, with accuracies from 67% to 79.83%. Deep learning models, i.e., LSTM, GRU, BiLSTM, and CNN, performed better, reaching up to 97.9% accuracy. The CNN with an attention mechanism and Grey Wolf Optimizer (GWO) achieved the highest accuracy of 97.83%. In comparison, the proposed CNN, BiLSTM, and Stacking Ensemble models each achieved 94% accuracy, recall, and F1-score, surpassing traditional methods and nearly matching the performance of advanced deep learning models.

As depicted in Figure 4, Figure 5 and Figure 6 show confusion matrices for sample deep learning (CNN-LSTM, CNN, LSTM, and BiLSTM) models, demonstrating their effectiveness in classifying text into four

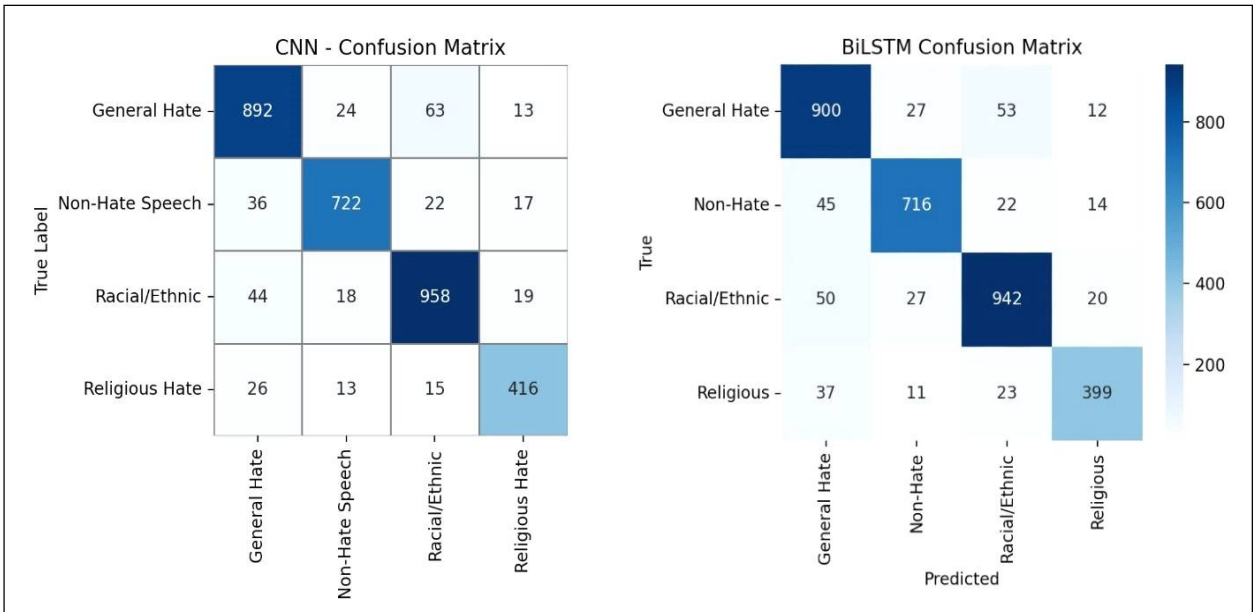


Figure 4: CNN and BiLSTM Confusion Matrix

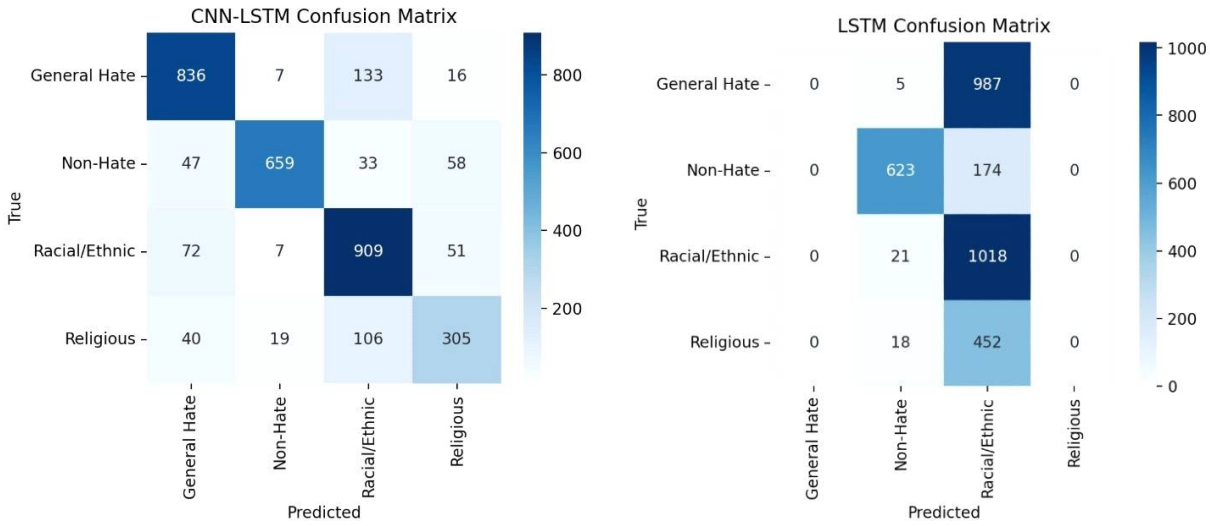


Figure 5: CNN and LSTM Confusion Matrix

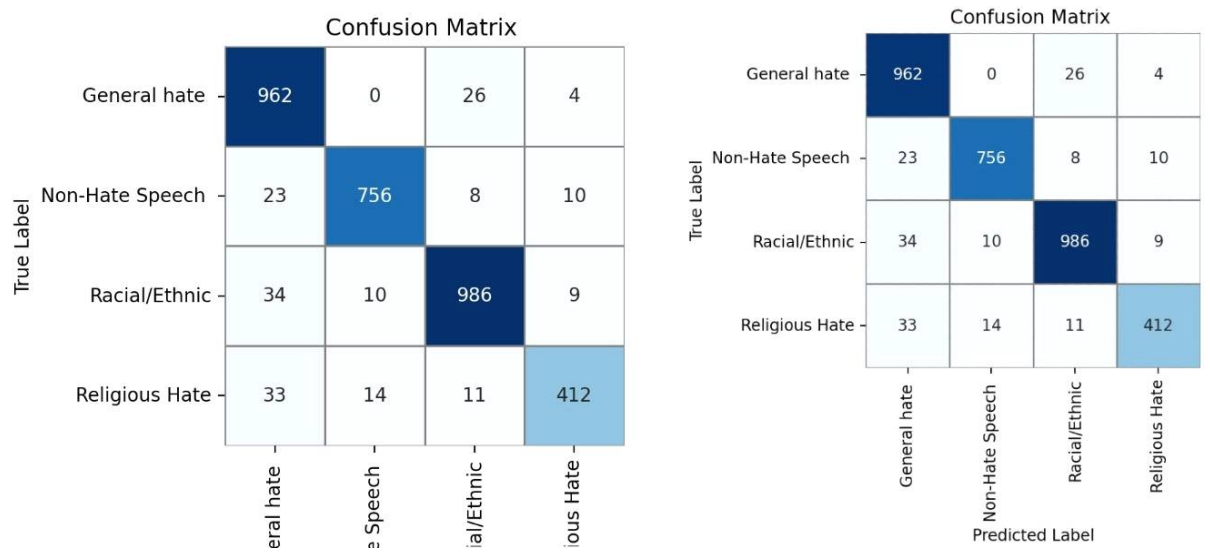


Figure 6: RF and SVM Confusion Matrix

categories: General Hate, Non-Hate Speech, Racial/Ethnic Hate, and Religious Hate. The CNN-LSTM model achieves strong accuracy in detecting General Hate, with 836 correct predictions, but shows noticeable misclassifications in the Non-Hate and Religious categories. Similarly, the CNN model performs well, correctly identifying 892 General Hate instances, though it also misclassifies several Non-Hate samples. The BiLSTM model improves contextual understanding, resulting in higher accuracy and fewer misclassifications, particularly for General Hate and Racial/Ethnic Hate. The LSTM model performs exceptionally well for racial/ethnic and general hate due to its strong ability to capture long-term dependencies, but it tends to overfit, reducing its accuracy for less frequent classes.

Generally, all models perform well in identifying General Hate, but their ability to distinguish among different hate speech types remains limited, highlighting the need for further fine-tuning to improve category separation and reduce overlap. Conversely, as shown in Table 3, the RNN and Naïve Bayes models performed worse than the others. This poor performance is mainly because they rely on simple word frequency methods, like Bag-of-Words or TF-IDF, which do not capture semantic context. Consequently, they often confuse categories. For instance, religious Hate speech is frequently misclassified as racial hate speech due to its overlapping features.

5. Discussion

The findings of this study provide a comprehensive understanding of the effectiveness of machine learning and deep learning models in multi-class mal-information detection. The experimental results indicated that the proposed models, particularly the proposed model using defined hyperparameters by CNN and Stacking Ensemble, effectively detect multi-class mal-information in Amharic, achieving an impressive accuracy of 94%. The result shows the superiority of deep learning approaches over traditional learning models, with models like CNN and BiLSTM adept at capturing complex linguistic patterns in low-resource languages as well as foreign languages. The scarcity of annotated datasets and a focus on binary classification in existing research problems are improved, and the inclusion of 985 Amharic stop words enhanced model performance, underscoring the need for localized linguistic resources.

In this study, a total of eleven (11) learning models were analyzed and evaluated using performance evaluation metrics. Figure 7 illustrates multi-class classification of mal-information into four sections: General Hate, encompassing broad derogatory expressions; Non-Hate Speech, representing positive and constructive dialogue; Racial/Ethnic Hate, focusing on derogatory language targeting specific racial or ethnic groups; and Religious Hate, which includes insults aimed at individuals based on their religious beliefs. The different colors and text styles highlight the prevalence and severity of these phrases, underscoring the need for awareness and the promotion of respectful communication. This visual serves as an educational tool to address the impact of mal-information and encourage inclusivity.

As presented in Figure 7, word clouds were generated for each speech category, visually highlighting the most frequent and salient terms that characterize mal-information, and normal content. In each word cloud, larger words represent higher frequency, revealing dominant linguistic patterns within the category

Figure 7: Word Clouds on Different Categories of Speech

6. Conclusion and Recommendations

Detecting mal-information such as hate speech, offensive speech, and harassment in Amharic social media texts is challenging due to limited linguistic resources and processing models. The anonymity of social platforms enables unchecked hate speech, prompting government interventions. This study addresses these issues by creating a 13,683-sample dataset and developing a robust multi-class model that classifies hate speech into general, racial/ethnic, religious, and non-hate categories. A key contribution of this study is the creation of a specialized word indicator, incorporating 985 commonly used stop words, which enhances the model's semantic understanding. Additionally, the study provides an exhaustive analysis of various machine learning and deep learning models, evaluating their performance in multi-class classification tasks. The experimental results show that CNN, Stacking Ensemble, and Random Forest achieved the highest accuracy of 94%, followed by BiLSTM (93%), SVM (89%) and LSTM (83%), indicating the superior performance of hybrid and ensemble models for multi-class classification. These contributions advance natural language processing for low-resource languages such as Amharic and lay the groundwork for effective hate speech detection on social media. They also hold significant implications for policy development, technological integration, and comprehensive data collection. It is recommended that future research adopt multimodal and multilingual approaches, supported by scalable annotated datasets, to further improve hate speech detection and combat misinformation within Ethiopia's digital landscape. Additionally, Generative AI's raises global relevance, but there's a lack of research and datasets tailored to the Ethiopian context, highlighting the need for a comprehensive understanding. Hence, it is imperative that future studies address this gap.

Acknowledgment

The first author gratefully acknowledges the support of the Ethiopian Ministry of Education for sponsoring the Ph.D. studies through a government scholarship at the Faculty of Computing and Software Engineering, Arbaminch University (AMiT), Arbaminch, Ethiopia.

Conflicts of Interest

The authors declare that they have no known financial or personal conflicts of interest that could have influenced the research and findings presented in this paper.

Funding

The authors received no funding for this work.

References

Aïmeur, E., Amri, S. and Brassard, G. (2023). Fake News, Disinformation, and Misinformation in Social Media: A Review. *Social Network Analysis and Mining*, 13(1), 30.

Ababu, T.M. and Woldeyohannis, M.M. (2022). Afaan {O}romo Hate Speech Detection and Classification on Social Media. in N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk and S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 6612-6619, European Language Resources Association. https://aclanthology.org/2022.lrec-1.712

Ababu, T.M., Woldeyohannis, M.M. and Getaneh, E.B. (2025). Bilingual Hate Speech Detection on Social Media: Amharic and Afaan Oromo. *Journal of Big Data*, 12(1). https://doi.org/10.1186/s40537-024-01044-y

Abate, S.T., Tachbelie, M.Y., Melese, M., Abera, H., Abebe, T., Mulugeta, W., Assabie, Y., Meshesha, M., Afnafu, S. and Seyoum, B.E. (2020). Large Vocabulary Read Speech Corpora for Four {E}thiopian Languages: {A}mharic, {T}igrigna, {O}romo and {W}olaytta. in N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk and S. Piperidis (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4167-4171, European Language Resources Association. https://aclanthology.org/2020.lrec-1.513/

Abebaw, Z., Rauber, A. and Atnafu, S. (2022). Multi-Channel Convolutional Neural Network for Hate Speech Detection in Social Media. *Advances of Science and Technology: 9th EAI International Conference, ICAST 2021, Hybrid Event, Bahir Dar, Ethiopia*, August 27-29, 2021, Proceedings, Part I, 603-618.

Al-Hassan, A. and Al-Dossari, H. (2022). Detection of Hate Speech in Arabic Tweets Using Deep Learning. *Multimedia Systems*, 28(6), 1963-1974.

Alatawi, H.S., Alhothali, A.M. and Moria, K.M. (2021). Detecting White Supremacist Hate Speech Using Domain Specific Word Embedding with Deep Learning and BERT. *IEEE Access*, 9, 106363-106374. https://doi.org/10.1109/ACCESS.2021.3100435

Alhejaili, R. (2025). Machine Learning Approaches for Sentiment Analysis on Social Media. in W.M.S. Yafooz and Y. Al-Gumaei (Eds.), *AI-Driven: Social Media Analytics and Cybersecurity*, 21-43, Springer Nature, Switzerland. https://doi.org/10.1007/978-3-031-80334-5_2

Ali, R., Farooq, U., Arshad, U., Shahzad, W. and Beg, M.O. (2022). Hate Speech Detection on Twitter Using Transfer Learning. *Computer Speech & Language*, 74, 101365. https://doi.org/https://doi.org/10.1016/j.csl.2022.101365

Aljohani, A., Alharbe, N., Al Mamlook, R.E. and Khayyat, M.M. (2024). A Hybrid Combination of CNN Attention with Optimized Random Forest with Grey Wolf Optimizer to Discriminate between Arabic Hateful, Abusive Tweets. *Journal of King Saud University-Computer and Information Sciences*, 36(2). https://doi.org/10.1016/j.jksuci.2024.101961

Alpaydin, E. (2020). Introduction to Machine Learning. MIT Press.

Aluru, S.S., Mathew, B., Saha, P. and Mukherjee, A. (2020). Deep Learning Models for Multilingual Hate Speech Detection. 1-16. http://arxiv.org/abs/2004.06465

Anderson, L. and Barnes, M. (2023). Hate Speech. in E.N. Zalta and U. Nodelman (Eds.), *The {Stanford} Encyclopedia of Philosophy* ({F}all 202), Metaphysics Research Lab, Stanford University.

Araque, O. and Iglesias, C.A. (2022). An Ensemble Method for Radicalization and Hate Speech Detection Online Empowered by Sentic Computing. *Cognitive Computation*, 14(1), 48-61. https://doi.org/10.1007/s12559-021-09845-6

Arshad, M.U., Ali, R., Beg, M.O. and Shahzad, W. (2023). UHated: Hate Speech Detection in Urdu Language Using Transfer Learning. *Language Resources and Evaluation*, 57(2), 713-732.

Ayele, A.A., Belay, T.D., Yimam, S.M., Dinter, S., Asfaw, T.T. and Biemann, C. (2022). Challenges of Amharic Hate Speech Data Annotation Using Yandex Toloka Crowdsourcing Platform. *Proceedings of the The Sixth Widening NLP Workshop (WiNLP) Co-Located with EMNLP 2022*. https://aclanthology.org/2022.winlp-1.0

Bakala Defersha, N. and Kekeba Tune, K. (2021). *Indian Journal of Science and Technology Detection of Hate Speech Text in Afaan Oromo Social Media using Machine Learning Approach*, 2567-2578. https://www.indjst.org/

Bockting, C.L., Van Dis, E.A.M., Van Rooij, R., Zuidema, W. and Bollen, J. (2023). Living Guidelines for Generative AI—Why Scientists Must Oversee its Use. *Nature*, 622(7984), 693-696.

Chhabra, A. and Vishwakarma, D.K. (2023). A Literature Survey on Multimodal and Multilingual Automatic Hate Speech Identification. *Multimedia Systems*, 29(3), 1203-1230. https://doi.org/10.1007/s00530-023-01051-8

Chiche, A. and Yitagesu, B. (2022). Part of Speech Tagging: A Systematic Review of Deep Learning and Machine Learning Approaches. *Journal of Big Data*, 9(1). https://doi.org/10.1186/s40537-022-00561-y

Cohen-Almagor, R. and Stamile, N. (2021). Freedom of Expression v. Social Responsibility on the Internet: Vivi Down Association v. Google.

De Santis, E., Martino, A., Ronci, F. and Rizzi, A. (2025). From Bag-of-Words to Transformers: A Comparative Study for Text Classification in Healthcare Discussions in Social Media. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 9(1), 1063-1077. https://doi.org/10.1109/TETCI.2024.3423444

Debele, A.G. and Woldeyohannis, M.M. (2022). Multimodal Amharic Hate Speech Detection Using Deep Learning. *2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, 102-107.

Ertel, W. (2018). Introduction to Artificial Intelligence. Springer.

Faris, H., Aljarah, I., Habib, M. and Castillo, P.A. (2020). Hate Speech Detection Using Word Embedding and Deep Learning in the Arabic Language Context. *ICPRAM*, 453-460.

Ganfure, G.O. (2022). Comparative Analysis of Deep Learning Based Afaan Oromo Hate Speech Detection. *Journal of Big Data*, 9(1), 76.

Gebremichael, H.T., Mengistu, T.M., Beyene, M.M. and Mengistu, F.G. (2022). OCR System for the Recognition of Ethiopic Real-Life Documents. in M.L. Berihun (Ed.), *Advances of Science and Technology*, 559-574, Springer, International Publishing.

Geetanjali and Kumar, M. (2025). Exploring Hate Speech Detection: Challenges, Resources, Current Research and Future Directions. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-025-20716-2

Hutson, M. (2023). Rules to Keep AI in Check: Nations Carve Different Paths for Tech Regulation. *Nature*, 620(7973), 260-263.

Huu, Q.P., Trung, S.N. and Pham, H.A. (2019). Automated Hate Speech Detection on Vietnamese Social Networks.

Islam, M.R., Liu, S., Wang, X. and Xu, G. (2020). Deep Learning for Misinformation Detection on Online Social Networks: A Survey and New Perspectives. *Social Network Analysis and Mining*, 10(1), 82.

Jahan, M.S. and Oussalah, M. (2023a). A Systematic Review of Hate Speech Automatic Detection Using Natural Language Processing. *Neurocomputing*, 546, 126232. https://doi.org/https://doi.org/10.1016/j.neucom.2023.126232

Jahan, M.S. and Oussalah, M. (2023b). A Systematic Review of Hate Speech Automatic Detection Using Natural Language Processing. *Neurocomputing*, 546, 126232. https://doi.org/10.1016/j.neucom.2023.126232

Kocoñ, J., Figas, A., Gruza, M., Puchalska, D., Kajdanowicz, T. and Kazienko, P. (2021). Offensive, Aggressive, and Hate Speech Analysis: From Data-Centric to Human-Centered Approach. *Information Processing & Management*, 58(5), 102643.

Megersa, T. and Minaye, A. (2023). *Ethiopian Journal of Social Sciences*, 9(1), 31-50.

Mehmood, F., Ghafoor, H., Asim, M.N., Ghani, M.U., Mahmood, W. and Dengel, A. (2024). Passion-Net: A Robust Precise and Explainable Predictor for Hate Speech Detection in Roman Urdu Text. *Neural Computing and Applications*, 36(6), 3077-3100.

Mohd Nafis, N.S. and Awang, S. (2021). An Enhanced Hybrid Feature Selection Technique Using Term Frequency-Inverse Document Frequency and Support Vector Machine-Recursive Feature Elimination for Sentiment Classification. *IEEE Access*, 9(Ml), 52177-52192. https://doi.org/10.1109/ACCESS.2021.3069001

Mossie, Z., Wang, J.-H. and Others. (2018). Social Network Hate Speech Detection for Amharic Language. *Computer Science & Information Technology*, 41-55.

Navya, K., Sabaha, H., Rajiakodi, S. and Sivagnanam, B. (2025). Detecting Homophobic and Transphobic Comments on Social Media in Malayalam and English Languages. *Procedia Computer Science*, 258, 2479-2489. https://doi.org/10.1016/j.procs.2025.04.510

Neshir, G., Rauber, A. and Atnafu, S. (2021). Meta-Learner for Amharic Sentiment Classification. *Applied Sciences*, 11(18). https://doi.org/10.3390/app11188489

Rascão, J.P. (2020). Freedom of Expression, Privacy, and Ethical and Social Responsibility in Democracy in the Digital Age. *International Journal of Business Strategy and Automation (IJBSA)*, 1(3), 1-23.

Ruwandika, N.D.T. and Weerasinghe, A.R. (2018). Identification of Hate Speech in Social Media. *2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer)*, 273-278.

Sevani, N., Soenandi, I.A., Wijaya, J. and Others. (2021). Detection of Hate Speech by Employing Support Vector Machine with Word2Vec Model. *2021 7th International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*, 1-5.

Tesfaye, S.G. and Kakeba, K. (2020). Automated Amharic Hate Speech Posts and Comments Detection Model Using Recurrent Neural Network.

The Federal Democratic Republic of Ethiopia. (2020). Hate Speech and Disinformation Prevention and Suppression Proclamation. *Federal Negarit Gazette*, 39(I), 8205-8234.

UnitedNations & StrategyandPlanofActiononHateSpeech. (2023). *Understanding-Hate-Speech*. https://doi.org/https://doi.org/10.1016/j.is.2024.102378

Vasist, P.N., Chatterjee, D. and Krishnan, S. (2024). The Polarizing Impact of Political Disinformation and Hate Speech: A Cross-Country Configural Narrative. *Information Systems Frontiers*, 26(2), 663-688. https://doi.org/10.1007/s10796-023-10390-w

Vo, C.N., Huynh, K.B., Luu, S.T. and Do, T.H. (2025). ViTHSD: Exploiting Hatred by Targets for Hate Speech Detection on Vietnamese Social Media Texts. In *Journal of Computational Social Science*, 8(2), Springer Nature, Singapore. https://doi.org/10.1007/s42001-024-00348-6

Cite this article as: Andualem Woldegiorgis, Mohammed Abebe, Durga Prasad Sharma and Worku Jimma (2026). Modeling Multi-Class Mal-Information Detection: A Comparative Analysis of Machine Learning and Deep Learning Approaches. *International Journal of Artificial Intelligence and Machine Learning*, 6(1), 1-17. doi: 10.51483/IJAIML.6.1.2026.1-17.