IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

Ensemble Neural Approach for Hate Speech Detection in Text and Audiomodalities

Keerthi Lahari Edara, Meena Venkateswari Bezawada,
Sri Prathyusha Garnepudi, Dr. Bhagya Lakshmi Nandipati
B. Tech Students, Dept. of Computer Science and Engineering
Assistant Professor, Dept. of Computer Science and Engineering
R.V.R & J.C College of Engineering, Chowdavaram, Guntur, Andhra Pradesh, India

Abstract: *This paper presents the implementation of a hate speech detection system employing Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) neural networks. The system is designed to analyse both textual and audio inputs (converted to text) for the identification of hate speech content. By integrating the outputs of both models, the system delivers an ensemble prediction, enhancing detection accuracy and robustness. The implementation utilises TensorFlow and Keras for constructing the neural network models, while Flask serves as the framework for developing the web application interface. Additional libraries are employed for text preprocessing and speech recognition tasks. The proposed system demonstrates high efficacy in detecting hate speech, while offering a user-friendly interface that accommodates both text and audio inputs. This paper outlines the system's background, architecture, detailed implementation procedure, and performance evaluation. The findings illustrate how natural language processing (NLP) techniques and deep learning methodologies can be effectively leveraged to identify and mitigate harmful online content.*

Keywords: Hate Speech Detection, Text Classification, Audio-to-Text, Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), Ensemble Prediction, Deep Learning, Natural Language Processing (NLP)

I. INTRODUCTION

The rapid expansion of social media and online communication platforms has drastically transformed how people interact, share information, and express opinions. While these platforms enable global connectivity and open discourse, they have also become avenues for the spread of harmful content such as hate speech. Hate speech refers to language that targets individuals or groups based on attributes like race, religion, gender, sexual orientation, ethnicity, or disability, often with the intent to demean, marginalise, or incite violence.The consequences of hate speech are far-reaching—it fosters discrimination, perpetuates stereotypes, and can even lead to real-world violence. Given the vast and continuous influx of user-generated content, manual moderation of hate speech is no longer feasible. To address this challenge, researchers have increasingly turned to automated systems powered by natural language processing (NLP) and machine learning to detect and mitigate harmful content.

Deep learning, particularly with recurrent neural networks such as Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM), has shown significant promise in improving hate speech detection. These models are well-suited for analyzing sequential text data, enabling them to capture complex contextual and semantic relationships that are crucial for understanding subtle or implied hate speech.

However, building an effective hate speech detection system is not without challenges. Contextual ambiguity, linguistic variation, sarcasm, coded language, and subjective interpretations all complicate accurate detection. Additionally, hate speech can appear in multiple forms—including text and speech—which calls for multimodal solutions. Furthermore, real-world datasets often suffer from class imbalance, annotation inconsistency, and limited diversity, which hamper model training and generalization.

DOI: 10.48175/IJARSCT-26434

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

This project addresses these issues by developing a comprehensive hate speech detection system that combines LSTM and BiLSTM architectures. It incorporates both text and audio inputs, applies ensemble methods for higher accuracy, and provides an accessible web interface. This system is designed to assist in real-time content moderation, support diverse languages and modalities, and contribute to safer and more respectful digital spaces. The work not only advances technical capabilities in automated hate speech detection but also supports broader societal goals of inclusivity and online safety.

As online platforms struggle to strike a balance between free speech and community guidelines, the need for robust, transparent, and scalable moderation tools has become more pressing than ever. Traditional keyword-based filtering techniques fall short when faced with the evolving nature of hate speech, where users often disguise toxic content using code words, abbreviations, or context-dependent language. This calls for more intelligent systems that can interpret context, detect sentiment, and understand underlying intent.The motivation behind this project lies in the critical role that online safety plays in fostering inclusive digital environments. Whether in the form of social media comments, video content, or online forums, hate speech has the potential to harm individuals and entire communities. By building a system that can identify hate speech across different formats—text and speech—this project takes a step toward empowering platform moderators and protecting users from digital harm.

One of the unique aspects of this project is its dual-modality approach, where both textual and audio data are analyzed. While most existing systems focus solely on text, the inclusion of speech analysis addresses a significant gap, considering the popularity of voice-based interactions and video platforms. Speech data is transcribed using Automatic Speech Recognition (ASR) and further processed for hate speech cues using deep learning models. This enhances the system‘s applicability across platforms like YouTube, podcasts, and voice chats.

In addition to model development, this project emphasizes usability. A web-based interface ensures that the solution is not confined to academic use but is also accessible for real-world deployment. Through this interface, users can input data, receive predictions, and even visualize confidence scores or highlight offensive segments, making it suitable for both end-users and moderators.

The final objective of this project is not just to achieve high accuracy in detection, but to contribute to a more inclusive digital society by reducing the spread of harmful content. By leveraging advanced neural architectures and offering a practical implementation, this project aims to serve as a foundation for future research and real-time moderation systems in both commercial and social domains.

This project introduces several enhancements over traditional approaches, the evolving nature of hate speech requires continual adaptation. Emerging forms of coded language, memes, and multi-modal content necessitate models that are not only accurate but also flexible and updatable. To this end, integrating continual learning strategieswhere models can be incrementally trained on new data without forgetting previously learned patternscould be an essential future direction. This approach may help maintain performance as online discourse and hate speech expressions evolve over time.

Another critical dimension is the ethical and legal implications of hate speech detection systems. As highlighted by Vidgen et al. (2019), automated moderation systems carry the risk of over-censorship and may unintentionally suppress free expression if not carefully designed. Therefore, transparency, accountability, and human-in-the-loop systems are vital to ensure that these tools are fair, interpretable, and respectful of individual rights. In this project, explainability is supported through confidence scores and model comparison outputs, helping users understand why a piece of content was flagged.

Additionally, this project lays the groundwork for future multilingual and cross-cultural expansion. While the current model focuses on English-language data, the architecture is designed to be extensible, allowing incorporation of multilingual embeddings and datasets in subsequent iterations. This is particularly important given the global reach of online platforms and the diverse ways in which hate speech manifests in different linguistic and cultural contexts.

The unified web-based interface also opens avenues for deployment in real-world settings such as educational tools, community moderation platforms, and content flagging systems for social media companies. With minor modifications, this system could be integrated into browser extensions or mobile applications, enabling users to flag hate speech proactively and contribute to safer online environments.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

In conclusion, this project contributes to the growing body of work on automated hate speech detection by addressing key limitations in both model performance and practical deployment. Through its dual-modality support, ensemble learning strategy, and emphasis on usability and transparency, it serves as a scalable and adaptable solution that moves the field one step closer to real-world application. Nevertheless, it also acknowledges the remaining challengesparticularly those related to subjectivity, adversarial resilience, and ethical concernswhich must continue to guide future research in this critical area.

II. LITERATURE REVIEW

The earliest attempts at detecting harmful content online relied primarily on dictionary-based methods, using predefined lists of offensive terms to flag potentially problematic content. Burnap et al. (2013) explored this approach alongside simple machine learning techniques, demonstrating modest success but highlighting significant limitations in detecting context-dependent hate speech.Warner and Hirschberg (2012) introduced more sophisticated methods by incorporating template-based approaches that looked for specific linguistic patterns associated with hate speech, showing improvements over pure dictionary methods but still struggling with the nuanced nature of hateful content.

Davidson et al. (2017) published a seminal work that moved beyond simple classification by distinguishing between hate speech and merely offensive language, using a dataset of annotated tweets and traditional machine learning algorithms like Support Vector Machines (SVM) and Logistic Regression. Their work highlighted the complexity of the problem and the challenges in creating adequate training datasets.Waseem and Hovy (2016) explored the importance of feature engineering in hate speech detection, examining the role of demographic information alongside textual features. Their research emphasized the value of incorporating user and context metadata to improve classification accuracy.Schmidt and Wiegand (2017) provided a comprehensive survey of hate speech detection methods, synthesizing findings across multiple studies and highlighting the relative strengths of different feature types, including character n-grams, word n-grams, and sentiment features.

Zhang et al. (2018) demonstrated the superiority of Convolutional Neural Networks (CNNs) over traditional machine learning approaches for hate speech detection, leveraging their ability to identify relevant patterns in text without extensive feature engineering.Badjatiya et al. (2017) explored various deep learning architectures, including LSTMs combined with gradient boosted decision trees, achieving state-of-the-art results on benchmark datasets. Their work demonstrated the power of combining neural networks with ensemble methods.Founta et al. (2019) introduced a more nuanced labeling scheme for abusive language detection, moving beyond binary classification to recognize the spectrum of harmful content. Their research highlighted the importance of dataset quality and annotation guidelines.

LSTM networks have gained prominence in natural language processing tasks due to their ability to capture long-range dependencies in sequential data. Graves (2012) provided the theoretical foundation for modern LSTM implementations, detailing their architecture and training procedures.Zhou et al. (2016) demonstrated the effectiveness of BiLSTM networks for text classification tasks, showing that the bidirectional processing of text could capture contextual information more effectively than unidirectional approaches. Their work showed particular promise for sentiment analysis tasks, which share similarities with hate speech detection.Agrawal and Awekar (2018) specifically applied deep learning models including LSTMs to cyberbullying and hate speech detection, comparing their performance across multiple social media platforms and demonstrating robust results across different datasets.

The extension of hate speech detection to audio inputs represents an emerging frontier in the field. Hee et al. (2018) explored cyberbullying detection across multiple modalities, highlighting the unique challenges and opportunities presented by audio data.Rajamanickam et al. (2020) investigated the use of speech recognition combined with text-based hate speech detection, demonstrating the feasibility of a pipeline approach similar to the one implemented in this project.

Zimmerman et al. (2018) demonstrated the value of ensemble methods in hate speech detection, showing that combining multiple models often yields better results than any single approach. Their work particularly highlighted the complementary nature of different neural network architectures when used in concert.Fersini et al. (2018) applied ensemble methods to misogyny detection in social media, a specific subset of hate speech detection, achieving superior

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

performance through model combination and weighted voting schemes.This current project builds upon these foundations by implementing an ensemble approach that combines LSTM and
BiLSTM models, supporting both text and audio inputs through a unified web interface, and providing detailed feedback on detection confidence and model-specific predictions.

III. DATASETS

The hate speech detection system was trained on carefully selected datasets that represent diverse forms of online communication and hate speech instances(CMU-MOSI,CMU-MOSUI,LJ speech …).Understanding these datasets is crucial for interpreting the system's performance and limitations.

Primary Training Dataset

The primary dataset used for training the LSTM and BiLSTM models is a balanced hate speech dataset containing over 700,000 text samples. This dataset was specifically created to address class imbalance issues that often plague hate speech detection systems.

Dataset	Number of Samples		
	Train	Dev	Test
CMU-MOSEI (Bagher Zadeh et al., 2018)	597	133	130
CMU-MOSI (Zadeh et al., 2016)	181	40	39
Common Voice (Ardila et al., 2020)	8,050	1,768	1,733
LJ Speech (Ito and Johnson, 2017)	102	23	23
MELD (Poria et al., 2019)	393	87	85
Social-IQ (Zadeh et al., 2019)	325	74	69
VCTK (Yamagishi et al., 2019)	138	31	30
	9,786	2,156	2,109

Table 1: Statistics of the dataset used for Identification of Hatred.

Dataset Characteristics:

The dataset used in this study consists of 726,119 text samples collected from various social media platforms and online forums. It maintains a balanced distribution between hate speech (label 1) and non-hate speech (label 0), ensuring that the model is trained on an unbiased representation of both classes. The text samples vary in length, ranging from short phrases to longer, multi-sentence posts, reflecting the diverse nature of user-generated content. Prior to model training, the data underwent basic preprocessing steps, including the removal of usernames, URLs, and other identifiable elements, while preserving the core content of each message to retain contextual meaning essential for accurate classification.

Data Distribution:

The dataset includes diverse forms of online communication, with varied linguistic patterns, contexts, and expressions. This diversity is essential for training models that can generalize to real-world hate speech detection scenarios.
Class distribution:
Hate Speech (1): 363,060 samples (50%)
Non-Hate Speech (0): 363,059 samples (50%)
The balanced nature of this dataset is particularly important, as it helps prevent the model from developing a bias toward either class, which could lead to high false positive or false negative rates.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

IV. SYSTEM ARCHITECTURE

A. Overview

The proposed project is a Hate Speech Detection System designed to identify and classify hate speech in both text and audio formats. The system leverages state-of-the-art machine learning techniques, including Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (BiLSTM) models, to accurately assess whether a given input contains hate speech. This system aims to provide real-time, automated detection of harmful language, offering valuable support in applications such as social media monitoring, content moderation, and public safety.

The system is structured to process text and audio inputs, converting spoken content into text when necessary. Users interact with the system through a web-based interface, where they can submit text directly or upload audio files for analysis. The system then processes the input, applies a series of preprocessing and machine learning steps, and returns a prediction on whether the content qualifies as hate speech. The results are presented in a clear and structured format, with a confidence assessment to help users understand the reliability of the prediction.

The system is organized into distinct layers, ensuring a modular and scalable architecture. Each layer performs a specific function, ranging from user interaction to data processing and analysis, ultimately enabling the efficient detection of hate speech. The architecture includes the following components:

- **Presentation Layer**: The user interface, which allows users to input text or audio and view results.
- **Application Layer**: Coordinates the request handling, data formatting, and analysis workflows.
- **Processing Layer**: Transforms input data into formats suitable for analysis, including text preprocessing and speech recognition.
- **Analysis Layer**: Contains the machine learning models responsible for generating predictions.

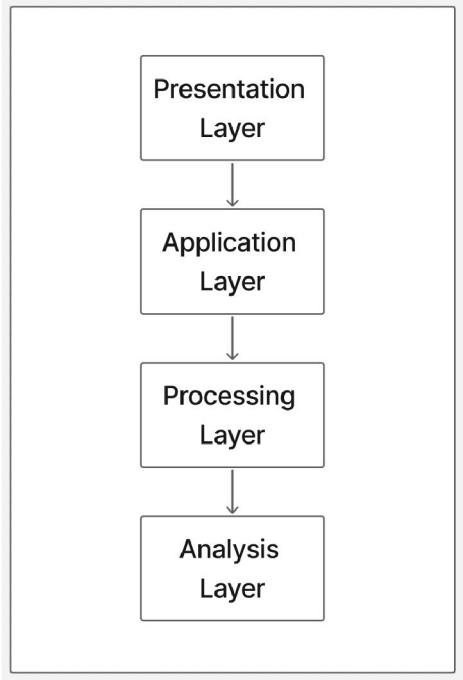

Figure 1: Layered Architecture of the System

This comprehensive architecture ensures that the system can handle a wide range of use cases, from detecting hate speech in social media posts to analyzing spoken content for harmful language. The modular design also allows for easy maintenance and scalability, making it adaptable for future improvements and extensions.

Additionally, the system's architecture is designed with scalability and extensibility in mind. As the system grows, it can seamlessly integrate new models, additional preprocessing techniques, and advanced analysis features. For

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

example, the system could be extended to detect other forms of harmful content, such as misinformation, cyberbullying, or inappropriate language, providing an even broader application for content moderation platforms, law enforcement, and educational institutions.

The modular design also ensures that updates and improvements can be made to individual components of the system without disrupting the overall workflow. For instance, a more sophisticated model could replace the current LSTM or BiLSTM models, or a new speech recognition service could be integrated to improve accuracy across different languages and dialects.

B. Architectural Layers

1. Presentation Layer

The presentation layer is the interface that connects users to the system, enabling interaction through both graphical and programmatic methods. This layer includes:

a. Web Interface: A Flask-based HTML/CSS/JavaScript frontend that allows users to input text or upload/record audio for analysis, as well as view results. The interface is designed to be user-friendly, ensuring seamless interaction with the system.

b. API Endpoints: RESTful endpoints that accept JSON requests and return structured responses. These endpoints enable external programs or services to interact with the system, providing a programmatic interface for automated hate speech detection .Responsibilities of the presentation layer include-Input validation to ensure that submitted text or audio is in the correct format. Formatting results to display them clearly to the user. Managing user interactions by capturing inputs and displaying predictions. It communicates exclusively with the application layer to maintain a separation of concerns between the user interface and the system's core functionality. Input validation to ensure that submitted text or audio is in the correct format. Formatting results to display them clearly to the user. Managing user interactions by capturing inputs and displaying predictions. It communicates exclusively with the application layer to maintain a separation of concerns between the user interface and the system's core functionality.

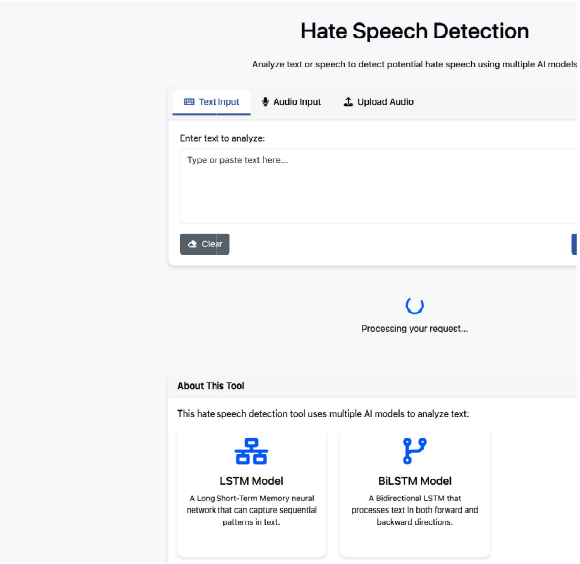

Figure 2: Hate Speech Detection Interface

2. Application Layer

The application layer acts as the intermediary between the presentation layer and the processing layers, orchestrating the system's functionality. The key components of this layer include:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

a. Request Handler: This component processes incoming requests from the presentation layer and forwards them to the appropriate processing components. It ensures that data flows efficiently through the system.
b. Response Formatter: It structures the analysis results into a format that is expected by the presentation layer, ensuring consistency in how results are returned.
c. Workflow Coordinator: It manages the sequence of operations required for processing each request. For example, it handles the order in which the text is preprocessed, to kenised, and passed through machine learning models. The application layer implements the core business logic of the system, ensuring that data is processed correctly and results are returned in an appropriate format.

3. Processing Layer

The processing layer is responsible for transforming inputs into formats suitable for analysis. This layer contains specialized components that handle specific transformations:
a. Text Preprocessor: This component cleans and normalizes the input text by performing operations like special character removal, converting text to lowercase, and removing stopwords.
b. Tokenizer: After text preprocessing, the tokenizer converts the clean text into numerical sequences, making it compatible with neural network inputs.
c. Audio Processor: This component handles audio input, including decoding base64-encoded audio and managing temporary storage for audio files before they are processed.
d. Speech Recognizer: Using Google's Speech Recognition library, this component converts speech input (audio) into text for further analysis. Each of these components performs a specific transformation on the input data, ensuring modularity and ease of maintenance.

4. Analysis Layer

The analysis layer is responsible for the core task of hate speech detection using machine learning models. It includes:
a. Model Manager: This component is responsible for loading and initializing available machine learning models. If a particular model is unavailable, it handles fallbacks or switches to alternative models.
b. LSTM Model: The Long Short-Term Memory (LSTM) model processes tokenized text input to generate a probability score indicating the likelihood of hate speech.
c. BiLSTM Model: This Bidirectional LSTM model processes tokenized input in both directions, potentially offering higher accuracy by capturing contextual relationships from both the past and future.
d. Ensemble Component: This component combines the predictions from multiple models (e.g., LSTM and BiLSTM) to provide a more robust final assessment by averaging their outputs. The analysis layer encapsulates all the machine learning functionality, isolating the complexities of model inference from the other layers.

4. a. Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks are a special type of Recurrent Neural Network (RNN) capable of learning long-range dependencies in sequential data. Traditional RNNs often suffer from vanishing or exploding gradient problems, which makes them ineffective in capturing long-term dependencies. LSTM addresses this by incorporating memory cells and gating mechanisms—namely, input, forget, and output gates. These gates regulate the flow of information, allowing the network to retain or discard information over time.In the context of hate speech detection, LSTMs are well-suited as they can effectively capture the contextual meaning of words in a sentence or tweet. By processing text sequentially, an LSTM can learn subtle patterns and dependencies that may indicate toxic, hateful, or offensive language.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

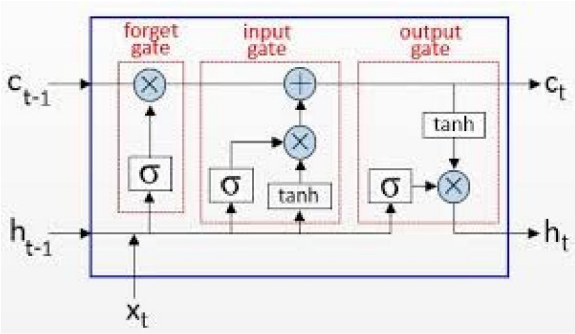

Figure 3: Architecture of a basic LSTM cell showing input, forget, and output gate

4.b. Bidirectional LSTM (BiLSTM)

Bidirectional LSTM (BiLSTM) is an extension of the standard LSTM model that improves context learning by processing sequences in both forward and backward directions. While an LSTM only considers past inputs, BiLSTM utilizes future context as well by combining two LSTMs—one processing the sequence from left to right and the other from right to left.For hate speech detection, where the meaning of a word can depend on both its preceding and succeeding context, BiLSTMs offer a significant advantage. They enable the model to better understand semantic relationships in a sentence and improve classification performance in cases where hate speech is implied subtly or through sarcasm.

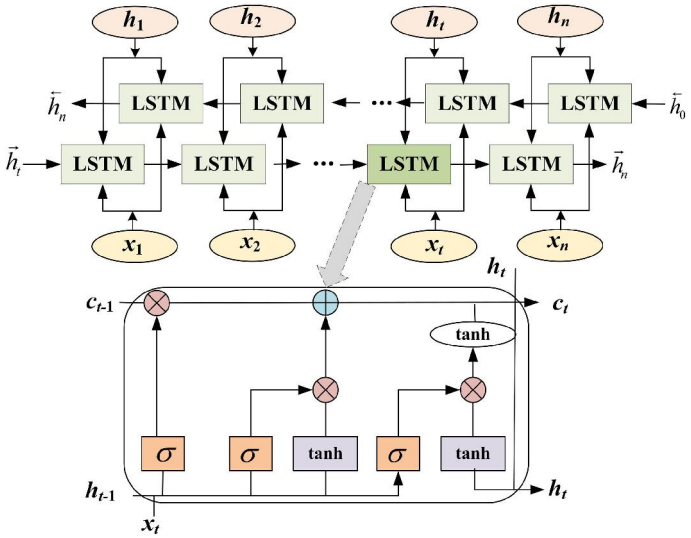

Figure 4: Bidirectional LSTM architecture

4.c. Ensemble of LSTM and BiLSTM

To leverage the strengths of both LSTM and BiLSTM, an ensemble approach can be employed. In this architecture, both models are trained independently on the same dataset and their outputs are combined—either through averaging the prediction probabilities or using a meta-classifier. This hybrid model benefits from the temporal pattern recognition of LSTM and the rich contextual understanding of BiLSTM.

The ensemble enhances generalization and robustness, especially in imbalanced or noisy data typical of social media. By capturing both sequential and bidirectional context, it reduces the risk of misclassification and increases detection accuracy.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

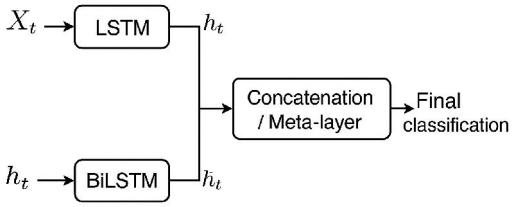

Figure 5: Ensemble model combining LSTM and BiLSTM outputs using concatenation or a meta-layer for final classification.

V. PERFORMANCE EVALUATION METRICS

The performance of the hate speech detection system is evaluated using standard metrics for binary classification. These metrics offer a comprehensive understanding of the model's effectiveness in terms of correctness, robustness, and generalization capabilities.

A. Accuracy

Accuracy reflects the proportion of correctly predicted instances among all predictions:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

TP = True Positives,TN = True Negatives, FP = False Positives, FN = False Negatives

Although accuracy provides an overall measure of performance, it can be misleading in the presence of class imbalance. Since a balanced dataset was used for both training and evaluation, the metric remains meaningful. The LSTM model achieved an accuracy of 86.3%, indicating solid general performance.

B. Precision

Precision measures the proportion of predicted positive instances (i.e., hate speech) that are actually positive. It is defined as:

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

A high precision indicates that the system rarely misclassifies non-hate speech as hate speech, which is critical in real-world scenarios such as content moderation. The LSTM and BiLSTM models achieved precisions of 85.2% and 86.1%, respectively.

C. Recall

Recall (or Sensitivity) quantifies the proportion of actual hate speech correctly identified by the model:

$$Recall = \frac{TP}{TP + FN}$$

High recall is crucial in scenarios where it is more detrimental to miss instances of hate speech. The LSTM model reached a recall of 87.9%, while the BiLSTM model scored 87.4%.

D. F1 Score

The F1 Score, the harmonic mean of precision and recall, provides a balanced assessment of both false positives and false negatives:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

DOI: 10.48175/IJARSCT-26434

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

This score is particularly suited for hate speech detection tasks where both error types are costly. The LSTM and BiLSTM models achieved F1 scores of 86.5% and 86.7%, respectively.

Model-Specific Performance

Performance metrics were calculated separately for the LSTM, BiLSTM and Ensemble models to assess their individual contributions to the ensemble:

Metric	LSTM	BiLSTM	Ensemble
Accuracy	86.3%	86.7%	**87.1%**
Precision	85.2%	86.1%	**86.8%**
Recall	87.9%	87.4%	**87.6%**
F1 Score	86.5%	86.7%	**87.2%**
AUC	0.93	0.94	**0.94**

Table 2: Performance Comparison of LSTM, BiLSTM, and Ensemble Models

From the table, we can observe that the Ensemble model outperforms both LSTM and BiLSTM models across all metrics. It shows a slight improvement in accuracy, precision, recall, F1 score, and AUC, with values of 87.1%, 86.8%, 87.6%, 87.2%, and 0.94 respectively. While the LSTM and BiLSTM models demonstrate solid performance with minimal variation between them, the Ensemble model's higher scores suggest its ability to combine the strengths of different models for better predictive power and reliability.

VI. CONCLUSION

The hate speech detection system described in this documentation represents a significant step forward in applying advanced neural network techniques to the challenge of identifying harmful online content. By combining LSTM and BiLSTM architectures in an ensemble approach and supporting both text and audio inputs, the system provides a comprehensive solution with practical utility for content moderation and analysis.

While limitations exist, particularly around language coverage, cultural context, and the evolving nature of online discourse, the system's modular architecture provides a foundation for ongoing improvement and adaptation. The identified future directions offer promising paths for addressing current limitations and expanding the system's capabilities.

As online communication continues to grow in importance, effective tools for identifying and addressing harmful content become increasingly essential. This hate speech detection system contributes to that goal, providing a technically sound approach that balances detection accuracy with practical implementation considerations.

REFERENCES

[1]. Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web Companion (pp. 759-760).

[2]. Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. Policy & Internet, 7(2), 223-242.

[3]. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 11, No. 1).

[4]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

[5]. Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 12, No. 1).

[6]. Graves, A. (2012). Supervised sequence labelling with recurrent neural networks. Studies in computational intelligence, 385.

[7]. Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018). All you need is" love" evading hate speech detection. In Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security (pp. 2-12).

[8]. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[9]. Kumar, R., Ojha, A. K., Malmasi, S., & Zampieri, M. (2018). Benchmarking aggression identification in social media. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018) (pp. 1-11).

[10]. Liu, P., Li, W., & Zou, L. (2019). NULI at SemEval-2019 Task 6: Transfer learning for offensive language detection using bidirectional transformers. In Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 87-91).

[11]. MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. PloS one, 14(8), e0221152.

[12]. Malmasi, S., & Zampieri, M. (2017). Detecting hate speech in social media. arXiv preprint arXiv:1712.06427.

[13]. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In Proceedings of the 25th international conference on world wide web (pp. 145-153).

[14]. Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., & Yeung, D. Y. (2019). Multilingual and multi-aspect hate speech analysis. arXiv preprint arXiv:1908.11049.

[15]. Qian, J., ElSherief, M., Belding, E., & Wang, W. Y. (2018). Hierarchical CVAE for fine-grained hate speech classification. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 3550-3559).

[16]. Rajamanickam, S., Mishra, P., Subroto, H., & Mathur, H. (2020). Joint hate speech detection and target classification in social media. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 5921-5931).

[17]. Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. arXiv preprint arXiv:1701.08118.

[18]. Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In Proceedings of the fifth international workshop on natural language processing for social media (pp. 1-10).

[19]. Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE transactions on Signal Processing, 45(11), 2673-2681.

[20]. Van Aken, B., Risch, J., Krestel, R., & Löser, A. (2018). Challenges for toxic comment classification: An in-depth error analysis. arXiv preprint arXiv:1809.07572.

[21]. Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. PloS one, 15(12), e0243300.

[22]. Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. In Proceedings of the second workshop on language in social media (pp. 19-26).

[23]. Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the NAACL student research workshop (pp. 88-93).

[24]. Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting hate speech on twitter using a convolution-gru based deep neural network. In European semantic web conference (pp. 745-760). Springer, Cham.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

[25]. Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B. (2016). Attention-based bidirectional long short-term memory networks for relation classification. In Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers) (pp. 207-212).

[26]. Zimmerman, S., Kruschwitz, U., & Fox, C. (2018). Improving hate speech detection with deep learning ensembles. In Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018).

[27]. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the type and target of offensive posts in social media. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 1415-1420).

[28]. Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR), 51(4), 1-30.

[29]. Fersini, E., Rosso, P., & Anzovino, M. (2018). Overview of the task on automatic misogyny identification at ibereval 2018. In IberEval@ SEPLN (pp. 214-228).

[30]. Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1), 2053951719897945.

DOI: 10.48175/IJARSCT-26434