AI BASED AUDIO SURVEILLANCE SYSTEM FOR DETECTING OFFENSIVE SPEECH IN PUBLIC SPACES

1st Arjun K
Dept. of CSE
St. Thomas College of Engineering and Technology
Mattannur, India
arju0og@gmail.com

2nd Ashika Raveendran
Dept. of CSE
St. Thomas College of Engineering and Technology
Mattannur, India
ashikaraveendran00@gmail.com

3rd Aysha P P
Dept. of CSE
St. Thomas College of Engineering and Technology
Mattannur, India
aysha313.ma@gmail.com

4th Karthik K
Dept. of CSE
St. Thomas College of Engineering and Technology
Mattannur, India
7604karthik@gmail.com

5th Dr. Shinu Mathew John
Dept. of CSE
St. Thomas College of Engineering and Technology
Mattannur, India
principal@stthomaskannur.ac.in

***Abstract*—The AI-Based Audio Surveillance System for Detecting Offensive Speech in Public Spaces is an application proposed to detect the offensive speech and alert it to the respected authority. This overcomes the challenges of long audio recordings which cannot be manually segmented or analyzed easily, making timely identification of abusive or threatening speech difficult. The proposed system provides an efficient and practical solution by automatically detecting harmful, offensive or panic-inducing language in real world environments. This ensures faster responses and enhances public safety through proactive audio based monitoring. The system is structured into three main modules the Admin Module, which managesauthorities, reports and system settings the Authorities Module, which receives real time alerts, accesses audio evidence and locates incidents through geotagging and the AI Module, which powers the core intelligence of the system through speech recognition, natural language processing and machine learning. Together these modules coordinate seamlessly to provide an intelligent and efficient surveillance framework. The app functions through these integrated stages continuous audio capture from the environment, speech-to-text conversion using ASR, semantic and sentiment analysis through NLP, harmful speech classification via machine learning and finally automated alert generation with audio evidence and location data. By minimizing human monitoring while enhancing accuracy and response time, the system delivers a scalable, real time and AI-driven solution for modern public safety needs.**

Keywords— Hate Speech Detection, ASR, NLP, Alert Manage.

I. INTRODUCTION

Public spaces are essential to civic life but can also become locations where harmful verbal behavior such as hate speech, harassment, and verbal aggression occurs. Traditional surveillance systems mainly rely on visual monitoring and are unable to detect offensive speech.

Numerous studies have contributed to the advancement of harmful speech detection, ASR, and multimodal analysis, forming the foundation for the proposed system. The MAHGA framework introduces multi-aspect heterogeneous graph analysis for harmful speech detection, modeling explicit keywords, implicit semantics, and emotional signals separately to improve contextual understanding [1]. SSL-GAN-RoBERTa is a semi-supervised model designed to detect anti-Asian COVID-19 hate speech on social media by combining generative adversarial networks with transformer-based language models for enhanced robustness [2]. Emotion decoding approaches using NLP applied to speech data demonstrate that sentiment analysis tools can effectively identify emotional markers such as anger and hostility from ASR outputs [3].

The LibriSpeech corpus provides a large-scale dataset for ASR research derived from public-domain audiobooks, serving as a benchmark resource for training and evaluating speech recognition systems [4]. Double-layer hybrid CNN–RNN models improve hate speech detection on imbalanced datasets by combining deep architectures with oversampling techniques [5]. Word embeddings and deep learning methods have been applied successfully for hate speech detection in Arabic, highlighting the effectiveness of neural approaches across different languages [6].

Applications of artificial intelligence in broadcasting and hosting systems show high ASR accuracy, demonstrating the feasibility of real-time speech transcription in practical environments [7]. Large-scale semi-supervised learning frameworks like BigSSL expand ASR performance using extensive unlabeled data [8]. Augmentation techniques that transform adult speech into child-like samples help address

dataset scarcity and improve ASR robustness [9]. Personalized fine-tuning methods further enhance recognition accuracy by adapting models to individual speakers [10].

Systematic reviews of hate speech detection using NLP provide comprehensive analyses of existing methods and research challenges [11]. Multimodal hate speech detection emphasizes the integration of textual and visual cues for improved performance [12], as seen in tasks such as the Hateful Memes Challenge, which highlights the complexity of detecting hateful content when text and visual elements interact [13]. Explainable hate speech detection is supported by datasets like HateXplain, which incorporate human rationales to promote transparency and interpretability in classification models [14]. Finally, heterogeneous graph attention networks capture complex relationships in structured networks, supporting advanced contextual modeling approaches [15].

A. Problem Statement

The current public surveillance system primarily relies on CCTV cameras and other vision-based tools. While these systems are effective for monitoring physical actions and behaviors, they are unable to detect offensive or harmful speech. This limitation creates a significant gap in public safety, as verbal threats, hate speech, and harassment can go unnoticed in crowded areas.

The problem addressed by this project is the lack of an intelligent audio surveillance system capable of detecting offensive speech in real time within public spaces. By implementing speech recognition, natural language processing, and machine learning techniques, the proposed system can identify harmful speech as it occurs, providing timely alerts to authorities. This approach not only enhances the ability to respond to verbal threats but also helps prevent potential escalation of conflicts. Overall, the system aims to strengthen public safety by offering a comprehensive solution to monitor, detect, and manage harmful speech, thereby improving community security and ensuring a safer environment for everyone.

II. Literature Survey

Previous studies on offensive and harmful speech detection have mainly concentrated on text based content gathered from social media platforms using natural language processing and various machine learning methods. This lack of focused work highlights the need for the proposed AI based audio surveillance framework that merges speech recognition natural language processing and machine learning for reliable offensive speech detection in public environments.

A. Artificial Intelligence Technology in the Field of Broadcasting and Hosting

The proposed system presents an intelligent and fully automated framework designed for broadcasting and hosting by combining several advanced technologies, including automatic speech recognition, natural language processing, and real time content generation. Its purpose is to remove the reliance on manual script preparation and human voice delivery while still ensuring strong levels of clarity, personalization, and audience engagement.

To ensure modularity, adaptability, and straightforward maintenance, the system follows a layered architectural design that is represented in figure 1. The Application Layer offers the interface through which users provide input, receive output, and adjust settings. The Service Layer directs data movement and computational processes between different parts of the system. The Model Layer performs operations including speech recognition, transcription checking, natural language processing based script creation, and optional text to speech conversion. The Data Layer stores inputs, logs metadata, retains user preferences, and organizes archived outputs. Each individual layer works independently and can be upgraded or replaced when improved models or services appear.

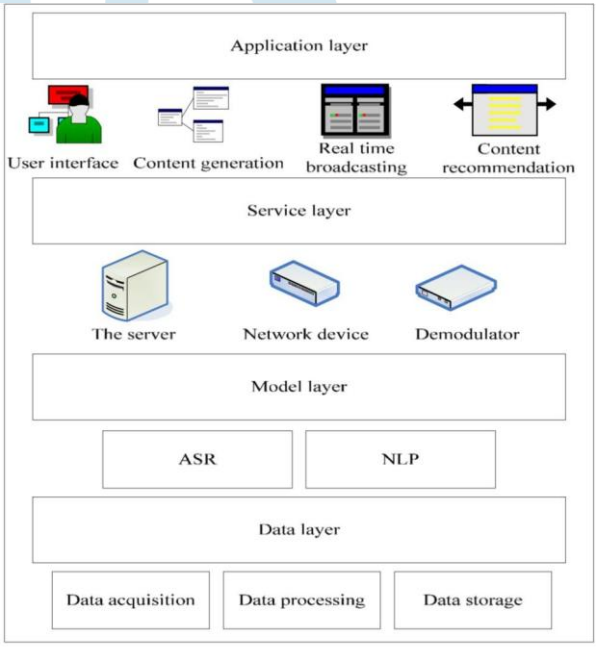

Fig. 1. Framework of Intelligent Broadcasting and Hosting System

B. Emotion Decoding of Speech Using NLP Analysis Approach

Recent advancements in artificial intelligence, natural language processing, and acoustic signal processing have accelerated the growth of emotion recognition from speech. Existing studies in speech emotion analysis have explored improvements in speech recognition, transcription evaluation, sentiment analysis, and emotion visualization, all of which form the basis for the current project on emotion decoding using NLP [11]. Preprocessing reduces noise and normalizes audio levels, while feature extraction methods such as MFCCs and log mel spectrograms convert sound signals into numerical forms suitable for modeling. Acoustic models, often based on RNN, LSTM, or Transformer architectures, interpret sequential audio patterns to predict phonemes or subword units before

the decoding stage converts these predictions into text. This evaluation step improves the quality of input data for emotion analysis, leading to more accurate detection performance [10].

C. MAHGA: Multi-Aspect Heterogeneous Graph Analysis for Harmful Speech Detection on Social Networks

Graph based methods offer an organized way to detect harmful speech on social media by mapping the connections between users, their posts, and the surrounding context. The figure 2 illustrates how MAHGA models different entities such as users, posts and contextual features as nodes, while interactions like replies, mentions and shared topics are represented as edges in a heterogeneous graph.

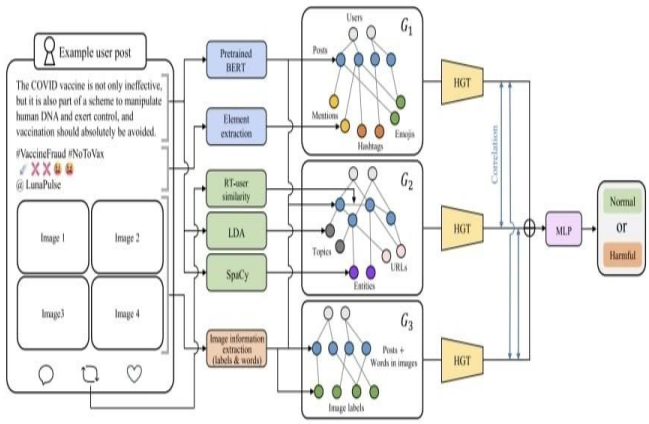

Fig. 2. MAGHA Framework

By applying Graph Neural Networks (GNNs) on this multi-aspect structure, the model learns complex social and linguistic dependencies that traditional classifiers fail to capture. Multi aspect heterogeneous graph [15] frameworks such as MAHGA address these limitations by integrating structural linguistic and behavioral data into a unified analysis process.

III. METHODODLOGY

Advancements in artificial intelligence and speech processing have significantly improved public safety monitoring, particularly in detecting harmful and abusive speech in crowded environments. For the proposed Smart Audio Surveillance System, multiple processing stages including Automatic Speech Recognition, Natural Language Processing, and Machine Learning classification are used.

A. Automatic Speech Recognition : Converts real-time audio input into accurate textual data under noisy conditions.

B. Natural Language Processing : Performs text cleaning, tokenization, and semantic analysis for understanding speech intent.

C. Random Forest Classification : Uses ensemble learning to improve detection accuracy and reduce false alerts.

D. Django-Based Web Module : Provides secure admin control, report management, and system monitoring.

E. Flutter Mobile Application : Enables authorities and users to receive alerts, record audio, and manage profiles.

F. Database Management : Stores audio, transcripts, and evidence in structured format for analysis and review.

G. Alert Generation System : Automatically forwards verified threat information to authorities for immediate action.

IV. PROPOSED SYSTEM

The proposed Smart AI-Based Audio Surveillance System aims to transform traditional public security monitoring into an intelligent, automated, and data-driven solution. Automatic Speech Recognition (ASR) technology converts the audio into text, which is then analyzed using Natural Language Processing (NLP) techniques. The processed data is stored in an SQL database in a structured CSV format for training, analysis, and future improvement. Harmful speech detection is carried out using a Random Forest machine learning algorithm, trained on labeled datasets containing both abusive and non-abusive speech samples.

The system follows the Model-View-Template (MVT) architecture and is implemented using the Django framework for the admin web portal. The frontend is developed using HTML, CSS, and JavaScript to provide a responsive and interactive user interface. Secure data communication between modules is achieved using GET and POST methods, enabling real-time updates and reliable data transfer.

Overall, the proposed system operates with minimal human intervention and offers a secure, reliable, and scalable solution for modern public safety management by integrating Artificial Intelligence, Machine Learning, Django-based web applications, Flutter mobile applications, and structured database management.

V. FEASIBILITY STUDY

The feasibility study for this system is divided into the following major aspects: technical feasibility, operational feasibility, economic feasibility, legal feasibility, and social/environmental feasibility.

A. Technical Feasibility

The technical feasibility of the proposed system is high, as it relies on well-established and widely used technologies such as speech recognition, natural language processing, machine learning algorithms, and web and mobile development frameworks.The backend system is developed using Python and Django framework in PyCharm, following the MVT architecture. The frontend uses HTML, CSS, and JavaScript, while mobile applications are developed using Flutter in Android Studio. The SQL database manages audio records, transcripts, user data, and alert logs.

Machine learning models based on Random Forest algorithms are trained using CSV datasets to classify harmful speech efficiently. Data communication between modules is achieved through secure GET and POST requests.

B. Operational Feasibility

From an operational perspective, the proposed system is designed for automation, ease of use, and minimal manual monitoring.The Admin dashboard allows administrators to manage users, authorities, reports, feedback, and system settings efficiently. Authorities can view alerts, access location information, verify evidence, and update incident status through mobile applications. Users can submit recordings and manage personal details using the Flutter-based interface.

C. Economic Feasibility

Economic feasibility evaluates whether the benefits of the proposed system are worth the costs required for its development, deployment, and maintenance. The initial investment includes expenses for audio capturing devices such as microphones, servers, software development, AI model training, cloud infrastructure, and personnel training.

VI. PROPOSED SYSTEM DESIGN

The system architecture of the proposed Smart AI-Based Audio Surveillance System is designed to bring together audio capturing devices, AI processing modules, database systems, and user interfaces into a single, integrated framework for intelligent public safety monitoring. The architecture explains how audio data collected from public environments is processed at a central server and then made accessible to administrators, authorities, and users through web dashboards and mobile applications.

Figure 3 illustrates the overall system architecture and the interaction between key components such as the Admin module, Authority module, AI module, Alert Management, and Notification Management. User authentication and account management are handled through centralized services, allowing administrators to control authority access and monitor system activities. The AI module processes real-time audio input, converts it into text, and analyzes it to detect potential threats. When harmful speech is identified, the alert management module generates notifications and forwards them to the concerned authorities.

Authorities can review alerts, access supporting evidence and reports, and provide feedback through the review management system. This architecture ensures secure access control, real-time alert delivery, and smooth communication among all system components, enabling effective and reliable public safety management.

The application layer includes both web-based and mobile-based interfaces for administrators, authorities, and users. The admin interface supports user management, report viewing, system monitoring, and feedback control. The authority interface provides access to alerts, recordings, location information, notifications, and incident status updates. Additionally, the system is designed with scalability, data privacy, and reliability in mind. Backup and logging mechanisms further enhance system reliability by maintaining records of system activities and preserving critical data in case of failures, thereby ensuring continuous and uninterrupted surveillance operations.

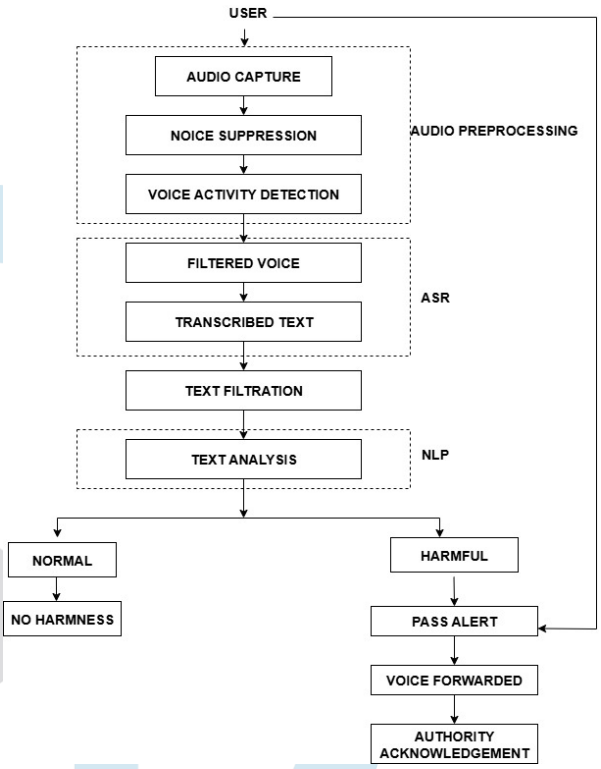

Fig. 3. Architecture diagram

VII. RESULT

The system shows that the hate speech detection platform works effectively across user, authority, and admin modules. The system successfully captures voice input, processes it using AI models, and accurately classifies speech as hate or non-hate in real time. Users can record speech easily, authorities can monitor live data and receive alerts, and admins can manage users and authorities through a structured dashboard. Detected hate speech is automatically flagged and forwarded to the concerned authority module for action. The integrated design ensures smooth coordination between detection, monitoring, and response, making the system reliable for real-time harmful speech detection and control.

The fig.4 shows the home page of the admin dashboard. It explains the purpose of SoundWatch in detecting hate speech in public places. The message highlights the impact of words on society. The illustration represents public speech being monitored and analyzed by the system.

The fig. 5 shows the hate speech reports page on the admin website. It displays detected speech data in a table format. Each entry shows the speech text and whether it is classified as hate or not. This page helps the admin review, analyze, and manage hate speech incidents effectively.

The fig. 6 show the Authority Details page shows a list of all authorities who has been registered by the admin. Each row shows the authority's department, email, mobile number, and city. There are Edit and Delete buttons at the end of each row. The admin can click Edit to update an authority's details or

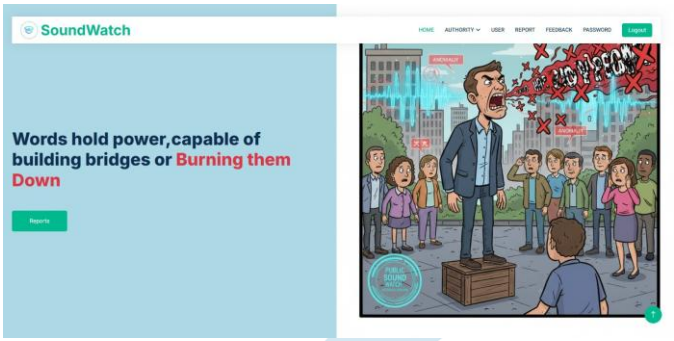

Fig. 4. Admin Interface

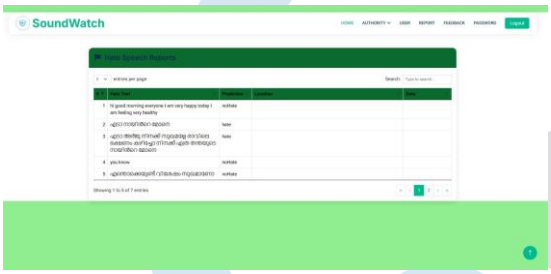

Fig. 5. Hate Speech Reports

Delete to remove the authority from the list. The table also has a search box at the top and pagination controls at the bottom.

The fig. 7 shows the registered user page which lists all registered users in a table for the admin to monitor. Each row has the user's name, email address, mobile number, profile photo, place, gender, and date of birth. The table also has an index column (number) on the left. The admin can scan this list to view user details. There is also a search box to find a specific user and pagination to navigate through the entries.

The fig. 8 is the user application of the hate speech detection system. The home screen welcomes the user and shows the Speech Monitoring System with a short description of real-time AI-based detection. A Start Detection button allows the user to begin the speech analysis process. The app has a simple bottom navigation bar for Home, Voice Detection, and Profile. This fig. 9 is the authority application of the SoundWatch hate speech detection system. The home screen welcomes the authority user (police) and shows AI-powered detection

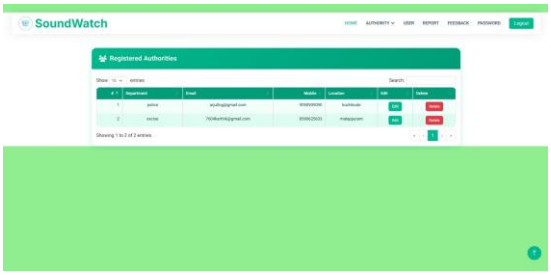

Fig. 6. Authority Details

Fig. 7. User Details

for real-time monitoring. It displays live monitoring* and highlights features like 24/7 protection, real-time analysis, and high accuracy. The home page also provides quick access to alerts and system features through simple navigation icons.

Fig. 8. User Interface

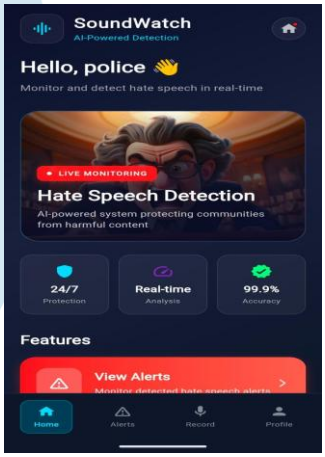

Fig. 9. Authority Interface

The alert generation module continuously monitors the output produced by the classification model to identify instances of harmful or hate speech. Whenever a potential threat is detected, the system automatically generates a structured alert containing the speech transcript, confidence score, timestamp, and corresponding location details. This alert is securely stored in the backend database for logging, reporting, and future reference Fig. 10. Authorities can acknowledge alerts, update the status of incidents , add remarks, and provide feedback, enabling systematic incident management and traceability.

Figure 11 presents the user hate speech detection results and demonstrates the effectiveness of the proposed model. In the first example, the input contains safe and neutral content, which the system correctly classifies as safe, indicating a low probability of harmful intent. In the second example, the input contains offensive or hate-related expressions, and the system accurately classifies it as not safe. The results highlight the model's ability to differentiate between normal communication and harmful speech with high reliability. This accurate classification ensures that only genuine threats trigger alerts, thereby reducing false positives while enabling timely intervention and appropriate corrective action when necessary.

Fig. 10. Hate Speech Alert

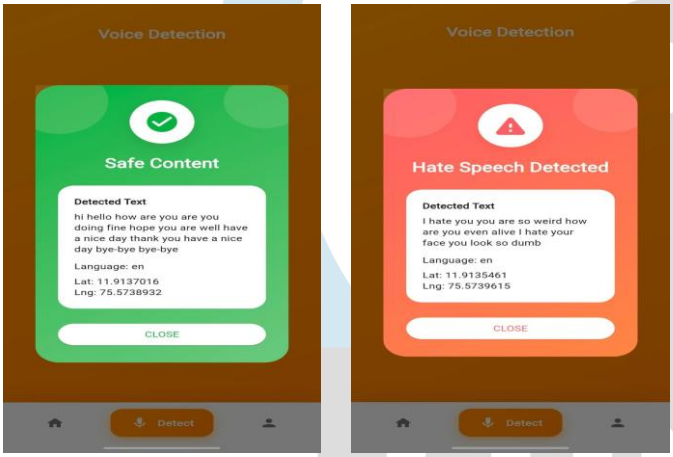

Fig. 11. User Hate Detection

VIII. CONCLUSION

The proposed Smart AI-Based Audio Surveillance System is designed to improve public safety by continuously monitoring audio and automatically detecting potential threats in real time. The system works by capturing audio, converting speech into text, and analyzing it using natural language processing and machine learning techniques such as Random Forest classification. Overall, the system provides a reliable, secure, and scalable solution for modern public safety, supporting early threat detection and helping create safer communities.

IX. FUTURE SCOPE

The proposed real-time audio-based hate speech surveillance system is designed to enhance public safety by automatically detecting harmful and aggressive speech in diverse environments. Future improvements may include multilingual and code-mixed speech recognition, context-aware NLP models, and emotion analysis to reduce false detections and improve accuracy. The system can also be extended with speaker identification, hotspot mapping, and noise-robust audio processing to strengthen monitoring in crowded or outdoor settings.

REFERENCES

[1] Yoshida, Ryo, Soh Yoshida, and Mitsuji Muneyasu. ”MAHGA: Multi-Aspect Heterogeneous Graph Analysis for Harmful Speech Detection on Social Networks.” IEEE Access 13 (2025): 106673-106687. DOI:10.1109/ACCESS.2025.3581214

[2] X. Su, Y. Li, P. Branco, and D. Inkpen, “SSL-GAN-RoBERTa: A robust semi-supervised model for detecting anti-asian COVID-19 hate speech on social media,” Natural Lang. Eng., vol. 30, no. 6, pp. 1161–1180, Nov. 2024.

[3] Bobbili, Charitha, Sri Varsha Dhushetti, Harshith Y, J. Shanmugapriyan, Shanmugasundaram Hariharan, and Bandar Vaishnavi. ”Emotion Decoding of Speech Using NLP Analysis Approach.” 2025 International Conference on Intelligent Control, Computing and Communications (IC3). IEEE, 2025. DOI:10.1109/IC363308.2025.10956754

[4] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2015, pp. 5206–5210.

[5] Slamet Riyadi, Annisa Divayu Andriyani, and S. N. Sulaiman, “Improving Hate Speech Detection Using Double-Layers Hybrid CNN-RNN Model on Imbalanced Dataset,” IEEE Access, pp. 1–1, Jan. 2024, doi: https://doi.org/10.1109/access.2024.3487433.

[6] H. Faris, I. Aljarah, M. Habib, and P. Castillo, “Hate speech detection using word embedding and deep learning in the arabic language context,” in Proc. 9th Int. Conf. Pattern Recognit. Appl. Methods. Valletta, Malta: SCITEPRESS, 2020, pp. 453–460, doi: 10.5220/0008954004530460.

[7] Wang, Chongchong. ”Artificial Intelligence Technology in the Field of Broadcasting and Hosting.” 2024 International Conference on Advances in Electrical Engineering and Computer Applications (AEECA). IEEE, 2024. DOI:10.1109/AEECA62331.2024.00068

[8] Yu, Daniel S. Park, Wei Han, James Qin, Anmol Gulati, Joel Shor, et al. ”Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition.” IEEE Journal of Selected Topics in Signal Processing 16.6 (2022): 1519-1532. DOI: 10.1109/JSTSP.2022.3182537

[9] M. Y. Yiwere, A. Barcovschi, R. Jain, H. Cucu, and P. Corcoran, “Augmentation Techniques for Adult-Speech to Generate Child-Like Speech Data Samples at Scale,” IEEE Access, vol. 11, pp. 109066–109081, 2023, doi: https://doi.org/10.1109/access.2023.3317360.

[10] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin,R. Leary, J. Li, and Y. Zhang, “Enhancing Automatic Speech Recognition With Personalized Models: Improving Accuracy Through Individualized Fine-Tuning” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), , pp. 6124–6128,May 2020

[11] M. S. Jahan and M. Oussalah, “A systematic review of hate speech automatic detection using natural language processing,” Neurocomputing, vol. 546, May 2023, Art. no. 126232.

[12] K. Perifanos and D. Goutsos, “Multimodal hate speech detection in Greek social media,” Multimodal Technol. Interact., vol. 5, no. 7, p. 34, Jun. 2021.

[13] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, “The hateful memes challenge: Detecting hate speech in multimodal memes,” in Proc. Adv. Neural Inf. Process. Syst., Jan. 2020, pp. 2611–2624.

[14] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, “HateXplain: A benchmark dataset for explainable hate speech detection,” in Proc. AAAI Conf. Artif. Intell., May 2021, vol. 35, no. 17, pp. 14867–14875.

[15] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P. S. Yu, “Heterogeneous graph attention network,” in Proc. Int. World Wide Web Conf., 2019, pp. 2022–2032.