# Toxic Speech Classification Using Deep Learning and Machine Learning: A Survey

**Monarch Jain, Chandra Prakash Singar, Puja Gupta**

[1]Research Scholar,[1,2]Asst Professor

Department of Information Technology , Shri G.S. Institute of Technology & Science, Indore, M.P.

***Abstract:-*** The exponential increase in toxic speech has significantly jeopardized the creation of an inclusive environment for all individuals. Though attempts have been taken to detect and restrict such information online, this is still difficult to discover. Deep learning-based methods have pioneered toxic speech detection. The context-dependent characteristics of poisonous speech, user intention, unwanted biases, etc., render this procedure overcritical. We provide a hierarchical architecture of automated hazardous speech detection difficulties in this study to fully examine them. We examine machine learning and deep learning toxic speech recognition difficulties. At the top, we differentiate data, model, and human issues. We analyze each hierarchical level in detail using examples. This poll will help toxic speech detection researchers create better solutions. This survey paper presents an extensive literature review of deep learning and machine learning methods towards the automatic identification of toxic speech, considering recent technological advancements. A multitude of algorithms and architectures have been evaluated in this context. This paper will assess the positive and negative aspects of various recognition and categorization models regarding speech expressed vocally in multilingual contexts. Additionally, there will be an analysis of occurrences of code-mixing. To demonstrate the impact of these techniques on the overall effectiveness of the model, additional analysis will be performed on the methods employed in feature selection during toxic speech detection.

***Keywords****: Toxic speech challenges, Toxic speech detection, Natural Language Processing.*

## 1. Introduction

The exponential proliferation of communication platforms has had various beneficial benefits on people lives. The increasing popularity of digital communication among the general populace is attributed to people expressing their ideas and opinions without hesitation. All individuals have the right to articulate their thoughts and feelings without the apprehension of retaliation. This privilege is now used to justify discriminatory behaviors and attacks, both physical and verbal, against others under the guise of free speech. This kind of discrimination is termed toxic speech. Toxic speech[1] is a communicative expression of hostility directed against a person or group based on their race, color, faith, gender, nationality, disability, or sexual orientation. This is only one definition of toxic speech; yet, it is a generally acknowledged and often used one. This often results in the propagation of violent and detrimental material, regardless of whether it was shared with purpose. A particular vocabulary is used to abuse individuals or organizations based on their unique attributes, such as gender, race, national origin, disability, and similar factors. Toxic speech refers to this kind of rhetoric. The changeable nature of the material complicates the regulation of its transmission on the internet. Moreover, the mental and physical health of the one addressed may be profoundly affected by a single hostile remark. Belief in the liberty of speech within the internet media landscape has deteriorated over time, complicating the ability to interact freely with all persons [2]. Consequently, it is essential to preemptively discover a resolution to this problem.

The principal obstacle to restricting harmful content on the internet is the lack of a universally accepted definition of hate speech. While people have a basic understanding of toxic speech, that does not equip them with the requisite knowledge to fully comprehend it. Furthermore, many limitations and rules have been instituted by social network platforms, including Twitter, YouTube, among Facebook, to govern its distribution on their sites [24]. Nonetheless, the immense volume of data collected poses management challenges and requires considerable focus. The ratio of hostile material on the internet is much lower than that of positive or neutral content [15], which therefore biases toxic speech identification algorithms in different ways. We want to examine and clarify these issues comprehensively in this study.

Concept behind Toxic Voice Identification:

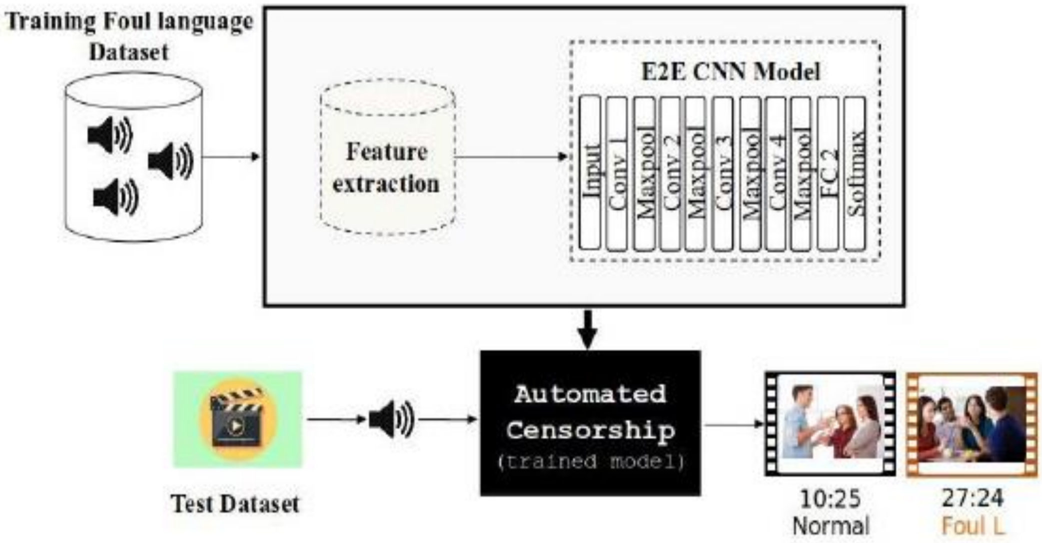

Figure 1: Framework of Proposed Toxic System based on voice[18]

The figure 1 illustrates the overall framework for toxic voice identification. Audio samples from a foul-language training dataset are processed with feature extraction and passed through an end-to-end CNN architecture comprising multiple convolutional and pooling layers for classification. Once trained, the model performs automated censorship by analyzing audio from test videos. The system outputs time-stamped segments labeled as normal or containing toxic/foul speech.

Concept behind toxic words identification

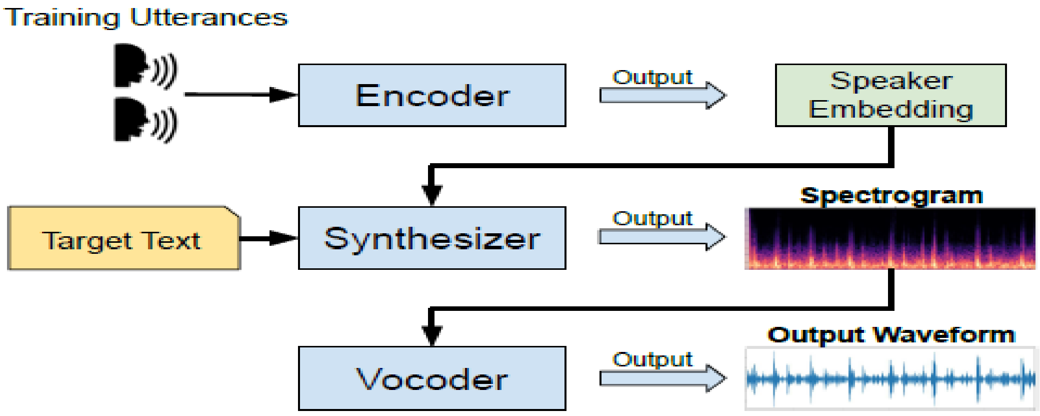

Figure 2: Framework of Proposed Toxic System based on words[30]

The figure 2 illustrates a three-stage audio processing pipeline used for detecting toxic or abusive words in speech. Training utterances are first encoded to create speaker embeddings that capture important voice characteristics. The synthesizer then converts the target text into a spectrogram, providing a detailed view of the speech features. Finally, the vocoder generates the corresponding audio waveform, enabling the system to analyze and classify the spoken content for toxicity. This paper comprises the following section after introduction, literature review, comparison with various methods then conclusion.

## 2. Literature review

The literature review presented below explores various research on multimodal systems leveraging text, audio, and video data, also delivering a grounds for understanding current approaches and research trends.

Based on the Text, various authors proposed

Herwanto et al. (2019) worked on developing a toxic speech and abusive language classification model for Indonesian with the help of deep learning approach. A dataset compiled from three previous studies, including the tweets from Twitter, Facebook, and YouTube was used by the researchers. The model utilized the fastText algorithm with a continuous bag-of-words (CBOW) representation and was trained with and without pre-trained data from Wikipedia. The highest F1-score for binary classification was 0.873 with pre-trained data, 50 epochs, and no sub-words, whereas the lowest came upto 0.764 without pre-trained data and 5 epochs with sub-words. It was found that models that used pre-trained vectors from Wiki, outperformed the ones that did not.[3]

Kalaivani et al. (2021) focused on multilingual hate speech and offensive language detection in English, Hindi (code-mixed), and Marathi. This study was directed to address both binary classification (HOF vs. NOT) and multi-class classification (HATE, OFFN, PRFN)for all the three languages. The researchers investigated wide range of approaches, including traditional machine learning models, ULMFiT, and transformer-based architectures such as BERT, RoBERTa, ALBERT, DistilBERT, and mBERT. Experimental results reflect that RoBERTa performs best for English Subtask A, BERT works best for English Subtask B, and mBERT does well for Hindi and Marathi tasks. The system observed macro F1-scores of 0.7919 (English A), 0.6242 (English B), 0.7320 (Hindi A), 0.5110 (Hindi B), and 0.8223 (Marathi A). The results present the effectiveness of multilingual transformer models for hate speech detection in low-resource and code-mixed languages and also note challenges related to handling of sarcasm and dataset imbalance.[4]

Soykan et al. (2022) conducted a study on Turkish profanity detection in search engine queries which was a challenging task particularly due to the agglutinative language structure and the short length queries. They built a dataset of around 400,000 queries labeled - profane or not profane, with 16.4% of the data categorized as profane. Several classical machine learning and deep learning methods, some of which being Logistic Regression, LinearSVC, and transformer-based models like BERT and ELECTRA, were then compared and it was established that the best performance was achieved using the ELECTRA model, with a score of 0.93 F1. The study also showed that LinearSVC too performed almost as well with a score of 0.92 F1 score.[5]

Kim et al. (2022)used an Artificial Neural Network (ANN) to conduct a study analyzing the effect of profanity on sentiment analysis in Korean. They used a movie review dataset called 'NaverSentimentMoviecorpusv 1.0'. The researchers put two models to training: one that included profanity and the other with profanity removed at the data preprocessing stage. The model with profanity showed an accuracy of 83.4%, whereas the model with profanity removed reflected an accuracy of 81.6%. The findings thus suggested that profanity may not always relate to noise data and in fact in this context, can improve the accuracy of sentiment analysis.[6]

Maity et al. (2023) followed a study focused on detection of toxic speech in Malay language, a low-resource language with datasets being narrowly available to public. In line with the study, they created a new benchmark dataset called HateM, consisting over 4,892 manually annotated tweets. A two-channel deep learning model named XLCaps was developed where One channel used XLNet language model followed by a capsule network, and the other one used FastText embedding with a Bi-GRU network. As a result of the study, the XLCaps model outperformed the baseline models, with and overall accuracy of 80.69% and F1 score of 80.41%.[7]

Saleh et al. (2023) took up the challenge to detecting online toxic speech, where coded language is used in order to avoid detection. The study was meant to probe two approaches for toxic speech detection: First being the bidirectional LSTM-based deep model with domain-specific word embedding and second being a fine-tuned BERT language model. A combined dataset from existing toxic speech datasets were used for the purpose which included Davidson-ICWSM, Waseem-EMNLP, and Waseem-NAACL. These experiments resulted in the Bi-LSTM model with domain-specific word embeddings achieving a 93% f1-score, while BERT reflecting a score of 96% f1 on the combined balanced dataset. The study thus concluded that even though BERT outperformed in terms of score, the domain-specific embedding approach came out to be more effective when detecting intentionally misspelled or coded toxic words.[8]

Omran et al. (2023) conducted a comparative analysis of machine learning algorithms for toxic speech detection on an English-language Twitter dataset of 24,783 tweets. He aimed at finding simple, efficient, and high-performing algorithmic combination for real-world deployment. Several models, including SVM, KNN, Random Forest, Naïve Bayes, and Decision Trees were compared by the researchers and it was found that a combination of Naïve Bayes and Decision Trees produces best results, reaching accuracy of 0.887 and an F1-score of 0.885. It was also noted that while other algorithms like SVM and Logistic Regression performed well, the computational requirements could prove a challenge for large datasets. The paper also presents a system design for real-time toxic speech detector which uses visual progress bar and cautionary pop-up message to notify users about the potentially harmful content before it is posted.[9]

Gutha et al. (2023) did a study focusing on detection of toxic speech in low-resource Indian languages, especially Bengali, Bodo, and Assamese, as part of the HASOC 2023-Task 4 competition. The researchers used dataset of tweets and other social media posts for binary classification of content as "toxic and Offensive" (HOF) or "Not Hate-Offensive" (NOT). The models that were explored included LSTM and BiLSTM together with CNN, and pre-trained BERT-based models like IndicBERT and MuRIL. It was found that for languages rich in resource, specialized BERT models were most effective, with IndicBERT achieving an F1 score of 69.726% for Assamese and Bengali MuRIL achieving 71.955% for Bengali. For low-resource Bodo language, a BiLSTM model with an additional Dense Layer produced best results, an F1 score of 83.513%. The study noted that tailoring NLP methodologies to specific resources of a language is highly important, and for low-resource languages like Bodo, neural network-based approaches may prove more effective in comparison to pre-trained BERT models.[10]

Awal et al. (2023) proposed a new framework named HateMAML addressing the issue of multilingual toxic speech detection in low-resource languages. The model-agnostic meta-learning (MAML)-based approach used a self-supervision approach to beat data scarcity and enabled quick adaptation to new languages and domains. Researchers conducted experiments on five datasets across eight low-resource languages. The results reflected that HateMAML outperformed state-of-the-art fine-tuning baselines by over 3% in cross-domain multilingual transfer settings. The study resulted in favor of the opinion that meta-training can prove to be an effective alternative to standard fine-tuning, providing high-end performance by learning a good initial model that adapts rapidly to new tasks.[11]

Singh et al. (2024) worked on developing an automated framework used to identify toxic speech on social media in low-resource Indian languages. The researchers proposed a federated learning approach known as MultiFED, trying to solve the issue of data scarcity and user privacy by training the models on client devices without sharing sensitive data with a central server. The study used a combined dataset of around 300,000 texts from 13 Indian languages and English, gathered from platforms such as Sharechat, YouTube, and Facebook. The MultiFED approach, which used fair client selection and pre-trained models like XLM-ROBERTa and Indic-BERT, performed better in comparison to state-of-the-art centralized baselines by over 8% in accuracy and 12% in F1-score. This research also noted that performance of federated learning models is good on diverse datasets and can be scaled while guarding user privacy.[12]

Abdellaoui et al. (2024) worked on offensive language detection in Moroccan Darija. The researchers created a dataset of more than 20,000 phrases from social media platforms, and labelled 37.8% of them as offensive. They fine-tuned various language models on this dataset, concluding that Darija RoBERTa-based model performed with 90% accuracy and an F1 score of 85% and proves to be best-prforming. This study also evaluated the robustness and fairness of this model by using metamorphic testing and adversarial attacks. It was noted that the model was vulnerable to following attacks: inserting dots (29.4% success rate) and spaces (24.5%), and modifying characters (18.3%). Fairness tests observed a 7% success rate for the attacks that targeted entities subject to discrimination, reflecting a bias in some cases. The authors drew a conclusion that it is not sufficient to evaluate machine learning systems solely on offline metrics like F1 score and accuracy.[13]

Mnassri et al. (2024) proposed a partially supervised generative adversarial approach to identify multilingual toxic speech and offensive language. The model, known as SS-GAN-PLM, combines Generative Adversarial Networks (GANs) with pretrained language models (PLMs) such as mBERT and XLM-ROBERTa. It used only 20% of the labeled data from HASOC2019 Indo-European corpora (English, German, and Hindi), and the SS-GAN-mBERT model outperformed a baseline semi-supervised mBERT model by an average F1 score of 9.23% with an increased accuracy of 5.75%. This study reflected that this approach is effective in multilingual, zero-shot crosslingual, and monolingual training scenarios, successfully mitigated the issue of data scarcity in toxic speech detection.[14]

Spiesberger et al. (2025) worked on inspecting abusive speech detection in audio recordings of 10 Indic languages while using only acoustic and prosodic features. These used the ADIMA dataset, containing 11,775 audio recordings from real-life conversations on ShareChat. Models were trained in both multilingual and cross-lingual settings and it was found that it was possible to classify abusive and non-abusive content using only paralinguistic features. The Random Forest (RF) classifier performed exceptionally well on the extracted EGEMAPS feature set, achieving Unweighted Average Recall (UAR) scores between 0.70 and 0.84 in multilingual settings, and between 0.66 and 0.84 in cross-lingual settings. Features related to loudness, mean F1, F2, and F3 amplitude, and mean spectral flux, were found to be most effective aligning with features of angry speech. The study concluded that relying on emotional and acoustic cues is a feasible approach for low-resource languages where text-based methods may not work due to lack of transcribed data or bad quality of audio.[15]

## Comparison Table on  Text-based paper

| Ref | Paper | Year | Modality | Language(s) | Models / Key technique | Dataset(s) | Key metrics/results (reported) | Strengths |
|---|---|---|---|---|---|---|---|---|
| [3] | ToxicSpeech and Abusive Language Classification (fastText) | 2019 | Text | Indonesian | fastText (CBOW + subwords) | Combined 3 Indonesian datasets (Twitter etc.) | Binary F1 up to 0.873 (with wiki pre-trained) | Simple and effective; pre-trained vectors help |
| [4] | Multilingual toxic speech and Offensive language detection in English, Hindi, and Marathi languages | 2021 | Text | English, Hindi, Marathi | BERT, Bi-directional LSTM | HASOC 2019 dataset, OffensEval 2019 | F1 scores of 0.7919 (English), 0.7320 (Hindi), and 0.8223 (Marathi) | Compares performance across multiple languages and models in a competition setting |
| [5] | A Comparison of ML Techniques for Turkish | 2022 | Text | Turkish | LR, LinearSVC, MultinomialNB, LSTM, BERT, ELECTRA | ~400k search queries (custom) | Electra/ BERT F1 ≈0.93; LinearSV | Large dataset; comparison across many |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Profanity | | | | | | C ≈0.92 | models |
| [6] | A Study of Profanity Effect in Sentiment Analysis (ANN) | 2022 | Text | Korean | LSTM (single-layer) | Naver movie corpus (train 1500; test 50k) | Accuracy incl. profanity 83.4 vs 81.6 w/o | Shows that profanity can improve sentiment tasks |
| [7] | A Deep Learning Framework for Malay toxicSpeech (XLCaps) | 2023 | Text | Malay | XLNet+Capsule; FastText+BiGRU+Attention (two-channel) | HateM (4,892 tweets) | Acc 80.69; F1 80.41 | Handles OOV & noisy Malay; beats baselines |
| [8] | Detection of toxicSpeech using BERT and toxicSpeech Word Embedding with Deep Model | 2023 | Text | English | BERT and Bidirectional LSTM with domain-specific word embeddings | Combined balanced dataset from available toxic speech datasets | BERT achieved 96% F1-score; BiLSTM achieved 93% F1-score | Highlights the importance of domain-specific embeddings; BERT performs exceptionally well on this task |
| [9] | A comparative analysis of machine learning algorithms for toxicspeech detection in social media | 2023 | Text | English | Naïve Bayes and Decision Tree | RAVDESS and SAVEE | Accuracy of 0.887 and F1-score of 0.885 | Balances simplicity and efficiency with strong performance |
| [10] | Multilingual toxicSpeech and Offensive Language Detection of Low-Resource Languages | 2023 | Text | Multi (Bengali, Bodo, Assamese) | IndicBERT, MuRIL, BiLSTM with CNN | HASOC 2023 dataset (Task 4) | F1 score of 83.513% for Bodo; 71.955% for Bengali; 69.726% for Assamese | Addresses low-resource Indian languages; demonstrates strong performa |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | nce with various models |
| [11] | Model-Agnostic Meta-Learning for Multilingual toxicSpeech | 2023 | Text | Multi | HateMAML (a MAML-based framework) | Founta-EN, HatEval 19, SemEval 20, HASOC 20, HaSpeedDe | Outperforms state-of-the-art baselines by >3% in cross-domaintransfer | Effectively handles low-resource languages and domain generalization; good for few-shot cross-lingual learning |
| [12] | Investigating Offensive LanguageDetection in a Low-Resource Setting with a Robustness Perspective | 2024 | Text | Moroccan Darija | Darija RoBERTa-based model | Human-labeled dataset of Darija text from social media | Accuracy 90%, F1 score 85% | Novel human-labeled dataset for a low-resource language; evaluates robustness and fairness |
| [13] | Generalizable Multilingual toxicSpeech Detection on Low-Resource Indian Languages using Fair Selection in Federated Learning | 2024 | Text | 13 Indic languages | MultiFED (federated approach) with BERT models | 13 Indic datasets | Outperforms baselines by 8% in accuracy and 12% in F-Score | Federated learning approach is suitable for decentralized data; improves generalization across datasets |
| [14] | Multilingual toxicSpeech Detection: A Semi-Supervised Generative Adversarial | 2025 | Text | Multi (English, German, Hindi) | Semi-supervised GAN with mBERT and XLM-RoBERTa | HASOC 2019 dataset | F1 score boost of 9.23% and accuracy increase of 5.75% over baseline | Leverages unlabeled data to improve performance; |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Approach | | | | | | mBERT | effective for multilingual and zero-shot cross-lingual tasks |
| [15] | Hate Speech Detection Using Large Language Models: A Comprehensive Review | 2025 | Text | Multi | LLMs for toxicspeech | N/A | N/A | Broad overview of LLM methods, challenges, and future directions for hate speech detection |

Based on Audio Video, various authors proposed

Hyder et al. (2017) worked on developing a system for Acoustic Scene Classification (ASC) using CNN-SuperVector (CNN-SV) approach combining auditory and spectrogram image features. This research used DCASE 2016 ASC challenge dataset, containing 30-second audio sections from 15 different indoor and outdoor locations. The researchers compared various features including linear-scaled, log-scaled, and Mel-scaled spectrograms, and noted that log-scaled and Mel-scaled spectrograms performed best, observing an average accuracy rate of 80%. The proposed CNN-SV approach, using activations from the final CNN layer to form a high-dimensional vector for a Probabilistic Linear Discriminant Analysis (PLDA) classifier, performed better than traditional CNN and GMM-SuperVector systems consistently. A merged score of multiple systems yielded a 7% improvement relatively in overall accuracy when compared to the baseline CNN system.[16]

Ghaleb et al. (2020) suggested a novel multimodal temporal deep network for enhancement of emotion recognition from audio-video clips. Two streams of audio-visual networks, connected incrementally via LSTM cells to model temporal dependencies were used. The method was evaluated on two datasets, CREMA-D and RAVDESS, achieving state-of-the-art performance on both. On CREMA-D, the model achieved an accuracy of 74.0%, outperforming human perception and other baselines. This study concluded that multimodal perception is a time function, with emotion recognition rates increasing over the duration of the clip, and that positive emotions are easier to identify and that too with more accuracy than the negative ones.[17]

Wazir et al. (2020) worked to mitigate the issue of foul language censorship in audio and video content manually, where possibilities of human error or inefficiency rises. This research suggessted an automated and robust detection model using deep Convolutional Neural Networks (CNNs) trained on spectrogram images derived from words that are spoken by an individual. Foul language dataset was gathered with 2-class (Foul vs. Normal) and 10-class annotation issues. Among the models tested, the Resnet50 architecture achieved the best performance, with low error rate of around 1.24% and high F1-score of 98.54% for the 2-class problem. It was thus noted as a result of such study that deep CNNs are practical and effective solution for classifying speech spectral images for censor purposes.[18]

Alcântara et al. (2020) addressed the issues related offensive video detection by making and publishing a dataset of 400 YouTube videos in Portuguese, named OffVidPT. This dataset includes textual and statistical charecteristics such as video descriptions, tags, titles, transcriptions, and view counts. The experiment was carried out with Classic, Deep Learning, and Transfer Learning models. The research thus reflected that while Deep Learning classifiers with word embeddings and CNN architectures were found to be performing best on average, the textual features were also sufficient for detection of offensive videos. All in all, the best result was obtained as an AUC of 0.78, with an F1 score of 0.74, putting it within the range of close competitions.[19]

Ba Wazir et al. (2021) advanced a deep learning-based system for foul language recognition in speech to help censor films. The researchers applied two different end-to-end deep neural network (DNN) architectures: a Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) cells. They produced a novel foul language dataset called the MMUTM foul language dataset, containing nine indecent words and samples of normal conversations. The dataset was labeled manually and exaggerated to increase its size and robustness to noise. The CNN model outperformed the RNN model in identifying pre-segmented foul language samples. The suggested CNN model observed a 96.92% F1-score for the foul class and a 98.39% F1-score for the normal class in a two-class problem. It also performed better than state-of-the-art pre-trained neural networks, reflecting a 1.91% improvement in F1-score and a 1.57% reduction in False Negative Rate (FNR) when compared to the best baseline model. The study put focus on the fact that the model is lightweight, with only 57k parameters. This makes it suitable for real-time applications with nominal computational cost.[20]

A. Chaudhari et al. (2021) advanced a system to detect indecency and accordingly remove it from videos using machine learning. The approach is carried out in two staged pipeline: an automatic speech recognition (ASR) system to transcribe audio to text, followed by a text-based profanity detection model. A dataset of 50 videos collected from various social media platforms, totaling approximately 29 minutes of footage was produced. The system uses a Speech-to-Text library for transcription and verifies the output text against an indecency check-list. The model observed an accuracy rate of around 85.03% on their dataset. This research concluded a practical and efficient method for automated content moderation, aimed at reducing the need for manual screening.[21]

Xia et al. (2022) worked on Speech Emotion Recognition (SER) which used a Deep Convolutional Neural Network (DCNN) with a data augmentation technique known as Random Circular Shift (RCS). The process involved using a time-frequency representation of the speech signal as input to the DCNN, along with its delta and delta-delta features. They experimented with DCNN architectures like Alexnet, Resnet-152, and Inception-v3, trained already on the ImageNet dataset. The study used two publicly available datasets: eNTERFACE05 and EMO-DB. The results reflected that Alexnet, when combined with RCS, observed the highest accuracy rate of 91.25% on the eNTERFACE05 dataset. This approach outperformed a more complex state-of-the-art method based on Discriminant Temporal Pyramid Matching (DCNN-DTPM) on the same dataset, which achieved an accuracy rate of 79.25% in its turn. However, their model had a little lower accuracy rate of 81.82% on the EMO-DB dataset compared to DCNN-DTPM's rate of 87.31%. It was concluded from this study that RCS significantly improves classification results and that for audio classification using time-frequency representations, Alexnet is a better choice.[22]

Thakran et al. (2023) put their focus on audio abuse detection in a multilingual social media context, with an assumption that abusive behavior produces distinct acoustic cues which may be detected without transcription. A framework called ACMAD using two modalities: the underlying emotions expressed and the language features of the audio, was used by the researchers. They used the ADIMA benchmark dataset, containing an audio of 65 hours from ShareChat in 10 Indic languages. This ACMAD approach observed a state-of-the-art test accuracy rate of 96% with an F1 score of 0.9579 on the test set, performing better than existing models by a well enough margin. It was concluded from the study that using large pre-trained acoustic or language models alone would not be sufficient for this task. Instead, the success of ACMAD is aligned with its careful selection of models for each modality, such as the IndicWav2Vec-Base model for language features and the XLS-R 300M model fine-tuned on RAVDESS for emotion features.[23]

Garg et al. (2024) comprehensively review the current state of Hate Speech Detection (HSD) using Large Language Models (LLMs). The paper highlights the evolution of HSD from traditional machine learning to deep learning and, most recently, to LLMs. Several LLM-based HSD techniques, including fine-tuning, zero-shot and few-shot learning, and in-context learning, were examined. This review covers wide range of datasets used for HSD, such as the HateXplain and Gab datasets, and addresses the difficulty posed by different languages, including code-mixed and low-resource languages. The researchers address key challenges mainly data scarcity, the subjective nature of toxic speech, and the inherent biases of LLMs, providing a futuristic roadmap for research in this area.[24]

Maitya et al. (2024) developed a multimodal multitask framework for the detection of toxic content in code-mixed Hindi-English videos. They proposed the ToxCMM dataset, which covers 931 YouTube videos annotated for toxicity, severity, and sentiment. This framework supports transformer-based models across text, audio, and visual modalities with a gated fusion mechanism. The results of this investigation reflect that multimodal learning performs better than unimodal approaches, achieving a weighted F1 score of 94.35% for toxicity detection. This study highlights the effectiveness of multimodal transformers for toxic content analysis in low-resource, code-mixed settings.[25]

Costa-jussà et al. (2024) worked on developing MuTox, a highly multilingual dataset and classifier to address the lack of multilingual audio-based toxicity detection research. The dataset includes 30 languages from 14 linguistic families, with 20,000 audios for English and Spanish, and 4,000 for the other 28 languages. Human annotators with specific guidelines to label the audio for various types of toxicity, including profanities, toxic speech, pornographic language, and physical violence or bullying language were used by the researchers for the purpose of experiment. The MuTox classifier is a simple architecture with a SONAR encoder and a three-layer binary classifier. It enables zero-shot toxicity detection across a wide range of languages. It was noted that the classifier expands language coverage more than tenfold and performs equivalent to existing text-based trainable classifiers. Also in comparison to a wordlist-based classifier with similar language coverage, MuTox enhances the F1-score by an average of 100%.[26]

Sankaran et al. (2024) addressed cross-lingual audio abuse detection in low-resource Indian languages using Few-Shot Learning (FSL). The researchers put to work a Model-Agnostic Meta-Learning (MAML) framework, focused on pre-trained audio features from Whisper and Wav2Vec models. The model was trained and evaluated on the ADIMA dataset which was a collection of audio clips across 10 Indian languages. In the 100-shot setting, Whisper with L2-Norm feature normalization observed the best accuracy scores, with a range between 78.98% to 85.22%. This study reflected that this few-shot approach is effective in helping pre-trained models to generalize and capture language-specific patterns, while also giving required insights for low-resource abuse detection.[27]

Bentaleb et al. (2024) studies and performed a detailed survey on the evolution of low-latency live media streaming systems with a focus on end-to-end (E2E) latency reduction in IP-based streaming architectures. This paper was meant to analyze live streaming workflows, protocols, and latency sources at following stages: media preparation, delivery and consumption. It examines traditional as well as modern streaming protocols namely DASH, HLS, WebRTC, and budding technologies like Media over QUIC (MOQ), highlighting low-latency extensions which also includes LL-DASH and LL-HLS. This survey also put emphasis on key enabling technologies such as chunked encoding, CMAF, adaptive bitrate algorithms, and playback buffer management. Furthermore, major challenges in achieving scalable and robust low-latency streaming, of which encoding complexity, network variability, and QoE optimization are also part, are also addressed by this study. This helped in positioning this survey work as a reference for future research and development in low-latency live streaming systems.[28]

Arya et al. (2024) introduced a multimodal framework for detection of toxic speech in memes by leveraging Contrastive Language-Image Pre-training (CLIP). The study addresses the challenge of implicit toxic speech in memes, which often relies on the interplay between text and image. The proposed model was evaluated on two datasets: the Hateful Memes Challenge (HMC) and Fakeddit. This framework achieved a state-of-the-art accuracy rate of 83.1% on the HMC dataset by combining a pre-trained vision-language model with a novel multi-task learning approach. As a result of the study, it was noted that jointly analyzing image and text is more effective as compared to using just unimodal methods for identifying toxic speech in multimodal content.[29]

Zhou et al. (2024) developed a survey of recent advances in Speech Language Models (SLMs). The research categorizes SLMs based on their architectures and training objectives, with distinguishes between models for speech recognition, emotion recognition, and acoustic event detection. It details various applications, including ASR and spoken language understanding, also discussing the challenges of building effective SLMs. One of such issue being the high computational cost and the need for large amounts of data. The survey puts lights on the shift from traditional methods to end-to-end deep learning models like Wav2Vec 2.0 and Whisper, which have significantly evolved the field by learning representations directly from raw audio waveforms.[30]

Zhang et al. (2025) proposed MultiTec, a data-driven deep learning framework for detecting healthcare misinformation in short-form videos on TikTok. The system jointly analyzes visual, acoustic, and textual information by learning caption-guided visual representations, acoustic-aware speech features, and cross-modal interactions through a dual-attentive fusion mechanism. The model was evaluated on two real-world TikTok datasets related to COVID-19 disease and COVID-19 vaccines. Experimental results show that MultiTec consistently outperforms state-of-the-art unimodal and multimodal baselines, achieving higher performance across multiple metrics including F1 score, Cohen’s Kappa, and AUC on both datasets. Ablation studies further demonstrate the contribution of each modality and the effectiveness of the dual-attentive fusion strategy. The study concludes that modality-aware multimodal learning is highly effective for identifying healthcare misinformation in short video platforms.[31]

Warren et al. (2025) suggested a better approach focused at detecting audio deepfakes by emphasising on acoustic prosodic analysis, which refers to the high-level linguistic features of human speech, such as pitch, intonation, and jitter. The authors created a detector based on six prosodic features: mean fundamental frequency (F0), jitter, shimmer, and harmonic-to-noise ratio (HNR), and their respective standard deviations. The model was trained and tested on the ASVspoof2021 dataset, achieving a 93% accuracy and an Equal Error Rate (EER) of 24.7%.It was noted from the experiment that this linguistic feature-based approach is robust and provides explanation by an attention mechanism, which identifies jitter, shimmer, and mean-F0 as the most influential features on the model's decision. Also, the model was proven resistant to a simple $L\infty$ norm attack causing a 99.3% accuracy degradation in other baseline models. This in turn reflected the superiority of this model.[32]

Shang et al. (2025) introduced a multimodal framework called MultiTec. It was designed to identify healthcare misinformation in short videos from TikTok. This model focuses on mitigating challenges such as misleading visual content and complex inter-modality dependencies by employing a visual learning module with caption guidance and a dual-attentive transformer mechanism. Two real-world datasets related to COVID-19 disease and COVID-19 vaccines were used to test this system. MultiTec maintained consistency and performed better than state-of-the-art baselines, with an increase of 6.90% and 4.73% in F1-score on the two datasets, respectively. It was also noted that the Visual-speech co-attention mechanism was a critical component for improving detection capabilities, and this was further confirmed by performing an ablation study.[33]

## Comparison Table on Audio Video-based paper

| Ref | Paper | Year | Modality | Language(s) | Models / Key technique | Dataset(s) | Key metrics/results (reported) | Strengths |
|---|---|---|---|---|---|---|---|---|

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| [16] | Acoustic Scene Classification (CNN-SuperVec) | 2017 | Audio | General | CNN-SV + PLDA; score fusion | DCASE2016 | Best fused acc 88.46% | Fusion & hybrid modeling lessons are transferable |
| [17] | Multimodal Attention for Temporal Emotion Recognition | 2019 | Multimodal (audio+video) | English | 3D-CNN for video + SoundNet; triplet mining + gating | CREMA-D, RAVDESS | AV fusion outperforms AO/VO; e.g., 74% on CREMA-D | Temporal multimodal fusion demonstration |
| [18] | Spectrogram-based classification of spoken foul language (Deep CNN) | 2020 | Audio | English | AlexNet, VGG16, GoogLeNet, ResNet50 on spectrograms | MMUTM foul dataset (3105 foul, 5100 normal) | 2-class F1 97.0–98.1; 10-class F1 up to 94.2 (ResNet50) | High performance on the curated dataset |
| [19] | Offensive Video Detection Dataset & Baselines | 2020 | Video | English | Dataset paper + baselines | Offensive Video Detection dataset | Provides dataset & baselines | Valuable dataset resource |
| [20] | Design & Implementation of Fast Spoken Foul Language | 2021 | Audio | English | End-to-end CNN & RNN (LSTM) on spectrograms /MFCC | MMUTM (augmented) | 2-class F1 ~96.9; 10-class avg F1 ~96.1 (CNN) | Lightweight models; low latency |
| [21] | Profanity Detection & Removal in Videos (Chaudhari et al.) | 2021 | Multimodal (audio+vision) | English | STT + lexicon; HOG+SVM face detection; lip pixelation | 50 videos (1734s) | End-to-end accuracy 82.35% | Practical pipeline for mute+pixelate |
| [22] | Speech Emotion Recognition using Time-frequency Random Circular Shift and | 2022 | Audio | English | DCNN with time-frequency random circular shift (RCS) | ENTERFACE05, EMO-DB | Achieved 85.33% accuracy on ENTERFACE05 (better than | Simple yet efficient method; RCS improves accuracy; lightweig |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Deep Neural Networks | | | | | | DCNN-DTPM); 81.82% on EMO-DB | ht architecture (Alexnet) |
| [23] | Investigating Acoustic Cues for Multilingual Abuse Detection (ACMAD) | 2023 | Audio | 10 Indic languages | IndicWav2Vec-base; XLS-R; stack features; 1D-IncNet (14k params) | ADIMA (~65 hours ShareChat) | 96% accuracy (ADIMA test) | Lightweight, high accuracy across languages |
| [24] | Abusive Speech Detection in Indic Languages Using Acoustic Features | 2024 | Audio | 10 Indic languages | RF, SVM, XGBoost using OPENSMILE EGEMAPS / COMPARE | ADIMA (11,775 clips) | UAR 0.70–0.84 (multilingual) | Interpretable paralinguistic features; SHAP explainability |
| [25] | ToxVidLM: Multimodal Multitask Learning Toxic Cont Detection | 2024 | Text + Audio + Video | Hindi–English (code-mixed) | HingRoBERTa, Whisper, VideoMAE; gated multimodal fusion; multitask learning | ToxCMM | Weighted F1: 94.35% (toxicity), 86.84% (severity), 83.42% (sentiment) | Effective multimodal fusion; strong performance on low-resource, code-mixed video data |
| [26] | Multilingual Audio Toxicity Dataset & Zero-shot | 2024 | Audio | 30 languages | SONAR encoders + 3-layer FF; Whisper for cascades | MuTox dataset (20k En/Es; 4k others) | Large F1 improvement vs wordlists; Hindi AUC 0.77–0.84 | Massive language coverage; zero-shot |
| [27] | Towards Cross-Lingual Audio Abuse Detection (Few-shot) | 2024 | Audio (few-shot) | 10 Indian languages | MAML + Whisper/Wav2Vec features + ANN | ADIMA (ShareChat) | Few-shot acc 79–85% across languages | Good cross-lingual few-shot generalization |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| [28] | Towards Low-Latency Offensive Content Detection for Real-Time Streaming | 2025 | Audio/Text/Streaming | Multi | Low-latency live (LLL) streaming technologies: CMAF, chunked delivery, LL-DASH, LL-HLS | AWS Elemental Media Services with open-source players (hls.js, dash.js, Shaka, ExoPlayer) | E2E latency of 5.11s with well-tuned parameters; improvements in bitrate and rebuffering | Operational focus on streaming; detailed analysis of latency sources across the entire pipeline |
| [29] | Multimodal toxicSpeech Detection in Memes Using Contrastive Language-Image Pre-Training | 2024 | Multimodal | English | CLIP model fine-tuned with prompt engineering | Facebook toxicMeme Dataset (~10k memes) | Accuracy of 87.42% | Zero-shot learning capability; effectively fuses vision and language |
| [30] | Recent Advances in Speech Language Models: A Survey | 2024 | Audio/Text | Multi | Speech LMs (Whisper, Wav2Vec, etc.) | N/A | N/A | Provides a comprehensive overview of SpeechLM architecture, training, and challenges |
| [31] | Deep DL-based Detection for Film Censorship (foul language) | 2025 | Audio | English | Lightweight CNN on log-Mel; compared to MobileNet /Inception /ResNet | MMUTM, TAPAD, continuous samples (6 real-world) | Macro AUC 93.85; weighted AUC 94.58; high F1; realtime ~0.46s/sec | Real-time; outperforms ASR-based systems |
| [32] | Pitch Imperfect: | 2025 | Audio | English | LSTM on prosodic | ASVspoof 2021 | Accuracy 93%; EER | Explainable |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Audio Deepfake Detection (prosody) | | | | features + attention | | 24.7% | prosodic features; robust to adversary |
| [33] | MultiTec: Multimodal TikTok Misinformation Detector | 2025 | Multimodal | English | CVRL (caption-guided visual rep), HuBERT ASRL, co-attention DMVC | TikTok COVID datasets (1,053; 1,137 videos) | F1 gain 4.7–6.9% over baseline | Strong multimodal fusion |

## 3. Conclusion

The domain of automated content moderation has significantly progressed, evolving from rudimentary lexical filters to sophisticated deep learning systems. An examination of 31 contemporary research underscores significant tendencies and efficacious strategies for various moderation issues. In text-based tasks, fine-tuned transformer models regularly outperform. BERT and ELECTRA get elevated F1 scores, with a BERT-based model achieving 96% on English hazardous speech. A Darija RoBERTa model attained 90% accuracy in a resource-constrained environment. These models include context and intricate patterns, making them proficient for delicate examination. Audio-native approaches are often more effective for spoken material, particularly in real-time applications. Efficient and accurate analysis of spectrograms and MFCCs is provided by lightweight CNNs and RNNs. A CNN system for foul language identification attained an F1 score of 98.1% in a binary classification framework, operating at a real-time pace of around 0.46 seconds per second of audio. This mitigates transcription-related mistakes and utilizes paralinguistic signals. Multimodal techniques are becoming more essential as internet material integrates several modalities. CLIP achieved 87.4% accuracy in identifying hazardous speech in memes by integrating visual and linguistic elements, underscoring the significance of cross-modal thinking.

Challenges persist, especially regarding robustness. Even robust models are susceptible to hostile manipulations such as the inclusion of extraneous spaces or symbols. Addressing this requires adversarial training and validation. Enhancing support for low-resource languages and advancing cross-lingual generalization are essential. Meta-learning has potential in adjusting models with little data, advancing towards inclusion. Research indicates a method that integrates modular technologies, merging precision with real-time performance, underpinned by ethical assessments to ensure fairness. Emphasizing robustness, inclusiveness, and openness will enhance the effectiveness, scalability, and accountability of future moderating systems.

## Refrences

[1] Isasi, A.C. and Juanatey, A.G., 2017. Hate speech in social media: a state-of-the-art review. Erişim Adresi: https://ajuntament. barcelona. cat.

[2] Kovács, G., Alonso, P. and Saini, R., 2021. Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources. SN Computer Science, 2(2), p.95.

[3] Herwanto, G.B., Ningtyas, A.M., Nugraha, K.E. and Trisna, I.N.P., 2019, December. toxicspeech and abusive language classification using fastText. In 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) (pp. 69-72). IEEE.

[4] Kalaivani, A. and Thenmozhi, D., 2021. Multilingual toxicspeech and Offensive language detection in English, Hindi, and Marathi languages. (FIRE 2021).

[5] Soykan, L., Karsak, C., Elkahlout, I.D. and Aytan, B., 2022, June. A comparison of machine learning techniques for Turkish profanity detection. In Proceedings of the Second International Workshop on Resources and Techniques for User Information in Abusive Language Analysis (pp. 16-24).

[6] Kim, C.G., Hwang, Y.J. and Kamyod, C., 2022. A study of profanity effect in sentiment analysis on natural language processing using ann. Journal of web engineering, 21(3), pp.751-766.

[7] Maity, K., Bhattacharya, S., Saha, S. and Seera, M., 2023. A deep learning framework for the detection of Malay toxicspeech. IEEE Access, 11, pp.79542-79552.

[8] Saleh, H., Alhothali, A. and Moria, K., 2023. Detection of toxicspeech using bert and toxicspeech word embedding with deep model. Applied Artificial Intelligence, 37(1), p.2166719.

[9] Omran, E., Al Tararwah, E. and Al Qundus, J., 2023. A comparative analysis of machine learning algorithms for toxicspeech detection in social media. Online Journal of Communication and Media Technologies, 13(4), p.e202348.

[10] Gutha, A.R., Adarsh, N.S., Alekar, A. and Reddy, D., 2023. Multilingual toxicSpeech and Offensive Language Detection of Low Resource Languages. In FIRE (Working Notes) (pp. 445-458).

[11] Awal, M.R., Lee, R.K.W., Tanwar, E., Garg, T. and Chakraborty, T., 2023. Model-agnostic meta-learning for multilingual toxicspeech detection. IEEE Transactions on Computational Social Systems, 11(1), pp.1086-1095.

[12] Singh, A. and Thakur, R., 2024, June. Generalizable multilingual toxicspeech detection on low resource indian languages using fair selection in federated learning. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 7204-7214).

[13] Abdellaoui, I., Ibrahimi, A., El Bouni, M.A., Mourhir, A., Driouech, S. and Aghzal, M., 2024. Investigating offensive language detection in a low-resource setting with a robustness perspective. Big Data and Cognitive Computing, 8(12), p.170.

[14] Mnassri, K., Farahbakhsh, R. and Crespi, N., 2024. Multilingual toxicspeech detection: a semi-supervised generative adversarial approach. Entropy, 26(4), p.344.

[15] Albladi, A., Islam, M., Das, A., Bigonah, M., Zhang, Z., Jamshidi, F., Rahgouy, M., Raychawdhary, N., Marghitu, D. and Seals, C., 2025. toxicspeech detection using large language models: A comprehensive review. IEEE Access.

[16] Hyder, R., Ghaffarzadegan, S., Feng, Z., Hansen, J.H. and Hasan, T., 2017, August. Acoustic scene classification using a CNN-SuperVector system trained with auditory and spectrogram image features. In Interspeech (pp. 3073-3077).

[17] Ghaleb, E., Popa, M. and Asteriadis, S., 2019, September. Multimodal and temporal perception of audio-visual cues for emotion recognition. In 2019 8th international conference on affective computing and intelligent interaction (ACII) (pp. 552-558). IEEE.

[18] Wazir, A.S.B., Karim, H.A., Abdullah, M.H.L., Mansor, S., AlDahoul, N., Fauzi, M.F.A. and See, J., 2020, September. Spectrogram-based classification of spoken foul language using deep CNN. In 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP) (pp. 1-6). IEEE.

[19] Alcântara, C., Moreira, V. and Feijo, D., 2020, May. Offensive video detection: dataset and baseline results. In Proceedings of the Twelfth Language Resources and Evaluation Conference (pp. 4309-4319).

[20] Ba Wazir, A.S., Karim, H.A., Abdullah, M.H.L., AlDahoul, N., Mansor, S., Fauzi, M.F.A., See, J. and Naim, A.S., 2021. Design and implementation of fast spoken foul language recognition with different end-to-end deep neural network architectures. Sensors, 21(3), p.710.

[21] Chaudhari, A., Davda, P., Dand, M. and Dholay, S., 2021, January. Profanity detection and removal in videos using machine learning. In 2021 6th International Conference on Inventive Computation Technologies (ICICT) (pp. 572-576). IEEE.

[22] Xia, S., Fourer, D., Audin-Garcia, L., Rouas, J.L. and Shochi, T., 2022, May. Speech emotion recognition using time-frequency random circular shift and deep neural networks. In Speech Prosody 2022 (pp. 585-589).

[23] Thakran, Y. and Abrol, V., 2023, January. Investigating Acoustic Cues for Multilingual Abuse Detection. In Proc. Annu. Conf. Int. Speech. Commun. Assoc (pp. 3642-3646).

[24] Spiesberger, A.A., Triantafyllopoulos, A., Tsangko, I. and Schuller, B.W., 2024. Abusive Speech Detection in Indic Languages Using Acoustic Features. arXiv preprint arXiv:2407.20808.Social Systems, 11(1), pp.1086-1095.

[25] Maity, K., Poornash, A.S., Saha, S. and Bhattacharyya, P., 2024. Toxvidlm: A multimodal framework for toxicity detection in code-mixed videos. arXiv preprint arXiv:2405.20628.

[26] Costa-jussà, M.R., Meglioli, M.C., Andrews, P., Dale, D., Hansanti, P., Kalbassi, E., Mourachko, A., Ropers, C. and Wood, C., 2024. Mutox: Universal multilingual audio-based toxicity dataset and zero-shot detector. arXiv preprint arXiv:2401.05060.

[27] Bentaleb, A., Lim, M., Akcay, M.N., Begen, A.C., Hammoudi, S. and Zimmermann, R., 2025. Toward one-second latency: Evolution of live media streaming. IEEE Communications Surveys & Tutorials.

[28] Sankaran, A.N., Farahbakhsh, R. and Crespi, N., 2024. Towards Cross-Lingual Audio Abuse Detection in Low-Resource Settings with Few-Shot Learning. arXiv preprint arXiv:2412.01408.

[29] Arya, G., Hasan, M.K., Bagwari, A., Safie, N., Islam, S., Ahmed, F.R.A., De, A., Khan, M.A. and Ghazal, T.M., 2024. Multimodal toxicspeech detection in memes using contrastive language-image pre-training. IEEe Access, 12, pp.22359-22375.

[30] Cui, W., Yu, D., Jiao, X., Meng, Z., Zhang, G., Wang, Q., Guo, Y. and King, I., 2024. Recent advances in speech language models: A survey. arXiv preprint arXiv:2410.03751.

[31] Shang, L., Zhang, Y., Deng, Y. and Wang, D., 2025. MultiTec: a data-driven multimodal short video detection framework for healthcare misinformation on TikTok. IEEE Transactions on Big Data.

[32] Warren, K., Olszewski, D., Layton, S., Butler, K., Gates, C. and Traynor, P., 2025. Pitch Imperfect: Detecting Audio Deepfakes Through Acoustic Prosodic Analysis. arXiv preprint arXiv:2502.14726.

[33] Shang, L., Zhang, Y., Deng, Y. and Wang, D., 2025. MultiTec: a data-driven multimodal short video detection framework for healthcare misinformation on TikTok. IEEE Transactions on Big Data..

[34] Gupta, P., Sharma, V. and Varma, S., 2022. A novel algorithm for mask detection and recognizing actions of human. Expert Systems with Applications, p.116823.

[35] Singh, U., Gupta, P. and Shukla, M., 2022. Activity detection and counting people using Mask-RCNN with bidirectional ConvLSTM. Journal of Intelligent & Fuzzy Systems, 43(5), pp.6505-6520

[36] Singh, U, Gupta, P., Shukla, M., Sharma, V., Varma, S. and Sharma, S.K., 2023. Acknowledgment of patient in sense behaviors using bidirectional ConvLSTM. Concurrency and Computation: Practice and Experience, 35(28), p.e7819.

[37] Gupta, P., Arya, N., Singar, C.P., Chaudhari, A., Singh, U. and Gupta, S., 2025. Safety of Pedestrians in AI-Optimized VANETs for Autonomous Vehicles via Real-Time Vehicle-to-Vehicle Communication. In AI-Driven Transportation Systems: Real-Time Applications and Related Technologies (pp. 169-181). Cham: Springer Nature Switzerland.

[38] Gupta, P. and Singh, U., 2025. Evaluation of several yolo architecture versions for person detection and counting. Multimedia Tools and Applications, pp.1-24.

[39] Gupta, P., Shukla, M., Arya, N., Singh, U. and Mishra, K., 2022. Let the Blind See: An AIIoT-Based Device for Real-Time Object Recognition with the Voice Conversion. In Machine Learning for Critical Internet of Medical Things (pp. 177-198). Springer, Cham

[40] Gupta, P. and Kulkarni, N., 2013. An introduction of soft computing approach over hard computing. International Journal of Latest Trends in Engineering and Technology (IJLTET), 3(1), pp.254-258

[41] Kushwaha, U., Gupta, P., Airen, S. and Kuliha, M., 2022, December. Analysis of CNN Model with Traditional Approach and Cloud AI based Approach. In 2022 International Conference on Automation, Computing and Renewable Systems (ICACRS) (pp. 835-842). IEEE

[42] Gupta, P., Saxena, R., Singar, C.P., Kaur, J., Sharma, V. and Prasad, D.K., 2025, October. An exhaustive analysis of security concerns, risks, and treatments in cloud computing. In AIP Conference Proceedings (Vol. 3343, No. 1, p. 030004). AIP Publishing LLC.