Journal of Deep Learning, Computer Vision and Digital Image Processing
E-ISSN: 2986-8939 P-ISSN: 2986-8920
Vol. 4, No. 2, June 2026
DOI. https://doi.org/10.61255/decoding.v4i2.1540

Automatic Detection of Toxic Content in Short Videos Using Deep Learning-Based Text and Audio Feature Integration

Heri Agus Supriyanto✉
(Janabadra University, Yogyakarta, Indonesia)

Ryan Ari Setyawan
(Janabadra University, Yogyakarta, Indonesia)

Jemmy Edwin Bororing
(Janabadra University, Yogyakarta, Indonesia)

OPEN ACCESS

ARTICLE HISTORY
Received: April 29, 2026
Revised: June 15, 2026
Accepted: June 30, 2026

KEYWORDS
Audio-text classification;
BiLSTM;
Multimodal fusion;
Short-form video;
Toxic video detection.

ABSTRACT

Purpose – This study develops and evaluates a multimodal classification model combining audio and text features via configurable weighted late fusion to detect toxic content in keyword-retrieved Indonesian short videos containing profanity-related contexts, addressing text-only detection limitations.

Methods – The pipeline utilized FFmpeg for audio extraction, MFCC for audio features, and Google Speech Recognition for text. Three fusion configurations (Audio-Text: 40%–60%, 50%–50%, and 60%–40%) combining a DNN and BiLSTM were evaluated on 1,484 manually labeled Indonesian short videos.

Findings – The Audio 60%–Text 40% configuration achieved the numerically highest test accuracy of 93.94% with a 95% confidence interval of 90.62%–96.13%, using a decision threshold of 0.60 selected from the validation set. The model obtained an F1-score of 0.95 for the toxic class. Compared with unimodal baselines, all fusion models achieved higher accuracy, indicating the benefit of integrating audio and text features.

Research implications – The findings suggest that multimodal audio-text fusion can improve Indonesian short-form video toxicity detection compared with audio-only or text-only models. However, the differences among the three fusion weighting schemes were not statistically significant based on McNemar's test, so the audio-dominant configuration should be interpreted as the numerically best configuration in this dataset rather than as a universally superior setting.

Originality – This study systematically compares unimodal baselines and configurable audio-text late-fusion weighting strategies for Indonesian short-form video toxicity detection. The study also applies validation-based threshold selection, confidence intervals, and pairwise McNemar testing to provide a more reliable evaluation of multimodal model performance.

Correspondence Author: ✉heri.agus5555@gmail.com
To cite this article: Supriyanto, H. A., Setyawan, R. A., & Bororing, J. E. (2026). Automatic detection of toxic content in short videos using deep learning-based text and audio feature integration. Journal of Deep Learning, Computer Vision and Digital Image Processing, 4(2), 221–235. https://doi.org/10.61255/decoding.v4i2.1540

INTRODUCTION

The rapid growth of short-form video platforms has fundamentally changed the way people consume, produce, and share digital content, particularly in Indonesia. Over the past few years, platforms such as YouTube Shorts, TikTok, and Instagram Reels have attracted hundreds of millions of active users, making Indonesia one of the largest markets for short-form video consumption globally. The shift from static text-based posts or long-form videos to bite-sized, highly dynamic visual content is driven by the widespread availability of high-speed mobile internet and the increasing affordability of smartphones. These platforms provide unprecedented opportunities for creative expression, entertainment, digital marketing, and rapid information sharing. They empower ordinary users to become content creators with minimal equipment, democratizing the digital media landscape.

However, while these technological advancements offer immense benefits, they have also become significant channels for the rapid spread of toxic, abusive, and harmful content[1]. The underlying structural design of short-form video ecosystems inadvertently exacerbates this issue. The combination of a very short duration (typically ranging from 15 to 60 seconds), an exceptionally high user engagement rate, and highly aggressive algorithmic recommendation systems has accelerated the dissemination of toxic material. Algorithms on these platforms are engineered to maximize watch time and user retention, often prioritizing sensational, controversial, or emotionally charged content. As a result, toxic videos can achieve viral status within hours, reaching millions of viewers before manual moderation teams can intervene. This phenomenon raises serious concerns regarding user safety, the psychological well-being of minors, and online social dynamics, as it fosters cyberbullying, polarization, and the normalization of verbal violence[2].

Detecting toxic content in short-form videos presents unique challenges that differ substantially from those encountered in traditional text-based social media or longer video formats. In short videos, toxicity is often conveyed not only through explicit lexical items but also through tone of voice, emotional intensity, prosody, and contextual cues[3]. Automatic speech recognition systems frequently struggle to accurately transcribe informal language, slang, regional dialects, and audio recorded in noisy environments, which are common characteristics of user-generated short videos. Consequently, approaches that rely solely on textual features tend to achieve suboptimal performance when applied to this type of content. These limitations highlight the importance of developing more robust detection systems that can effectively leverage both audio and textual information[4],[5].

In recent years, significant progress has been made in the application of deep learning techniques for hate speech and toxic content detection. Early studies primarily focused on text-based classification using models such as Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), Convolutional Neural Networks, and transformer-based architectures such as BERT and its variants[6]. These text-centric methods have demonstrated strong performance on social media comments and posts. However, their effectiveness tends to decline when applied to short-form video content due to the inherent limitations of automatic transcription and the critical role of non-verbal cues in conveying harmful intent.

To overcome the shortcomings of unimodal approaches, researchers have increasingly adopted multimodal learning frameworks that integrate information from multiple modalities, including text, audio, and visual features. Multimodal models have shown superior performance across various tasks such as emotion recognition, sentiment analysis, and hate speech detection. For example, Gumelar et al. [7] proposed a multimodal LSTM-based model that combines speech features and transcribed text to detect toxic speech in Indonesian YouTube confrontation videos. Their findings demonstrated that the integration of audio and text modalities significantly improved detection accuracy compared to unimodal baselines. Similarly, Sharma et al. [8] developed a multimodal deep learning framework for hate speech detection in social media videos by fusing audio-visual and textual features, achieving notable performance gains over text-only models.

Despite these advancements, several important gaps remain in the existing literature. Most multimodal studies apply either equal weighting or text-heavy fusion strategies without systematically examining how different weight distributions between audio and text modalities influence detection performance[9]. Furthermore, limited research has specifically addressed short-form videos, which differ from longer videos and static text posts in terms of duration, linguistic characteristics, and acoustic properties. In short videos, audio features that capture prosodic and emotional information may provide complementary information to textual features, particularly when automatic speech recognition quality is affected by informal language, slang, background noise, or expressive speech. Therefore, the relative contribution of audio and text modalities should be examined empirically rather than assumed in advance.

Another critical consideration is the context of low-resource languages such as Indonesian. While substantial progress has been achieved in hate speech detection for high-resource languages, research focusing on Indonesian and other low-resource languages remains relatively limited. Most existing studies on Indonesian hate speech detection have concentrated on text-based approaches using social media comments and tweets[10]. There is still a lack of comprehensive research exploring multimodal approaches for toxic content detection in Indonesian short-form videos. This limitation is particularly concerning given the high volume of short video content consumed daily in Indonesia and the potential negative societal impacts of unchecked toxic content.

Motivated by these gaps, the present study aims to develop and evaluate a multimodal deep learning model for toxic short-form video classification by combining MFCC audio features and BiLSTM text representations through configurable weighted late fusion[11]. This research systematically compares unimodal audio-only and text-only baselines with three different audio-text weighting configurations: Audio 40%–Text 60%, Audio 50%–Text 50%, and Audio 60%–Text 40%. The primary objective is to identify whether multimodal fusion improves performance over single-modality models and to examine the relative effect of fixed modality weighting schemes in Indonesian short-form video toxicity detection.

The main contributions of this study are fourfold. First, it proposes a reproducible end-to-end pipeline for extracting audio and text features from Indonesian short-form videos, incorporating a safe caching mechanism to support efficient and iterative experimentation. Second, it compares unimodal audio-only and text-only baselines with three configurable weighted late-fusion models. Third, it applies validation-based threshold selection to avoid test-set optimization bias. Fourth, it reports confidence intervals and pairwise McNemar tests to support a more cautious interpretation of performance differences among model configurations[12].

The remainder of this paper is organized as follows. Section II describes the research methodology, including dataset construction, feature extraction processes, model architecture, and experimental setup. Section III presents the experimental results along with a detailed discussion of the findings. Finally, Section IV concludes the paper and outlines potential directions for future research.

METHOD

This study employed an experimental computational research design to develop and evaluate a multimodal deep learning model for toxic short-form video classification. The pipeline consists of four main stages: dataset construction, feature extraction, data preparation, and model development.

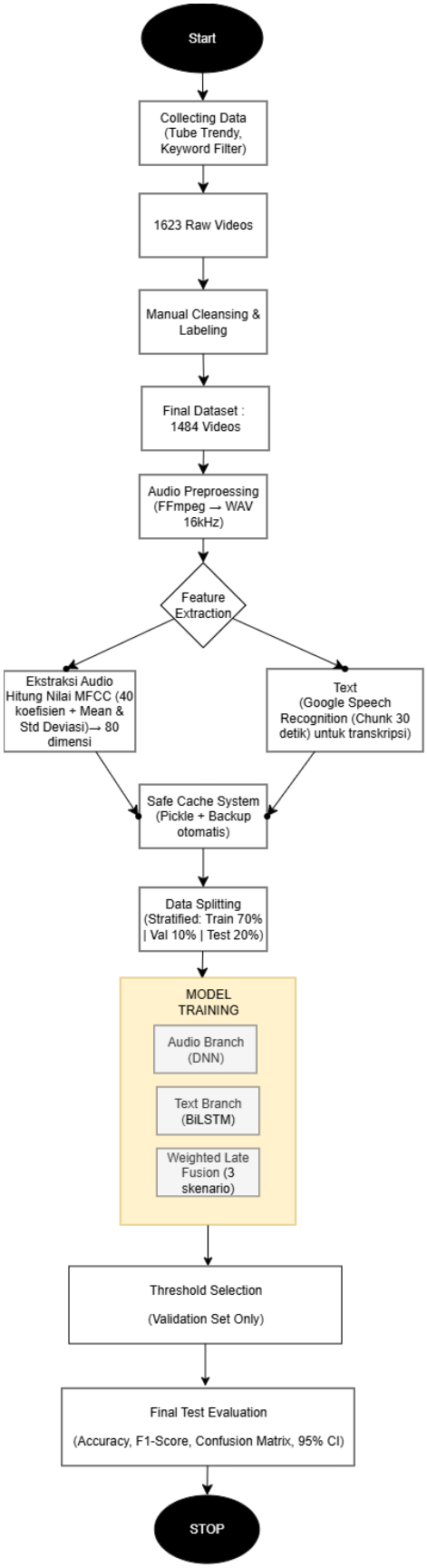

Figure 1. Overall Research Pipeline

Dataset Construction and Manual Curation

The dataset was constructed through automated video collection using Tube Trendy, a YouTube downloading tool. To target potentially toxic content, a set of common Indonesian profane keywords was used: Anjing, Babi, Bajingan, Bangsat, Goblok, Jancuk, Tai, and Tolol. Because the initial retrieval used profanity-related keywords, the safe class in this study should be understood as non-toxic videos within profanity-related contexts rather than ordinary safe videos sampled from the general short-video population. This framing is important for interpreting model generalizability and the role of audio-text features.

Following collection, a comprehensive manual curation and labeling procedure was conducted. Videos were labeled as "kasar" (toxic) if they contained explicit rude words, hate speech, harassment, threats, or offensive intent [13]. Conversely, videos utilizing profane words in humorous, educational, or musical contexts without toxic intention were labeled as "aman" (safe). Duplicate clips, non-Indonesian content, and samples with severely degraded audio were systematically removed. The final curated dataset comprised 1,484 short videos (583 "aman" and 901 "kasar").

Audio and Text Feature Extraction

All retained videos underwent a standardized preprocessing pipeline. Each video was converted into a mono-channel WAV file with a 16 kHz sampling rate using FFmpeg. Acoustic features were extracted using the Librosa library. Specifically, 40 Mel-frequency cepstral coefficients (MFCCs) were computed across time frames. To yield a uniform input shape, the mean and standard deviation of these coefficients were calculated, generating an 80-dimensional audio feature vector per video [14].

Textual transcripts were generated via the Google Speech Recognition API with Indonesian language settings ("id-ID"). To maximize transcription stability across varying video durations, a chunk-based strategy was applied where the audio was segmented into 30,000 ms consecutive intervals before concatenation [15]. Unintelligible speech segments were recorded as "kosong". To support reproducibility, a robust serialization mechanism cached extracted features into pickle format with automated backups every 25 videos.

Data Preparation and Model Development

The final dataset consisting of 1,484 videos was partitioned using stratified random sampling to ensure that the original class distribution was preserved across all subsets. As shown in Table 1, the overall dataset contained 583 “aman” videos (39.29%) and 901 “kasar” videos (60.71%). The training set consisted of 1,038 videos (408 aman and 630 kasar), the validation set consisted of 149 videos (58 aman and 91 kasar), and the test set consisted of 297 videos (117 aman and 180 kasar). The validation set was used for threshold selection and training monitoring, while the test set was reserved exclusively for final performance evaluation[16].

Table 1. Summary of Dataset, Class Distribution, and Preprocessing Parameters

Parameter	Value	Description
Total Videos	1,484	Final dataset after manual cleansing and labeling
Total Aman (0)	583 (39.29%)	Safe/non-toxic class
Total Kasar (1)	901 (60.71%)	Toxic class
Training Set	1,038 (408 Aman; 630 Kasar)	Used for model training
Validation Set	149 (58 Aman; 91 Kasar)	Used for threshold selection, hyperparameter tuning, and early stopping
Test Set	297 (117 Aman; 180 Kasar)	Held-out final evaluation only
Audio Sampling Rate	16 kHz	Standardized audio format
Number of MFCC Coefficients	40	Extracted using Librosa
Audio Feature Dimension	80	Mean + Standard Deviation of MFCCs
Text Transcription Tool	Google Speech Recognition	Language setting: id-ID
Audio Chunk Size for Transcription	30 seconds	Applied to improve transcription stability
Maximum Sequence Length	120 tokens	Padding and truncation length
Vocabulary Size	10,000	Tokenizer vocabulary limit
Data Splitting Method	Stratified Random Sampling	Maintains class distribution across splits

To prevent data leakage and ensure a fair assessment of model generalization, audio feature normalization was performed using a StandardScaler fitted solely on the training data [17]. The fitted scaler was then applied to transform the validation and test sets. This practice is essential in machine learning pipelines to avoid incorporating information from the validation or test distributions during the preprocessing stage. For the textual modality, transcripts were tokenized using a Tokenizer with a vocabulary size limited to 10,000 words. Out-of-vocabulary tokens were mapped to a special "<OOV>" token. All sequences were padded or truncated to a fixed maximum length of 120 tokens to ensure uniform input dimensions for the neural network.

A late-fusion multimodal neural network architecture was designed to process audio and textual features through two independent branches before combining their representations. The audio branch consisted of two fully connected dense layers with 256 and 128 units, respectively[18]. Each dense layer was followed by batch normalization to stabilize and accelerate the training process, as well as dropout layers with rates of 0.4 and 0.3 to mitigate overfitting. The text branch began with an embedding layer that converted tokenized sequences into dense vector representations. This was followed by two stacked bidirectional LSTM layers with 128 and 64 units, respectively, which enabled the model to capture contextual information from both forward and backward directions in the text sequences. A dropout layer with a rate of 0.4 was applied after the LSTM layers to further regularize the model.

The outputs of the audio and text branches were scaled using Lambda layers according to the chosen modality weights before being concatenated. This weighting mechanism allowed explicit control over the contribution of each modality during the fusion process. Three different weighting configurations were systematically evaluated in this study: Audio 40%–Text 60%, Audio 50%–Text 50%, and Audio 60%–Text 40%. The fused representation was then passed through additional dense layers with batch normalization and dropout before producing a final output through a sigmoid activation function for binary classification.

To address the need for unimodal comparison, two baseline models were also implemented. The audio-only baseline used the same DNN architecture as the audio branch and received the 80-dimensional MFCC feature vector as input [19]. The text-only baseline used the same BiLSTM architecture as the text branch and received tokenized transcript sequences as input. All baseline and fusion models were trained using the same data split, preprocessing procedure, class-weighting strategy, random seed, and validation-based threshold selection procedure to ensure fair comparison.

The modality weights used in this study were predefined hyperparameters selected solely for experimental comparison and remained constant throughout both the training and inference stages. Consequently, the proposed fusion strategy represents a configurable weighted late fusion approach rather than an input-adaptive weighting mechanism. The objective of comparing multiple fixed weighting configurations was to systematically investigate the influence of modality contribution on toxic content detection performance and to identify the most effective configuration for Indonesian short-form videos within the constraints of this dataset.

All models were trained using the Adam optimizer with an initial learning rate of 1e-4 and binary cross-entropy as the loss function. To address class imbalance in the dataset, class weights were incorporated during training to penalize misclassification of the minority class more heavily. The training process was monitored using two key callbacks: EarlyStopping with a patience of 10 epochs based on validation loss, and ReduceLROnPlateau to dynamically reduce the learning rate when validation loss plateaued. Model checkpoints were utilized to save the weights corresponding to the lowest validation loss observed during training.

After model training, decision-threshold selection was performed exclusively on the validation set. Threshold values ranging from 0.10 to 0.90 with increments of 0.05 were evaluated, and the threshold that produced the best validation performance was selected for each model. The selected threshold was then applied once to the held-out test set for final evaluation. The test set was not used during model training, hyperparameter tuning, or threshold selection, thereby reducing the risk of

optimistic performance estimation. In addition to accuracy, precision, recall, F1-score, and confusion matrices, The 95% confidence intervals for test accuracy were calculated using the Wilson score interval for binomial proportions [20]. Pairwise McNemar tests were also conducted on test-set predictions to examine whether differences between model configurations were statistically significant at the 0.05 level.

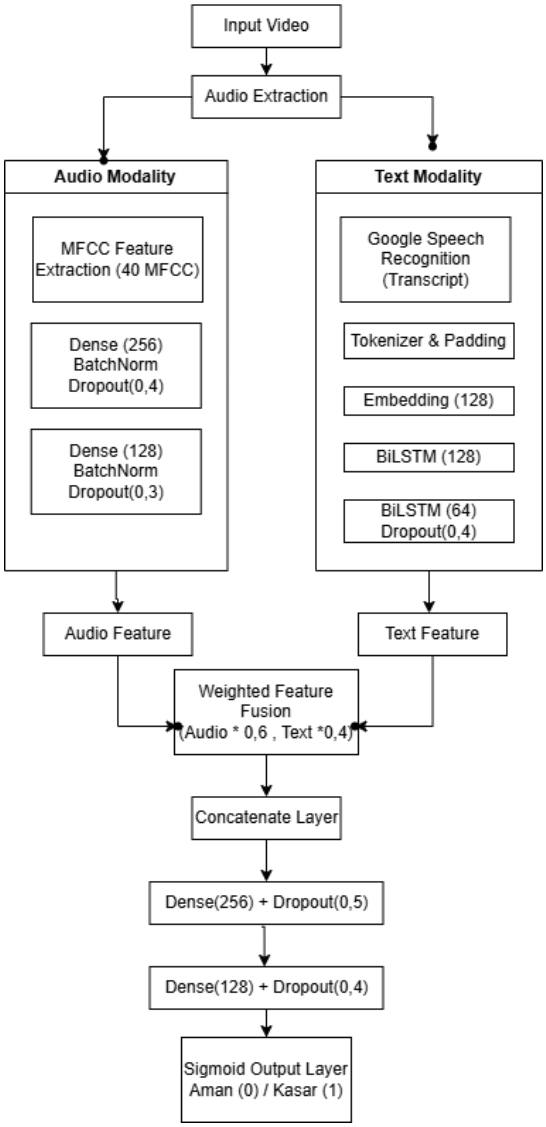

Figure 2. Architecture of the Proposed Audio-Text Classification Model

The proposed model adopts a late-fusion architecture consisting of two parallel branches designed to process audio and textual features independently before combining them. As illustrated in Figure 2, the audio branch receives an 80-dimensional feature vector obtained from MFCC extraction[21]. This branch comprises two fully connected dense layers with 256 and 128 units, respectively. Each dense layer is followed by batch normalization to stabilize the learning process and dropout layers with rates of 0.4 and 0.3 to prevent overfitting.

The text branch processes tokenized transcripts through an embedding layer with an output dimension of 128, followed by two stacked bidirectional LSTM layers with 128 and 64 units. These layers enable the model to capture contextual dependencies in both forward and backward directions within the text sequences. A dropout layer with a rate of 0.4 is applied after the LSTM layers to further regularize the network [22].

The outputs from both branches are then passed through Lambda layers that apply modality-specific weights. In the best-performing configuration, the audio features are scaled by a factor of 0.6, while

the text features are scaled by 0.4. This weighted late fusion mechanism allows the model to explicitly control the contribution of each modality. The weighted representations are subsequently concatenated and fed into two additional dense layers with 256 and 128 units, followed by dropout rates of 0.5 and 0.4, respectively. The final layer employs a sigmoid activation function to produce a probability score for binary classification, where a value closer to 1 indicates toxic ("kasar") content and a value closer to 0 indicates safe ("aman") content.

This architecture was chosen to balance model complexity and performance while enabling systematic experimentation with different audio-text weighting schemes. The late-fusion approach also provides flexibility in adjusting modality contributions without requiring retraining of the entire feature extraction pipeline.

RESULTS AND DISCUSSION

Two unimodal baselines and three audio-text weighting configurations were trained and evaluated using the same held-out test set consisting of 297 short videos. The decision threshold for each model was selected using the validation set and then applied once to the test set. The overall performance of the baseline and fusion models is summarized in Table 2

Table 2. Model Performance Comparison with Unimodal Baselines and Validation-Based Thresholds

Model	Threshold from Validation	Test Accuracy	95% CI	F1-Score (Safe)	F1-Score (Offensive)	Overall Interpretation
Audio-only DNN	0.45	0.8586	0.8146–0.8932	0.83	0.88	Baseline
Text-only BiLSTM	0.65	0.8687	0.8255–0.9024	0.80	0.90	Baseline
Audio 40% + Text 60%	0.70	0.9226	0.8862–0.9478	0.91	0.93	Fusion
Audio 50% + Text 50%	0.60	0.9226	0.8862–0.9478	0.91	0.93	Fusion
Audio 60% + Text 40%	0.60	0.9394	0.9062–0.9613	0.92	0.95	Numerically highest

As shown in Table 2, all multimodal fusion models outperformed the unimodal baselines. The audio-only DNN achieved a test accuracy of 85.86%, while the text-only BiLSTM achieved 86.87%. In comparison, the three fusion models achieved accuracies above 92%. The Audio 60%–Text 40% configuration obtained the numerically highest test accuracy of 93.94% with a 95% confidence interval of 90.62%–96.13% and an F1-score of 0.95 for the toxic class. These results indicate that integrating audio and text features provides a more effective representation than relying on a single modality.

Table 3. Confusion Matrix Statistics for Baseline and Fusion Models

Model	TN	FP	FN	TP	Correct	Error
Audio-only DNN	104	13	29	151	255	42
Text-only BiLSTM	79	38	1	179	258	39
Audio 40% + Text 60%	110	7	16	164	274	23
Audio 50% + Text 50%	110	7	16	164	274	23
Audio 60% + Text 40%	109	8	10	170	279	18

The results demonstrate that multimodal fusion improved toxic content detection compared with the unimodal baselines. The Audio 60%–Text 40% model achieved the highest number of correct predictions (279 out of 297) and the lowest overall error count (18). Compared with the balanced and text-dominant fusion models, the audio-dominant model reduced false negatives from 16 to 10 but slightly increased false positives from 7 to 8. Therefore, its advantage lies in a better overall balance between safe and toxic classification rather than a complete reduction of all error types.

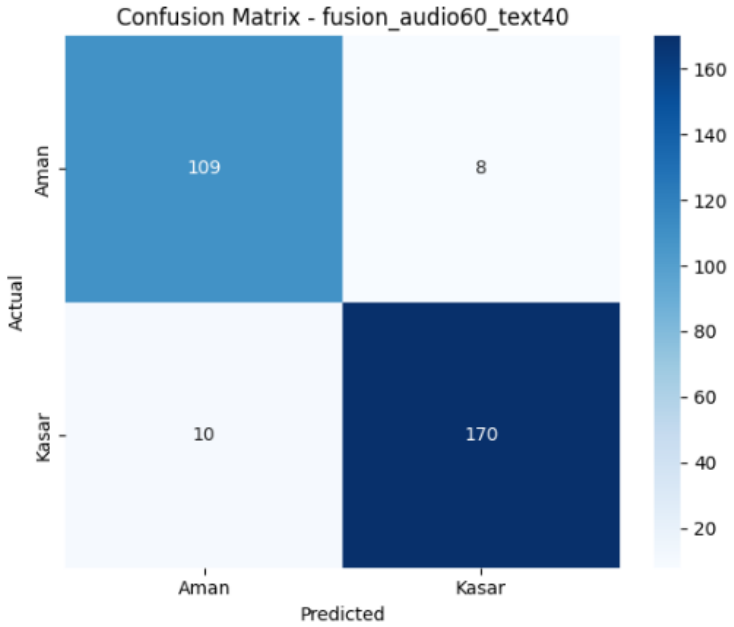

Figure 3. Confusion Matrix of the Audio 60%–Text 40% Configuration

The confusion matrix of the best-performing model, shown in Figure 3, provides further insight into the model's classification behavior. Out of 117 videos labeled as "aman", the model correctly classified 109 videos and misclassified 8 videos as toxic. Out of 180 videos labeled as "kasar", the model successfully detected 170 videos and produced 10 false negatives. Although the text-only baseline produced only one false negative, it also generated 38 false positives, indicating that the fusion model provided a more balanced classification performance.

Table 3 further shows that the Audio 60%–Text 40% configuration achieved stronger overall performance than the other models. However, the performance differences among the three fusion configurations should be interpreted cautiously because their accuracy confidence intervals overlap. This outcome suggests that audio features may provide useful complementary cues for toxic content detection in this dataset, particularly when combined with textual information. However, because the differences among fusion configurations were not statistically significant, the result should not be interpreted as definitive evidence that audio-dominant weighting is universally superior [23]. This consideration is further examined through the performance comparison in Figure 4 and the pairwise McNemar test in Table 4.

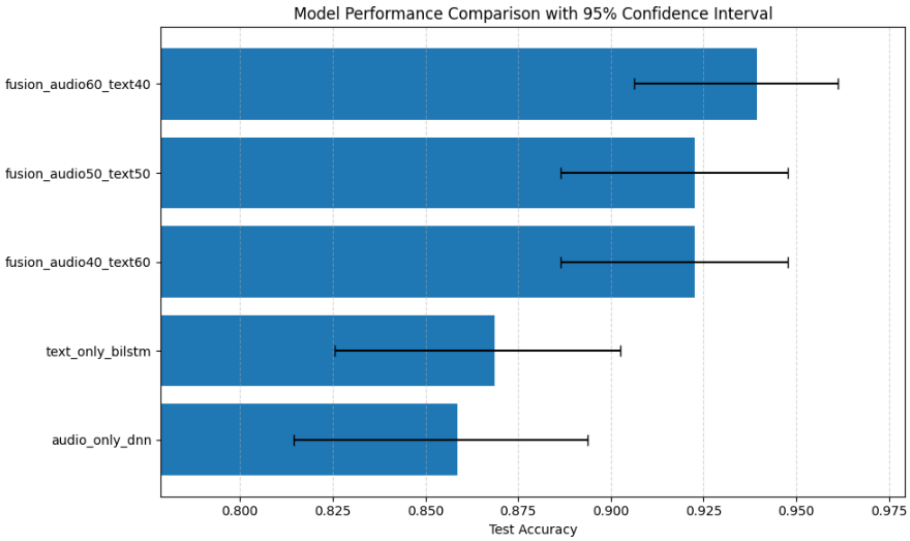

Figure 4. Model Performance Comparison with 95% Confidence Intervals

Figure 4 illustrates that the three fusion models achieved higher test accuracy than the unimodal baselines. However, the confidence intervals among the three fusion configurations overlap, suggesting that the observed differences among fusion weights should be interpreted cautiously.

Table 4. Pairwise McNemar Test for Selected Model Comparisons

Comparison	p-value	Significant at 0.05	Interpretation
Audio-only DNN vs Audio 60% + Text 40%	0.000439	Yes	Fusion significantly better
Text-only BiLSTM vs Audio 60% + Text 40%	0.001456	Yes	Fusion significantly better
Audio 40% + Text 60% vs Audio 60% + Text 40%	0.332306	No	Difference not significant
Audio 50% + Text 50% vs Audio 60% + Text 40%	0.266846	No	Difference not significant

Pairwise McNemar tests indicate that the best fusion model differed significantly from the audio-only and text-only baselines. However, the differences between the Audio 60%–Text 40% model and the other two fusion configurations were not statistically significant at the 0.05 level. These results suggest that multimodal fusion provides a statistically supported improvement over unimodal baselines, while the specific superiority of one weighting scheme over another remains inconclusive. The training process of the best-performing fusion model was monitored through accuracy and loss curves to assess convergence and potential overfitting. Figure 5 illustrates the training and validation accuracy across epochs for the Audio 60%–Text 40% configuration.

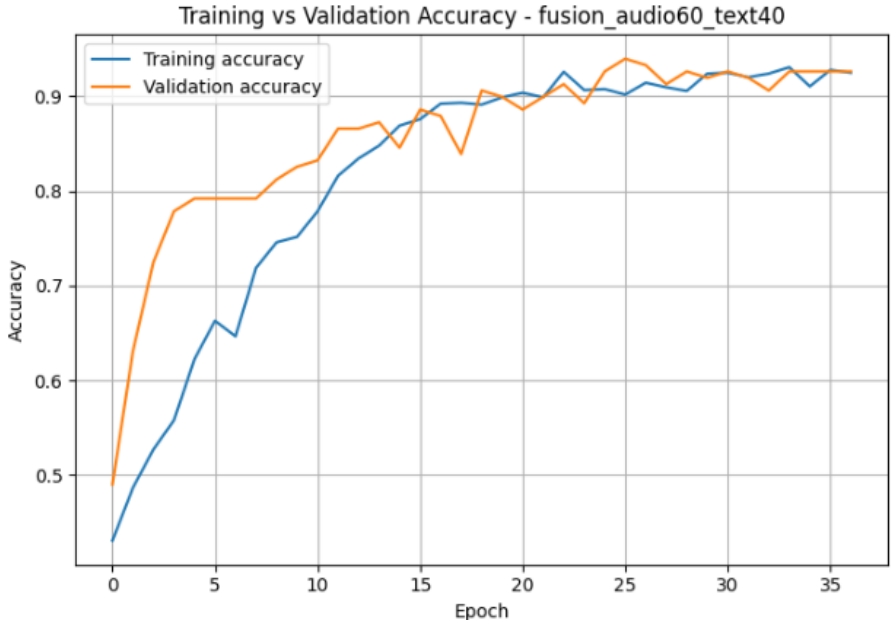

Figure 5. Training and Validation Accuracy Across Epochs

As shown in Figure 5, both training and validation accuracy increased steadily throughout the training process. The validation accuracy closely followed the training accuracy without significant divergence, indicating that the model did not suffer from severe overfitting. The final validation accuracy stabilized above 90%, demonstrating stable and effective learning.

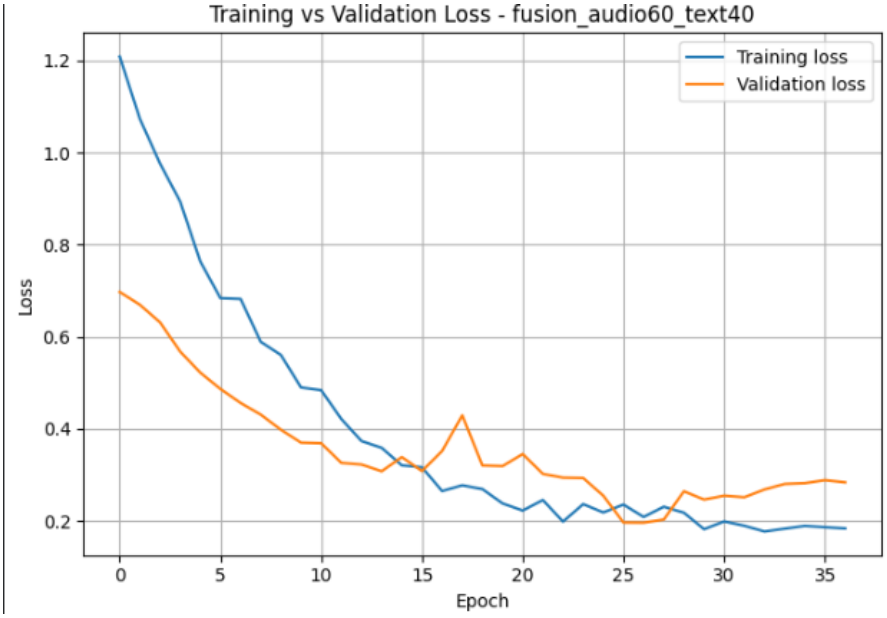

Figure 6. Training and Validation Loss Across Epochs

The training and validation loss curves, presented in Figure 6, further confirm the model's stable optimization. Both curves show a consistent downward trend, with the validation loss remaining relatively close to the training loss throughout most training epochs. The application of early stopping and ReduceLROnPlateau callbacks helped prevent severe overfitting and supported stable convergence.

Overall, the results from Tables 2, 3, and 4 and Figures 3, 4, 5, and 6 demonstrate that multimodal fusion achieved stronger performance than unimodal baselines. The Audio 60%–Text 40% model achieved the numerically highest performance, but the nonsignificant McNemar results among fusion configurations indicate that the advantage of this weighting scheme should be interpreted cautiously.

Discussion

The results provide evidence that multimodal fusion is more effective than using audio or text alone for toxic short-form video classification in this dataset. The audio-only DNN and text-only BiLSTM baselines achieved 85.86% and 86.87% accuracy, respectively, whereas all fusion configurations achieved accuracies above 92%. This demonstrates the benefit of integrating acoustic and textual information for Indonesian short-form video toxicity detection.

Although the Audio 60%–Text 40% model achieved the numerically highest accuracy of 93.94%, the pairwise McNemar tests showed that its differences from the Audio 40%–Text 60% and Audio 50%–Text 50% configurations were not statistically significant. Therefore, the result should not be interpreted as definitive evidence that audio-dominant fusion is universally superior. Instead, the finding suggests that increasing the audio contribution can be beneficial in this dataset, but the optimal weighting may depend on dataset characteristics, transcription quality, and acoustic conditions.[24].

One possible explanation for the strong performance of fusion models is that audio and text capture different aspects of toxic expression. Textual transcripts provide lexical information, while MFCC-based audio features capture acoustic and prosodic cues such as intensity, tone, and speech dynamics. These cues may be particularly useful when automatic transcription is affected by slang, informal expressions, noisy background, or fast-paced speech.

When comparing the three weighting schemes, the Audio 60%–Text 40% model achieved the best numerical accuracy and F1-score for the toxic class. However, the differences among fusion configurations were relatively small and statistically nonsignificant. This means that the evidence more strongly supports the benefit of multimodal fusion over unimodal classification than the superiority of one fixed fusion weight over another.

These findings are consistent with and extend previous multimodal studies. Gumelar et al. [7] demonstrated that combining audio and text features improved toxic speech detection in Indonesian YouTube videos. However, their study did not explore the effect of varying modality weights. Similarly, Sharma et al. reported performance gains from multimodal fusion in social media video analysis. The present study contributes additional empirical evidence by comparing unimodal baselines, fixed weighted fusion strategies, validation-based thresholds, confidence intervals, and McNemar significance tests.

From a practical standpoint, the results have meaningful implications for the development of content moderation systems on short-video platforms in Indonesia. The performance gains of fusion models over audio-only and text-only baselines suggest that moderation tools could benefit from integrating acoustic and textual analysis, especially when dealing with noisy audio, informal speech, or imperfect transcription. However, deployment decisions should still consider platform-specific data, computational constraints, and the potential cost of false positives and false negatives.

Despite the promising results, this study has several limitations that should be acknowledged. First, the reliance on Google Speech Recognition introduces both financial costs and potential variability in transcription quality. The performance of the speech recognition service may vary depending on audio quality, speaker accent, and background noise, which could affect the consistency of text-based features. Second, although the dataset was manually curated with careful attention to labeling quality, it was collected from a single source. This may limit the generalizability of the findings to other platforms or regions with different linguistic characteristics and content styles. Third, the binary classification approach does not capture the varying degrees of harmfulness present in real-world toxic content, such as hate speech, harassment, misinformation, or sexually explicit material. A more granular classification scheme could provide more actionable insights for content moderation[25].

Another limitation concerns the keyword-based retrieval strategy used during dataset construction. Since videos were initially collected using profanity-related search terms, the resulting safe class may not fully represent ordinary short-form videos but rather videos containing similar lexical contexts without toxic intent. This sampling strategy may introduce selection bias and should be considered when interpreting the superiority of audio features.

Another limitation concerns the annotation process. The dataset was manually labeled by a single researcher. Although the labeling process followed predefined criteria and considered both lexical content and vocal delivery, subjective interpretation may still occur, especially in videos involving sarcasm, humor, quotation, or ambiguous aggressive tone. Because no second annotator was available during this study, inter-annotator agreement metrics such as Cohen's kappa could not be calculated. Future research should involve multiple annotators and report agreement scores to strengthen label reliability.

Future research could build upon the findings of this study in several directions. Replacing Google Speech Recognition with more robust open-source models, such as Whisper, could improve transcription accuracy and reduce dependency on commercial services. Incorporating visual features from video frames or on-screen text may further enhance model performance, especially for content where toxicity is conveyed through visual cues. Expanding the dataset through crowdsourced annotation across multiple platforms and regions would improve the generalizability of the model. Future studies should also include multiple annotators, calculate inter-annotator agreement, evaluate naturally sampled safe videos, and explore attention-based or transformer-based multimodal fusion architectures.

In summary, this study demonstrates that empirically evaluated multimodal fusion can achieve strong performance in toxic short-form video detection. The findings support the value of combining audio and text features, while also showing that claims about the superiority of a specific fusion weight should be made cautiously when statistical differences among fusion configurations are not significant.

CONCLUSION

This study developed and evaluated a multimodal deep learning model for detecting toxic content in keyword-retrieved Indonesian short videos containing profanity-related contexts by combining MFCC audio features and BiLSTM-based text representations through configurable weighted late fusion. In the revised evaluation procedure, decision thresholds were selected using the validation set and then applied once to the held-out test set. The Audio 60%–Text 40% configuration achieved the numerically highest test accuracy of 93.94% with a 95% confidence interval of 90.62%–96.13% and an F1-score of 0.95 for the toxic class.

The comparison with unimodal baselines showed that all fusion models outperformed audio-only and text-only models, confirming the benefit of integrating audio and text features. However, McNemar's test indicated that the differences among the three fusion weighting schemes were not statistically significant. Therefore, the audio-dominant configuration should be interpreted as the best-performing configuration in this dataset rather than as a universally superior weighting strategy.

Future research is recommended to involve multiple annotators, report inter-annotator agreement, expand the dataset beyond profanity-keyword retrieval, incorporate naturally sampled safe videos, explore more robust speech recognition models, and investigate visual or transformer-based multimodal architectures. Overall, this study shows that validation-based evaluation and multimodal baseline comparison provide a more reliable basis for assessing toxic short-form video classification models.

ACKNOWLEDGMENT

The author would like to express sincere gratitude to Mr Ryan Ari Setyawan and Mr Jemmy Edwin Bororing for the invaluable guidance, continuous support, and constructive feedback throughout the completion of this research. The author also appreciates the technical support from Google Colab, which enabled the execution of computationally intensive tasks during model training and evaluation. Finally, heartfelt thanks go to family and friends for their unwavering encouragement and understanding during the course of this study.

AUTHOR CONTRIBUTION STATEMENT

HS was responsible for data collection, manual cleansing and labeling, feature extraction pipeline implementation, model development, conducting experiments, and writing the original draft. RS and JB contributed to the study conceptualization, methodology design, validation of results, supervision, and review and editing of the final manuscript.

AI DISCLOSURE STATEMENT

The author utilized AI-assisted language tools to improve the clarity and structure of the manuscript. All aspects of the research, including study design, data collection and curation, model development, experimental implementation, data analysis, interpretation of results, and final manuscript decisions, were conducted and validated by the author. The AI tool was employed solely for language refinement and was not used to generate scientific content or substitute the author's accountability for the research findings.

REFERENCES

[1] T. I. Sari, Z. N. Ardilla, N. Hayatin, and R. Maskat, "Abusive comment identification on Indonesian social media data using hybrid deep learning," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 3, pp. 895–904, Sep. 2022, doi: 10.11591/ijai.v11.i3.pp895-904.

[2] M. O. Ibrohim and I. Budi, "Hate speech and abusive language detection in Indonesian social media: Progress and challenges," Aug. 01, 2023, *Elsevier Ltd*. doi: 10.1016/j.heliyon.2023.e18647.

[3] A. Chhabra and D. K. Vishwakarma, "A literature survey on multimodal and multilingual automatic hate speech identification," *Multimed. Syst.*, vol. 29, no. 3, pp. 1203–1230, Jun. 2023, doi: 10.1007/s00530-023-01051-8.

[4] M. Subramanian, V. Easwaramoorthy Sathiskumar, G. Deepalakshmi, J. Cho, and G. Manikandan, "A survey on hate speech detection and sentiment analysis using machine learning and deep learning models," Oct. 01, 2023, *Elsevier B.V.* doi: 10.1016/j.aej.2023.08.038.

[5] G. A. Koushik, D. Kanojia, and H. Treharne, "Towards a Robust Framework for Multimodal Hate Detection: A Study on Video vs. Image-based Content," in *WWW Companion 2025 - Companion Proceedings of the ACM Web Conference 2025*, Association for Computing Machinery, Inc, May 2025, pp. 2014–2023. doi: 10.1145/3701716.3718382.

[6] H. S. Wicaksana, K. Huda, and G. Airlangga, "Improved Text Classification for Indonesian Hate Speech Detection: FastText-LSTM Model with Easy Data Augmentation," *Jurnal Sistem Komputer dan Informatika (JSON)*, vol. 7, no. 3, pp. 1074–1083, Mar. 2026, doi: 10.30865/json.v7i3.9637.

[7] A. B. Gumelar, E. M. Yuniarno, A. Nugroho, D. P. Adi, I. Sugiarto, and M. H. Purnomo, "An Improved Toxic Speech Detection on Multimodal Scam Confrontation Data Using LSTM-Based Deep Learning," *International Journal of Intelligent Engineering and Systems*, vol. 17, no. 6, pp. 880–904, 2024, doi: 10.22266/ijies2024.1231.67.

[8] R. Sharma and A. Arya, "MMHFND: Fusing Modalities for Multimodal Multiclass Hindi Fake News Detection via Contrastive Learning," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 11, Nov. 2024, doi: 10.1145/3686797.

[9] K. Yousaf and T. Nawaz, "A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos," *IEEE Access*, vol. 10, pp. 16283–16298, 2022, doi: 10.1109/ACCESS.2022.3147519.

[10] R. Prabhu and V. Seethalakshmi, "A comprehensive framework for multi-modal hate speech detection in social media using deep learning," Scientific Reports, vol. 15, Art. no. 13020, Apr. 2025, doi: 10.1038/s41598-025-94069-z.

[11] H. Zennou, R. Ouadad, M. Ouhda, and M. Baslam, "Real-Time Speech Emotion Recognition with a CNN-BiLSTM-Attention Deep Learning Model," *Engineering, Technology & Applied Science Research*, vol. 16, no. 3, pp. 35047–35055, Jun. 2026, doi: 10.48084/etasr.17495.

[12] F. Andayani, L. B. Theng, M. T. Tsun, and C. Chua, "Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files," *IEEE Access*, vol. 10, pp. 36018–36027, 2022, doi: 10.1109/ACCESS.2022.3163856.

[13] M. Das, R. Raj, P. Saha, B. Mathew, M. Gupta, and A. Mukherjee, "HateMM: A Multi-Modal Dataset for Hate Video Classification," Zenodo, Jun. 2023, doi: 10.5281/zenodo.7799469.

[14] M. K. Gourisaria, R. Agrawal, M. Sahni, and P. K. Singh, "Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques," *Discover Internet of Things*, vol. 4, no. 1, Dec. 2024, doi: 10.1007/s43926-023-00049-y.

[15] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in Proceedings of the 40th International Conference on Machine Learning, vol. 202, pp. 28492–28518, Jul. 2023.

[16] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," Scientific Reports, vol. 14, Art. no. 6086, Mar. 2024, doi: 10.1038/s41598-024-56706-x.

[17] J. An, W. Lee, Y. Jeon, J. Ok, Y. Kim, and G. G. Lee, "An Investigation into Explainable Audio Hate Speech Detection," in Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2024, pp. 533–543, doi: 10.18653/v1/2024.sigdial-1.45.

[18] M. Alfaro-Contreras, J. J. Valero-Mas, J. M. Iñesta, and J. Calvo-Zaragoza, "Late multimodal fusion for image and audio music transcription," *Expert Syst. Appl.*, vol. 216, Apr. 2023, doi: 10.1016/j.eswa.2022.119491.

[19] L. H. Moeslich and C. B. Santoso, "Pengembangan Sistem Media Intelligence ESG Berbasis NLP Bahasa Indonesia Menggunakan TF-IDF dan IndoBERT," JSAI: Journal Scientific and Applied Informatics, vol. 9, no. 2, pp. 213–220, Jun. 2026, doi: 10.36085/jsai.v9i2.10586.

[20] N. M. D. Sikiandani, I. M. A. Dwi Suarjaya, and Y. Perdana Putra, "Browser-Based Detection of Harmful Content with Deep Learning Model," Journal of Applied Informatics and Computing, vol. 9, no. 4, pp. 1800–1811, Aug. 2025, doi: 10.30871/jaic.v9i4.9804.

[21] P. Rawat, M. Bajaj, S. Vats, and V. Sharma, "A comprehensive study based on MFCC and spectrogram for audio classification," *Journal of Information and Optimization Sciences*, vol. 44, no. 6, pp. 1057–1074, 2023, doi: 10.47974/jios-1431.

[22] W. Mu, B. Yin, X. Huang, J. Xu, and Z. Du, "Environmental sound classification using temporal-frequency attention based convolutional neural network," *Sci. Rep.*, vol. 11, no. 1, Dec. 2021, doi: 10.1038/s41598-021-01045-4.

[23] S. M. Saleem Abdullah, S. Y. Ameen, M. A. M. Sadeeq, and S. R. M. Zeebaree, "Multimodal Emotion Recognition using Deep Learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 1, pp. 73–79, Jan. 2021, doi: 10.38094/jastt20291.

[24] L. L. Manihuruk, D. Ginting, S. Manik, and L. W. Manurung, "Slang Words Used by the Millennial Indonesian Generation in TikTok," *Journal of Applied Linguistics*, vol. 4, no. 2, pp. 302–308, Jan. 2025, doi: 10.52622/joal.v4i2.362.

[25] E. S. Alshahrani and M. S. Aksoy, "Adversarially Robust Multitask Learning for Offensive and Hate Speech Detection in Arabic Text Using Transformer-Based Models and RNN Architectures," *Applied Sciences (Switzerland)*, vol. 15, no. 17, Sep. 2025, doi: 10.3390/app15179602.