International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Systematic Literature Review of Deep Learning-Based Multimodal Approaches for Detecting Abusive Language in Short Videos

Ryan Ari Setyawan[1], Herman Dwi Surjono[2], Ratna Wardani[3]

Submitted:14/03/2024 **Revised**: 29/04/2024 **Accepted**: 06/05/2024

Abstract: This research aims to design and implement a comprehensive deep learning-based multimodal framework for accurately detecting abusive language in video content on social media platforms. The framework seeks to leverage the integration of visual, audio, and textual modalities to capture and convey the context within the videos. By combining insights from these modalities, the research aims to enhance the precision, recall, and overall reliability of abusive language detection systems. The optimization of multimodal fusion techniques will be central to this research, involving the testing of various fusion architectures to identify the most effective configuration for real-world applications. One significant challenge addressed by this research is the issue of imbalanced datasets. To tackle this, Generative Adversarial Networks (GANs) will be employed for synthetic data generation, producing realistic and diverse abusive content samples. This approach will improve the model's ability to generalize across different contexts and content types. The proposed framework incorporates state-of-the-art architectures, such as Long Short-Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT), which will be used to model the temporal dependencies within video sequences. This paper focuses on multimodal representation visual, audio, and text as deep modalities in the hate speech classification process. We divide the data into visual (image) data, audio data, and text data, and optimize the model by enhancing transformer fusion to achieve maximum results. We chose a generative approach because it is more optimal compared to other models. Finally, we suggest that future studies combine Generative Adversarial Networks (GANs) with BERT and LSTM for more effective abusive language detection.

Keywords: *BERT, Generative, LSTM, Multimodal, Short Video.*

1. Introduction

Short video content on social media is currently popular. Short video content on social media certainly cannot be separated from content that contains abusive language[1]. On several social media platforms, short video content does not yet have a filtering process. Therefore, a process is needed to filter content that contains abusive language. Video content filtering solution using multimodal[2]. Video is a form of multimodal data consisting of visuals, audio and text [3]. Many multimodal methods have been applied to help detect abusive language content. Multimodal combination by carrying out image feature extraction, audio feature extraction and text feature extraction which then uses the final ensemble results to determine the output results[4].

A multimodal approach with an enhanced fusion model was also carried out for automatic hatespeech identification via two image and text components[5]. Multimodal fusion enhancement collaboration technique with Biddirectional Encoder Representation from Transformers (BERT) and Convolutional Neural Network (CNN) for the hatespeech classification process[6]. Multimodal the used of fusion levels in audio-visual as modalities can be used for multimodal development[7]. Multimodal interpretation process with a model of the effectiveness of capturing socio-cultural context through textual features [8].

Various multimodal models were carried out to obtain optimal results for detecting abusive language, one of which was using an artificial intelligence algorithm[9]. Artificial intelligence algorithms that can be used use machine learning or deep learning. Machine learning for multimodal can use Glove, BERT, Contrasive Language-Image Pre-training (CLIP), or Multimodal bi-transormer models[10]. Deep learning model techniques are also used to improve multimodal performance with deep neural networks (DNN), generative adversarial networks (GAN)[11].

Multimodal representation to handle visual, audio and text according to real datasets by performing deep learning. Optimization of multimodal fusion improvements to obtain the right ensemble for the final result of those modalities. Deep learning for the process of classifying the results of the modalities that have been carried out.

1.1. Research Objective

The objective of this research is to design and implement a comprehensive deep learning-based multimodal framework aimed at accurately detecting abusive language

[1] *Doctoral Student of Engineering Science, State University of Yogyakarta,*
[2]*Dept of Informatics, Janabadra University. Indonesia.*
ORCID ID : 0000-0001-9782-250X
[3] *Dept of Engineering Science, State University of Yogyakarta, Indonesia.*
ORCID ID : 0000-0002-2720-2206
[4] *Dept of Engineering Science, State University of Yogyakarta, Indonesia.*
ORCID ID : 0000-0002-7680-6487
** Corresponding Author Email: ryan@janabadra.ac.id*

in video content shared on social media platforms[12][13][14]. This framework seeks to leverage the integration of visual, audio, and textual modalities, recognizing that these elements often work in tandem to convey context in videos. By combining insights from these modalities, the research aims to improve the precision, recall, and overall reliability of abusive language detection systems, addressing the complexity and subtlety often associated with hate speech and abusive content[15][2][3].

Central to this research is the optimization of multimodal fusion techniques. Advanced transformer-based methods will be explored to effectively integrate the diverse modalities, ensuring that the combined data representations capture meaningful relationships between audio, visual, and text components[16][17][18][19]. This approach aims to overcome the traditional limitations of single-modality systems, which often miss contextual nuances present in multimodal data. The optimization process will involve testing various fusion architectures, including early, late, and hybrid fusion techniques, to identify the most effective configuration for real-world applications[20][21].

1.2. Significance of the Research

Use Additionally, the research addresses a significant challenge in abusive language detection: the issue of imbalanced datasets. Abusive language datasets often have a disproportionate number of non-abusive samples compared to abusive ones, which can bias model performance[22][23]. To counteract this, Generative Adversarial Networks (GAN) will be utilized for synthetic data generation. GANs will produce realistic and diverse abusive content samples to balance the dataset, improving the model's ability to generalize across different contexts and content types[24][11].

The proposed framework also incorporates state-of-the-art architectures like Long Short-Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT). LSTM will be used to model temporal dependencies within video sequences, particularly in audio and text data, where the context of words or sounds over time is critical. BERT, on the other hand, will provide advanced contextual embeddings for textual data, ensuring that the meaning of words and phrases is accurately captured, even in complex or multilingual content[25][26]. By combining these architectures with GAN, the research aims to build a robust system that can effectively detect abusive language with high accuracy and low latency.

Ultimately, the framework is designed to be scalable and capable of real-time operation, making it suitable for deployment on social media platforms where content is generated and shared at high volumes[27]. This solution not only contributes to the academic field of multimodal deep learning but also offers practical applications for improving content moderation and creating safer online environments.

This research holds immense significance in fostering safer and more inclusive online communities by providing social media platforms with a powerful, AI-driven tool to moderate harmful video content. With the rapid proliferation of user-generated content on platforms like YouTube, TikTok, and Instagram, the prevalence of abusive language in multimedia content poses a growing threat to online safety and well-being[28]. This study addresses this issue by developing a sophisticated detection system capable of identifying and mitigating abusive content in real-time, contributing to a healthier and more respectful digital environment.

The study significantly advances the field of multimodal learning by integrating data from visual, audio, and text modalities for abusive language detection. Unlike traditional approaches that rely on single modalities, this research introduces an innovative framework that leverages the complementary strengths of multiple data types. For example, while text alone may capture explicit abusive phrases, visual and audio cues can reveal implicit or contextual forms of abuse, such as sarcasm, tone, or imagery. By fusing these modalities, the proposed model achieves a deeper understanding of abusive content, making it more robust and versatile in diverse scenarios.

A major focus of the research is improving the accuracy of detection systems by reducing false positives (incorrectly flagging non-abusive content) and false negatives (failing to detect abusive content). These errors can have serious consequences, such as suppressing legitimate expression or allowing harmful content to proliferate. By optimizing transformer-based fusion techniques and leveraging state-of-the-art architectures like BERT and LSTM, the proposed model significantly enhances precision and recall. This ensures a higher level of reliability and accuracy compared to existing methods, making it a valuable asset for content moderation teams.

Additionally, the research addresses one of the most pressing real-world challenges: dataset imbalance. Abusive language datasets often suffer from a disproportionate number of non-abusive samples, which can lead to biased and ineffective models. To overcome this, the study employs Generative Adversarial Networks (GAN) to generate synthetic abusive content, thereby balancing the dataset and improving the model's generalizability. This approach not only enhances the framework's adaptability to diverse datasets but also ensures consistent performance across different types of abusive content and varying social media platforms.

Furthermore, the insights and outcomes from this research open new avenues for AI-driven content moderation. The multimodal framework's design is inherently scalable and adaptable, allowing it to be extended to multilingual and culturally diverse contexts. This is particularly important in a globalized digital space where abusive language can vary significantly across languages, dialects, and cultural norms. By addressing these complexities, the proposed framework sets the stage for future advancements in AI moderation, including its application in detecting abusive language, misinformation, and other harmful behaviors beyond abusive language. In summary, this research not only makes a substantial contribution to the academic field of multimodal deep learning but also provides a practical solution to a critical societal issue.

2. Critical Requirements

The critical requirements for the research on deep learning-based multimodal approaches for detecting abusive language in short videos focus on enabling effective visual content analysis to detect abusive or harmful language.

2.1. Robust and Diverse Dataset

A high-quality dataset of video content with annotated visual components is essential. The dataset must include a wide variety of scenarios, such as explicit imagery, offensive gestures, hate symbols, and text embedded in visuals. Diverse representations of abusive content across different cultural, geographical, and linguistic contexts are critical for generalizability. Data augmentation techniques, such as scaling, cropping, and color adjustments, should be applied to enhance model robustness.

2.2. Preprocessing and Feature Extraction

Effective preprocessing techniques are required to isolate key visual elements in frames. This includes object detection, scene segmentation, and text recognition for detecting embedded offensive language in images or videos.Techniques like Optical Character Recognition (OCR) are necessary to extract and analyze abusive text content present in visuals. Noise reduction and enhancement methods must be applied to ensure that low-quality or compressed video frames are still interpretable.

2.3. Advanced Deep Learning Models

The integration of computer vision techniques, such as Convolutional Neural Networks (CNNs), is essential for feature extraction from image frames within videos. Pretrained models like ResNet, EfficientNet, or Vision Transformers (ViT) can be fine-tuned for the task of identifying visual cues associated with abusive language. Multimodal transformer-based models must incorporate visual data alongside audio and text for a unified understanding of video content.

2.4. Temporal Analysis for Video Frames

Videos consist of sequential frames, so the model must account for temporal dependencies. Techniques like 3D CNNs, Recurrent Neural Networks (RNNs), or Long Short-Term Memory (LSTM) networks should be used to process temporal changes. Keyframe extraction algorithms are needed to focus on the most relevant visual data, reducing computational load while preserving critical abusive visual context.

2.5. Multimodal Fusion

A seamless fusion mechanism is necessary to integrate visual features with audio and textual data. This includes early, late, and hybrid fusion strategies to combine insights from multiple modalities effectively. Transformers or attention mechanisms must be designed to give appropriate weight to visual inputs when determining abusive content.

3. Building Deep Learning Based Multimodal Techniques

The capabilities of deep learning-based multimodal systems, particularly in the context of analyzing short video content. Short videos, which typically feature a combination of visuals, audio, and text, present unique challenges in terms of content interpretation. The integration of computer vision allows for a comprehensive and accurate understanding of such multimedia data[29].

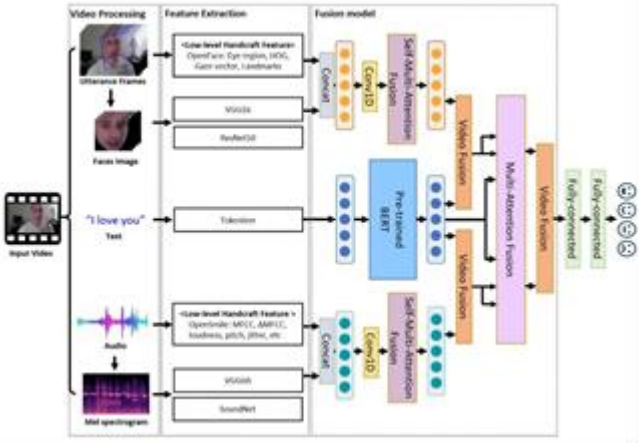

Fig. 1 Multimodal Technique for Video [30]

Figure 1. illustrates the architecture of a system designed to process multimodal data, including image, text, and audio, with the ultimate goal of performing tasks such as classification or analysis [30].

3.1. Visual Content Interpretation

Object and Scene Recognition **:** Computer vision enables the system to identify objects, people, scenes, and key elements within video frames. This is especially important for short videos, where the context can quickly shift, and visual cues may provide crucial information for interpreting the overall message of the video. For instance,

offensive gestures, symbols, or objects in the video can be detected using visual recognition models. Contextual Understanding **:** Visual data often provides context that may be missed in audio or text alone. For example, a video might contain abusive language in its audio track, but a specific visual context (like a threatening gesture or a hate symbol) can amplify the harmful nature of the content. Computer vision models identify these elements to build a full understanding of the content.

3.2. Enhanced Multimodal Fusion

Combining Visual, Audio, and Text**:** in short video analysis, the ability to integrate data from multiple sources (visual, audio, and text) is essential for accurate content understanding. Computer vision helps create detailed visual embeddings that can be combined with audio and text features for a unified analysis. This fusion improves the detection of complex phenomena, such as identifying abusive language when it is paired with harmful visuals or offensive gestures. Transformer-Based Fusion **:** advanced fusion techniques, such as those based on transformers, allow for dynamic interaction between visual features and other modalities. Computer vision contributes critical visual embeddings that are combined with audio and textual inputs, allowing the system to focus on important visual cues while simultaneously understanding spoken or written content.

3.3. Temporal Dynamics and Action Recognition

Frame-by-Frame Analysis : short videos often consist of rapid scene changes, actions, or dialogues. Computer vision allows for the analysis of each frame and the identification of dynamic elements like movement, gestures, or changes in facial expressions. Recognizing temporal relationships across frames is essential for understanding context, such as detecting escalating aggression or changes in tone that may indicate abusive behavior.

Action Recognition **:** in short videos, understanding actions and interactions between individuals is key to identifying inappropriate behavior. Computer vision-based models can detect actions like threatening gestures or violent movements, which can be used alongside audio and text for a comprehensive understanding of potentially abusive content.

Computer vision is indispensable in the development of deep learning-based multimodal systems for short video analysis. Its ability to process and interpret visual data ranging from identifying objects and gestures to recognizing actions and context enables these systems to better understand and detect abusive language. By combining computer vision with audio and text modalities, this approach creates a powerful framework for real-time, scalable, and accurate detection of harmful content in short videos. This multimodal fusion not only enhances the effectiveness of content moderation but also contributes to a safer and more inclusive online environment.

4. Result and Discussion

The detection of abusive language in social media video content is a multifaceted challenge that involves integrating various data modalities, such as text, audio, and visual information. A deep learning-based multimodal approach is essential to capture the complexity and nuances of abusive content in these videos.

4.1. Advancing Multimodal Fusion Techniques

Improved Fusion Architectures**:** One of the key challenges is optimizing how different modalities (visual, audio, and text) are fused. Research should focus on developing more advanced fusion strategies, such as hierarchical, attention-based, or cross-modal transformers, to allow for better interaction and integration between these modalities[31]. These techniques could enhance the system's ability to identify nuanced abusive language when it appears in combination with visual or auditory cues. Real-Time Multimodal Fusion: Current models often struggle with real-time processing due to the complexity of multimodal fusion. Future research could focus on lightweight models and efficient fusion techniques that reduce computational complexity while maintaining high accuracy, allowing for real-time content moderation on social media platforms[32].

4.2. Deep Learning Architectures for Multimodal Data

Generative Models for Data Augmentation : the use of Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) to generate synthetic abusive content can help create more balanced datasets. This will be especially useful for handling rare or subtle forms of abusive language, such as covert hate speech or implicit aggression. Transformer and Attention Models : transformer-based architectures like BERT for text and ViT (Vision Transformers) for images have shown great promise in various NLP and computer vision tasks. Research should focus on enhancing these models for multimodal fusion tasks, especially in real-time scenarios. Attention mechanisms can be employed to prioritize important features from each modality (text, audio, and visuals) for effective detection of abusive language[33].

A Generative Adversarial Network (GAN) is a revolutionary deep learning framework introduced by Ian Goodfellow in 2014, designed to generate synthetic data that mimics real-world data. GANs consist of two neural networks the generator and the discriminator that work in a competitive, adversarial manner. In Figure 2, the generator creates fake data from random noise, while the discriminator evaluates whether the data is real (from the

training dataset) or fake (generated by the generator). The two networks are trained simultaneousl, with the generator improving its ability to create realistic data, while the discriminator becomes better at distinguishing fake data from real data[24]. This back-and-forth process continues until the generator produces data so realistic that the discriminator cannot tell the difference between real and fake data. The strength of GANs lies in this adversarial relationship, pushing both networks to continuously improve until they reach an optimal state[34].

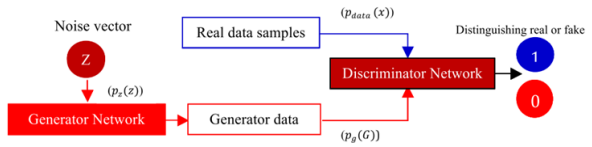

Fig. 2 GAN Architecture [11]

The generator is responsible for producing synthetic data, such as images, text, or videos, starting from random noise. Over time, it learns to produce data that mimics the real data distribution as closely as possible. The discriminator, on the other hand, serves as the evaluator, learning to differentiate between real data and the synthetic data generated by the generator[11]. It provides feedback to the generator, which helps it adjust its output to become more realistic. The training process continues as the generator and discriminator improve, leading to the generation of increasingly convincing data.

GANs have a wide array of applications, from generating realistic images to improving data augmentation in machine learning tasks. One of the most notable applications is in image generation, where GANs can produce high-quality, photorealistic images from random noise[35]. These images can be used for various purposes, including art generation, fashion design, and gaming. GANs have also proven useful in data augmentation, particularly in situations where real-world data is scarce or difficult to obtain. For example, GANs can generate synthetic samples of rare events, improving the performance of models trained on imbalanced datasets, such as abusive language detection, where abusive content is underrepresented.

Furthermore, GANs have become a powerful tool in image-to-image translation tasks, such as transforming images from one domain to another (e.g., from a photograph to a painting) without requiring paired training data[36]. CycleGANs, a variant of GANs, excel in this type of task by learning to map data between domains without the need for direct correspondences between input-output pairs. Additionally, conditional GANs (cGANs) have been developed to condition the generator on specific inputs, such as class labels or text descriptions, to generate specific types of data, such as generating images based on given categories or captions.

Despite their remarkable capabilities, GANs come with challenges, primarily around training stability. The generator and discriminator must be balanced: if one network becomes too powerful, the other fails to improve. This can result in mode collapse, where the generator produces a limited variety of outputs, or the discriminator becomes too strong, preventing the generator from making meaningful progress. To address these challenges, variants like Wasserstein GANs (WGANs) have been developed, which modify the loss function to improve stability and provide a more meaningful training signal. Progressive GANs, another popular variant, start the training process at lower resolutions and progressively increase the resolution, allowing the model to generate high-quality images more efficiently.[37]

The applications of GANs extend beyond image generation. In text generation, GANs can generate coherent sentences and even entire paragraphs, based on textual descriptions or context. In video generation, GANs can produce realistic short video clips from random noise or transform existing videos into new styles. However, training GANs for such complex tasks requires significant computational resources, as generating high-quality content, particularly in high dimensions like video, requires powerful hardware and large amounts of data[36].

In addition to their practical applications, GANs present a variety of ethical challenges. The ability to generate realistic data, including images and videos, can be misused for creating deepfakes or manipulating digital content in harmful ways. As a result, ethical guidelines are crucial in guiding the development and deployment of GANs, ensuring they are used responsibly and do not contribute to the spread of misinformation or harm[38].

Generative Adversarial Networks (GANs) represent a powerful and versatile tool in modern machine learning. They have revolutionized fields like image generation, data augmentation, and unsupervised learning, offering new possibilities in creative industries and AI development. However, challenges like training instability and ethical concerns must be addressed as GANs continue to evolve. As research progresses, the potential for GANs to generate increasingly realistic and diverse data across various domains will continue to expand, making them a foundational technology in AI[39].

4.3. Combining GAN, BERT, and LSTM

We obtained a pattern of approach new model GAN adding LSTM and BERT. GAN is a part of deep learning that specifically works with generative. This GAN process has logic like a game player, where there are two players, namely the Generator (G) and the Discriminator (D)[40].

The Generator process will generate data through a generator network which becomes a synthetic data pattern. This synthetic data resembles the shape of the original data. The Discriminator functions to process incoming real sample data, or original data. The Discriminator (D) determines the shape of the synthetic data pattern and this real data becomes the determinant.

$$Min_G\, Max_D\, V\,(G,D) = E_{x\sim Pdata(x)}\,[Log\, D\,(x)] + E_{z\sim p(z)}\,[Log\,(1 - D(G(z)))] \quad (1)$$

Equation 1 is a general formula form of GAN, where is Min_G the Generator process working to produce synthetic data, and Max_D is the Discriminator process working for real sample data, where for the V process it is the form of Generator and Discriminator which works with $E_{x\sim Pdata(x)}$ real sample data x up to the limit of data x which has been determined by $[Log\, D\,(x)]$ added with synthetic data $E_{z\sim p(z)}$ up to the synthetic data limit, where the discriminator process will perform the function with$[Log\,(1 - D(G(z)))]$[11].

$$Min_{G(\alpha,\beta)}\, Max_{D(\alpha,\beta)}\, V\,(G,D) = E_{x(\alpha,\beta)\sim Pdata(x(\alpha,\beta))}\,[Log\, D\,(x)] + E_{z(\alpha,\beta)\sim p(z(\alpha,\beta))}\,[Log\,(1 - D(G(z(\alpha.\beta))))] \quad (2)$$

The challenge is that the GAN pattern is added with LSTM and BERT. LSTM and BERT act as patterns to reduce imbalances that occur in the training dataset. LSTM and BERT work in the real data sample process which will enter the Discriminator (D) and Generator (G) processes which produce synthetic data (resembling the original shape pattern of the sample data). The GAN used is Conditional GAN (CGAN) where the Generator and Discriminator with input c[41]. The c input is embedded with LSTM then BERT in series. When added to the mathematical model shown in Equation. (2). LSTM is exemplified by α and BERT is exemplified by β [42].

$$\beta_{input\,(x,y)} = BPEE_{(x,y)} + SE_{(x,y)} + PE_{(x,y)} \quad (3)$$

BERT (β) is first split with data from real data (x) and synthetic data (y) to input the BPPE token[43].

$$O = W_x + b \quad (4)$$

For fine tuning in equation 4 using the procedure and analysis of the weight model to be used as parameters and polling (b)[44].

$$P\,(c|x), \theta = \frac{\exp(o_c)}{\sum_{c\in abusive,not-hate} exp_{Oc}} \quad (5)$$

Equation 5, it goes from single layer to multiple layers with a weight matrix, and adds b as the polling output x for the final result to get a model with classification probability P (c | x). This research focuses on the process of adding these two components to process real datasets and synthetic data with the aim of embedding and classification models using the LSTM and BERT model approaches to the GAN structure.

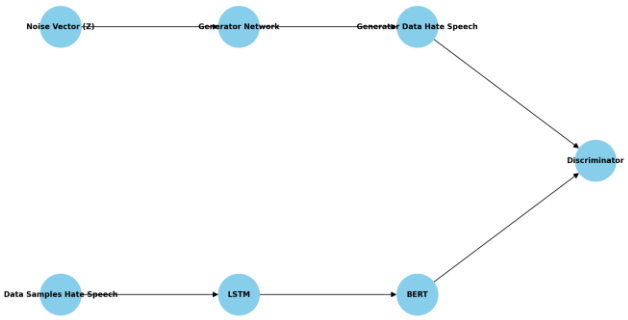

Fig. 3 Proposed Model GAN combining LSTM and BERT

Figure 3 illustrates a system for generating and evaluating synthetic hate speech data. It begins with a Noise Vector (Z), which serves as random input for the Generator Network. This network produces generated data abusive language, which is then evaluated by a discriminator to differentiate between real and synthetic data. Meanwhile, real abusive language data samples are processed through an LSTM (Long Short-Term Memory) network for sequential modeling and passed to BERT (Bidirectional Encoder Representations from Transformers) for contextual understanding. The Discriminator simultaneously evaluates the processed real data and the generated synthetic data, completing the adversarial framework. This setup resembles a GAN (Generative Adversarial Network) architecture, enhanced with NLP models to improve the quality and accuracy of synthetic text data.

4.4. Main Research

The research aims to develop an advanced Deep Learning-based multimodal framework to effectively detect abusive language in social media video content. As social media platforms like YouTube, Facebook, TikTok, and Twitter increasingly rely on video content, identifying harmful behaviors such as hate speech, cyberbullying, harassment, and discriminatory language in videos has become essential for improving online safety. The challenge in this task lies not only in detecting abusive language in text but also in understanding the nuanced cues conveyed through audio and visual modalities. To address this, the proposed system integrates three key modalities: text, audio, and visual signals, which are processed through distinct deep learning models and combined to produce a comprehensive understanding of the context in which abusive language is used.

Textual Data : The first modality involves extracting text from multiple sources within a video. This includes speech-to-text transcription of any spoken dialogue, as well as text extracted from captions, subtitles, or video descriptions. To process this data, the system utilizes advanced Natural Language Processing (NLP) techniques,

including pre-trained transformer-based models such as BERT or RoBERTa. These models are fine-tuned to classify the presence of toxic language, offensive slurs, or other markers of abuse in the text. Sentiment analysis is also incorporated to assess whether the language conveys aggression, hostility, or negativity[45].

Auditory Data : In addition to textual data, the system analyzes audio signals for emotional tone, vocal intonation, and speech patterns that may indicate aggression or hostility. Features such as pitch, tone, speaking rate, and volume fluctuations are extracted using techniques like Mel-frequency cepstral coefficients (MFCCs) and prosody analysis. These features are then input into a Recurrent Neural Network (RNN) or Long Short-Term Memory (LSTM) network, which is adept at modeling temporal sequences and identifying auditory cues associated with abusive speech[46]. The system can detect raised voices, sarcasm, or other vocal markers that are often linked to abusive or aggressive language.

Visual Data : The third modality focuses on visual cues from the video, including the analysis of facial expressions, body language, and gestures. Computer vision techniques such as Facial Action Coding System (FACS) or Convolutional Neural Networks (CNNs) are employed to identify emotions expressed through facial movements (e.g., anger, contempt, disgust)[47]. Body language analysis is used to assess physical gestures like aggressive postures or hostile movements. Additionally, the system uses action recognition models to detect behaviors that may accompany abusive language, such as threatening gestures or violent actions[48][49].

Model Integration and Fusion : Once the individual modalities are processed, the outputs from each model are fused into a unified decision-making process. This multimodal fusion approach is typically done using techniques like late fusion (where predictions from each modality are combined) or early fusion (where features from all modalities are combined at the input level)[50][51]. A deep neural network (DNN) or multimodal transformer model is trained on these fused features to make a final prediction on whether the video contains abusive language[52]. This model is designed to handle complex interdependencies between text, audio, and visual data, ensuring that contextual signals are accurately integrated for robust detection[53][54].

Training and Evaluation : The deep learning models are trained on large-scale datasets of annotated social media videos, which include examples of both abusive and non-abusive content across various domains such as personal vlogs, news, and entertainment. The dataset includes a variety of languages and cultural contexts to ensure the model's generalizability. The system is evaluated using standard metrics such as accuracy, precision, recall, and F1-score, with particular emphasis on minimizing false positives and false negatives—critical for ensuring that the system is both effective and fair[55][56].

Impact and Application: The proposed multimodal approach has the potential to significantly improve content moderation on social media platforms. By combining text, audio, and visual data, the system is more capable of understanding context and identifying subtle instances of abusive language that may be missed by traditional text-only classifiers. Moreover, it can operate in real-time, providing timely alerts and enabling moderators to take action more quickly. The system can also assist in automated flagging and content removal, while ensuring that free speech is respected, by accurately distinguishing between offensive content and legitimate expression[57][58].

Additionally, the research could contribute to the development of personalized content filtering systems that allow users to set preferences for the types of content they wish to avoid, while ensuring that the algorithms do not inadvertently censor non-abusive content.

5. Conclusion

The literature review of the study presents the design and implementation of a deep learning-based multimodal framework aimed at detecting abusive language in video content, specifically on social media platforms. This framework integrates three key modalities visual, audio, and textual data intending to capture the full context of the content within the videos. By leveraging these diverse data sources, the system aims to improve the precision, recall, and overall effectiveness of abusive language detection. One of the central challenges addressed in the study is the issue of imbalanced datasets, which can undermine the performance of detection models. To mitigate this, the authors propose using Generative Adversarial Networks (GANs) to generate synthetic, diverse samples of abusive content, thereby enhancing the model's ability to generalize across various contexts. Additionally, the research incorporates advanced architectures such as Long Short-Term Memory (LSTM) networks and Bidirectional Encoder Representations from Transformers (BERT) to model the temporal dependencies in video sequences.

Ultimately, the study aims to optimize multimodal fusion techniques and identify the best configuration for deploying this technology in real-world applications.

References

[1] J. L. Jaxonlangloislrcahotmailcom, N. St-pierre, and M. Hollis, "Short Video Recommendation through Multimodal Feature Fusion with Attention Mechanism Short Video Recommendation through Multimodal Feature Fusion with Attention

Mechanism," pp. 0–6, 2023.

[2] A. Chhabra and D. K. Vishwakarma, "Engineering Applications of Artificial Intelligence Multimodal hate speech detection via multi-scale visual kernels and knowledge distillation architecture," *Eng. Appl. Artif. Intell.*, vol. 126, no. PB, p. 106991, 2023, doi: 10.1016/j.engappai.2023.106991.

[3] H. M. Sayed, H. E. Eldeeb, and S. A. Taie, "Bimodal variational autoencoder for audiovisual speech recognition," *Mach. Learn.*, vol. 112, no. 4, pp. 1201–1226, 2023, doi: 10.1007/s10994-021-06112-5.

[4] F. T. Boishakhi, "Multi-modal Hate Speech Detection using Machine Learning," no. 2017, 2018.

[5] F. Yang and G. Predovic, "Exploring Deep Multimodal Fusion of Text and Photo for Hate Speech Classification," no. 2017, pp. 11–18, 2019.

[6] T. Caselli, V. Basile, J. Mitrović, and M. Granitzer, "HateBERT: Retraining BERT for Abusive Language Detection in English," *WOAH 2021 - 5th Work. Online Abus. Harms, Proc. Work.*, pp. 17–25, 2021, doi: 10.18653/v1/2021.woah-1.3.

[7] A. Iqbal, B. Nag, and S. Roy, "Knowledge-Based Systems Deep learning based multimodal emotion recognition using model-level fusion of audio – visual modalities," *Knowledge-Based Syst.*, vol. 244, p. 108580, 2022, doi: 10.1016/j.knosys.2022.108580.

[8] P. Vijayaraghavan, H. Larochelle, and D. Roy, "Interpretable Multi-Modal Hate Speech Detection," 2019.

[9] S. Dowlagar and R. Mamidi, "HASOCOne@FIRE-HASOC2020: Using BERT and multilingual BERT models for hate speech detection," *CEUR Workshop Proc.*, vol. 2826, pp. 180–187, 2020.

[10] Z. Wang, Y. Zhao, X. Cheng, H. Huang, J. Liu, and L. Tang, "Connecting Multi-modal Contrastive Representations," no. NeurIPS, pp. 1–16, 2023.

[11] M. Farajzadeh-Zanjani, R. Razavi-Far, M. Saif, and V. Palade, "Generative Adversarial Networks: A Survey on Training, Variants, and Applications," *Intell. Syst. Ref. Libr.*, vol. 217, pp. 7–29, 2022, doi: 10.1007/978-3-030-91390-8_2.

[12] R. Cao, R. K. W. Lee, and T. A. Hoang, "DeepHate: Hate Speech Detection via Multi-Faceted Text Representations," *WebSci 2020 - Proc. 12th ACM Conf. Web Sci.*, pp. 11–20, 2020, doi: 10.1145/3394231.3397890.

[13] E. Mahajan, H. Mahajan, and S. Kumar, "EnsMulHateCyb : Multilingual hate speech and cyberbully detection in online social media," *Expert Syst. Appl.*, vol. 236, no. May 2023, p. 121228, 2024, doi: 10.1016/j.eswa.2023.121228.

[14] R. Rajalakshmi, S. Selvaraj, R. Faerie Mattins, P. Vasudevan, and M. Anand Kumar, "HOTTEST: Hate and Offensive content identification in Tamil using Transformers and Enhanced STemming," *Comput. Speech Lang.*, vol. 78, no. October 2022, p. 101464, 2023, doi: 10.1016/j.csl.2022.101464.

[15] A. Aggarwal *et al.*, "BERT base model for toxic comment analysis on Indonesian social media," *Int. J. Inf. Manag. Data Insights*, vol. 1, no. 1, pp. 714–721, 2021, doi: 10.1016/j.jjimei.2020.100004.

[16] F. Yousif and A. Anezi, "applied sciences Neural Networks," 2022.

[17] R. Pan, J. A. García-díaz, and M. Á. Rodríguez-garcía, "Computer Standards & Interfaces Spanish MEACorpus 2023 : A multimodal speech – text corpus for emotion analysis in Spanish from natural environments," vol. 90, no. March, 2024.

[18] G. Valle-cano, L. Quijano-sánchez, F. Liberatore, and J. Gómez, "SocialHaterBERT : A dichotomous approach for automatically detecting hate speech on Twitter through textual analysis and user profiles," *Expert Syst. Appl.*, vol. 216, no. October 2021, p. 119446, 2023, doi: 10.1016/j.eswa.2022.119446.

[19] A. Velankar, H. Patil, and R. Joshi, "A Review of Challenges in Machine Learning based Automated Hate Speech Detection," pp. 1–9, 2022, [Online]. Available: http://arxiv.org/abs/2209.05294.

[20] R. Cao and R. K.-W. Lee, "HateGAN: Adversarial Generative-Based Data Augmentation for Hate Speech Detection," pp. 6327–6338, 2021, doi: 10.18653/v1/2020.coling-main.557.

[21] H. Sohn and H. Lee, "MC-BERT4HATE: Hate speech detection using multi-channel bert for different languages and translations," *IEEE Int. Conf. Data Min. Work. ICDMW*, vol. 2019-Novem, pp. 551–559, 2019, doi: 10.1109/ICDMW.2019.00084.

[22] K. Abainia, "The Online Behaviour of the Algerian Abusers in Social Media Networks," vol. 2011, pp. 1–13, 1945, [Online]. Available: http://www.dailymail.co.uk/.

[23] S. Sharifirad, "Boosting Text Classification Performance on Sexist Tweets by Text Augmentation and Text Generation Using a Combination of Knowledge Graphs," *2nd Workshop on Abusive Language Online - Proceedings of the Workshop, co-located with EMNLP 2018*. pp. 107–114, 2018, [Online]. Available: https://api.elsevier.com/content/abstract/scopus_id/85

122034405.

[24] M. Gen-recsys *et al.*, *A Review of Modern Recommender Systems Using Generative*, vol. 1, no. 1. Association for Computing Machinery.

[25] H. Fan *et al.*, "Social media toxicity classification using deep learning: Real-world application uk brexit," *Electron.*, vol. 10, no. 11, pp. 1–18, 2021, doi: 10.3390/electronics10111332.

[26] S. Khan *et al.*, "BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 7, pp. 4335–4344, 2022, doi: 10.1016/j.jksuci.2022.05.006.

[27] M. Das, R. Raj, P. Saha, B. Mathew, M. Gupta, and A. Mukherjee, "HateMM : A Multi-Modal Dataset for Hate Video Classification," no. Icwsm, 2023.

[28] P. Aggarwal and B. Mathew, *HateProof : Are Hateful Meme Detection Systems really Robust ?*, vol. 1, no. 1. Association for Computing Machinery.

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

[30] S. Lee and D. K. Han, "Multimodal Emotion Recognition Fusion Analysis Adapting BERT With Heterogeneous Feature Unification," *IEEE Access*, vol. 9, pp. 94557–94572, 2021, doi: 10.1109/ACCESS.2021.3092735.

[31] E. Festus and Ö. Özgöbek, "An inter-modal attention-based deep learning framework using unified modality for multimodal fake news , hate speech and offensive language detection," *Inf. Syst.*, vol. 123, no. November 2023, p. 102378, 2024, doi: 10.1016/j.is.2024.102378.

[32] A. Mandal *et al.*, "Attentive Fusion: A Transformer-based Approach to Multimodal Hate Speech Detection."

[33] X. Zhao, Y. Liao, Z. Tang, and Y. Xu, "Integrating audio and visual modalities for multimodal personality trait recognition via hybrid deep learning," no. January, pp. 1–11, 2023, doi: 10.3389/fnins.2022.1107284.

[34] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 2255–2264, 2018, doi: 10.1109/CVPR.2018.00240.

[35] G. Zhe *et al.*, "Triangle Generative Adversarial Networks," *Adv. Neural Inf. Process. Syst.*, vol. 30, no. Nips, 2017.

[36] A. Anaissi, Y. Jia, A. Braytee, M. Naji, and W. Alyassine, "Damage GAN: A Generative Model for Imbalanced Data," *Commun. Comput. Inf. Sci.*, vol. 1943 CCIS, pp. 48–61, 2024, doi: 10.1007/978-981-99-8696-5_4.

[37] N. Jaafar and Z. Lachiri, "Multimodal fusion methods with deep neural networks and meta-information for aggression detection in surveillance," vol. 211, no. August 2022, 2023.

[38] Y. Qu, J. Nathaniel, and P. Gentine, "Deep Generative Data Assimilation in Multimodal Setting."

[39] A. Nalamothu, "Computer Engineering Commons, and the Computer Sciences Commons Repository Citation Repository Citation Nalamothu, Abhishek," 2019, [Online]. Available: https://corescholar.libraries.wright.edu/etd_allhttps://corescholar.libraries.wright.edu/etd_all/2094.

[40] D. Yan, L. Qi, V. T. Hu, M.-H. Yang, and M. Tang, "Training Class-Imbalanced Diffusion Model Via Overlap Optimization," 2024, [Online]. Available: http://arxiv.org/abs/2402.10821.

[41] Z. Wu, Q. Zhang, D. Miao, K. Yi, W. Fan, and L. Hu, "HyDiscGAN : A Hybrid Distributed cGAN for Audio-Visual Privacy Preservation in Multimodal Sentiment Analysis," 2023.

[42] C. Breazzano, D. Croce, and R. Basili, "MT-GAN-BERT: Multi-Task and Generative Adversarial Learning for sustainable Language Processing," *CEUR Workshop Proc.*, vol. 3015, 2021.

[43] M. A. Ibrahim, N. T. M. Sagala, S. Arifin, R. Nariswari, N. P. Murnaka, and P. W. Prasetyo, "Separating Hate Speech from Abusive Language on Indonesian Twitter," *2022 Int. Conf. Data Sci. Its Appl. ICoDSA 2022*, pp. 187–191, 2022, doi: 10.1109/ICoDSA55874.2022.9862850.

[44] H. Karayiğit, "Homophobic and Hate Speech Detection Using Multilingual-BERT Model on Turkish Social Media," *Inf. Technol. Control*, vol. 51, no. 2, pp. 356–375, 2022, doi: 10.5755/j01.itc.51.2.29988.

[45] S. Nuggehalli, J. Zhang, L. Jain, and R. Nowak, "DIRECT: Deep Active Learning under Imbalance and Label Noise," 2023, [Online]. Available: http://arxiv.org/abs/2312.09196.

[46] R. Alshalan and H. Al-Khalifa, "A deep learning approach for automatic hate speech detection in the saudi twittersphere," *Appl. Sci.*, vol. 10, no. 23, pp. 1–16, 2020, doi: 10.3390/app10238614.

[47] M. B. Shaikh and D. Chai, "Multimodal fusion for audio-image and video action recognition," *Neural Comput. Appl.*, vol. 36, no. 10, pp. 5499–5513, 2024, doi: 10.1007/s00521-023-09186-5.

[48] H. Faris, I. Aljarah, M. Habib, and P. A. Castillo, "Hate speech detection using word embedding and deep learning in the Arabic language context," *ICPRAM 2020 - Proc. 9th Int. Conf. Pattern Recognit. Appl. Methods*, no. January, pp. 453–460, 2020, doi: 10.5220/0008954004530460.

[49] B. U. Patil, A. D. Virupakshappa, A. Prakash, and B. Vijaya, "Optimized multi-layer self-attention network for feature-level data fusion in emotion recognition," vol. 13, no. 4, pp. 4435–4444, 2024, doi: 10.11591/ijai.v13.i4.pp4435-4444.

[50] S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren, "Multimodal deep learning for biomedical data fusion : a review," vol. 23, pp. 1–15, 2022.

[51] C. Lv, L. Fan, H. Li, J. Ma, W. Jiang, and X. Ma, "Biomedical Signal Processing and Control Leveraging multimodal deep learning framework and a comprehensive audio-visual dataset to advance Parkinson ' s detection," *Biomed. Signal Process. Control*, vol. 95, no. PA, p. 106480, 2024, doi: 10.1016/j.bspc.2024.106480.

[52] J. Jang, Y. Kim, K. Choi, and S. Suh, "Sequential targeting: A continual learning approach for data imbalance in text classification," *Expert Syst. Appl.*, vol. 179, no. November 2020, p. 115067, 2021, doi: 10.1016/j.eswa.2021.115067.

[53] S. Ramiah, T. Y. Liong, and M. Jayabalan, "Detecting text based image with optical character recognition for English translation and speech using Android," *2015 IEEE Student Conf. Res. Dev. SCOReD 2015*, pp. 272–277, 2015, doi: 10.1109/SCORED.2015.7449339.

[54] R. M. O. Cruz, W. V. de Sousa, and G. D. C. Cavalcanti, "Selecting and combining complementary feature representations and classifiers for hate speech detection," *Online Soc. Networks Media*, vol. 28, no. February, 2022, doi: 10.1016/j.osnem.2021.100194.

[55] H. Nguyen and J. M. Chang, "Synthetic Information towards Maximum Posterior Ratio for deep learning on Imbalanced Data," *IEEE Trans. Artif. Intell.*, 2023, doi: 10.1109/TAI.2023.3330949.

[56] S. Tuarob, M. Satravisut, P. Sangtunchai, S. Nunthavanich, and T. Noraset, "FALCoN : Detecting and classifying abusive language in social networks using context features and unlabeled data," *Inf. Process. Manag.*, vol. 60, no. 4, p. 103381, 2023, doi: 10.1016/j.ipm.2023.103381.

[57] Y. Zong, O. Mac Aodha, T. Hospedales, and S. Member, "Self-Supervised Multimodal Learning : A Survey," pp. 1–25.

[58] Y. Chen and X. Chen, "Exploration of Deep Semantic Analysis and Application of Video Images in Visual Communication Design Based on Multimodal Feature Fusion Algorithm," vol. 15, no. 8, pp. 1051–1061, 2024.