Multimodal Hate Content Detection using Deep Learning

M.Tech Thesis

by

Anukriti Bhatnagar

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

May 2025

Multimodal Hate Content Detection using Deep Learning

A THESIS

Submitted in partial fulfillment of the

requirements for the award of the degree

of

Master of Technology

by

Anukriti Bhatnagar

2302101007

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

May 2025

INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled **Multimodal Hate Content Detection using Deep Learning** in the partial fulfillment of the requirements for the award of the degree of **Master of Technology** and submitted in the **Department of Computer Science and Engineering, Indian Institute of Technology Indore,** is an authentic record of my own work carried out during the period from July 2023 to July 2025 under the supervision of Dr. Nagendra Kumar, Indian Institute of Technology Indore, India.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

AKriti
05/16/2025
Signature of the Student with Date
(Anukriti Bhatnagar)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Nagendrakumar
16/05/2025
Signature of Thesis Supervisor with Date
(Dr. Nagendra Kumar)

Anukriti Bhatnagar has successfully given his M.Tech. Oral Examination held on **30th April, 2025**.

Nagendrakumar
16/05/2025
Signature(s) of Supervisor(s) of M.Tech. thesis
Date: 16/05/2025

Subhra Mazumdar
Signature of Chairman, PG Oral Board
Date: 18.05.2025

Signature of HoD
Date: 18-May2025

ACKNOWLEDGEMENT

I would like to take this opportunity to extend my sincere gratitude to all those who have supported me throughout this journey. Their guidance, encouragement, and contributions have been instrumental in shaping my academic and personal growth, making this experience both rewarding and memorable.

First and foremost, I express my deepest thanks to my supervisor, **Dr. Nagendra Kumar**, for his invaluable mentorship and consistent support. His insightful advice, expert guidance, and unwavering encouragement have been pivotal in navigating the challenges of this research. I am profoundly grateful for his dedication and commitment, which have inspired and motivated me throughout this work. I am also deeply thankful to my lab mates for their collaboration and constructive feedback during various stages of my research. Their suggestions, camaraderie, and shared enthusiasm have been invaluable in improving the quality of my work and making this journey more enjoyable. I am equally grateful to Dr. Ranveer Singh, Head of the Department of Computer Science and Engineering, for his guidance and support, which have significantly contributed to the successful completion of this project.

I wish to extend my heartfelt gratitude to Prof. Suhas Joshi, Director of the Indian Institute of Technology Indore, for fostering a conducive academic environment and providing the opportunity to explore my research potential. The resources and encouragement offered by this institution have greatly enhanced my learning and growth. Above all, I owe my deepest gratitude to my parents for their unwavering love, patience, and encouragement. Their faith in my abilities and constant support have been a cornerstone of my achievements. Their sacrifices and dedication have enabled me to pursue my aspirations, and I remain forever grateful for their presence in my life. Lastly, I extend my appreciation to everyone who has contributed to this journey, whether directly or indirectly. Your guidance, support, and encouragement have played a vital role in making this endeavor a success. Thank you for being part of this significant milestone in my academic life.

Anukriti Bhatnagar

Dedicated to My Family

ABSTRACT

Over the past two decades, social media platforms have revolutionized global communication by enabling billions of users to share and consume content in real-time across geographic and cultural boundaries. With the rise of video-first platforms, such as TikTok, Instagram, and YouTube, communication has increasingly become multimodal, combining text, images, and audio into complex and expressive formats. While this evolution enriches user interaction, it also complicates content moderation, particularly in detecting subtle and context-dependent forms of hate speech. Implicit hate speech, unlike its explicit counterpart, often relies on coded language, cultural references, sarcasm, or multimodal cues, making it significantly harder to detect using conventional, unimodal systems.

In this thesis, we address this critical and underexplored problem by introducing a novel task of hate speech detection in videos. To facilitate research in this domain, we present a new dataset curated specifically for this task, consisting of approximately 2,000 annotated videos. Each video is enriched with aligned modality-specific inputs, including textual transcripts, extracted audio features, and visual frames, along with auxiliary features such as sentiment scores, emotion cues, and image captions.

We propose a multimodal framework that harnesses the strengths of pretrained encoders (BERT for text, ViT for images, and Wav2Vec2 for audio) to obtain robust feature representations. These are augmented with handcrafted sentiment, emotion, and caption embeddings, and fused using a hierarchical attention mechanism to capture the interplay between modalities. To better separate implicit hate signals from benign or overtly toxic content, we employ supervised contrastive learning over concatenated multimodal embeddings, encouraging intra-class compactness and inter-class distinctiveness in the representation space.

Our approach achieves strong performance on both the proposed dataset and the existing HateMM dataset, outperforming prior methods and setting a benchmark for implicit hate detection in multimodal video content. This work offers critical insights into the design of socially responsible AI systems for content moderation in today's multimedia-driven digital landscape.

Contents

List of Figures

List of Tables

Acronyms

Acc	Accuracy
BERT	Bidirectional Encoder Representations from Transformers
CAAF	Context-Aware Self-Attention Fusion
CMHFM	Cross Modal Hierarchical Fusion Multimodal
CNNs	Convolutional Neural Networks
GPT	Generative Pre-trained Transformer
LLaMA	Large Language Model Meta AI
LLM	Large Language Model
MCMF	Multilevel Correlation Mining Framework
MSA	Multimodal Sentiment Analysis
MulT	Multimodal Transformer
NLP	Natural Language Processing
NRC	National Research Council
OFA	One-For-All
Pre	Precision
Rec	Recall
RNN	Recurrent Neural Network
SCL	Supervised Contrastive Learning
SOTA	State of the Art
SVM	Support Vector Machine
VADER	Valence Aware Dictionary for sEntiment Reasoning
VGG	Visual Geometry Group
ViT	Vision Transformer

ViViT	Video Vision Transformer
WAV	Waveform Audio File Format

Chapter 1

Introduction

Over the past twenty years, social media platforms have fundamentally transformed how individuals communicate, share information, and form communities. Platforms like as Twitter (now X), Instagram, Facebook, YouTube, and TikTok enable users to produce and consume vast amounts of content in real-time, transcending geographic, linguistic, and cultural boundaries. With billions of users across the globe, social media has become a powerful tool not only for social interaction but also for influencing public opinion, mobilizing movements, and driving the global flow of information.

The world-wide scale and penetration of social media are staggering. The total count of social media users globally has grown from just 970 million in 2010 to a roughly 5.24 billion users as of January 2025, representing 63.9% of the global population. Among audiences aged 18 and above, this penetration is even higher, reaching 86.1% globally[1] as illustrated in Figure 1.1.

The accessibility and user-driven structure of social networks have allowed people around the world to create and share content freely, allowing anyone with an internet connection to broadcast their thoughts, emotions and experiences on an unprecedented scale. However, this democratization also presents significant challenges, particularly when it

[1] https://backlinko.com/social-media-users

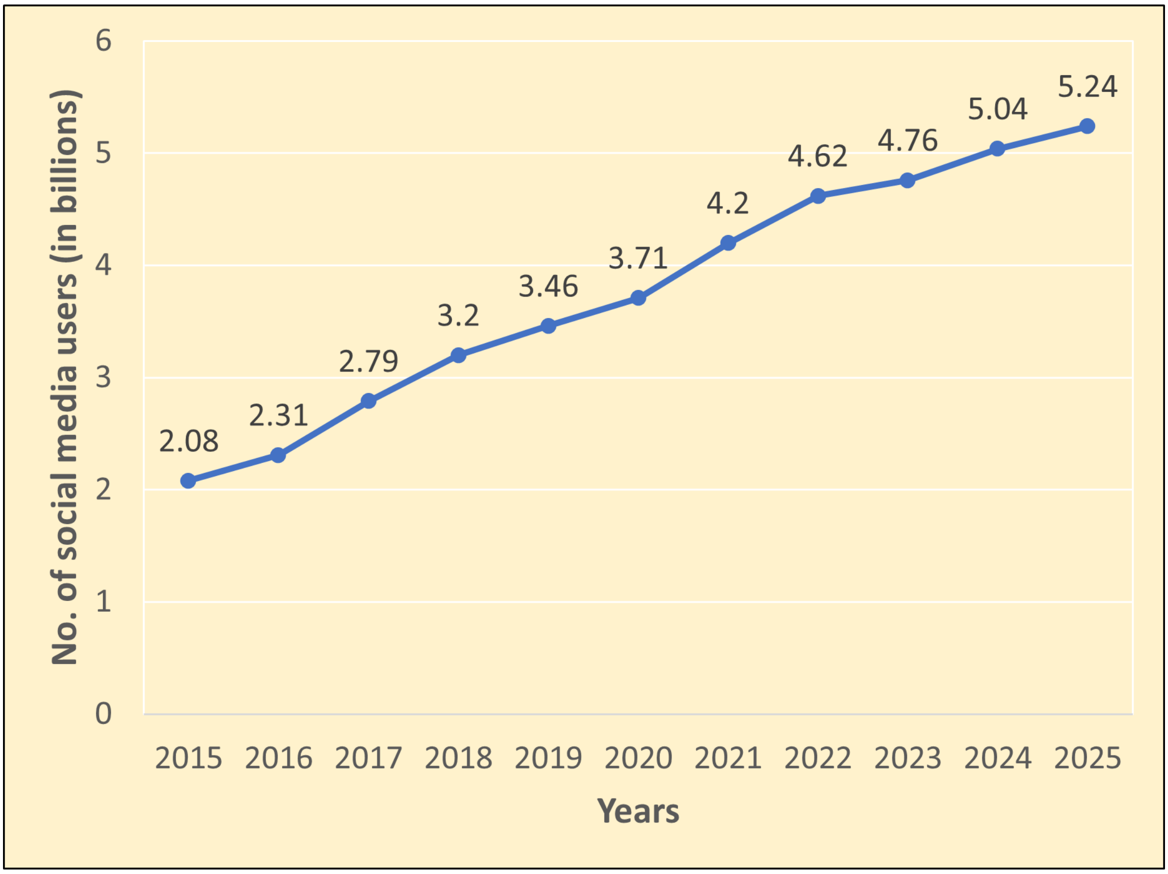


Figure 1.1: Social Media Users over the Years

comes to regulating harmful or offensive content. As social media continues to evolve, so does the complexity of managing the information landscape it creates, making it a critical area of research and intervention. Addressing these challenges is essential to maintaining a safe and responsible digital environment.

Social media today thrives on a dynamic and diverse ecosystem of multimodal content, with each format catering to distinct user behaviors and modes of expression. Text posts allow users to quickly share thoughts, opinions, or updates in a compact form. Images, ranging from personal photos to infographics and memes, communicate visually compelling messages that often carry emotional or cultural significance. Videos combine visuals, motion, and sound, making them particularly effective in capturing attention and conveying complex ideas or narratives. Stories, as seen on platforms like Instagram and Snapchat,

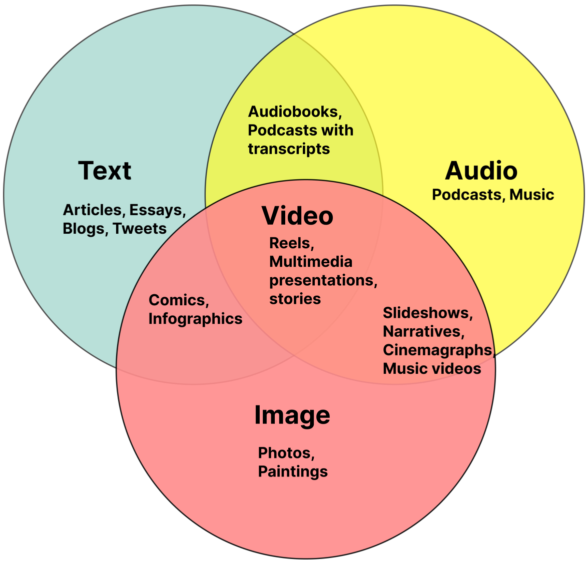


Figure 1.2: Different Content Formats on Social Media

blend short videos, images, and text overlays into time-sensitive content designed for real-time sharing and spontaneous interaction. Meanwhile, live streams offer immersive, unfiltered communication, enabling creators to connect with audiences through a real-time synthesis of video, audio, and interactive feedback. The various formats of social media content have also been shown in Figure 1.2.

Among these formats, video content has emerged as the most dominant and engaging, reshaping how information is shared and consumed online. Its ability to convey emotion, intent, and context more vividly than static text or images has made it a preferred medium for storytelling, education, marketing, and entertainment. This prominence is reinforced by platform algorithms that prioritize video content, boosting visibility and audience engagement. Entire platforms like TikTok have centered their ecosystems around short, viral videos, while YouTube remains a cornerstone for long-form content across a wide array of topics.

However, this proliferation of video sharing also introduces significant challenges. Unlike text or images in isolation, videos encompass multiple modalities (visuals, audio cues, facial expressions, speech, and embedded text), making them inherently complex to analyze. Moreover, videos can be selectively edited, misrepresented, or stripped of context, amplifying their potential for misuse. This complexity complicates content moderation efforts, especially in detecting nuanced forms of harmful content such as hate speech, harassment, and misinformation, where meaning often emerges from the interaction of several modalities rather than a single channel alone.

While social media has revolutionized communication and community-building, it has also become a breeding ground for harmful and hateful content. The very openness that enables global participation and free expression also facilitates the spread of racism, xenophobia, misogyny, homophobia, religious intolerance, and other forms of discriminatory speech. Hate content on these platforms can take many forms, ranging from overt slurs and explicit threats to more subtle expressions like dog whistles, coded language, or biased memes, making its identification and moderation especially challenging.

The viral nature of social media exacerbates the impact of such content. Harmful posts can rapidly reach millions, reinforcing negative stereotypes, inciting violence, or targeting vulnerable communities. In response, social media companies have developed content moderation policies that combine human review with automated tools like keyword filtering, machine learning classifiers, and computer vision systems. However, the scale of user-generated content means that moderation often lags behind, allowing harmful messages to spread before being flagged or removed.

Moreover, moderating hate speech is far from straightforward, especially across languages, cultures, and contexts. What is considered hateful in one region might not be recognized as such elsewhere, and the intent behind a post often depends on cultural cues, tone, or multimodal signals. This makes automated detection of hate content a deeply complex task, requiring models to not only analyze individual components like text or images but also understand their interaction within broader communicative contexts.

To effectively moderate online content and assess its potential harm, sentiment analysis has become a widely used computational tool. At its core, sentiment analysis involves the automated detection and classification of emotional tone in textual data typically categorizing content as positive, neutral, or negative. It has a critical role in understanding public opinion, gauging user attitudes, and flagging emotionally charged or aggressive discourse.

In the social network context, sentiment analysis has been widely applied to monitor user feedback, assess brand perception, and increasingly, to aid in the detection of toxic or hateful content. For instance, highly negative sentiment combined with specific keywords can indicate potential hate speech or harassment. Emotion-based cues such as anger, disgust, or fear, especially when directed toward particular individuals or communities, can serve as early indicators of harmful intent, particularly in cases where traditional keyword-based detection methods fall short. Despite ongoing efforts by online platforms to regulate such content through AI-based detection systems, manual moderation, and community reporting mechanisms [1], hateful content continues to persist. This is largely due to the sheer volume of data generated every day across platforms [2, 3, 4], which poses a major challenge for effective and timely moderation.

However, traditional sentiment analysis models often struggle with the complexity and nuance of user-generated content. Sarcasm, irony, slang, and implicit bias can lead to misclassifications, while context such as the relationship between users or the broader thread of a conversation is frequently overlooked. These limitations become even more pronounced when dealing with content that spans multiple modalities, where the sentiment may be expressed not only through words but also through tone of voice, facial expression, or visual symbolism.

This gap between conventional sentiment analysis and the reality of modern social media has resulted in the development of multimodal sentiment analysis techniques, which aim to incorporate a richer set of features beyond just text.

Multimodal sentiment analysis (MSA) extends traditional sentiment analysis by incorporating multiple sources of information such as text, images, audio, and video to better capture the emotional and semantic content of social media posts. This approach acknowledges that users rarely communicate using a single modality; instead, they combine expressive cues from language, facial expressions, tone of voice, gestures, and visual context to convey sentiment more effectively.

For example, a sarcastic comment accompanied by a laughing emoji or a video clip with a specific intonation may completely alter the perceived sentiment compared to the text alone. Similarly, memes or reaction GIFs often carry strong emotional or cultural meanings that cannot be decoded by analyzing captions alone. Multimodal models aim to learn from the interactions between modalities, thereby achieving a more holistic and accurate understanding of sentiment, intent, and emotional tone.

Recent deep learning developments and transformer-based architectures have substantially boosted the capabilities of multimodal sentiment analysis systems. These models are increasingly adept at aligning and fusing heterogeneous data, such as matching spoken words with facial expressions or correlating image features with text tone. By integrating such signals, MSA can improve the detection of complex or implicit sentiment, especially in scenarios involving sarcasm, coded hate speech, or emotionally manipulative content.

As multimodal communication becomes the norm on platforms like TikTok, YouTube, and Instagram, the importance of sentiment analysis that goes beyond plain text is more critical than ever. Not only does MSA improve the accuracy of content moderation systems, but it also enhances our ability to monitor trends, understand public discourse, and design safer, more empathetic digital environments.

While explicit hate speech, marked by overt slurs or threats, is easier to detect, implicit hate speech presents a significant challenge. Implicit hate speech is defined as expressions that communicate discriminatory or prejudiced views indirectly, often through coded language, implied meanings, or contextual cues [5], which can easily evade traditional detection systems. Instead of using direct hate terms, individuals may rely on stereotypes, in-group/out-group language, or seemingly neutral phrases that carry discriminatory messages in specific contexts. The multimodal nature of social media further complicates detection, as these expressions often appear alongside images, memes, or videos that provide additional context.

The primary difficulty in detecting implicit hate speech lies in its context-dependency. A phrase that seems harmless in one context may be deeply offensive in another, making it challenging for automated systems to interpret accurately. Additionally, implicit hate speech varies across cultures and languages, with certain expressions carrying different meanings depending on regional or linguistic contexts. Sarcasm and irony also pose challenges, as these forms of expression rely on shared understanding and tone, which machines struggle to detect. Furthermore, the multimodal nature of social media requires systems to analyze both text and accompanying visuals or audio to fully understand the message, adding another layer of complexity.

As hate speech continues to spread, there is an urgent need for more sophisticated detection models capable of understanding nuanced language, context, and multimodal cues. Most existing work on hate speech detection has focused on textual content like tweets and comments [6, 7], or image-based hate speech, particularly in memes [8, 9, 10, 11, 12]. While some studies have explored hate detection in videos [13, 3, 14], implicit hate speech detection in videos remains an underexplored area. To the best of our knowledge, we are the first to work in implicit hate speech detection in videos.

In this work, we address the challenge of detecting hate content in videos, with a greater emphasis on implicit hate speech, which is subtle and context-dependent. We propose a contrastive learning approach to effectively model multimodal hateful content by training three modality-specific encoders for text, audio, and images. These encoders are optimized using contrastive loss, computed over concatenated feature representations. This approach allows for a more comprehensive capture of implicit hate speech, which traditional unimodal or basic fusion methods often miss. By aligning features from different modalities, our method leverages the complementary information to create a richer, more nuanced representation for hate speech detection in videos.

The major contributions of this thesis are summarized as follows:

- We introduce a new dataset specifically curated for implicit hate speech detection in videos. The dataset consists of approximately 2,000 videos and provides a valuable benchmark for future.
- We propose a contrastive learning approach to effectively model multimodal hateful content in videos. We train three modality-specific encoders(audio, text, and image) using contrastive loss, computed over concatenated feature representations.
- We evaluate our approach on both our newly curated dataset and the publicly available HateMM dataset. The results illustrate the effectiveness of our proposed multimodal contrastive learning framework in detecting hateful content in videos, particularly implicit hates peech.

The rest of this thesis is organized as follows. Chapter 2 presents a brief overview of the existing literature on multimodal hate content detection, highlighting key trends and research gaps. Chapter 3 describes the proposed dataset and methodology in detail, outlining the design choices and architectural components of the approach. Chapter 4 discusses the experimental setup and analyses the performance of the proposed method through extensive results and analysis. Finally, Chapter 5 concludes the thesis with a summary of findings and potential directions for future work.

Chapter 2

Literature Survey

2.1 Hate Speech Detection

Implicit hate speech poses a significant challenge for automated detection due to its subtle, indirect, and context-dependent nature. Recent studies have primarily focused on text-based approaches to address this complexity. For instance, the study by Elsherief *et al.* [5] introduced a benchmark grounded in social science, featuring a six-class taxonomy and annotated posts from U.S. hate groups. While models like SVM and BERT performed adequately on binary classification, they struggled with fine-grained categorization and explanation. To improve detection in other linguistic and cultural contexts, Guo *et al.* [15] proposed a BERT-based multi-task model (BMA) that integrates semantic, emotional, metaphorical, and fallacy-related features. Similarly, ImpCon by Kim *et al.* [16] has advanced generalizability by employing contrastive learning and multi-feature fusion, aligning post-implication pairs in representation space using contrastive and cross-entropy losses. Additional efforts like ToxiGen [17] have demonstrated the potential of machine-generated data in enhancing classifier robustness. ToxiGen used GPT-3 with demonstration-based prompting to create over 274k toxic and benign statements, significantly improving model performance when fine-tuned on existing classifiers such as HateBERT. Furthering the datacentric approach, a 2024 study introduced adversarial implicit hate speech generation using autoregressive models and demonstrated that retraining on the most challenging examples improves system robustness. Despite these advances, most existing work relies solely on textual information, neglecting the rich multimodal context inherent in video content. In contrast, our study is the first to explore implicit hate speech detection in videos, incorporating not only text but also visual and audio modalities to capture a more comprehensive spectrum of hateful content.

2.2 Feature Extraction

2.2.1 Textual Feature Extraction

Textual feature extraction is fundamental to enabling language models to understand and analyze text data. It plays a critical role in downstream NLP tasks by transforming raw text into meaningful numerical representations that models can process. Two prominent techniques for this are BERT and fastText, each offering distinct advantages. BERT [18] provides deep contextualized embeddings by processing text bidirectionally through a stack of Transformer layers. This architecture allows BERT to capture rich semantic relationships and word dependencies, making it especially powerful for tasks that require a deep understanding of context, such as stance detection, sentiment analysis, and natural language inference. On the other hand, fastText is a lightweight and efficient model developed by Facebook AI that represents words as bags of character-level n-grams. This design allows fastText to generate robust embeddings even for rare or misspelled words, making it suitable for fast and scalable applications. While BERT prioritizes depth and contextual nuance, fastText emphasizes speed and generalization, and each can be selected depending on the computational resources and task complexity involved.

2.2.2 Visual Feature Extraction

Extracting rich visual features is a critical step in multimodal learning pipelines, and a variety of architectures have been employed depending on the spatiotemporal nature of the input data. Vision Transformer (ViT) [19] adopts the transformer architecture for images by splitting an image into patches and encoding them as token sequences, allowing the model to capture long-range dependencies in a non-local, attention-driven manner. ViViT (Video Vision Transformer) [20] extends this approach to videos, introducing spatiotemporal attention mechanisms to process frame-level and temporal relationships without relying on convolutional operations. In contrast, 3D Convolutional Neural Networks (3D-CNNs) [21] extend traditional 2D convolutions by adding a temporal dimension, enabling them to directly model motion and appearance across successive video frames—making them well-suited for video classification tasks. Meanwhile, Inception V3 [22], a deep convolutional network built with optimized inception modules, focuses on efficient spatial feature extraction using multi-scale convolutions and factorization strategies to reduce computation. Together, these architectures offer a diverse toolkit: ViT and ViViT for transformer-based modeling, and Inception V3 and 3D CNNs for hierarchical and temporal feature extraction in static and dynamic visual content.

2.2.3 Audio Feature Extraction

To effectively represent audio signals for downstream multimodal tasks, a combination of traditional signal processing and deep learning approaches is often employed. Mel-Frequency Cepstral Coefficients (MFCC) [23] are a classical and widely used technique that captures the short-term power spectrum of audio by mimicking the human auditory system, producing compact and interpretable features suitable for tasks like speech and emotion recognition. In contrast, Wav2Vec2 [24] is a transformer-based model that learns contextualized representations directly from raw audio waveforms in a self-supervised manner, enabling robust performance even with limited labeled data, particularly for speech understanding. Additionally, AVGG19 (a variant of the VGG architecture adapted for spectrogram inputs) [25], utilizes deep convolutional layers to extract high-level temporal-frequency patterns from audio, making it an effective model for classification tasks. Each of these methods offers unique strengths depending on the complexity of the task, from lightweight handcrafted features (MFCC) to powerful self-supervised embeddings (Wav2Vec2) and deep CNN-based representations (AVGG19).

2.2.4 Emotion Recognition

Emotion recognition aims to identify emotional states like joy, anger, fear, and sadness in text, supporting tasks in user intent analysis, affective computing, and human-computer interaction. NRCLex is a lightweight Python library based on the NRC Emotion Lexicon, which maps words to eight core emotions and sentiment polarity. It provides interpretable emotion tagging by lexicon matching, without relying on training data or complex models, making it ideal for quick, domain-independent analysis of short texts such as tweets, captions, or messages.

2.2.5 Sentiment Detection

Sentiment detection aims to assess the emotional tone behind textual content, playing a key role in tasks such as opinion mining, user feedback analysis, and content moderation. The VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis tool is a rule-based model specifically designed for analyzing sentiments expressed in social media and other informal text. VADER leverages a combination of a sentiment lexicon and grammatical heuristics to detect positive, negative, neutral, and compound sentiment scores in a given sentence. It accounts for nuances such as punctuation, capitalization, degree modifiers (e.g., ”very”), negation, and even emoticons or slang, making it particularly effective for short and noisy text data. As a lexicon and rule-based method, VADER requires no pretraining or labeled datasets, offering a lightweight and interpretable solution for sentiment classification. Its effectiveness and ease of integration make it a popular baseline for sentiment analysis tasks, especially when rapid and domain-agnostic sentiment tagging is needed.

2.2.6 Caption Generation

Caption generation bridges vision and language by converting images into coherent text. The OFA (One For All) framework by Wang et al. [26] introduces a unified, task- and modality-agnostic sequence-to-sequence approach that treats image captioning as text generation guided by instructions. Using a Transformer-based encoder-decoder and a shared vocabulary, it tokenizes visual elements for seamless fusion with text. Despite being pretrained on only 20M image-text pairs, OFA achieves state-of-the-art results through multitask instruction-based learning across tasks like visual grounding and VQA, enabling zero-shot generalization without task-specific heads.

2.3 Hate Speech Detection in Videos

Hate speech detection in videos remains underexplored, with most studies focusing on explicit content. HateMM [3] and MultiHateClip [14] provide datasets of 1,083 and 2,000 videos, respectively, primarily targeting overt hate. Wu et al. [4] used speech-to-text transcription and trained models on extracted audio transcripts to classify hateful content. Similarly, Alcântara et al. [13] built a Portuguese dataset from YouTube using offensive keywords, applying both classical and deep learning models, with CNN and LSTM variants showing strong performance on word embeddings. While these works offer valuable baselines, they largely rely on surface-level text features, revealing a critical gap in detecting implicit hate and highlighting the need for multimodal, context-aware approaches.

Recent advances in multimodal sentiment analysis, emotion recognition, and sarcasm detection tackle challenges in fusion and modality inconsistencies through increasingly refined architectures. Li et al. [27] proposed a multi-level correlation mining framework with self-supervised label generation, using a Transformer-based model to capture low-to-high level interactions across modalities. Dixit et al. [28] demonstrated real-time emotion recognition with modality-specific models fused via bagging and stacking, showing strong generalization on CMU-MOSEI. Wang et al. [29] addressed noisy joint representations through a hierarchical pipeline combining unimodal feature learning, inter-modal interaction, and multi-tasking, using tools like BERT, BiGRU, and Tensor Fusion Networks. MulT [30] introduced directional crossmodal transformers for handling non-aligned sequences, enhancing modalities via temporal convolution and crossmodal attention. For sarcasm detection, Xue et al. [31] proposed a Context-Aware Self-Attention Fusion (CAAF) and Word Weight Calculation (WWC) to capture fine-grained modality inconsistencies, leveraging BERT, ResNet, and OpenSmile. These works highlight the field's shift toward modular, context-aware multimodal learning.

Chapter 3

Proposed Work

3.1 Overview of Proposed Method

The proposed method as shown in Figure 3.3 first extracts modality-specific features from text, images, and audio using dedicated feature extractors. It then applies attention to focus only on the critical content and then aligns these features in a embedding space through modality-specific encoders trained with supervised contrastive loss. Additionally, emotion, sentiment and caption features are also extracted.

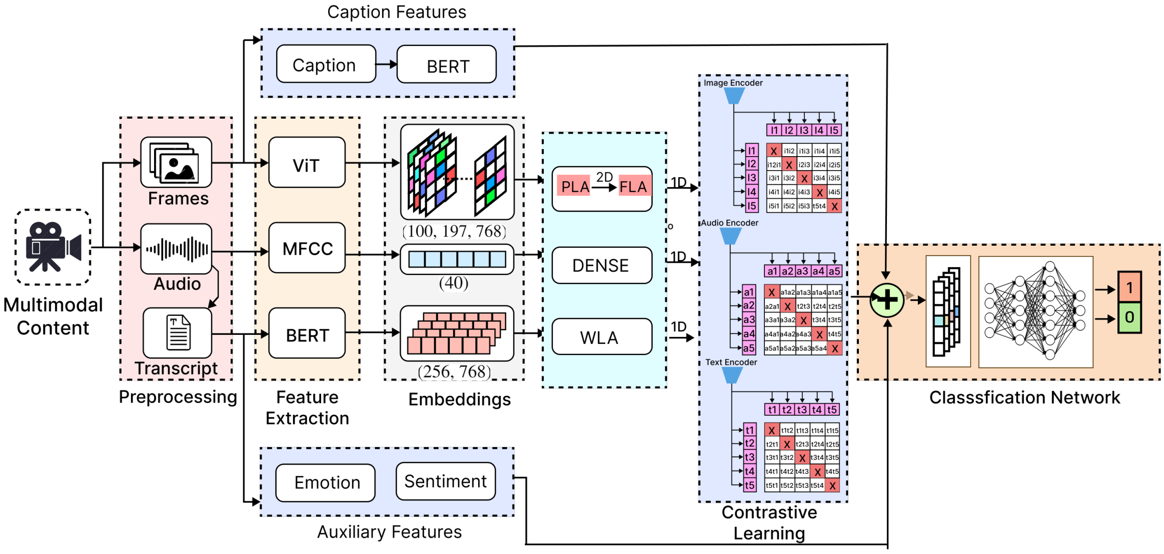


Figure 3.1: Proposed model architecture

3.2 Data Preprocessing

Videos are converted to WAV audio using FFmpeg [32] and transcribed via speech-to-text conversion. FFmpeg is a powerful multimedia processing tool that enables efficient extraction and conversion of audio from video files, supporting a wide range of formats and ensuring compatibility with downstream processing pipelines.

For the visual modality, individual frames are extracted from each video using OpenCV's VideoCapture [33], which allows frame-by-frame reading of video streams for downstream visual processing tasks. The frames are sampled per second. For further processing, 100 frames are uniformly sampled (with padding for videos having fewer frames) from each video, ensuring consistent input dimensions across samples. The entire process has also been illustrated in Figure 3.4.

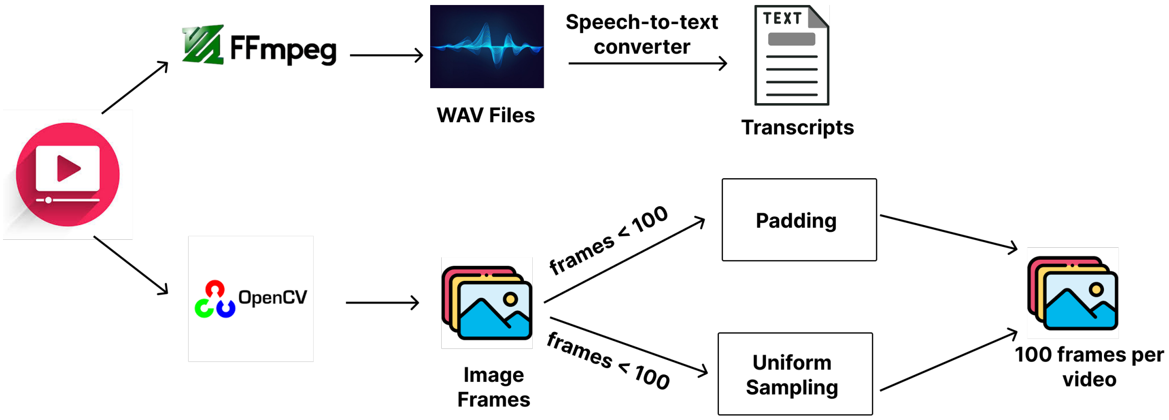


Figure 3.2: Data Preprocessing

3.3 Feature Extraction

We begin by extracting modality-specific features from the video content using dedicated encoders for each input type. For the **visual modality**, we use a Vision Transformer (ViT) to extract spatial features from individual video frames, resulting in a representation denoted by f_I. For the **text modality**, we process the speech-to-text transcriptions using a pre-trained BERT model to obtain contextualized token embeddings, represented as f_T. For the **audio modality**, we extract Mel-Frequency Cepstral Coefficients (MFCC) from the raw audio signal, capturing temporal and spectral features in a compact form denoted by f_A.

In addition to these core multimodal representations, we extract **complementary features** to enhance the semantic richness of our fused representation. From the transcribed text x_T, we compute emotion-related features using the `NRCLex` tool, yielding a vector $e \in \mathbb{R}^{d_e}$ representing scores across multiple emotional categories. We also derive an overall **sentiment score** $s \in \mathbb{R}$ using the `VADER` sentiment analysis tool. These are concatenated to form an emotion-sentiment feature vector:

$$f_{ES} = [e, s] \in \mathbb{R}^{d_e+1} \tag{3.1}$$

This representation enriches the textual features with affective information, which can be critical for stance classification.

Transcript	**Trust**	**Anger**	**Disgust**	**Sadness**	**Anticipation**	**Joy**	**Surprise**	**Fear**	**Positive**	**Negative**
I have got music on the ground now. Ain't he so full of life!	1	1	1	2	3	3	1	1	5	1
Niggers are black horrible ugly monkey looking dogs.	1	1	1	2	1	1	1	1	1	5

Table 3.1: Emotion Detection using NRCLex

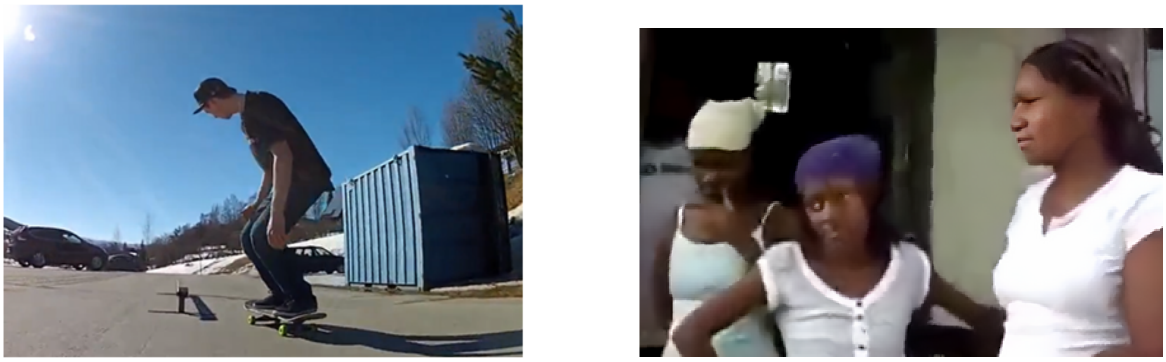
Similarly, for the visual modality, we generate **image captions** using a caption generation model applied to each frame or key frame x_I. The generated caption $c =$

Transcript	**Positive**	**Neutral**	**Negative**	**Compound Score**	**Verdict**
I beg you might like the pharaoh. No verdicts will make the thing more terrible for it is that there's been no friend who can live down the shore of the ship.	0.076	0.809	0.155	-0.3491	Negative
The sunset over the hills was breathtaking, and everyone enjoyed the peaceful evening together.	0.714	0.286	0.034	0.8432	Positive

Table 3.2: Sentiment Detection using VADER

$\texttt{caption\_gen}(x_I)$ is then embedded using BERT to obtain a textual representation:

$$f_{CP} = \texttt{BERT}(c), \quad f_C \in \mathbb{R}^{d_c} \tag{3.2}$$



(A) A young boy wearing a cap on a skateboard in the snow

(B) Three girls in white tops in front of a house

Figure 3.3: Captions generated by the OFA model for selected image frames.

This caption-based embedding captures high-level semantic cues from the visual content, enhancing interpretability and the discriminative strength of the multimodal features.

Together, the core and complementary features, f_I, f_T, f_A, f_{ES}, and f_{CP}, constitute a comprehensive and semantically aligned representation of the video, which serves as input for downstream stance detection.

3.4 Attention Mechanism

To effectively capture salient information from high-dimensional multimodal inputs, we utilize specialized attention mechanisms tailored to the structure of each modality. The attention mechanisms serve to reduce dimensionality while preserving the most informative elements in each feature set, thus enabling a compact and interpretable joint representation.

3.4.1 Image Modality: Patch-Level and Frame-Level Attention

For the visual modality, we extract frame-level features using a pre-trained Vision Transformer (ViT). Each video comprises 100 frames, and each frame is divided into 197 non-overlapping patches, each embedded into a 768-dimensional vector, resulting in a tensor of shape (100, 197, 768).

Patch-Level Attention: The first attention operation, Patch-Level Attention, is applied across the 197 patch embeddings within each frame. This mechanism computes attention scores for each patch, weighing them based on their contextual importance in the frame. Through this operation, less informative patches are down-weighted or ignored, allowing us to aggregate patch information into a single vector per frame. As a result, the shape is reduced from (100, 197, 768) to (100, 768), preserving one embedding per frame that encapsulates the most relevant patch information.

Let $X \in \mathbb{R}^{197\times768}$ represent the patch embeddings for a single frame. We compute self-attention as:

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V \tag{3.3}$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V \tag{3.4}$$

where $W^Q, W^K, W^V \in \mathbb{R}^{768 \times d}$ are learnable weight matrices and d is the attention dimension.

Frame-Level Attention: Next, we apply Frame-Level Attention over the 100 frame embeddings. This step allows the model to learn temporal importance, identifying which frames in the sequence carry the most significant information with respect to the stance classification task. Attention scores are assigned to each frame, and a weighted aggregation is performed to obtain a single video-level embedding of shape (768). This two-stage visual attention pipeline ensures that both spatial (within-frame) and temporal (across-frame) salient information is retained.

We compute frame-level attention as:

$$\alpha_i = \frac{\exp(f_i \cdot w)}{\sum_{j=1}^{100} \exp(f_j \cdot w)} \tag{3.5}$$

$$f_{\text{att}} = \sum_{i=1}^{100} \alpha_i f_i \tag{3.6}$$

where f_i is the feature vector of the i-th frame, and $w \in \mathbb{R}^{768}$ is a learnable context vector.

3.4.2 Audio Modality: Direct Use of MFCC Features

For the audio stream, we extract Mel-Frequency Cepstral Coefficients (MFCC) features, resulting in a single 40-dimensional vector per audio clip. Since the MFCCs are already compact and represent core spectral properties of the audio, we do not apply any attention mechanism in this modality. The features are directly fed into the downstream fusion and classification modules.

3.4.3 Text Modality: Word-Level Attention

Transcripts derived from the videos are tokenized using BERT, which produces an embedding of shape (256, 768), where 256 is the maximum token length and 768 is the embedding size. To distill the most informative tokens, we apply Word-Level Attention across the token sequence. This mechanism learns attention weights for each token, focusing on those that carry strong semantic or emotional signals relevant to stance. A weighted sum of token embeddings is computed based on their attention scores, yielding a final text representation of shape (768). This step ensures that the model concentrates on crucial segments of the text while suppressing less relevant information.

Let text embeddings be represented as $T \in \mathbb{R}^{256\times768}$. Word-level attention is computed as:

$$\beta_i = \frac{\exp(t_i \cdot u)}{\sum_{j=1}^{256} \exp(t_j \cdot u)} \tag{3.7}$$

$$t_{\text{att}} = \sum_{i=1}^{256} \beta_i t_i \tag{3.8}$$

where t_i is the token embedding for the i-th word, and $u \in \mathbb{R}^{768}$ is a learnable context vector.

Thus, by employing modality-specific attention mechanisms, we are able to create a compact, semantically rich representation for each modality. These representations are then used in the downstream multimodal fusion and stance classification stages.

3.5 Encoder training using Contrastive Learning

In this stage, we train modality-specific encoders for image, audio, and text using supervised contrastive learning. This approach helps the encoders learn feature representations that bring samples of the same class closer in the embedding space while pushing apart those of different classes. This is particularly effective in scenarios where multimodal information needs to be aligned semantically across diverse input forms.

The supervised contrastive loss is formulated as:

$$\mathcal{L}_{SCL} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\sum\limits_{k\in S_i^+}\exp\left(\frac{\text{sim}(z_i,z_k)}{\tau}\right)}{\sum\limits_{j=1,j\neq i}^{N}\exp\left(\frac{\text{sim}(z_i,z_j)}{\tau}\right)+\epsilon} \tag{3.9}$$

Here, z_i is the representation of the i^{th} sample, S_i^+ denotes the set of positive pairs for i, sim$(\cdot,\cdot)$ is the cosine similarity, τ is a temperature parameter, and ϵ is a small constant added for numerical stability. N is the number of samples in a batch.

The similarity function used is the cosine similarity, defined as:

$$\cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n}A_iB_i}{\sqrt{\sum\limits_{i=1}^{n}A_i^2}\sqrt{\sum\limits_{i=1}^{n}B_i^2}} \tag{3.10}$$

This loss is applied independently to each encoder (image, text, and audio) encouraging them to learn modality-specific, discriminative embeddings. These embeddings serve as the foundation for subsequent multimodal fusion layers.

By jointly optimizing these encoders using the contrastive loss, the framework effectively aligns multimodal data in a shared semantic space. This alignment enhances the model's ability to integrate complementary information across modalities, ultimately improving performance in downstream tasks such as stance detection or hate speech classification in videos.

3.6 Multimodal Classification

The learned representations of the three modalities are concatenated with the already fused emotion-sentiment feature f_{ES} and caption feature f_{CP} to form a unified feature F. This vector is passed through dense layers with ReLU activations and dropout regularization to produce the final prediction:

$$\begin{aligned} y = &\mathrm{softmax}\Big(W_4\,\text{Dropout}(\sigma(W_3\,\text{Dropout}(\sigma(W_2 \\ &\text{Dropout}(\sigma(W_1 F + b_1)) + b_2)) + b_3)) + b_4\Big) \end{aligned} \tag{3.11}$$

where $y \in \mathbb{R}^C$ represents the predicted probability distribution over C classes.

This streamlined framework effectively integrates multimodal data and contrastive learning to improve the detection of hateful content in videos.

Chapter 4

Experimental Evaluation and Results

4.1 Datasets

We introduce a novel multimodal dataset (Dataset-1) comprising approximately 2000 videos annotated for implicit hate, explicit hate, and non-hate content. The hate videos span a diverse range of targets, including misogyny, racism, terrorism, xenophobia, religious intolerance, anti-Semitism, and disparagement of individuals from developing countries. The dataset captures a broad spectrum of explicit hate cues such as direct slurs, threats, and overt derogatory language, as well as implicit hate cues like sarcasm, coded language, dog whistles, insinuation, and visual symbolism. By encompassing varied modalities and nuanced expressions of hate, this dataset aims to support more comprehensive and robust multimodal hate content detection. The videos have been collected from alternative video-sharing platforms that have gained popularity among creators seeking fewer restrictions on content. These platforms, known for minimal moderation and hosting politically controversial or removed content, serve as a valuable source of unfiltered multimodal data for studying online hate and extreme views.

In addition to the proposed dataset, we used HateMM [3] (Dataset-2), another multimodal dataset to evaluate the performance of our model. HateMM consists of 1,083 videos in total comprising 43.26 hours of multimodal content in English collected from BitChute and Odysee platforms. Out of the 1,083 videos, 431 videos have been labeled as hate videos and 652 videos have been labeled as non-hate videos. However, we used only 1,035 videos out of 1,083 for our study.

4.2 Experimental Setup

The experiments were conducted on a system with Intel(R) Xeon(R) Silver 4215R CPU @ 3.20GHz Processor, Tesla T4 and 384 GB RAM. We split both datasets into train, validation, and test sets containing the total videos, respectively. The HateMM dataset had 662, 166, and 210 samples in training, validation, and test sets, respectively, while our dataset had 1,283 samples in the training set, 325 samples in the validation set, and 401 samples in the test set. We experimented with several combinations of hyperparameters. 32 and 64 were used as batch sizes. Our learning rate was in the range 1e-3,1e-4,1e-5, and we trained our model for 30, 50, 75, and 100 epochs. Adam optimizer [34] was used.

4.2.1 Compared Methods

We have compared our proposed method with several unimodal methods to demonstrate the need for multimodality. These methods have been listed as follows:

- **BERT** [18]: It is a powerful Transformer-based language model designed for natural language understanding tasks, excelling in contextual text representation and classification. Widely used for textual modality analysis in various NLP applications.

- **GPT-4o** [35]: It is a large language model optimized for zero-shot learning via API access, capable of understanding and generating human-like text. Employed here for its strong generalization and language generation capabilities.

- **Llama 3.1-8b** [36]: It is a compact yet powerful large language model known for efficient fine-tuning and inference. Used with zero-shot prompting to evaluate textual modality performance.

- **ViT** [19] : It is a Vision Transformer model that applies Transformer architecture to image patches, enabling effective image classification and feature extraction for visual data.

- **ViViT** [20] : It is an extension of ViT for video analysis, combining spatial and temporal attention to capture video features for tasks such as action recognition and classification.

- **MFCC** [37] : Mel-Frequency Cepstral Coefficients are widely used audio features that capture the timbral texture of sound, essential for speech and audio modality analysis.

- **Wav2Vec2** [24] : It is a self-supervised model for learning robust audio representations directly from raw waveforms, excelling in speech recognition and audio classification tasks.

In addition, we have compared the performance of our proposed model with the following Vision Large Language Models:

- **GPT-4 (Vision-Language)** [1] : It is a state-of-the-art vision-language model combining image and text understanding, used here via API for zero-shot video content analysis and captioning.

[1] https://cdn.openai.com/papers/GPTV_System_Card.pdf

- **Llama-VL** [38] : It is a vision-language large model specialized in video understanding tasks, combining multi-modal inputs for enhanced video classification performance.

We have also compared our approach with other SOTA multimodal methods to demonstrate the effectiveness of our model. These methods have been listed as follows:

- **DeepCNN** [28]: It is a multimodal emotion recognition approach using modality-specific CNN architectures with ensemble fusion, enabling robust real-time emotion classification.
- **CMHFM** [29]: It is a hierarchical multimodal fusion model integrating unimodal feature learning and inter-modal interactions, optimized for sentiment and emotion recognition tasks.
- **CSID** [31]: It is a recent model focusing on attention mechanisms to capture complex cross-modal interactions for sentiment analysis and classification.
- **MCMF** [27]: It is a multi-level correlation mining framework with self-supervised label generation that leverages unimodal and multimodal features for enhanced sentiment detection.
- **MulT** [30] : It is a Transformer-based architecture that employs crossmodal attention to handle non-aligned multimodal sequences, improving emotion and sentiment recognition performance.

4.3 Evaluation Metrics

To assess the performance of our stance detection model, we utilize standard evaluation metrics commonly used in classification tasks: **accuracy**, **precision**, **recall**, **F1 score**, and **macro-F1 score**. These metrics provide a comprehensive evaluation of the model's effectiveness across different aspects of prediction quality.

Accuracy reflects the overall proportion of correctly classified instances and is given by:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

Precision measures the proportion of predicted positive instances that are truly positive:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.2}$$

Recall (or sensitivity) captures the proportion of actual positive instances that were correctly predicted:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.3}$$

The *F1 score* is the harmonic mean of precision and recall, balancing the trade-off between the two:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.4}$$

Given the potential for class imbalance in stance detection tasks, we report the *macro-F1 score*, which computes the unweighted average of F1 scores across all classes, ensuring each class contributes equally to the final score:

$$\text{Macro-F1} = \frac{1}{N}\sum_{i=1}^{N} \text{F1}_i \tag{4.5}$$

These metrics collectively enable a robust evaluation of the model's predictive capability and its performance across varying stance categories.

4.4 Effectiveness Comparison

4.4.1 Binary Classification

Modality	**Method**	**Dataset-1**				**Dataset-2**			
		Acc	F1	Prec	Rec	Acc	F1	Prec	Rec
Text	BERT	0.6907	0.6884	0.6907	0.6907	0.7350	0.6640	0.6750	0.6670
	GPT-4o	0.5312	0.1132	**1.0000**	0.0600	0.5652	0.1964	0.3793	0.1325
	Llama 3.1-8b	0.5312	0.2034	0.6667	0.1200	0.5459	0.2540	0.3721	0.1928
Image	ViT	0.7655	0.7684	0.7658	0.7656	0.7480	0.6720	0.6950	0.6560
	ViViT	0.4912	0.5255	0.4912	0.4914	0.5293	0.5176	0.5172	0.5182
	3D-CNN	0.7521	0.7438	0.7561	0.7390	0.6740	0.5710	0.6190	0.5470
	InceptionV3	0.7702	0.7561	0.7663	0.7515	0.7200	0.6430	0.6530	0.6370
Audio	MFCC	0.4987	0.6655	0.2493	0.5000	0.6750	0.6220	0.5930	0.6790
	Wav2Vec2	0.7531	0.7724	0.7610	0.7533	0.5810	0.5810	0.5270	0.5160
	AVGG19	0.7205	0.6920	0.7073	0.6790	0.6900	0.5755	0.5930	0.5590
Video	GPT-4	0.4988	0.6656	0.4988	**1.0000**	0.4010	0.5724	0.4010	**1.0000**
	LlamaVL	0.4010	0.5724	0.4010	**1.0000**	0.3800	0.5300	0.3700	0.9500
Multimodal	DeepCNN	0.7623	0.7800	0.7481	0.7933	0.5622	0.3065	0.4565	0.2307
	CMHFM	0.7922	0.7921	0.7860	0.7980	0.6057	0.5629	0.6184	0.6184
	CSID	0.8150	0.8154	0.8082	0.8233	0.7320	0.7140	0.7200	0.7230
	MCMF	0.8224	0.8220	0.8200	0.8240	0.5769	0.0435	0.5000	0.2422
	MulT	0.8352	0.8352	0.8320	0.8380	0.6571	0.5212	0.4318	0.6571
	Proposed Method	**0.8660**	**0.8645**	0.8600	0.8700	**0.9612**	**0.9570**	**0.9585**	0.9542

Table 4.1: Effectiveness Comparison for Binary Classification across Different Methods and Datasets

Table 4.1 highlights the performance improvements of our proposed multimodal method over the strongest unimodal and multimodal baselines across the Proposed and HateMM datasets.

BERT displays the strongest performance amongst the text-only models, delivering a balanced performance on both datasets with F1 scores of 68.84% and 66.40% on proposed dataset and HateMM dataset respectively. GPT-4o, when tested based only on transcripts-based prompts, performs very poorly with F1-scores of just 11.32% and 19.64%. Despite a perfect precision (100%) on the proposed dataset, the recall is extremely low (60%), indicating severe class imbalance issues. It likely predicts only one class. LLaMa 3.1–8B, another Large Language Model(LLM) tested, performed only marginally better than GPT-4o, but its performance was still underwhelming. F1 and Recall are low, showing lack of robustness in text classification for hate speech. Despite being powerful LLMs, GPT-4o and LLaMA fail at this task in their raw form, likely due to absence of fine-tuning. BERT, being task-specific, performs better.

ViT is the strongest performer in the visual modality with high and balanced metrics across both datasets (F1 = 76.84% and 67.20%). InceptionV3 performs slightly better than ViT in the proposed dataset and is competitive across datasets. 3D-CNN, though gives a reasonable performance, but generally trails behind ViT and InceptionV3. ViViT surprisingly displays weak performance with F1-score of 52.55%, especially considering it's a video transformer. Static image models like ViT and InceptionV3 outperform spatiotemporal ViViT and even 3D-CNN in this context, likely because frame-wise features are more informative than motion for hate speech cues.

Wav2Vec2is the best performing audio-only model with very strong performance on the proposed dataset (F1 = 77.24%), but much weaker results on HateMM (F1 = 58.10%).MFCC gives high F1-score on the proposed dataset (66.55%), but its precision is extremely low (24.93%), hinting at noisy predictions. AVGG19 shows a moderate performance overall which is neither very strong nor very weak. Inference: Wav2Vec2 confirms the benefit of deep pretrained speech models. MFCC's high recall and low precision indicate it detects many positives—but often they are false posotives.

When testing the LLMs on video-based prompts consisting of image frames and the associated transcripts, GPT-4, LLaMA-VL show extremely high recall(100%) but very low precision and accuracy. Inference: These models over-predict hate speech, possibly labeling nearly everything as hate-related, leading to recall inflation but precision collapse. This behavior makes them unreliable alone.

Coming to comparison with other SOTA models, our proposed method performs the best across all the metrics, on both datasets with F1-scores of 86.45% and 95.7% on the proposed dataset and HateMM dataset respectively. MulT is the second-best performer on the proposed dataset and gives moderate results on HateMM. CSID is consistently strong. CMHFM and DeepCNN are good on the proposed dataset, but degrade sharply on HateMM. MCMF has a decent F1-score on the proposed dataset (82.20%) but collapses on HateMM (4.35%), likely suffering from generalization issues.

4.4.2 Multiclass Classification

Modality	**Method**	**Non Hate Videos**				**Implicit Hate Videos**				**Explicit Hate Videos**				**Overall**
		Acc	F1	Prec	Rec	Acc	F1	Prec	Rec	Acc	F1	Prec	Rec	Macro-F1
Text	BERT	0.7195	0.7192	0.7122	0.7264	0.7107	0.2927	0.4138	0.2264	0.7207	0.5172	0.4348	0.6383	0.5907
	GPT-4o	0.5362	0.6804	0.5197	**0.9851**	0.7282	0.0684	0.3636	0.0377	**0.7880**	0.1748	**1.0000**	0.0957	0.3078
	Llama 3.1-8b	0.5771	0.5066	0.4514	0.5771	0.0189	0.0231	0.0299	0.0189	0.4043	0.4444	0.4935	0.4043	0.3247
Image	ViT	0.7805	0.7854	0.7703	0.8010	0.7307	0.4906	0.4906	0.4906	0.7706	0.4889	0.5116	0.4681	0.5883
	ViViT	0.5012	0.5745	0.5019	0.6716	0.6559	0.1786	0.2419	0.1415	0.6708	0.1951	0.2286	0.1702	0.3161
Audio	MFCC	0.5262	0.6769	0.5142	0.9900	**0.7357**	0.2563	0.3180	0.2146	0.7506	0.0741	0.2857	0.0426	0.2503
	Wav2Vec2	0.7781	0.7963	0.7357	0.8657	**0.7357**	0.3117	0.5000	0.2264	0.7930	**0.6066**	0.5470	**0.6809**	0.5716
Video	GPT-4	0.4938	0.6381	0.4972	0.8905	0.7082	0.0488	0.1765	0.0283	0.7556	0.1695	0.4167	0.1064	0.2855
	Llama-VL	0.4250	0.4800	0.4000	0.7800	0.2500	0.0250	0.1000	0.0150	0.3800	0.1500	0.3800	0.1000	0.2180
Multimodal	DeepCNN	0.7623	0.7512	0.7634	0.7398	0.6785	0.6345	0.6189	0.6512	0.6612	0.5803	0.5692	0.5917	0.6587
	CMHFM	0.7645	0.7534	0.7701	0.7405	0.6802	0.6372	0.6241	0.6509	0.6634	0.5814	0.5723	0.5922	0.6604
	CSID	0.7658	0.7556	0.7714	0.7437	0.6814	0.6394	0.6273	0.6534	0.6649	0.5834	0.5745	0.5938	0.6621
	MCMF	0.7661	0.7568	0.7735	0.7452	0.6819	0.6401	0.6289	0.6541	0.6652	0.5845	0.5751	0.5942	0.6625
	MulT	0.7667	0.7574	0.7741	0.7459	0.6823	0.6408	0.6296	0.6548	0.6658	0.5849	0.5756	0.5948	0.6627
	Proposed Method	**0.7935**	**0.7802**	**0.7910**	0.7697	**0.7020**	**0.6581**	**0.6455**	**0.6710**	**0.6884**	**0.6088**	**0.5981**	**0.6201**	**0.6824**

Table 4.2: Effectiveness Comparison for Multiclass Classification across Different Methods on Proposed Dataset

Table 4.2 presents the performance comparison of various unimodal and multimodal models for multiclass classification across Non-Hate, Implicit Hate, and Explicit Hate categories on the proposed dataset. Our proposed multimodal approach consistently outperforms all baselines—both unimodal and multimodal—across all metrics and categories.

The macro-F1 score is a crucial indicator in imbalanced multiclass settings. The score achieved by our proposed method is 68.24%, which is the highest among all methods. This represents a notable: 5.97% absolute improvement over the best unimodal method (ViT, 58.83%). 1.97% increase over the best performing multimodal baseline (MulT, 66.27%). This consistent improvement across macro-F1 underscores our model's ability to effectively generalize across all categories, particularly the challenging Implicit Hate class.

The Non-Hate class is the most frequently occurring and often the easiest to classify. Most models performed reasonably well for this class. Our proposed method achieves an F1-score of 78.02%, which is only slightly lower than Wav2Vec2 (79.63%) but higher than ViT (78.54%) and BERT (71.92%). However, our model achieves the highest precision (79.10%), indicating fewer false positives. The recall (76.97%) is also competitive, demonstrating the model's balanced ability to detect Non-Hate content without overfitting to the majority class.

Implicit Hate Class is the most subtle and challenging category, where context, tone, and multi-modal signals play a crucial role. Our proposed model achieves an F1-score of 65.81%, the highest across all models. It shows a 16.75% improvement over the best unimodal model (ViT, 49.06%) and 1.73% gain over the best multimodal baseline (MulT, 64.08%). The precision (64.55%) and recall (67.10%) indicate that our model is capable of identifying implicit hate with both sensitivity and specificity. This improvement can be attributed to our model's use of emotion and sentiment embeddings, which enhance the semantic understanding of subtle hate indicators, and contrastive cross-modal fusion, which effectively aligns weak but informative signals across modalities.

Explicit hate is typically characterized by strong language and overtly offensive content. While it is easier to identify than implicit hate, performance can still vary depending on how the modality encodes such cues. Our method achieves an F1-score of 60.88%, the highest overall. This represents a 12.99% improvement over the best unimodal model (Wav2Vec2, 60.66%) and 1.39% over the best multimodal method (MulT, 58.49%). It also maintains the highest precision (59.81%) and recall (62.01%), suggesting that our model detects explicit hate with greater consistency and fewer misclassifications.

For analyzing the overall performance, BERT performs the best among textual models (Macro-F1 = 59.07%), whereas GPT-4o and LLaMA 3.1-8b perform poorly, especially for implicit hate (F1 $<$ 7% for GPT-4o, $<$3% for LLaMA). This emphasizes the need for fine-tuning LLMs in this domain and their current limitations in zero-shot multimodal understanding. ViT significantly outperforms ViViT across all classes (Macro-F1 = 58.83% vs. 31.61%), indicating that temporal modeling in ViViT does not compensate for poor feature alignment in hate classification. Wav2Vec2 stands out (Macro-F1 = 57.16%) compared to MFCC (25.03%), showcasing the importance of learned speech representations over handcrafted features. Among multimodal baselines, all recent methods (e.g., CSID, MulT) perform similarly (Macro-F1 approximately 66.2%), indicating a performance plateau. Our method breaks this ceiling, suggesting that prior fusion strategies lack the ability to integrate fine-grained emotional and semantic cues effectively.

Based on these results we can conclude that multimodal fusion significantly boosts performance compared to unimodal approaches. The proposed method outperforms all, suggesting it integrates complementary cues effectively from image, text, and audio. Robustness across both datasets highlights its generalizability.

These consistent and sizable improvements across datasets underscore the effectiveness of our multimodal architecture in capturing complementary information from text, image, and audio modalities. Moreover, the incorporation of emotion-sentiment signals, caption information, and contrastive alignment enhances cross-modal interaction, ultimately leading to superior performance in implicit hate speech detection.

4.5 Ablation Analysis

4.5.1 Impact of Different Modalities

We evaluate the contribution of different modalities—text, image, and audio—across two classification settings: binary (hate vs. non-hate) and multiclass (non-hate, implicit hate, explicit hate), using both the Proposed Dataset and the HateMM benchmark. The results in Tables 4.3 and 4.4 provide insights into the discriminative power of each modality, as well as their synergistic combinations. The results have also been illustrated in Figures 4.1 and 4.2.

Modality	**Dataset-1**				**Dataset-2**			
	Acc	F1	Prec	Rec	Acc	F1	Prec	Rec
Text	0.8541	0.8365	**0.8601**	0.8129	0.6154	0.5987	0.6231	0.5770
Image	0.8489	0.8233	0.8573	0.7912	0.5935	0.5672	0.3251	0.4780
Audio	0.7705	0.7441	0.7812	0.7115	0.5952	0.5588	0.3178	0.4764
Text + Audio	0.8038	0.7742	0.7990	0.7512	0.5993	0.5755	0.7742	0.5341
Audio + Image	0.8150	0.7899	0.8205	0.7623	0.5932	0.5602	0.3200	0.4911
Text + Image	0.8535	0.8290	0.8590	0.8005	0.6175	0.5901	0.6842	0.5120
Proposed Method	**0.8660**	**0.8645**	0.8600	**0.8700**	**0.9612**	**0.9570**	**0.9585**	**0.9542**

Table 4.3: Impact of Different Modalities on Binary Classification across Different Datasets

For binary classification on the proposed dataset, text(F1 = 86.28%) and image(F1 = 86.03%) modalities achieve near-identical performance.These results suggest that textual and visual features independently capture strong cues for distinguishing between hate and non-hate content. Audio trails behind (F1 = 77.81%), indicating that acoustic cues alone are less discriminative in this binary setting. On the HateMM dataset, the performance of all unimodal models significantly drops.Moreover, image and audio modalities yield very low precision (29.95% approximately), highlighting poor discriminative capability when used in isolation on HateMM. For the proposed dataset, the combination Text+Image performs

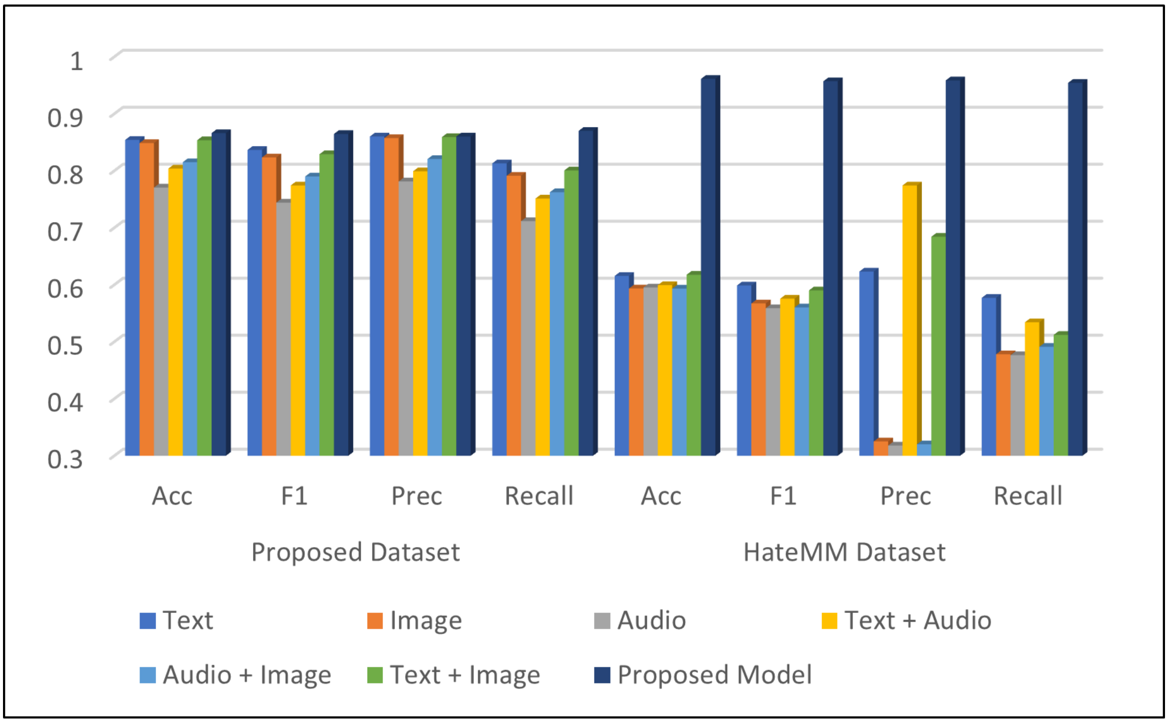


Figure 4.1: Impact of Different Modalities on Binary Classification

similarly to unimodal text and image, suggesting redundancy between these modalities in the binary setting. Text+Audio and Audio+Image offer moderate gains over unimodal audio but do not outperform text or image alone. On HateMM, Text+Image achieves the best binary performance among the baselines (F1 = 62.32%), primarily due to the robustness of the text modality. Text+Audio achieves the highest precision (80.24%) but at the cost of lower recall (51.2%), suggesting a conservative classifier. Our proposed multimodal model outperforms all baselines. This consistent superiority across datasets highlights the model's generalizability and robust cross-modal fusion strategy, capable of leveraging complementary information without being misled by weak or noisy features.

The multiclass setup is more challenging, especially due to the subtlety of implicit hate, which often lacks overt linguistic or visual markers. Here, we analyze performance per class and modality.

Textual modality is the best for Non-Hate (F1 = 84.24%, Recall = 89.05%), showing

Modality	**Non Hate Videos**				**Implicit Hate Videos**				**Explicit Hate Videos**				**Overall**
	Acc	F1	Prec	Rec	Acc	F1	Prec	Rec	Acc	F1	Prec	Rec	Macro-F1
Text	**0.8905**	**0.8424**	0.7991	**0.8905**	0.2264	0.3038	0.4615	0.2264	0.7447	0.6393	0.5600	0.7447	0.5951
Image	0.8607	0.8317	0.8046	0.8607	0.2453	0.3077	0.4127	0.2453	0.7234	0.6267	0.5587	0.7234	0.5587
Audio	0.7811	0.7677	0.7548	0.7811	0.4151	0.4583	0.5116	0.4151	0.6064	0.5672	0.5327	0.6064	0.5977
Text + Audio	0.8657	0.8208	0.7803	0.8657	0.3113	0.3750	0.4714	0.3113	0.6915	0.6435	0.6019	0.6915	0.6131
Audio + Image	0.8706	0.8413	**0.8140**	0.8706	0.2642	0.3394	0.4746	0.2642	**0.7872**	**0.6697**	0.5827	**0.7872**	0.6168
Text + Image	0.6816	0.6903	0.6995	0.6816	0.1604	0.2716	**0.8500**	0.1604	0.5532	0.3728	0.2811	0.5532	0.4448
Proposed Method	0.7935	0.7802	0.7910	0.7697	**0.7020**	**0.6581**	0.6455	**0.6710**	0.6884	0.6088	**0.5981**	0.6201	**0.6824**

Table 4.4: Impact of Different Modalities on Multiclass Classification on Proposed Dataset

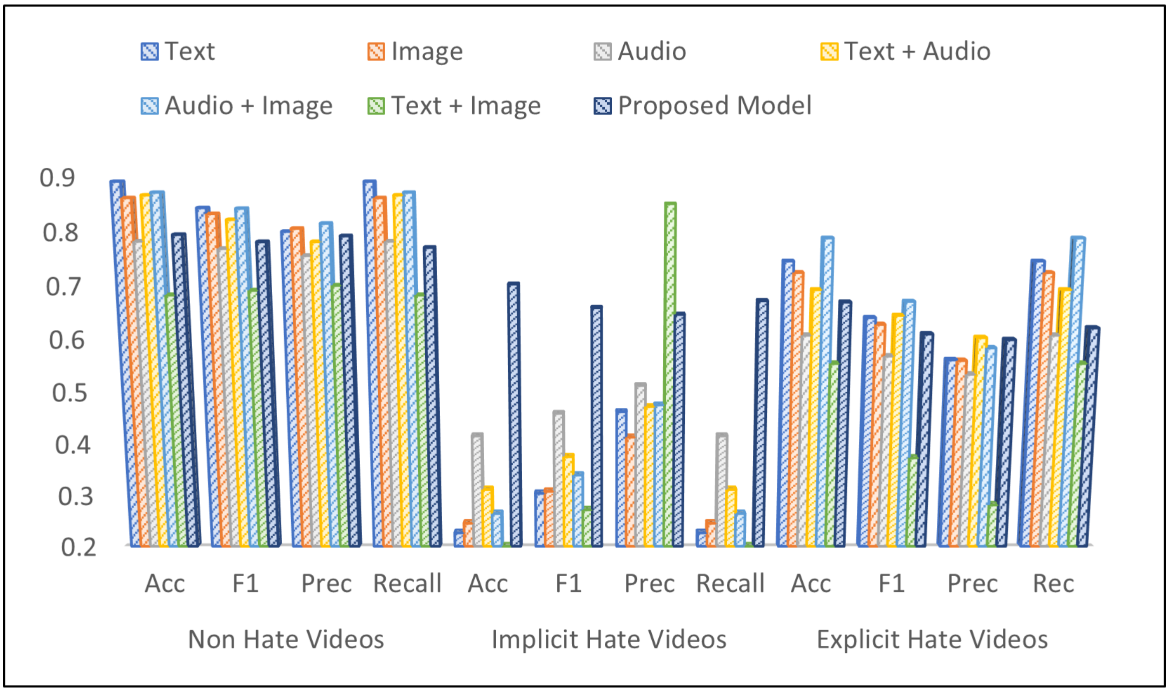


Figure 4.2: Impact of Different Modalities on Multiclass Classification

that textual content is particularly effective in detecting neutral or safe language. It performs poorly on Implicit Hate (F1 = 30.38%) and modestly on Explicit Hate (F1 = 63.93%), indicating difficulty in capturing nuanced or aggressive cues without context. Image modality slightly trails text for Non-Hate (F1 = 83.17%) but performs similarly on the other two classes. Implicit Hate remains difficult (F1 = 30.77%), and Explicit Hate is handled with moderate success (F1 = 62.67%). Audio modality performs significantly better on Implicit Hate (F1 = 45.83%), likely due to intonation and paralinguistic features. It also achieves decent scores for Explicit Hate (F1 = 56.72%), but lags for Non-Hate. This highlights that audio is crucial in distinguishing subtle emotional cues, especially for implicit expressions of hate, where textual or visual modalities may lack salience.

Text + Audio offers balanced performance across all classes, with an overall Macro-F1 = 61.31%. It improves implicit hate detection (F1 = 37.50%) over text-only modality. Audio + Image performs slightly better (Macro-F1 = 61.68%) with high precision (81.40%) for Non-Hate and decent performance across the board. Text + Image performs the worst among multimodal variants (Macro-F1 = 44.48%), primarily due to poor implicit hate recognition (F1 = 27.16%). Despite high precision (85%), recall is extremely low (16.04%), indicating the model is overly cautious, possibly due to noisy alignment between image and text signals.

Our proposed model significantly outperforms all baselines across all classes with Macro-F1 of 68.24%, a 6.56% absolute improvement over the next best multimodal system (Audio + Image, 61.68%). The model also exhibits balanced precision and recall, avoiding the trade-off seen in other models.

This performance stems from three core innovations. Attention-guided fusion (PLA, WLA, FLA) which allows the model to focus on modality-relevant features per class. Emotion and sentiment embeddings help detect subjective and context-rich cues critical for implicit hate. Contrastive alignment of cross-modal representations improves the model's ability to reconcile weak signals across modalities.

To summarize, Text dominates for Non-Hate detection, while audio proves essential for Implicit Hate. Image features, though moderately effective alone, boost performance when paired properly. Naive multimodal fusion strategies (e.g., Text+Image) underperform, particularly for subtle categories.

The proposed method consistently achieves the best performance, showing high adaptability across class types, strong generalization across datasets and balanced precision and recall. These results strongly support the necessity of fine-grained, semantically aware fusion methods and highlight the value of integrating emotional and contrastive learning signals in hate speech detection.

4.5.2 Impact of Emotion, Sentiment and Caption Features

Features	**Proposed Dataset**				**HateMM Dataset**			
	Acc	F1	Prec	Rec	Acc	F1	Prec	Rec
w/o Emotions	0.8310	0.8072	0.8341	0.7864	0.5961	0.5634	0.5375	0.4892
w/o Captions	0.8547	0.8298	0.8594	0.8021	0.6002	0.5811	0.7913	0.4984
w/o Sentiments	0.8392	0.8193	0.8435	0.7947	0.6063	0.5774	0.6635	0.5079
Proposed Method	**0.8660**	**0.8645**	0.8600	**0.8700**	**0.9612**	**0.9570**	**0.9585**	**0.9542**

Table 4.5: Impact of Different Features on Binary Classification across Different Datasets

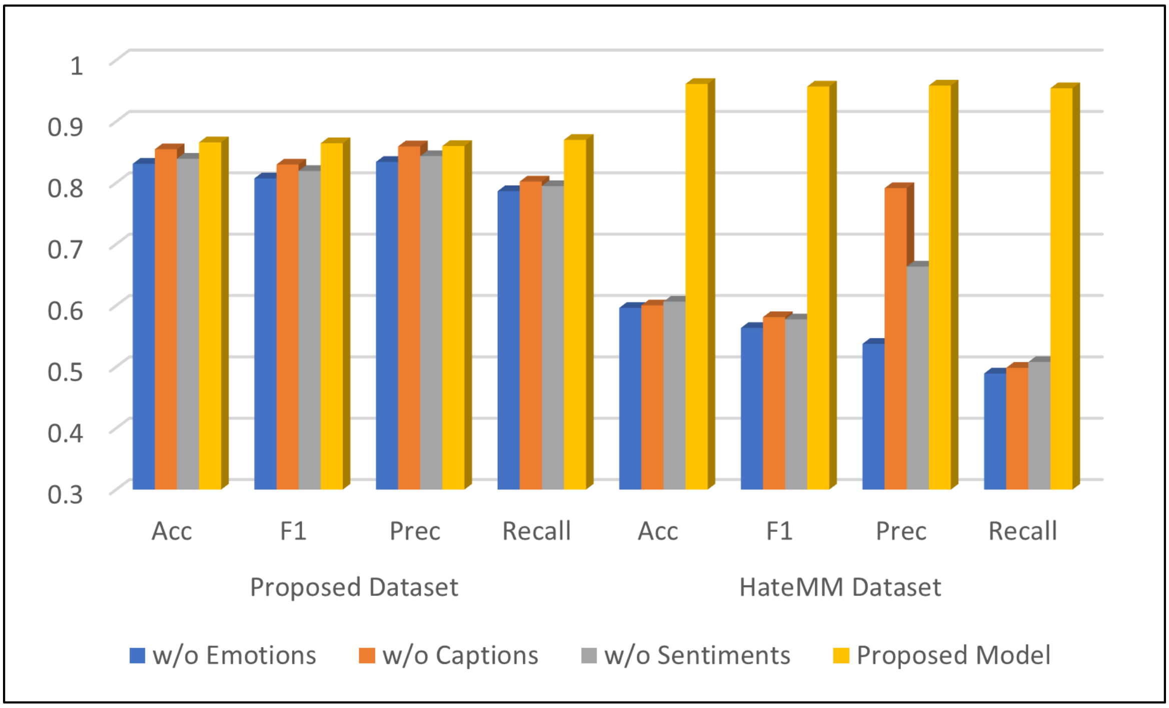


Figure 4.3: Impact of Different Features on Binary Classification

Features	**Non Hate Videos**				**Implicit Hate Videos**				**Explicit Hate Videos**				**Overall**
	Acc	F1	Prec	Rec	Acc	F1	Prec	Rec	Acc	F1	Prec	Rec	Macro-F1
w/o Emotions	0.8209	0.8312	**0.8418**	0.8209	0.6415	0.6507	0.6602	0.6415	0.5851	0.5609	0.5392	0.5851	0.6809
w/o Captions	0.8258	0.8217	0.8177	0.8258	0.6509	0.6448	0.6389	0.6509	0.5425	0.5542	0.5667	0.5425	0.6736
w/o Sentiments	**0.8408**	**0.8325**	0.8244	**0.8408**	0.6226	0.6438	**0.6667**	0.6226	0.5638	0.5548	0.5463	0.5638	0.6770
Proposed Method	0.7935	0.7802	0.7910	0.7697	**0.7020**	**0.6581**	0.6455	**0.6710**	**0.6884**	**0.6088**	**0.5981**	**0.6201**	**0.6824**

Table 4.6: Impact of Different Features on Multiclass Classification on Proposed Dataset

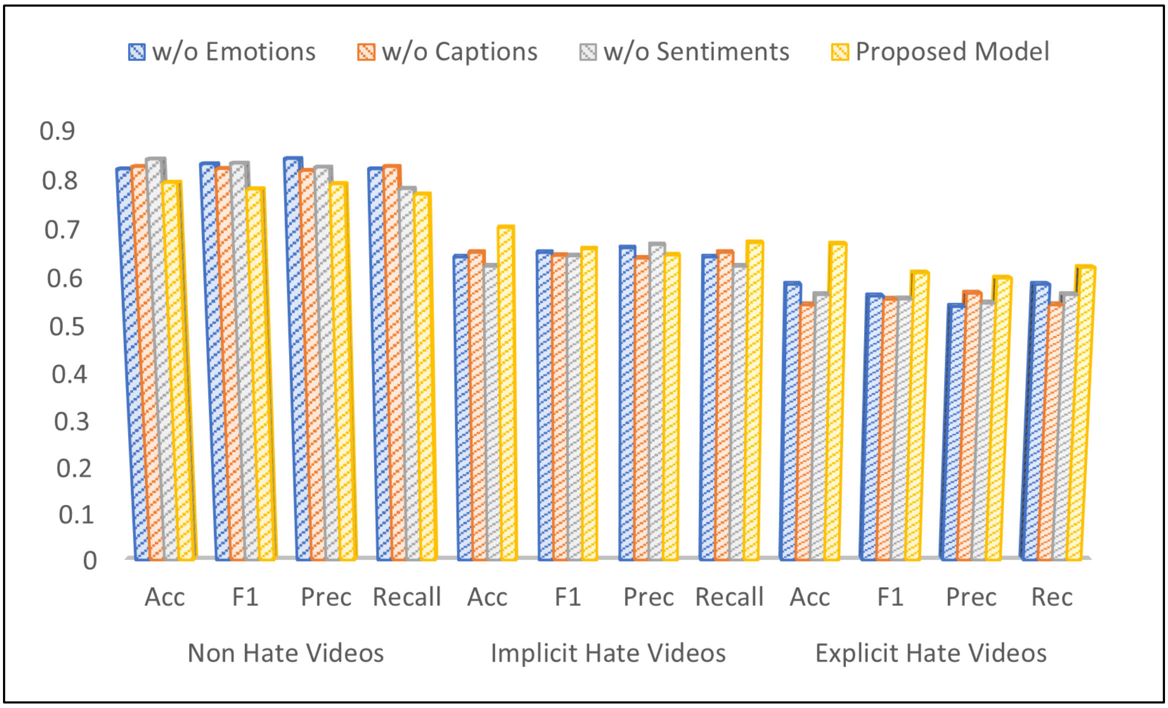


Figure 4.4: Impact of Different Features on Multiclass Classification

Table 4.5 and Figure 4.3 compare the performance of models trained with and without emotion, caption, and sentiment features, against the full Proposed Method on binary classification tasks using the Proposed Dataset and the HateMM Dataset. Notably, the full proposed model, which uses all features, achieves a balanced performance with an overall accuracy of 71.32% and a macro-F1 of 66.29%, outperforming the ablated versions.

For the proposed dataset, the proposed method performs the best with 86.6% accuracy and F1-score 86.45%. Removing Emotions causes the largest drop in performance. F1 drops by almost 2.4%, indicating emotions are critical for distinguishing hate and non-hate. Removing Sentiments causes a moderate performance drop. F1 falls to 84.54%, showing sentiments contribute but are slightly less crucial than emotions. Removing Captions has the least impact with F1-score of 86.03%, nearly matching the full model suggesting that captions might offer some redundancy with text or visual content.

For HateMM dataset, the proposed method again dominates with very high performance with accuracy of 96.12% and F1-score of 95.7%. Removing any feature reduces performance, but eliminating Emotions yields the worst drop, especially in Recall (95.42% → 52.79%), showing that emotional cues are crucial for detecting hate content. Sentiment removal still maintains the best precision (67.78%), but drops in F1 and recall. Caption removal slightly increases precision (80.1%), but harms recall (50.6%), suggesting captions may overfit toward specific surface-level cues.

Table 4.6 assesses how the removal of each feature affects classification into Non-Hate, Implicit Hate, and Explicit Hate, along with macro-F1.The overall Macro-F1 of the proposed method is 68.24% compared to 68.09%, 67.7% and 67.36% without Emotions, Sentiments and Captions respectively. Although the differences are relatively small, all feature types contribute, with captions showing the greatest performance loss when excluded, especially in Explicit Hate detection.

Classwise, the best performance for Non-Hate category is registered when sentiments are removed with accuracy of 84.08% and F1-score of 83.25%. This suggests that sentiments might add noise or ambiguity for clearly non-hateful content. The proposed method, howver, leads in both Implicit and Explict Hate detection with F1-scores of 65.81% and 60.88% respectively.

Removal of any feature reduces recall and F1, emphasizing the need for multiple features to capture subtle cues. Caption removal causes the sharpest drop (F1 = 55.42%), reinforcing that captions are especially important for detecting overt hate, likely due to explicit cues in generated or user-provided text. Emotions are essential for modeling intensity and affective tone of hateful content. Sentiments sometimes contribute noise to clear-cut cases but are still important for the overall results. Captions prove to be crucial for capturing direct, textual indicators of hate not present in base text.

Chapter 5

Discussion

This work explores a comprehensive multimodal framework for video-based hate speech detection, with a strong emphasis on implicit hate. To ensure consistent input dimensionality and efficient downstream learning, preprocessing tools such as FFmpeg, OpenCV, and uniform frame sampling are utilized. These steps standardize video inputs and enable batch processing, forming a reliable pipeline for multimodal analysis.

Multimodal fusion emerges as critical for achieving state-of-the-art performance. Unlike early fusion or naive concatenation techniques (e.g., DeepCNN, CMHFM), the proposed model employs advanced fusion strategies—including attention mechanisms and cross-modal transformers—which effectively align and integrate information from text, image, and audio. This facilitates the capture of complementary cues such as hateful tone in audio, offensive language in transcripts, and symbolic imagery.

The incorporation of cross-modal fusion greatly enhances the expressiveness of learned representations, which is particularly advantageous for detecting nuanced forms of hate such as implicit hate. While many existing models (e.g., MCMF, DeepCNN) exhibit dataset-specific strengths—performing well on the proposed dataset but poorly on HateMM—the proposed method demonstrates high precision and recall across both datasets. This suggests superior generalization, stronger regularization, and improved robustness to domain shifts, likely due to effective pretraining on diverse multimodal data.

Certain baseline models, such as GPT-4 (video) and MFCC-only (audio), display high recall but low precision, revealing a tendency to overpredict the hate class. In contrast, the proposed method maintains near-symmetric precision and recall, indicating a genuine capacity to distinguish hateful from non-hateful content beyond mere exploitation of class imbalances.

Large language models like GPT-4o and LLaMA 3.1 (text), or LLaMAVL (video), underperform in this task—likely due to the absence of task-specific fine-tuning. The proposed model, in contrast, is trained end-to-end and specifically optimized for hate speech detection, yielding higher adaptability and effectiveness.

The use of strong unimodal encoders (Vision Transformer (ViT) for images, Wav2Vec2 for audio, and BERT/transformers for text) enhances the model's ability to extract rich and discriminative features prior to fusion. Beyond these core representations, the proposed method introduces auxiliary semantic features such as sentiment and emotion embeddings derived from transcriptions, and caption-based cues extracted from video frames. These features provide additional context, particularly helpful for distinguishing aggressive but non-hateful language from subtle, covert hate.

The model also integrates a hierarchical attention mechanism for visual data using patch-level (intra-frame) and frame-level (interframe) attention, allowing it to localize salient features both spatially and temporally. Word-level attention for textual transcripts aids in highlighting stance-relevant keywords, while the no-attention strategy for audio leverages the compactness of MFCCs efficiently, reflecting domain-specific design choices.

Crucially, the model incorporates supervised contrastive loss at the modality level, complementing classification loss. This design promotes better semantic clustering of similar examples across modalities, leading to improved generalization, particularly in low-resource conditions.

The proposed system also achieves balanced F1-scores across all three classes—Non-Hate, Implicit Hate, and Explicit Hate—demonstrating its ability to avoid overfitting to majority classes. This balance reflects the model's holistic understanding of multimodal hate speech, making it a strong candidate for real-world deployment in nuanced detection scenarios.

Chapter 6

Conclusion

This work makes two primary contributions toward advancing multimodal hate speech detection.

First, we introduce a novel dataset specifically curated for detecting hate speech in videos. With approximately 2,000 samples, it stands as one of the first large-scale benchmarks focused on this challenge. By addressing a gap in the literature, this dataset provides a vital resource for future research. The proposed dataset also presents greater classification difficulty compared to HateMM, likely due to its inclusion of subtler hate cues and a more imbalanced label distribution. This makes it a more realistic and challenging benchmark for evaluating generalization.

Second, we propose a contrastive learning-based multimodal framework that effectively integrates textual, visual, and auditory features. Unlike existing models, our method is designed end-to-end with cross-modal attention, strong unimodal backbones, semantic auxiliary features, and supervised contrastive loss—all contributing to a robust and adaptable system. The result is a model that consistently outperforms baselines across multiple datasets while demonstrating balanced performance across class categories.

Together, these contributions offer a meaningful step toward building more accurate, generalizable, and context-aware systems for combating online hate in videos.

6.0.1 Limitations and Scope for Future Work

Despite the strong performance of the proposed approach, several limitations remain that warrant attention in future work.

First, the effectiveness of the multimodal framework is highly dependent on the quality and temporal alignment of data across modalities. Noisy or misaligned inputs, such as inaccurate transcripts, irrelevant image frames, or poor quality audio, can distort the joint embedding space, thereby reducing the accuracy of the classification.

Second, the model's reliance on pre-trained encoders, while beneficial for leveraging prior knowledge, also imposes constraints. These encoders may exhibit domain mismatch or fail to capture subtle, context-specific cues of hate speech, limiting their representational adequacy in complex scenarios.

Third, while supervised contrastive loss (SCL) enhances discriminative feature learning, it is sensitive to hyperparameters like the temperature value and the selection of positive and negative pairs. Suboptimal configurations can hinder the formation of coherent clusters, reducing its overall benefit.

Additionally, although the proposed method outperforms competing models such as MulT and CSID on the proposed dataset, the margin of superiority is not dramatic. This could indicate that the dataset poses higher complexity, or that baseline models may be overfitting to dataset-specific patterns while the proposed method prioritizes generalization and robustness.

A notable architectural limitation is the lack of temporal modeling in the audio stream. While MFCC features offer compact representations, they do not capture prosodic or sequential patterns over time. Incorporating temporal models—such as CNNs, RNNs, or

Transformers—could better exploit audio dynamics relevant to hate speech and stance detection.

Moreover, although the model uses supervised contrastive loss to improve unimodal representations, the encoders are trained independently. This may miss opportunities for cross-modal alignment unless explicitly reinforced in the fusion stage.

The final fusion mechanism is a static concatenation followed by dense layers. In contrast, dynamic fusion strategies, such as gated fusion, co-attention, or cross-modal transformers, might better handle modality-specific noise and enhance adaptive reasoning across modalities.

The model also lacks explicit cross-modal interactions before fusion. For example, there is no cross-attention between modalities (e.g., visual attending to text), which is a feature in some transformer-based fusion models that could enhance contextual understanding.

From a practical standpoint, the method is computationally intensive. Caption generation for sampled frames and BERT-based text processing introduce considerable overhead, potentially limiting scalability or real-time applicability in large-scale deployments.

Furthermore, the uniform sampling of 100 frames with zero-padding may introduce irrelevant information or dilute key temporal signals, especially for shorter or faster-paced videos. A dynamic frame selection strategy or temporal attention mechanism could offer a more adaptive solution.

Future research should aim to address these limitations by exploring models with dynamic and cross-modal attention mechanisms, incorporating temporal modeling across all modalities, and optimizing for efficiency in real-time applications. Extending the framework to support multilingual content and expanding the dataset to include a broader spectrum of hate speech phenomena—including cultural, political, and socio-linguistic variations—would also enhance the model's applicability and generalizability.

Bibliography

[1] Ayako Hatano. “Regulating Online Hate Speech through the Prism of Human Rights Law: The Potential of Localised Content Moderation”. In: *The Australian Year Book of International Law Online* 41.1 (Oct. 2023), pp. 127–156. ISSN: 2666-0229. DOI: 10.1163/26660229-04101017.

[2] Michael Ibanez et al. “Audio-Based Hate Speech Classification from Online Short-Form Videos”. In: *2021 International Conference on Asian Language Processing (IALP)*. IEEE, Dec. 2021, pp. 72–77. DOI: 10.1109/ialp54817.2021.9675250.

[3] Mithun Das et al. “HateMM: A Multi-Modal Dataset for Hate Video Classification”. In: vol. 17. Association for the Advancement of Artificial Intelligence (AAAI), June 2023, pp. 1014–1023. DOI: 10.1609/icwsm.v17i1.22209.

[4] Ching Seh Wu and Unnathi Bhandary. “Detection of Hate Speech in Videos Using Machine Learning”. In: *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*. 2020, pp. 585–590. DOI: 10.1109/CSCI51800.2020.00104.

[5] Mai ElSherief et al. “Latent Hatred: A Benchmark for Understanding Implicit Hate Speech”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021. DOI: 10.18653/v1/2021.emnlp-main.29.

[6] Paula Fortuna and Sérgio Nunes. “A Survey on Automatic Detection of Hate Speech in Text”. In: *ACM Computing Surveys* 51.4 (July 2018), pp. 1–30. ISSN: 1557-7341. DOI: 10.1145/3232676.

[7] Anna Schmidt and Michael Wiegand. “A Survey on Hate Speech Detection using Natural Language Processing”. In: *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics, 2017. DOI: `10.18653/v1/w17-1101`.

[8] Rui Cao et al. “Pro-Cap: Leveraging a Frozen Vision-Language Model for Hateful Meme Detection”. In: *Proceedings of the 31st ACM International Conference on Multimedia*. MM ’23. ACM, Oct. 2023, pp. 5244–5252. DOI: `10.1145/3581783.3612498`.

[9] Rui Cao et al. “Prompting for Multimodal Hateful Meme Classification”. In: (2022). DOI: `10.18653/v1/2022.emnlp-main.22`.

[10] Shivam Sharma et al. “Detecting and understanding harmful memes: A survey”. In: *arXiv preprint arXiv:2205.04274* (2022). Accessed on 2025-12-05. URL: `https://arxiv.org/abs/2205.04274`.

[11] Roshan Nayak et al. “Multimodal Offensive Meme Classification u sing Transformers and BiLSTM”. In: *International Journal of Engineering and Advanced Technology* 11.3 (Feb. 2022), pp. 96–102. ISSN: 2249-8958. URL: `http://dx.doi.org/10.35940/ijeat.c3392.0211322`.

[12] Ming Shan Hee, Wen-Haw Chong, and Roy Ka-Wei Lee. “Decoding the Underlying Meaning of Multimodal Hateful Memes”. In: IJCAI-2023 (Aug. 2023), pp. 5995–6003. DOI: `10.24963/ijcai.2023/665`.

[13] Cleber Alcântara, Viviane Moreira, and Diego Feijo. “Offensive Video Detection: Dataset and Baseline Results”. eng. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Accessed on 2025-12-05. Marseille, France: European Language Resources Association, May 2020, pp. 4309–4319. ISBN: 979-10-95546-34-4. URL: `https://aclanthology.org/2020.lrec-1.531/`.

[14] Han Wang et al. “MultiHateClip: A Multilingual Benchmark Dataset for Hateful Video Detection on YouTube and Bilibili”. In: *Proceedings of the 32nd ACM In-*

ternational Conference on Multimedia. MM '24. ACM, Oct. 2024, pp. 7493–7502. DOI: `10.1145/3664647.3681521`.

[15] Tengda Guo et al. "Implicit Offensive Speech Detection Based on Multi-feature Fusion". In: *Knowledge Science, Engineering and Management*. Springer Nature Switzerland, 2023, pp. 27–38. ISBN: 9783031402869. DOI: `10.1007/978-3-031-40286-9_3`.

[16] Youngwook Kim, Shinwoo Park, and Yo-Sub Han. "Generalizable Implicit Hate Speech Detection Using Contrastive Learning". In: *Proceedings of the 29th International Conference on Computational Linguistics*. Ed. by Nicoletta Calzolari et al. Accessed on 2025-12-05. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 6667–6679. URL: `https://aclanthology.org/2022.coling-1.579/`.

[17] Thomas Hartvigsen et al. "ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022. DOI: `10.18653/v1/2022.acl-long.234`.

[18] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: `10.18653/v1/N19-1423`.

[19] Alexey Dosovitskiy. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020). Accessed on 2025-05-12. URL: `https://arxiv.org/pdf/2010.11929/1000`.

[20] Anurag Arnab et al. "ViViT: A Video Vision Transformer". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2021, pp. 6816–6826. DOI: `10.1109/iccv48922.2021.00676`.

[21] Shuiwang Ji et al. "3D Convolutional Neural Networks for Human Action Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.1 (2013), pp. 221–231. DOI: 10.1109/TPAMI.2012.59.

[22] Christian Szegedy et al. "Rethinking the Inception Architecture for Computer Vision". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016, pp. 2818–2826. DOI: 10.1109/cvpr.2016.308.

[23] STEVEN B. DAVIS and PAUL MERMELSTEIN. "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences". In: (1990), pp. 65–74. DOI: 10.1016/b978-0-08-051584-7.50010-3.

[24] Alexei Baevski et al. "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". In: 33 (2020). Ed. by H. Larochelle et al. Accessed on 2025-05-12, pp. 12449–12460. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf.

[25] Shawn Hershey et al. "CNN architectures for large-scale audio classification". In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Mar. 2017, pp. 131–135. DOI: 10.1109/icassp.2017.7952132.

[26] Peng Wang et al. "Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework". In: *International conference on machine learning*. Accessed on 2025-12-05. PMLR. 2022, pp. 23318–23340. URL: https://proceedings.mlr.press/v162/wang22al.html.

[27] Zuhe Li et al. "Multi-level correlation mining framework with self-supervised label generation for multimodal sentiment analysis". In: *Information Fusion* 99 (2023), p. 101891. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2023.101891.

[28] Chhavi Dixit and Shashank Mouli Satapathy. “Deep CNN with late fusion for real time multimodal emotion recognition”. In: *Expert Systems with Applications* 240 (Apr. 2024), p. 122579. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2023.122579.

[29] Lan Wang et al. “A cross modal hierarchical fusion multimodal sentiment analysis method based on multi-task learning”. In: *Information Processing amp; Management* 61.3 (May 2024), p. 103675. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2024.103675.

[30] Yao-Hung Hubert Tsai et al. “Multimodal transformer for unaligned multimodal language sequences”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019. DOI: 10.18653/v1/p19-1656.

[31] Hongfei Xue et al. “Breakthrough from Nuance and Inconsistency: Enhancing Multimodal Sarcasm Detection with Context-Aware Self-Attention Fusion and Word Weight Calculation.” In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari et al. Accessed on 2025-12-05. Torino, Italia: ELRA and ICCL, May 2024, pp. 2493–2503. URL: https://aclanthology.org/2024.lrec-main.224/.

[32] FFmpeg Developers. *FFmpeg*. Accessed on 2025-12-05. 2024. URL: https://ffmpeg.org.

[33] Gary Bradski. “The OpenCV Library”. In: *Dr. Dobb’s Journal of Software Tools* (2000). Accessed on 2025-12-05. URL: https://www.elibrary.ru/item.asp?id=4934581.

[34] P Kingma Diederik. “Adam: A method for stochastic optimization”. In: *(No Title)* (2014). Accessed on 202-12-05. URL: https://arxiv.org/abs/1412.6980.

[35] Alec Radford. “Improving language understanding by generative pre-training”. In: (2018). Accessed on 2025-12-05. URL: `https://www.mikecaptain.com/resources/pdf/GPT-1.pdf`.

[36] Hugo Touvron et al. “Llama: Open and efficient foundation language models”. In: *arXiv preprint arXiv:2302.13971* (2023). Accessed on 2025-12-05. URL: `https://noticias.ai/wp-content/uploads/2023/02/333078981_693988129081760_4712707815225756708_n.pdf`.

[37] Shing-Yun Jung et al. “Efficiently classifying lung sounds through depthwise separable CNN models with fused STFT and MFCC features”. In: *Diagnostics* 11.4 (Apr. 2021), p. 732. ISSN: 2075-4418. DOI: `10.3390/diagnostics11040732`.

[38] Hang Zhang, Xin Li, and Lidong Bing. “Video-llama: An instruction-tuned audio-visual language model for video understanding”. In: *arXiv preprint arXiv:2306.02858* (2023). Accessed on 2025-12-05. URL: `https://arxiv.org/abs/2306.02858`.