university of groningen / campus fryslân

Can Multimodal Transformers Beat LLMs? A Cross-Attention Approach to Sarcasm Detection in Social Media Videos

Mohammadhossein Narang
S6028608

Supervisor: Xiyuan Gao

university of groningen

campus fryslân

University of Groningen

Can Multimodal Transformers Beat LLMs? A Cross-Attention Approach to Sarcasm Detection in Social Media Videos

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Xiyuan Gao (PhD Candidate, University of Groningen)
and
Prof. Shekhar Nayak (Assistant Professor Voice Technology, University of Groningen)

Mohammadhossein Narang (s6028608)

June 11, 2025

Acknowledgments

I want to start by sincerely thanking my supervisors, Xiyuan Gao and Shekhar Nayak. Their guidance, patience, and thoughtful feedback throughout this process have been incredibly valuable, and I've learned so much from working with them.

I'm also very grateful to all the professors in my program, who put in so much time and effort to teach and support us over the past year. A special thanks to Dr. Matt Coler, our program director and professor, for his support and for creating such a meaningful academic environment.

Thanks as well to the Faculty of Campus Fryslân at the University of Groningen for providing a supportive and inspiring place to study and grow.

To my family - thank you for always being there for me, not just during this past year but throughout my whole journey. Your support has meant everything.

And finally, I want to give a special thank you to Hiva Naazeri and Dr. Sepideh Yousefzadeh. Your encouragement, kindness, and belief in me have made a big difference, and I'm truly grateful.

Abstract

Detecting sarcasm in social media videos is a complex challenge for natural language processing, largely due to the inherent ambiguity and semantic incongruity of sarcastic expressions, where the intended meaning often contrasts with the literal words. Sarcasm often depends on subtle, unspoken cues such as exaggerated intonation, prosodic changes, or facial expressions that convey underlying attitudes. For example, a raised pitch or exaggerated tone when saying "What a fantastic plan!" may signal a negative or ironic sentiment beneath the surface meaning. While these characteristics complicate automatic detection, the multimodal nature of sarcasm, which includes textual, auditory, and visual signals, offers complementary information that can be exploited to enhance recognition accuracy. This thesis proposes a novel sarcasm detection system employing a transformer-based architecture augmented with cross-attention mechanisms, allowing the model to dynamically integrate and interpret synchronized inputs from text, speech, and facial expressions. Leveraging the MUStARD++ dataset, the model is trained to identify sarcasm in short video content typical of platforms like TikTok. Traditional sarcasm detection methods typically depend on textual cues alone, limiting their ability to capture the nuanced cues embedded in tone and facial expression. By incorporating cross-modal attention, the system dynamically prioritizes and aligns salient features across modalities, effectively capturing the complex interplay of conflicting cues such as a cheerful tone contrasted with negative words that are essential to recognizing sarcasm. Comparative experiments with large language models are conducted to benchmark the proposed model's performance against unimodal and text-only baselines, highlighting the advantages of multimodal integration for sarcasm detection. This research advances the field of affective computing and has practical applications in content moderation, recommendation systems, and social media analytics. Ethical considerations, including bias mitigation and user privacy, are addressed, with future work proposed to explore transfer learning for low-resource contexts and real-time deployment strategies.

Keywords: Sarcasm detection, multimodal sarcasm recognition, cross-attention mechanisms, transformer architecture, natural language processing, prosodic features, affective computing, video content analysis, social media analytics.

Declaration

I hereby affirm that this Master thesis was composed by myself, that the work herein is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification except as specified, nor has it been published. Where other people's work has been used (from any source: printed, internet or other), this has been carefully acknowledged and referenced. During the preparation of this thesis, I used ChatGpt-3.5, ChatGpt-o4-mini and Gemini (2.5 flash) for the following purposes:
sentence restructuring in chapters 1, 2, 3, 4, 5, 6, and 7, generating alternative explanations for technical concepts in chapters 3, 4, 5, creating initial code documentation templates, summarizing background literature for preliminary review. All content was subsequently reviewed, verified, and substantially modified by me.

Mohammadhossein Narang / 11 June 2025

Contents

1 Introduction

Sarcasm is a complex and often ambiguous form of communication where the speaker's intended meaning typically diverges from the literal interpretation of their words. In digital contexts such as social media, sarcasm is not merely linguistic; it is a multimodal performance, conveyed through combinations of textual expression, vocal tone, and facial cues. For example, a comment like "Great job breaking the printer again" may seem like praise if viewed in isolation as text, but when paired with a frustrated tone or an eye roll, it clearly expresses irritation. Such nuances are often missed by text-only models, which may incorrectly classify sarcastic remarks as genuine. While traditional natural language processing (NLP) systems have focused on textual features, recent research highlights that text alone is insufficient for accurately detecting sarcasm (Farabi, Ranasinghe, Kanojia, Kong, and Zampieri, 2024; Gao, Nayak, and Coler, 2024).

Multimodal sarcasm detection models address the challenge of interpreting sarcasm by integrating inputs from text, audio, and visual channels, capturing the nuanced cues that humans intuitively use to discern sarcastic intent. Studies have shown the importance of prosodic features in spoken sarcasm (Aguert, 2022) and demonstrated how facial expressions can enhance model performance (Bhosale et al., n.d.; Ray, Mishra, Nunna, and Bhattacharyya, n.d.). Recent research further supports this approach. Y. Zhang, Zhu, et al. (2025) proposed a progressive interaction model that aligns semantic differences across modalities to improve detection accuracy. Wang et al. (2024) introduced RCLMuFN, a framework that incorporates relational context learning and multiplex fusion. Jia, Xie, and Jing (2023) developed a contrastive learning method to reduce bias in multimodal sarcasm detection. The MO-Sarcation model by Mohit2b (2024) highlights the impact of modality order in identifying sarcasm effectively. These advances collectively reinforce the value of multimodal integration in capturing the subtle and often conflicting signals necessary for accurate sarcasm recognition. As short-form, expressive videos gain popularity on platforms like TikTok, the demand for robust multimodal sarcasm detection continues to grow.

Recent advancements in transformer-based architectures, particularly those employing cross-attention mechanisms, have shown strong potential in effectively fusing multimodal features (Qin, Luo, and Nong, 2024; Castro et al., 2019). Cross-attention enables the model to compute attention weights across different modalities by using one modality to guide the interpretation of another. This allows the system to dynamically align relevant features, such as associating a sarcastic tone or an eye-roll with the underlying meaning of text. By explicitly modeling the interaction between modalities, the mechanism helps capture subtle contradictions and context shifts that are characteristic of sarcasm. This targeted fusion improves the system's ability to resolve ambiguity and interpret intent more accurately. Despite these advancements, several challenges remain, including limitations in dataset size, synchronization of modalities, and the risk of overfitting (Pramanick, Roy, and Patel, 2022; Valliyammai, Monish Raaj, Athish, and Kumar, n.d.).

This thesis builds on existing research by designing a cross-attention transformer-based model for sarcasm detection, trained and evaluated on the MUStARD++ dataset[1]. The proposed system aims to outperform unimodal baselines by leveraging joint representations from all three channels.

[1] https://github.com/cfiltnlp/MUStARD_Plus_Plus

Through empirical experimentation, this project seeks to contribute to ongoing discourse in affective computing, with practical applications in sentiment analysis, content moderation, and online safety monitoring.

1.1 Research Question

The central research question for this thesis revolves around the ability of cross-attention mechanisms in multimodal transformers to effectively detect sarcasm in social media videos. The specific questions are as follows:

1. **How can cross-attention mechanisms effectively integrate text, audio, and facial features for sarcasm detection in multimodal content?**
   This question explores the role of cross-attention mechanisms in integrating different modalities—text, audio, and visual (facial expressions)—in a way that allows for more accurate sarcasm detection. It seeks to understand how each modality contributes to understanding sarcasm and how their interactions are captured within the model.

2. **To what extent does this multimodal approach outperform large language models (LLMs) in sarcasm detection, and is this advantage attributable to the model architecture, the multimodal data, or both?**
   This question compares the performance of multimodal transformer models with that of text-based LLMs, such as BERT or GPT-style architectures, in the context of sarcasm detection. It seeks to critically assess whether any observed performance gains stem solely from the use of multiple data modalities (text, audio, visual) or from the structural benefits of the cross-attention-based modeling approach. This comparison aims to distinguish the contribution of modality richness from architectural innovations.

1.2 Hypothesis

It is hypothesized that a deep learning-based approach utilizing cross-attention mechanisms for the fusion of multimodal inputs (text, audio, and facial expressions) can significantly improve sarcasm detection accuracy. Specifically:

A multimodal transformer model with separate encoders for text, audio, and facial features—combined using cross-attention layers that dynamically prioritize salient cues across modalities—will outperform large language models (LLMs) in detecting sarcasm.

This hypothesis is grounded in prior research emphasizing the unique contributions of different modalities to sarcasm detection (Castro et al., 2019; Aguert, 2022; Gao et al., 2024). For instance, shifts in vocal pitch or exaggerated facial expressions often signal sarcastic intent that may not be evident from text alone. Integrating these features within a unified multimodal framework allows the model to capture subtle sarcastic cues that emerge from the interaction of modalities.

The choice of cross-attention is motivated by its ability to explicitly model the interdependencies between modalities—such as aligning tone of voice with conflicting textual sentiment or associating

facial expressions with sarcastic phrasing. By computing attention weights across different input streams, cross-attention enables the model to identify the most informative modality (or combination) at each step, resulting in a more context-sensitive and accurate prediction.

Furthermore, existing literature (Farabi et al., 2024) highlights the limitations of text-only models in capturing nonverbal cues, reinforcing the need for multimodal modeling in real-world sarcasm detection scenarios (Li, Cao, Xia, and Song, 2023; X. Zhang, Chen, and Li, 2021). By leveraging both the richness of multimodal data and the dynamic alignment capability of cross-attention, the proposed model aims to advance the state-of-the-art in this domain (Karun and Adithya, 2025).

2 Literature Review

The challenge of sarcasm detection has become an important area of study in natural language processing (NLP) and multimodal machine learning. Sarcasm, a form of verbal irony, involves expressing the opposite of what is meant, often requiring nuanced understanding of context, tone, and other modalities. Social media platforms, where sarcastic content is abundant in videos, pose a unique challenge for models to correctly detect and interpret sarcastic intent. This section reviews the key developments in multimodal transformers, large language models (LLMs), and their application to sarcasm detection in videos.

2.1 Multimodal Transformers for Sarcasm Detection

The field of multimodal transformers has made significant strides in addressing tasks that require understanding across different modalities such as text, speech, and images. Transformers, a type of deep learning model, have become dominant in NLP and computer vision due to their ability to capture long-range dependencies within data. Integrating multimodal data into a transformer framework has been shown to significantly improve performance in tasks like emotion recognition, sentiment analysis, and sarcasm detection. (Tsai et al., 2019) proposed a multimodal transformer capable of learning both intra- and inter-modal dynamics through attention mechanisms, enabling effective modeling of unaligned language sequences. (Yoon, Byun, and Jung, 2018) demonstrated that fusing audio and text inputs through a multimodal architecture enhances speech emotion recognition, emphasizing the value of combined modality representations. More recently, (Tian, Xu, Zhang, and Mao, 2023) introduced the Dynamic Routing Transformer Network (DynRT-Net), which dynamically routes multimodal features to better capture sarcastic cues, achieving improved detection accuracy in multimodal sarcasm benchmarks (Farabi et al., 2024).

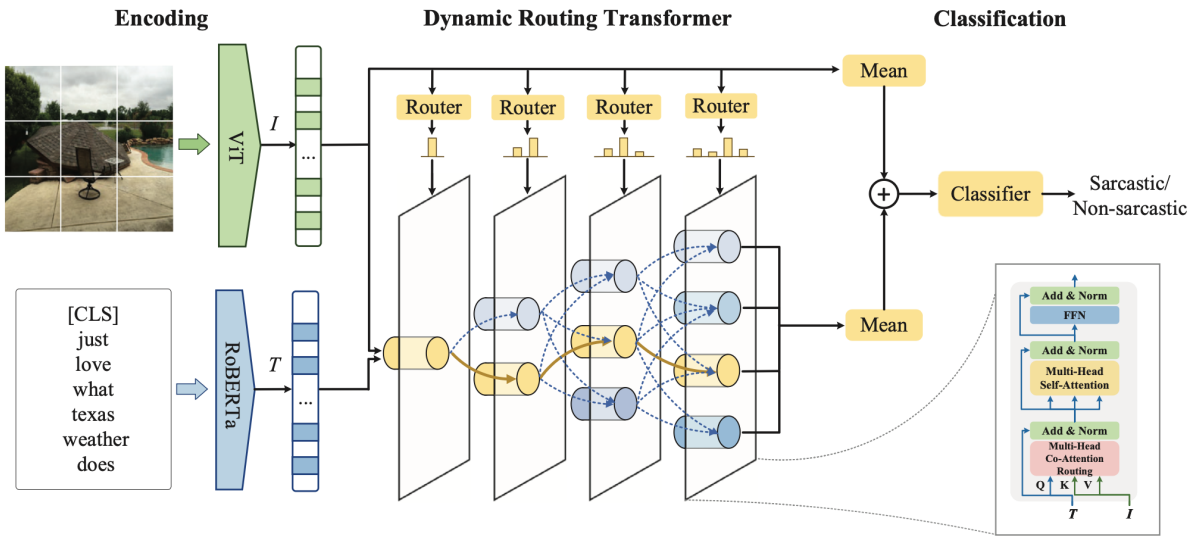


Figure 1: Architecture of the DynRT-Net for multimodal sarcasm detection from (Tian et al., 2023)

Building on this foundation, several methods have emerged that further enrich multimodal transfo rmer-based sarcasm detection. (Gupta, Shah, Shah, Syiemlieh, and Maurya, 2021) introduced a model combining co-attention and Feature-wise Linear Modulation (FiLM), enabling dynamic modulation of modality-specific features to emphasize sarcasm-related signals. (X. Zhang et al., 2021) leveraged contrastive attention to explicitly model inter-modal incongruities—for example, when a cheerful image is paired with a scathing caption. These methods underscore the role of attention-based fusion in detecting nuanced sarcastic intent across input streams.

(Pramanick et al., 2022) proposed MuLOT, a resource-efficient model that applies optimal transport theory to align and fuse modalities with minimal computational cost. This approach is particularly effective when computational budgets are constrained, yet modality alignment is essential. (Pandey, Aggarwal, and Vishwakarma, 2024) emphasized modeling visual semantics to capture semantic incongruity, highlighting how sarcasm often arises from contradiction between visual content and accompanying text.

A central innovation enabling this progress is the attention mechanism, particularly cross-attention. These mechanisms build upon self-attention, which allows the model to compute the relevance of each element in a sequence relative to others by assigning attention scores, thereby capturing dependencies and context within a single modality. Cross-attention extends this mechanism to interactions across different modalities. In a multimodal context, cross-attention enables the model to learn which modalities—such as text, audio, or visual inputs—provide the most informative signals at each step. For example, when analyzing a social media video, the model can prioritize intonational cues from speech or expressive facial gestures, depending on which more strongly indicates sarcasm in a given context. This dynamic alignment across modalities facilitates the detection of subtle sarcastic cues that would otherwise be missed in unimodal approaches.

(Pramanick et al., 2022) discuss how cross-attention allows models to shift focus between modalities in real time, strengthening multimodal representation learning. (Gao et al., 2024) also emphasize the benefits of cross-attention in sarcasm detection, especially in aligning incongruent signals across modalities. (Tsai et al., 2019; Tian et al., 2023; Yoon et al., 2018) all demonstrate how various cross-modal attention techniques enhance sarcasm-related tasks by leveraging modality-specific relevance.

Summary: This subsection reviewed the rise of multimodal transformers in sarcasm detection, emphasizing their capacity to model cross-modal dependencies through attention mechanisms. A variety of architectures, including DynRT-Net, MuLOT, and FiLM-enhanced co-attention models, demonstrate the power of fusing visual, textual, and auditory cues. These innovations collectively affirm that cross-attention is central to capturing sarcasm's contextually rich and subtle nature (Hasan et al., 2021).

2.2 Sarcasm Detection in Social Media Videos

Social media platforms are prime venues for sarcasm due to their informal, expressive, and user-driven nature. The combination of spontaneous speech, short-form text captions, and visual reac-

tions makes these platforms rich in multimodal content. Detecting sarcasm in such videos presents notable challenges because sarcastic intent can be conveyed not only through the literal meaning of words but also through tone of voice, facial expressions, gestures, and contextual visual cues. Earlier research focused mainly on text-based sarcasm detection, often struggling to capture the full spectrum of sarcastic expression. With the pervasive rise of video content across platforms like TikTok, YouTube, and Instagram, it has become increasingly evident that ignoring non-textual features—such as prosody in speech or eye-rolls and smirks in visuals—leads to significant performance limitations (Bhosale et al., n.d.; Castro et al., 2019). The inherent ambiguity of sarcasm, where literal meaning often contradicts intended meaning, is amplified in multimodal contexts, requiring models to process subtle cues across modalities (Aguert, 2022).

Multimodal models have been specifically proposed to address these limitations by integrating text, audio, and visual data to better capture the layered nature of sarcastic expression. These approaches leverage deep learning techniques to learn joint representations across modalities, significantly improving sarcasm detection accuracy compared to unimodal methods. For instance, early works recognized the importance of combining modalities, with studies demonstrating that models incorporating acoustic and textual features outperform those relying solely on text (Yoon et al., 2018). More advanced techniques now focus on intricate fusion mechanisms. Researchers have explored various fusion strategies, including early fusion (concatenating features before input to the model), late fusion (combining predictions from modality-specific models), and hybrid approaches that dynamically weight modalities based on context (Farabi et al., 2024; Karun and Adithya, 2025). The goal is to build models that can identify the incongruity often present in sarcastic expressions, such as a positive sentiment in text paired with a negative facial expression or an unusual tone of voice.

However, significant challenges persist in detecting sarcasm in social media videos. One major hurdle is the difficulty of temporal synchronization and alignment between modalities, as speech, gestures, and text may not always align perfectly in time. Another challenge is the complexity of interpreting nuanced sarcastic cues in diverse social media contexts, which can vary across cultures and individual communication styles (Qin et al., 2024). Data scarcity is also a critical issue; while text-based sarcasm datasets are relatively abundant, high-quality, large-scale multimodal datasets specifically annotated for sarcasm in video content are still limited. This lack of diverse and extensive multimodal data can hinder the generalization capabilities of models (Ray et al., n.d.). Moreover, dealing with code-mixed conversations, which are common in global social media, introduces additional linguistic complexities that unimodal or even basic multimodal models may struggle to process effectively (Bedi, Kumar, Akhtar, and Chakraborty, 2021). Furthermore, handling noise, varying video quality, and spontaneous, unscripted human behavior in social media videos adds another layer of complexity.

Despite these challenges, the ability to accurately detect sarcasm in social media videos holds significant practical value. It can profoundly enhance downstream applications such as sentiment analysis, making it more robust and accurate by correctly identifying cases where apparent positive or negative sentiment is actually ironic. It is also crucial for effective content moderation on social media platforms, helping to filter out toxic or misleading content that uses sarcasm as a ve-

hicle. Moreover, it contributes to the development of more sophisticated and socially aware AI systems, enabling virtual assistants, chatbots, and other conversational agents to better understand human communication nuances, leading to safer, more empathetic, and contextually aware digital communication environments (Valliyammai et al., n.d.; Yaghoobian, Arabnia, and Rasheed, 2023). Recent advancements continue to explore novel architectures and fusion techniques to overcome these challenges, pushing the boundaries of multimodal sarcasm detection in dynamic social media environments (Wang et al., 2024).

Summary: This subsection underscores the importance of multimodal approaches for detecting sarcasm in social media videos, given that sarcastic intent often manifests across textual, audio, and visual cues. It highlights how integrating these modalities through deep learning models improves accuracy by capturing incongruities. While advancements have been made in fusion techniques and joint representations, challenges remain regarding temporal alignment, contextual interpretation, data scarcity, and handling code-mixed content. Nevertheless, accurate sarcasm detection in this domain offers significant practical benefits for sentiment analysis, content moderation, and developing more socially aware AI systems.

2.3 Large Language Models (LLMs) and Sarcasm Detection

While multimodal transformers have shown promise in sarcasm detection, Large Language Models (LLMs) such as GPT-3, BERT, and similar architectures, which are also based on the transformer framework, have become central to many natural language processing tasks due to their exceptional ability to model linguistic patterns and generate human-like text. However, despite their remarkable success in general language understanding and generation, traditional text-based LLMs are inherently limited in their ability to detect sarcasm when non-verbal cues are essential. Sarcasm often relies on a complex interplay of contextual signals beyond the literal text, including tone of voice, facial expression, gestures, and body language, which are simply not accessible to purely text-based models. Their performance on sarcasm can be inconsistent, as the task often requires a deep understanding of implicit meaning, cultural nuances, and shared social references that may not be fully captured in their text-only training data (Global Radiance Review, 2025). Research suggests that sarcasm judgment is often an intuitive, holistic cognitive process, not strictly a step-by-step logical reasoning task, which can pose challenges for how LLMs typically process information sequentially (Dong, Gao, Tang, Yin, and Guo, 2024).

Although recent developments in visual-language models (VLMs) and multimodal LLMs (MLLMs), such as Flamingo, GPT-4 with vision capabilities, and GPT-4o, demonstrate a rapidly growing capacity to process and reason across modalities, standard LLMs remain less equipped for sarcasm detection in video-based content where speech and visual cues are critical. These advanced MLLMs are beginning to bridge the gap by integrating visual and auditory data directly into their architectures, allowing them to leverage the full context of a multimodal interaction (Tang, Lin, Yan, and Li, 2024). For instance, a novel generative multimodal sarcasm model leverages LLMs with visual instruction and demonstration retrieval to enhance detection accuracy, particularly for image-text pairs (Tang et al., 2024). Another innovative framework, Commander-GPT, decomposes sarcasm detection into sub-tasks, assigning specialized multimodal LLMs to each, demonstrating significant

performance improvements by fully integrating various data types like text, images, and potentially audio (Y. Zhang, Zou, Wang, and Qin, Y. Zhang, Zou, et al.). Furthermore, in-context learning frameworks like IRONIC are enabling MLLMs to achieve state-of-the-art performance on zero-shot multimodal sarcasm detection by analyzing referential, analogical, and pragmatic image-text linkages (Anantha Ramakrishnan, 2025).

Comparing the performance of text-only LLMs with dedicated multimodal transformers or MLL Ms is essential to determine whether integrating multiple modalities yields measurable improvements in sarcasm recognition. While some studies explore adapting LLMs through specific prompting techniques to improve their textual sarcasm detection, such as Chain-of-Contradiction or Graph-of-Cues (Dong et al., Dong et al.), these still operate within the textual domain. Multimodal architectures, on the other hand, are designed to simultaneously process and align textual, auditory, and visual information from the ground up, making them more suitable for nuanced tasks like sarcasm detection in social media videos, where meaning often emerges from the interaction of several expressive channels (Yaghoobian et al., 2023). The continued development of specialized multimodal LLM frameworks highlights the recognition that while powerful, general-purpose LLMs alone are insufficient for robust sarcasm detection in rich, real-world media, and require architectural adaptations or specialized prompting to handle the complexity of cross-modal incongruity.

Summary: This subsection discusses the role of Large Language Models (LLMs) in sarcasm detection, highlighting their strengths in linguistic patterns but inherent limitations with sarcasm due to its reliance on non-verbal and contextual cues. While traditional text-based LLMs struggle with the nuanced and often implicit nature of sarcasm, the emergence of multimodal LLMs (MLLMs) is addressing this gap by integrating visual and auditory data. Recent research demonstrates MLLMs' enhanced capabilities through generative models, task decomposition frameworks, and in-context learning, which allow them to process complex cross-modal incongruities. The section emphasizes that integrating multiple modalities is crucial for measurable improvements in sarcasm recognition, distinguishing MLLMs from text-only LLMs for robust detection in social media videos.

2.4 Challenges and Future Directions in Multimodal Sarcasm Detection

Despite the considerable progress in multimodal sarcasm detection, there remain significant challenges in the field that necessitate ongoing research and innovative solutions. First, the inherent heterogeneity and high variability of social media videos pose a formidable hurdle. These videos often feature diverse recording conditions, including inconsistent lighting, fluctuating background noise levels, and various camera angles, all of which can significantly affect the quality and interpretability of visual and audio cues. Furthermore, sarcasm is deeply embedded in cultural nuances and individual communication styles, meaning that effective sarcasm detection models must adapt to distinct language patterns, subtle gestural differences, and socio-cultural contexts across different regions and online communities (Farabi et al., 2024). While foundational approaches leveraging word embeddings and traditional language models have been instrumental in capturing linguistic patterns, their inherent limitations become pronounced when dealing with the complex, often non-literal and context-dependent nature of sarcasm.
These models frequently struggle to fully grasp the subtle pragmatic shifts and implicit meanings

that are crucial for accurate sarcasm interpretation, particularly in multimodal scenarios where verbal cues are only part of the message (Kumar and Sarin, 2020).

Additionally, the availability and integration of large-scale, high-quality multimodal datasets remain crucial for training robust and generalizable models. Many datasets currently used in sarcasm detection predominantly focus on textual data, or they provide limited, often imperfect, visual and audio annotations, hindering comprehensive multimodal learning. The process of manually annotating multimodal content for sarcasm is not only arduous but also costly, demanding expert knowledge of both linguistic and non-verbal cues. This significantly limits the scale and diversity of publicly available data, which in turn hampers the development of models that can truly generalize to the dynamic and unpredictable nature of real-world social media content. Future research should critically examine current datasets for potential biases and address the challenge of data imbalance, where certain types of sarcastic expressions are underrepresented. Focus should also be placed on refining the fusion techniques for combining multimodal data more effectively, aiming to reduce computational costs while enhancing model generalization to novel, unseen data. A comprehensive survey of the fundamental theories, various formulations, available datasets, and diverse detection methods provides crucial insights into the current state of automatic sarcasm detection and highlights significant opportunities for future advancements, particularly in bridging the gap between unimodal and truly multimodal understanding (Chen, Lin, Li, and Liu, 2023).

Beyond these immediate challenges, several promising avenues for future research exist. Enhancing the interpretability and explainability of multimodal sarcasm detection models is paramount. Understanding why a model predicts sarcasm based on specific textual, visual, or auditory cues would not only build user trust but also facilitate debugging and continuous refinement of these complex systems. Developing models that are robust to noise, capable of handling missing modalities, and efficient enough for real-time processing in live streams or rapidly evolving social media conversations represents a critical frontier for practical deployment. Furthermore, a personalized approach to sarcasm detection, where models can adapt to individual users' unique communication styles and their historical interaction patterns, could significantly improve accuracy and user experience. Finally, incorporating common-sense reasoning and a deeper understanding of pragmatic knowledge, possibly through hybrid approaches combining symbolic reasoning with advanced deep learning, could empower models to grasp the more profound contextual and cognitive processes underlying sarcastic expressions, moving beyond mere pattern recognition to a more human-like understanding.

Summary: This subsection outlines key challenges and future directions in multimodal sarcasm detection. Major challenges include the heterogeneity of social media videos (varying conditions, cultural nuances), and the limitations of traditional language models in capturing complex sarcasm. It also stresses the critical need for larger, high-quality multimodal datasets due to annotation difficulties and existing data limitations. Future research should prioritize refining multimodal fusion techniques, improving model interpretability, ensuring robustness, enabling real-time detection, developing personalized models, and integrating common-sense reasoning to achieve more sophisticated and contextually aware sarcasm detection.

Table 1: List of references for subsections 2.1-2.4, summarized

Reference	Brief Description	Subsection
Aguert (2022)	Critical review of prosody's role in verbal irony comprehension.	2.2
Bhosale et al. (n.d.)	Benchmarking and expansion for multimodal sarcasm detection.	2.2
Castro et al. (2019)	Pioneering work on multimodal sarcasm detection.	2.2
Farabi et al. (2024)	Comprehensive survey of multimodal sarcasm detection.	2.1, 2.2, 2.4
Gao et al. (2024)	Review of multimodal approaches to sarcasm detection in social media.	2.1
Gupta et al. (2021)	Model combining co-attention and FiLM for multimodal sarcasm detection.	2.1
Karun and Adithya (2025)	Applying cross-modal feature alignment and fusion for effective sarcasm detection.	2.2
Pramanick et al. (2022)	Discussion on cross-attention for real-time multimodal representation learning.	2.1
Qin et al. (2024)	Innovative CGL-MHA model for sarcasm sentiment recognition.	2.2
Ray et al. (n.d.)	Multimodal corpus for emotion recognition in sarcasm.	2.2
Tian et al. (2023)	Dynamic Routing Transformer Network (DynRT-Net) for multimodal sarcasm detection.	2.1
Tsai et al. (2019)	Multimodal Transformer for unaligned multimodal language sequences.	2.1
Valliyammai et al. (n.d.)	Cyberbullying detection with multimodal data using transfer learning.	2.2
Wang et al. (2024)	Relational Context Learning and Multiplex Fusion Network (RCLMuFN) for multimodal sarcasm detection.	2.2
Yaghoobian et al. (2023)	Advancements in multimodal sentiment analysis techniques and applications.	2.2, 2.3
Yoon et al. (2018)	Multimodal speech emotion recognition using audio and text.	2.1, 2.2
X. Zhang et al. (2021)	Leveraging contrastive attention for inter-modal incongruities in sarcasm detection.	2.1
Bedi et al. (2021)	Multimodal sarcasm and humor classification in code-mixed conversations.	2.2
Anantha Ramakrishnan (2025)	IRONIC: Coherence-Aware Reasoning Chains for Multi-Modal Sarcasm Detection.	2.3

Table 1: List of references for subsections 2.1-2.4, summarized

Reference	Brief Description	Subsection
Dong et al. (2024)	Investigating if sarcasm detection is a step-by-step reasoning process in LLMs.	2.3
Global Radiance Review (2025)	Discussion on whether Large Language Models can detect sarcasm.	2.3
Tang et al. (2024)	Leveraging generative LLMs with visual instruction for multimodal sarcasm detection.	2.3
Y. Zhang, Zou, et al. (2025)	Commander-GPT framework for unleashing multimodal sarcasm detection capabilities of MLLMs.	2.3
Kumar and Sarin (2020)	Word embedding and language model based sarcasm detection (WELMSD).	2.4
Chen et al. (2023)	Survey of automatic sarcasm detection: theories, datasets, methods, and opportunities.	2.4

3 Methodology

This section details the systematic approach undertaken to develop and evaluate a multimodal sarcasm detection model. Recognizing that sarcasm often relies on a complex interplay of verbal and nonverbal cues, this methodology focuses on integrating diverse data streams to enhance detection accuracy beyond what single modalities can achieve. Here, we outline the robust framework guiding our research, from data preparation to model validation.

Specifically, this section begins by describing the dataset selection and preprocessing techniques applied to prepare the multimodal data for analysis, emphasizing the critical steps for extracting relevant features from text, audio, and visual components. We then delve into the innovative model design and architecture, highlighting how a multimodal transformer, coupled with sophisticated cross-attention mechanisms, is employed to effectively fuse these distinct data streams. Following this, the comprehensive training and evaluation protocols are described, detailing the metrics used to rigorously assess the model's performance and compare it against established benchmarks. Finally, a clear timeline provides a structured overview of the project's phases, ensuring transparency and feasibility in its execution.

GitHub Repository: The code and implementation details for this project are publicly available at: `https://github.com/m-h-narang/MSc-Voice-Technology-Thesis`.

3.1 Dataset and Preprocessing

MUStARD++ is utilized in this study to evaluate the effectiveness of the proposed model. The dataset contains short video clips labeled for sarcasm, each including aligned text, audio, and visual components that enable multimodal sarcasm detection. It is a publicly available dataset intended for academic research and can be accessed freely at the URL listed in Table 2.

- **Text Data:** Textual information will be sourced directly from the transcripts provided in the Mustard++ dataset, which includes high-quality, manually annotated subtitles aligned with spoken content. This eliminates the need for additional speech-to-text processing and ensures accurate text input for sarcasm detection.

- **Audio Data:** Audio files will be processed using Mustard++ to extract features such as pitch, intonation, and speech patterns that are critical for sarcasm detection. Tools like Librosa or Mel-Frequency Cepstral Coefficients (MFCC) will be used to capture these features.

- **Visual Data:** Facial features and expressions will be extracted from video frames using OpenFace or Mustard++'s built-in facial analysis tool. This will allow for the detection of facial gestures, such as sarcasm-related cues (e.g., raised eyebrows, exaggerated smiles).

Table 2: MUStARD++ Dataset Summary

Property	**Description**
Total video duration	Approximately 1.4 hours of conversational clips
Number of utterances	1202 annotated instances (balanced between sarcastic and non-sarcastic)
Average utterance length	4.19 seconds
Number of speakers	Not explicitly specified, but from diverse sitcom casts
Modalities included	Text, Audio, Visual
Source shows	*Friends*, *The Big Bang Theory*, etc.
Labeling scheme	Binary classification: sarcastic vs. non-sarcastic
Dataset access	`https://github.com/cfiltnlp/MUStARD_Plus_Plus`

3.2 Dataset Availability

The MUStARD++ dataset used in this study is publicly available and openly licensed for academic research purposes. It can be freely accessed via its official GitHub repository[2].

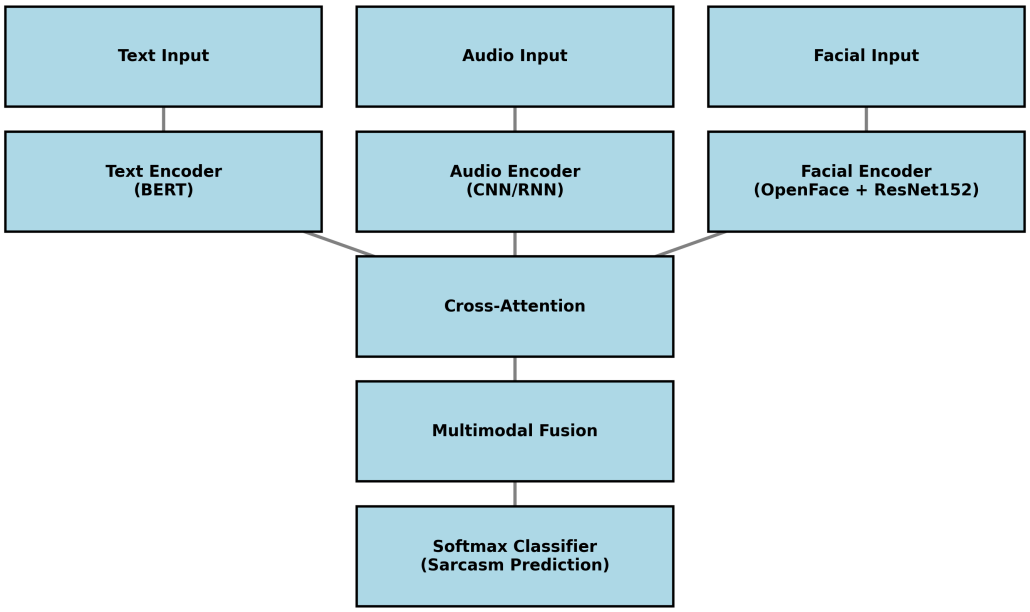
The dataset includes aligned text, and clips, along with sarcasm annotations. No user-identifiable or sensitive data is included. Its public release promotes reproducibility, comparability across models, and responsible research practices in multimodal sarcasm detection.

3.3 Model Design and Architecture

To improve sarcasm detection, a multimodal transformer architecture will be employed, using cross-attention mechanisms to integrate text, audio, and facial expression data. The methodology will follow these steps:

[2] `https://github.com/cfiltnlp/MUStARD_Plus_Plus`

Figure 2: An overview of the model architecture



- **Separate Modality Encoders:** Three separate encoders will be built for text, audio, and facial expressions.

  - **Text Encoder:** A pre-trained BERT model will be used for the textual input. This encoder will capture the contextual meaning of the text and detect sarcasm-specific language.

  - **Audio Encoder:** The MFCC features extracted from the audio will be passed through an RNN or CNN model designed to learn intonation and pitch variations associated with sarcasm.

  - **Facial Expression Encoder:** A pre-trained OpenFace model will be used to extract facial action units and other relevant facial gesture features. These extracted features will then be processed using a ResNet152 backbone, in alignment with Mustard++'s existing facial encoder design, which has demonstrated strong performance in capturing sarcasm-relevant expressions. The use of ResNet152 ensures robust feature extraction from visual data, avoiding the need to train a custom CNN from scratch.

- **Cross-Attention Mechanism:** A cross-attention layer will allow the model to dynamically learn the relationships between the different modalities, focusing on the most important features for sarcasm detection. This will be crucial to overcoming the challenges of integrating text, audio, and facial expressions.

- **Multimodal Fusion:** After processing the modalities individually, the features will be fused through a multimodal transformer. This model will combine the outputs of the separate encoders and pass them through a softmax classifier to predict whether the video content contains sarcasm.

3.4 Training and Evaluation

Training will follow these steps:

- **Dataset Preparation:** The MUStARD++ dataset will be used as the primary training source, consisting of videos annotated as sarcastic or non-sarcastic. Baseline models, including SVM and pre-trained LLMs (e.g., BERT), will be implemented to establish performance comparisons.

- **Model Training:** The model will be trained using supervised learning for binary classification (sarcastic or not sarcastic). The architecture will be optimized using the Adam optimizer with a binary cross-entropy loss function. Additional training parameters will include ReLU activation, a learning rate of 1e-4, and dropout regularization to prevent overfitting.

- **Evaluation:** Performance will be assessed using metrics such as accuracy, precision, recall, and F1 score. Evaluation will compare the proposed multimodal model with baseline approaches to determine performance improvements.

3.5 Timeline

- Month 1
  - **Data Collection:** Gather and preprocess sarcasm-labeled video datasets using Mustard++, speech-to-text algorithms, and facial expression analysis tools such as OpenFace.
  - **Model Design:** Begin implementing and testing separate modality encoders (text, audio, facial expression).
  - **Preliminary Testing:** Run initial tests on the audio and facial encoders with available datasets to ensure proper feature extraction.
- Month 2
  - **Cross-Attention Integration:** Implement the cross-attention mechanism and integrate the modality encoders.

  - **Model Training:** Begin training the multimodal transformer architecture with the labeled sarcasm dataset.

  - **Early Evaluation:** Evaluate initial results against baseline LLM models, focusing on identifying key performance differences.

- Month 3

  - **Final Model Refinement:** Optimize the model based on evaluation results, adjusting hyperparameters and fine-tuning for better performance.

  - **Final Evaluation and Comparison:** Perform comprehensive testing, comparing the multimodal model with existing LLM approaches.

  - **Thesis Writing:** Complete the writing of the thesis, including methodology, results, analysis, and conclusion. Prepare for the final submission.

3.6 Risk Mitigation

3.6.1 Data Scarcity

The MUStARD++ dataset is limited in size. To mitigate overfitting and enhance model generalizability, regularization techniques (e.g., dropout layers), data augmentation (e.g., adding noise to audio or shifting facial landmarks), and transfer learning via pre-trained models (e.g., BERT, ResNet) are employed.

3.6.2 Multimodal Alignment

Temporal misalignment between modalities can impair learning. To mitigate this, timestamp alignment provided in MUStARD++ will be verified, and synchronization techniques such as dynamic time warping (DTW) will be employed when necessary.

3.6.3 Computational Resources

Training multimodal transformers is computationally expensive. Resource usage is managed by using cloud-based GPUs when available, batching data efficiently, and limiting the depth of certain encoder layers where performance is not affected.

3.6.4 Overfitting

With relatively few training samples, overfitting is a risk. Aside from data augmentation, early stopping and cross-validation will be used to monitor model generalization.

3.6.5 Bias and Generalization

Since MUStARD++ consists of scripted sitcom dialogues, the model may not generalize well to spontaneous real-world conversations. To address this, results will be interpreted cautiously, and limitations will be clearly stated in the discussion chapter.

3.7 Ethical Issues

3.7.1 Privacy

No personally identifiable information is used or collected. The MUStARD++ dataset is fully anonymized and publicly released under research-friendly terms.

3.7.2 Bias and Stereotyping

Since the dataset is derived from Western sitcoms, it may encode cultural or linguistic biases. The model may inadvertently learn biases related to tone, facial expressions, or dialects. This limitation will be acknowledged in the analysis, and future work will be suggested to include more diverse datasets.

3.7.3 Misuse Potential

While the system is designed for academic purposes (e.g., analyzing communication), automated sarcasm detection could be misapplied in surveillance or manipulative media monitoring. The thesis will explicitly state its intended use and discourage misuse through clear documentation and licensing.

3.7.4 Transparency

All methods and results will be openly documented in the thesis and GitHub repository to ensure replicability, promote peer review, and support responsible AI practices.

4 Experimental Setup

This chapter outlines the experimental setup used to conduct the research and evaluate the proposed models. It begins with an overview of the tools, frameworks, and computational resources employed throughout development and experimentation. Subsequently, it describes the configuration of experiments, including preprocessing routines, model architectures, and hyperparameter tuning strategies. Finally, the chapter presents a detailed account of training performance across different unimodal and multimodal models, offering insights into their learning behavior and convergence patterns. This comprehensive setup ensures methodological consistency and forms the basis for the evaluations and analyses presented in the next chapter.

4.1 Tools and Technologies

All experiments were conducted on the University of Groningen's high-performance computing cluster, Hábrók, equipped with NVIDIA Tesla V100 GPUs. This HPC environment provided the necessary computational resources for efficient training of deep neural networks involved in multimodal sarcasm detection. The MUStARD++ dataset was utilized as the primary data source, containing multimodal video utterances labeled as sarcastic or non-sarcastic, with aligned text, audio, and visual features.

Textual features were derived from manual transcripts of MUStARD++ using the BERT-base-uncased pre-trained transformer to extract rich contextualized embeddings. Audio features were extracted using Mel-Frequency Cepstral Coefficients (MFCCs) and processed through convolutional and recurrent neural networks to capture prosodic elements such as pitch and intonation relevant to sarcasm. Visual features were obtained using the OpenFace toolkit, extracting frame-level facial Action Units (AUs), gaze, and head pose information. These features were aggregated via mean pooling and encoded using a ResNet-152 backbone or simple MLP classifiers to capture subtle nonverbal cues like eyebrow raises and smiles.

This suite of state-of-the-art tools and carefully designed encoders enabled comprehensive and robust multimodal feature extraction, forming the foundation for downstream sarcasm classification tasks.

Table 3: Summary of Tools, Technologies, Dataset, and Hardware

Category	**Details**
Computing System	University of Groningen Hábrók HPC Cluster
GPU	NVIDIA Tesla V100
Dataset	MUStARD++ multimodal sarcasm detection dataset
Text Processing Model	BERT-base-uncased pre-trained transformer
Audio Feature Extraction	Mel-Frequency Cepstral Coefficients (MFCC)
Audio Encoder	CNN and RNN architectures capturing pitch and intonation features
Visual Feature Extraction	OpenFace v2.2.0 extracting facial Action Units, gaze, and head pose
Visual Encoder	ResNet-152 backbone and MLP classifiers for expression encoding

4.2 Experimental Configurations and Hyperparameter Optimization

The experiments in this project are organized into two major phases. The first phase involves unimodal experiments designed to establish baselines and extract modality-specific embeddings. The second phase focuses on multimodal fusion, evaluating both naive and attention-based fusion architectures. This design allows for systematic comparison and component-level analysis of each modality and integration strategy.

1. **Text Modality Experiments**

   (a) **BERT Classifier Baseline:** A fine-tuned `bert-base-uncased` model trained directly on sarcasm-labeled text samples. The input text is tokenized and padded to a maximum length of 64 tokens using the HuggingFace tokenizer. The model uses the pooled CLS embedding to drive a two-layer feedforward classification head.

Table 4: BERT Classifier Baseline: Configuration and Training Details

Category	**Details**
Model Architecture	BERT-base-uncased pretrained transformer
Embedding Source	CLS token output (`pooler_output`)
Input Sequence Length	64 tokens (with padding and truncation)
Classifier Head	Linear (768→256) → ReLU → Dropout(0.4) → Linear (256→2)
Loss Function	Cross-entropy with class weights
Class Weights	Computed via `compute_class_weight` from scikit-learn
Optimizer	AdamW with differential learning rates
Learning Rates	5e−6 (BERT layers 1–6), 1e−5 (layers 7–12), 2e−5 (classifier head)
Scheduler	Linear decay (no warmup)
Batch Size	16
Training Epochs	6
Early Stopping	Enabled, patience = 2 (based on macro F1)
Evaluation Metrics	Macro F1-score, classification report
Model Saving	Best model checkpointed on validation performance

   (b) **SVM Baseline:** A lightweight traditional classifier using lexical features. The raw text is cleaned and vectorized using unigram and bigram TF-IDF features, with a vocabulary limited to the top 5000 terms. These features are used to train a linear Support Vector Machine (SVM), serving as a non-neural benchmark for sarcasm detection.

Table 5: SVM Baseline: Configuration and Training Details

Category	**Details**
Model Architecture	Linear Support Vector Machine (`LinearSVC`)
Text Cleaning	Lowercasing, removal of URLs and non-alphanumeric characters
Feature Representation	TF-IDF with unigrams and bigrams
Maximum Features	5000 most frequent tokens
Vectorization Library	`TfidfVectorizer` from `scikit-learn`
Train/Test Split	80/20 split (random seed = 42)
Target Labels	Binary: Sarcastic vs. Not Sarcastic
Loss Function	Hinge loss (default in `LinearSVC`)
Optimization Objective	Maximize margin between two classes
Evaluation Metrics	Accuracy, Precision, Recall, F1-score

(c) **Embedding Extraction:** To enable downstream multimodal fusion, we extract contextualized text embeddings using a pretrained BERT model. Specifically, the [CLS] token embeddings are obtained from the final hidden layer for each utterance. These embeddings are saved and later used as the text modality input for multimodal architectures.

Table 6: Text Embedding Extraction: Configuration and Processing Details

Category	**Details**
Pretrained Model	`bert-base-uncased` (from HuggingFace Transformers)
Embedding Type	Final hidden state of [CLS] token
Tokenization	`AutoTokenizer` with truncation and padding
Filtering Strategy	Utterances ending in `_u` with non-empty sentences
Batching	Sequential processing (one sentence at a time)
Output Format	Torch tensor: `text_features.pt`
Saved Metadata	Text features, sarcasm labels, and utterance keys
Embedding Dimension	768 (BERT base hidden size)
Device	CUDA (if available), else CPU
Downstream Use	Input for multimodal fusion models

2. **Audio Modality Experiments**

(a) **Audio Baseline:** Each utterance video is first processed to extract its audio waveform using MoviePy. The waveform is then passed through a pretrained Wav2Vec 2.0 encoder to generate dense frame-level acoustic embeddings. These features are used as input to an RNN-based classifier with an attention mechanism to predict sarcasm.

Table 7: Audio Baseline: Wav2Vec2 Feature Extraction and Classifier Configuration

Category	**Details**
Audio Extraction Tool	`moviepy` (converted to 16kHz WAV)
Pretrained Model	`torchaudio.pipelines.WAV2VEC2_BASE`
Feature Dimension	768 (per frame from Wav2Vec2 encoder)
Feature Truncation	Max sequence length = 300 frames
Padding Strategy	Zero-padding for shorter sequences
Classifier Model	Bidirectional GRU (2-layer) with attention
Classifier Head	GRU $\rightarrow$ LayerNorm $\rightarrow$ Attention $\rightarrow$ Dropout(0.5) $\rightarrow$ Linear(512$\rightarrow$2)
Loss Function	Cross-entropy with class weights
Class Weights	Computed via `compute_class_weight`
Optimizer	Adam
Learning Rate	1e$-$4
Scheduler	ReduceLROnPlateau (mode=max, patience=3, factor=0.5)
Batch Size	16 (standard), 32 (combined training variant)
Training Epochs	25 (or 10 for combined variant)
Evaluation Metrics	Accuracy, macro F1-score, classification report
Model Saving	Best checkpoint based on validation F1

(b) **Embedding Extraction:** Audio utterances are passed through a pretrained Wav2Vec2 encoder. The output hidden states are aggregated via mean pooling to obtain fixed-length embeddings representing each utterance. These are later used as audio features in downstream fusion or classification pipelines.

Table 8: Utterance-Level Audio Embedding Extraction (Wav2Vec2 + Mean Pooling)

Category	**Details**
Input Format	Mono 16kHz WAV files
Preprocessing Tools	`torchaudio` (resample, stereo-to-mono)
Pretrained Model	`facebook/wav2vec2-base-960h` (via HuggingFace Transformers)
Processor	`Wav2Vec2Processor` for input preparation
Embedding Dimension	768 (per token)
Utterance Embedding Strategy	Mean-pooling over time dimension ($\frac{1}{T}\sum_t h_t$)
Output Format	Torch tensor: `[num_samples, 768]`
Missing Audio Handling	Replaced with zero vector embedding
Device	GPU-accelerated embedding extraction
Saving Format	Serialized via `torch.save(...)`

3. **Visual Modality Experiments**

(a) **Video Baseline:** Facial features are extracted from video frames using the OpenFace toolkit. Each utterance is represented by a time-aggregated pseudo-image or tabular feature vector. Two types of models are explored: (1) a CNN-based classifier using ResNet-152 trained on reshaped OpenFace features treated as low-resolution RGB-like images; and (2) a simple MLP trained directly on raw OpenFace feature vectors. Both models are trained to predict sarcasm labels in a supervised fashion.

Table 9: Video Baseline: Facial Feature Classification

Category	**Details**
Feature Source	OpenFace (AU, gaze, head pose, etc.)
CNN Input Format	Reshaped OpenFace features as 3×8×8 pseudo-image
CNN Architecture	ResNet-152 (pretrained on ImageNet, finetuned)
CNN Classifier Head	Dropout + Linear(2048 $\rightarrow$ 2)
MLP Input Format	Raw flattened OpenFace features per utterance
MLP Architecture	512-hidden-layer MLP, penultimate features extracted for downstream use
Loss Function	CrossEntropyLoss
Optimizer	Adam (LR = 1e-4)
Train/Validation Split	Stratified 80/20 split using sklearn
Batch Size	32
Epochs	10
Feature Saving Format	Torch tensor (`.pt`), shape: [N, 512] for MLP features

(b) **Embedding Extraction:** Facial expression embeddings are generated from each video utterance using OpenFace. Frame-level features such as facial Action Units (AUs), head pose, and gaze vectors are averaged across all valid frames to produce a single feature vector per video. These vectors are then used as inputs to a ResNet-152-based CNN or a simple MLP classifier for downstream sarcasm prediction.

Table 10: Video Embedding Extraction Pipeline

Stage	**Details**
Face Tracking Tool	OpenFace v2.2.0
Input Format	MP4 utterance-level video clips
Output Format	Per-frame CSV files with 700+ facial features
Selected Features	AU intensities (e.g., `AU06_r`), gaze, and head pose
Aggregation Strategy	Mean pooling across valid frames
Missing or Corrupt Frames	Skipped; fallback = zero vector if file is empty
Final Embedding Format	PyTorch tensor (shape: [N, D]), saved as `.pt`
Fallback for Missing Files	Excluded or replaced with zero-filled vectors
Usage	Used as input to CNN (ResNet152) or MLP models for classification

4. **Multimodal Fusion Experiments**

   (a) **CrossModalFusionModel:** A neural network that performs early fusion by concatenating modality-specific embeddings from text, audio, and video. Each modality is first projected to a common hidden dimension using fully connected layers with ReLU activations, LayerNorm, and dropout. The concatenated vector is passed through another fully connected layer before classification. This architecture enables the model to jointly reason over multiple modalities and learn cross-modal interactions in a simple yet effective manner.

Table 11: CrossModalFusionModel Architecture

Component	**Details**
Input Modalities	Text (BERT), Audio (wav2vec 2.0), Video (OpenFace features)
Text Input Dimension	768
Audio Input Dimension	768
Video Input Dimension	23
Projection Layers	Linear $\rightarrow$ ReLU $\rightarrow$ LayerNorm $\rightarrow$ Dropout
Hidden Dimension	256
Fusion Strategy	Concatenation of projected text, audio, and video embeddings
Fusion Layer	Linear $\rightarrow$ ReLU $\rightarrow$ LayerNorm $\rightarrow$ Dropout
Classifier	Linear $\rightarrow$ Softmax over 2 classes (sarcastic vs. non-sarcastic)
Loss Function	CrossEntropyLoss
Optimizer	Adam (lr=$1e^{-4}$, weight decay=$1e^{-5}$)
Regularization	Dropout (p=0.3) + Early Stopping (patience=3 epochs)
Evaluation Metrics	Accuracy, Weighted F1, Confusion Matrix

(b) **CrossModalAttentionFusionModel:** A cross-attention-based fusion architecture that dynamically attends across modalities. This model enables the text representation to attend to both audio and video signals using multi-head attention, allowing it to learn intermodal dependencies and enhance contextual understanding. Each modality is first projected to a common hidden space, followed by cross-attention and nonlinear fusion layers. The final representation is passed to a classification layer.

Table 12: CrossModalAttentionFusionModel Architecture

Component	**Details**
Input Modalities	Text (BERT), Audio (wav2vec 2.0), Video (OpenFace)
Text Input Dimension	768
Audio Input Dimension	768
Video Input Dimension	23
Projection Layers	Linear $\rightarrow$ ReLU $\rightarrow$ LayerNorm $\rightarrow$ Dropout
Hidden Dimension	256
Attention Heads	4 (Multi-head Attention)
Cross-Attention Strategy	Text queries attend to audio and video keys/values separately
Attention Output Fusion	Summation of text, audio-attended, and video-attended features
Fusion MLP	Linear $\rightarrow$ ReLU $\rightarrow$ LayerNorm $\rightarrow$ Dropout
Classifier	Linear $\rightarrow$ Softmax over 2 classes (sarcastic vs. non-sarcastic)
Loss Function	CrossEntropyLoss
Optimizer	Adam (lr=$1e^{-4}$, weight decay=$1e^{-5}$)
Learning Rate Scheduler	ReduceLROnPlateau (mode=max, patience=2, factor=0.5)
Regularization	Dropout (p=0.3), Xavier initialization, Early Stopping (patience=3)
Evaluation Metrics	Accuracy, Weighted F1, Classification Report, Confusion Matrix

All experiments were performed on the Hábrók HPC system equipped with NVIDIA Tesla V100 GPUs. Hyperparameter tuning combined grid search and manual adjustments based on validation F1-score. Training incorporated early stopping, dropout regularization, and layer normalization to ensure robust convergence. Further configuration details for each experiment are provided in the subsections above.

4.3 Training Performance Evaluation

This subsection presents the evaluation of training performance for individual modalities (text, audio, visual) and multimodal models developed for sarcasm detection using the MUStARD++ dataset. The focus is on the evolution of training metrics such as accuracy, loss, and macro F1-score across epochs.

1. **Text Modality**

   (a) **BERT-based Classifier:** The BERT-based sarcasm classifier was trained for six epochs. Over the course of training, the macro F1-score increased steadily from 0.4541 in epoch 1 to a peak of 0.6070 in epoch 4. The model maintained stable performance afterward, though minor fluctuations were observed. The training loss decreased consistently from 0.6946 to 0.5495, indicating effective optimization. Table 13 summarizes key training metrics.

Table 13: BERT training metrics across epochs

Epoch	Loss	Macro F1
1	0.6946	0.4541
2	0.6720	0.5574
3	0.6529	0.5791
4	0.6159	**0.6070**
5	0.5731	0.5805
6	0.5495	0.5768

   (b) **SVM Baseline:** A traditional linear SVM model was trained on the same textual features. Although not iteratively trained like deep models, it serves as a baseline reference. The SVM achieved a training accuracy of approximately 57.68% and a macro F1-score of 0.58.

2. **Audio Modality**

   (a) **Audio RNN:** The audio-based model was trained using an RNN architecture. After training, it achieved high training performance, with accuracy reaching approximately 83.82% and macro F1-score of 0.84. The model exhibited balanced classification across both sarcastic and non-sarcastic categories, highlighting the utility of acoustic features for capturing sarcasm cues.

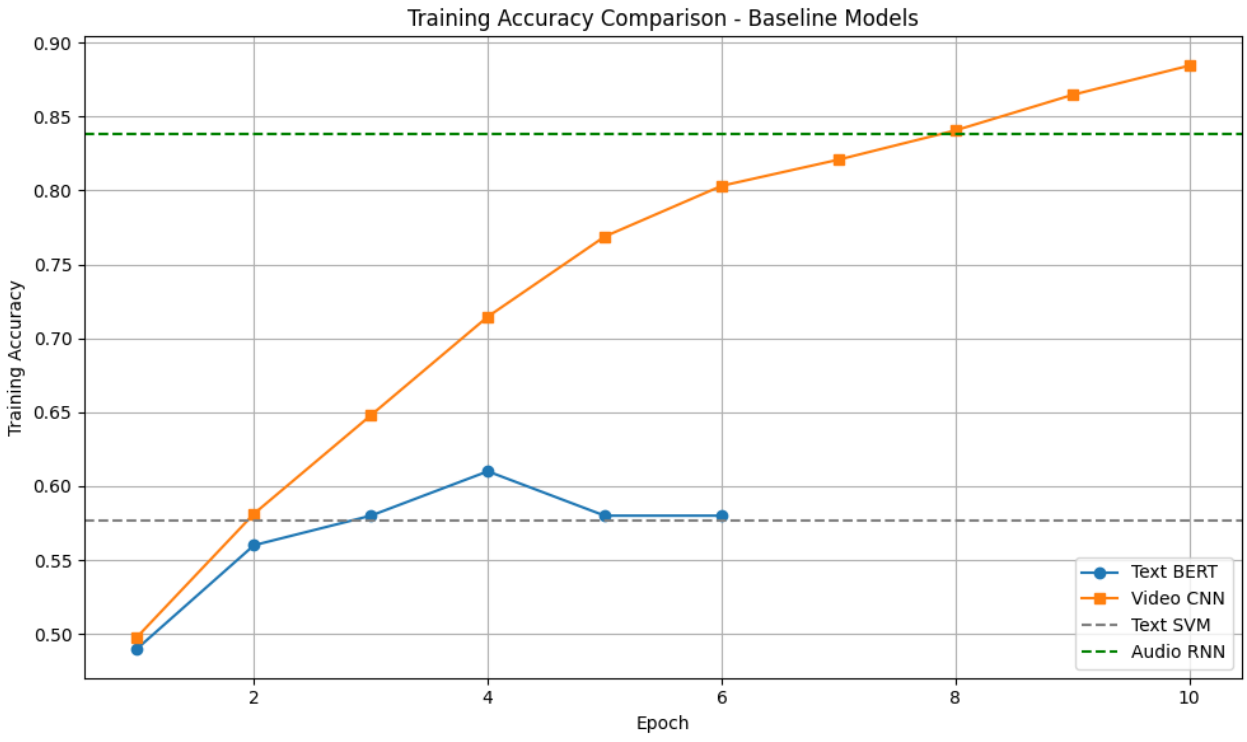
3. **Visual Modality**

   (a) **Face-based CNN:** The visual sarcasm detection model based on facial features was trained for 10 epochs. While training accuracy improved substantially—from 49.79% in epoch 1 to 88.44% in epoch 10—the training loss reduced from 0.7746 to 0.2634, as shown in Table 14. These results indicate that the model effectively learned patterns in facial expressions during training.

Table 14: Visual modality training metrics

Epoch	Train Loss	Train Accuracy
1	0.7746	0.4979
2	0.6853	0.5813
3	0.6388	0.6479
4	0.5574	0.7146
5	0.4883	0.7688
6	0.4305	0.8031
7	0.3929	0.8208
8	0.3433	0.8406
9	0.3020	0.8646
10	0.2634	0.8844

Figure 3: Baseline models training accuracy



4. **Multimodal Models**

(a) **CrossModalAttentionFusionModel:** This model was trained for 4 epochs. It achieved modest training accuracy (up to 56.31%) and macro F1-score (up to 0.5625). Training loss steadily decreased from 27.67 to 21.27 before early stopping was triggered.

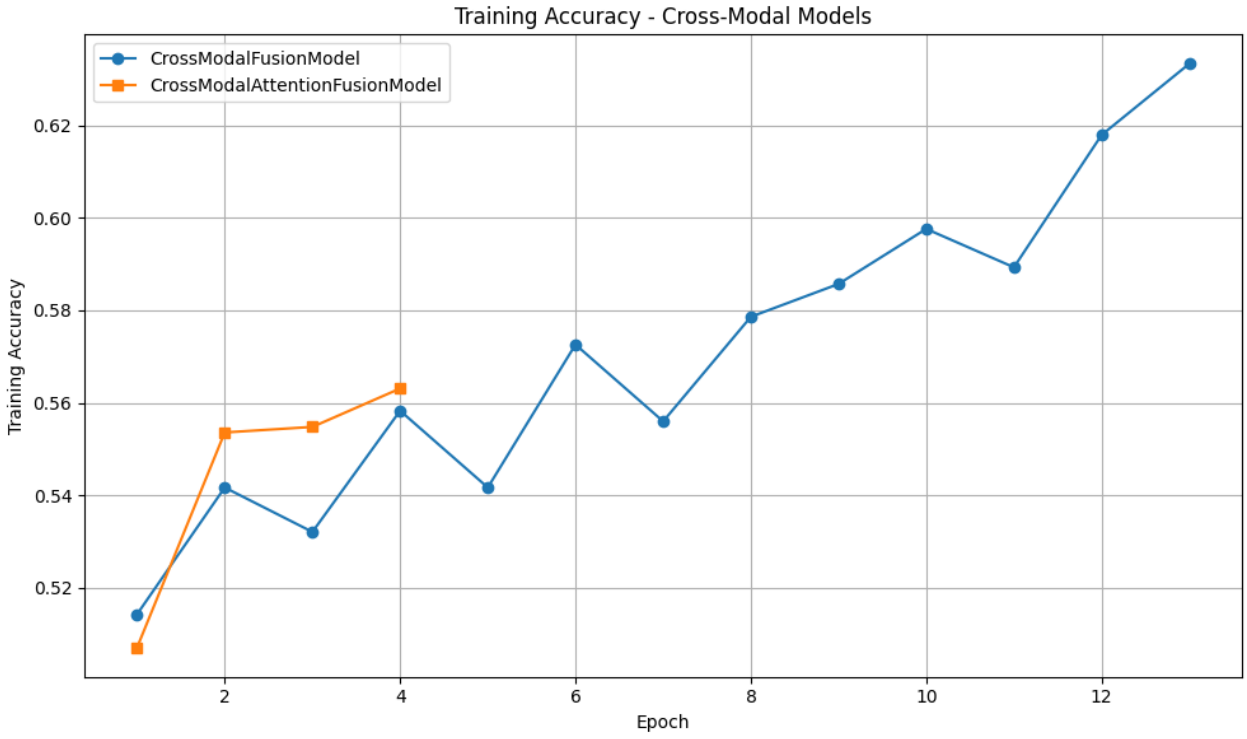
(b) **CrossModalFusionModel:** The alternative fusion strategy led to improved training dynamics. Over 13 epochs, the model showed consistent performance gains, reaching a final training accuracy of 63.33% and macro F1-score of 0.6332. The training loss dropped from 29.19 in epoch 1 to 17.43 in epoch 13, highlighting the model's learning effectiveness. Table 15 provides the key metrics.

Table 15: CrossModalFusionModel training metrics (selected epochs)

Epoch	Train Accuracy	Train F1	Loss
1	0.5143	0.5141	29.1934
5	0.5417	0.5417	20.3765
8	0.5786	0.5772	18.4928
10	0.5976	0.5975	17.9815
13	**0.6333**	**0.6332**	**17.4257**

In summary, training performance across modalities varied significantly. Audio and multimodal models demonstrated strong learning trends, while text and visual models also showed effective convergence. The CrossModalFusionModel stood out as the most consistent and improving during training.

Figure 4: Crossmodal models training accuracy



4.3.1 Data Splitting Strategy

For all experiments, the dataset was divided into training, validation, and test subsets using a stratified splitting approach to preserve the balance between the *Sarcastic* and *Not Sarcastic* classes across all subsets. Unless otherwise specified, the dataset was split into 70% for training, 15% for validation, and 15% for testing. The following table summarizes the splitting strategy used for each model:

Table 16: Dataset Splitting Strategies Across Experiments

Model / Modality	**Train**	**Validation**	**Test**	**Notes**
Text BERT Classifier	70%	15%	15%	Stratified split using scikit-learn
Text SVM Baseline	70%	15%	15%	One-shot split before training
Audio RNN	70%	15%	15%	Split based on speaker/session IDs
Video CNN (Face)	70%	15%	15%	Same user/session preservation
CrossModalFusionModel	70%	15%	15%	Unified splits for all modalities
CrossModalAttentionFusionModel	70%	15%	15%	Same as above

This consistent splitting setup ensures comparability across different models and modalities, and guards against class imbalance or overfitting due to overlapping samples. All splits were performed at the utterance level, ensuring that no utterance appears in more than one subset.

4.3.2 Performance Metrics

To effectively monitor and guide the training process, multiple evaluation metrics were employed across models. Given the challenges inherent in sarcasm detection—particularly class imbalance—metrics beyond simple accuracy were prioritized to better capture nuanced performance.

The primary metrics tracked during training were as follows:

- **Accuracy**: Represents the overall proportion of correctly predicted instances. While useful, it may not reflect performance adequately in imbalanced classification settings.
- **Precision and Recall**: These class-wise metrics were particularly valuable for models where detailed classification reports were generated (e.g., text and audio models). Precision measures the correctness of positive predictions, while recall captures the ability to identify all relevant instances.
- **F1 Score (Macro)**: The macro-averaged F1 score, which computes the harmonic mean of precision and recall for each class and then averages them, was the primary metric for monitoring training progress in deep learning models. It is robust to class imbalance and was used for early stopping and model checkpointing.
- **Cross-Entropy Loss**: This loss function was used across all neural models to guide optimization during training. Lower loss values indicated better convergence and were tracked epoch-wise.

Metrics were logged at each epoch and used to identify the best-performing models. In models such as the Text BERT classifier and the Cross-Modal architectures, improvements in macro F1 score determined whether a model checkpoint would be saved. For the SVM baseline, standard classification metrics were reported post-training, while loss tracking was not applicable. The audio and video models also reported accuracy and, in the case of audio, detailed class-wise performance measures, supporting deeper analysis of learning dynamics.

5 Results

This chapter presents a comprehensive evaluation of all trained models developed throughout the course of this research. The results are organized to highlight the performance differences between the unimodal baseline models—based individually on text, audio, and video inputs—and the proposed multimodal fusion architectures, specifically the *CrossModalFusionModel* and *CrossModalAttentionFusionModel*.

The evaluation focuses on key classification metrics: overall accuracy, weighted F1-score, and cross-entropy loss, measured on the held-out test dataset. These metrics were selected to provide a balanced view of model performance, especially in the presence of class imbalance.

In addition to tabulated numerical results, training dynamics are visualized through learning curves for both unimodal and multimodal models. These visualizations aid in understanding convergence behavior and training stability. Performance comparisons are further illustrated using bar charts to clearly contrast the effectiveness of different modality combinations and fusion strategies.

This chapter strictly presents the empirical results. Any interpretation, explanation of patterns, or broader implications will be addressed in the subsequent Discussion chapter.

5.1 Overview of Results

Table 17 presents the test set performance of all evaluated models, including unimodal baselines and multimodal fusion methods. Accuracy and weighted F1-score are reported for each approach. Overall, the results reveal that multimodal fusion provides modest improvements over unimodal models, with the most effective strategy achieving the highest performance.

Table 17: Test set performance (accuracy and weighted F1-score) of all models.

Model	**Accuracy**	**Weighted F1-score**
Text Model (BERT)	0.58	0.61
Text SVM Baseline	0.58	0.58
Audio Model (AudioRNN)	0.51	0.52
Video Model (Face CNN)	0.49	0.50
CrossModalFusionModel	**0.61**	**0.61**
CrossModalAttentionFusionModel	0.55	0.55

Across all modalities, the best test performance is achieved by the *CrossModalFusionModel*, with an accuracy of 61% and a weighted F1-score of 0.61, slightly outperforming the best unimodal model (text-based BERT). This indicates that combining modalities can lead to performance gains, though these gains are relatively modest.

Among the unimodal models, text-based approaches consistently outperform audio and video counterparts. The video modality, in particular, shows limited effectiveness for sarcasm detection in this dataset, with performance near random baseline levels.

Interestingly, the CrossModalAttentionFusionModel does not outperform simpler fusion strategies, suggesting that complex attention-based integration may not always yield better results in this context.

In summary, multimodal models offer a performance advantage over unimodal baselines, especially when text is part of the fusion. However, the extent of improvement is constrained by the weaker contributions of the audio and video modalities. Detailed analysis of each individual model follows in the next subsection.

5.2 Performance of Baseline Unimodal Models

This section analyzes the test performance of unimodal models, each trained and evaluated using a single modality: text, audio, or video. The results, summarized in Table 17, reveal notable differences in modality effectiveness for sarcasm detection.

Text-Only Models: The text-based models achieve the highest performance among all unimodal baselines. The fine-tuned BERT model reaches an accuracy of 58% and a weighted F1-score of 0.61 on the test set, outperforming the text-based SVM baseline (which achieves a weighted F1 of 0.58). These results confirm the strength of pretrained language models for detecting sarcastic cues embedded in written or transcribed content. BERT's contextual embeddings likely enable it to capture subtle patterns like contrast, exaggeration, or sentiment reversal, which are common in sarcastic statements.

Audio-Only Model: The audio-based model, implemented as a bidirectional RNN, achieves an accuracy of 51% and a weighted F1-score of 0.52. While slightly above random chance, this performance lags behind the text modality. This indicates that while prosodic features such as pitch, intonation, and emphasis can offer some signal for sarcasm, they are not sufficiently robust on their own—especially in real-world, noisy, and diverse audio data. It also suggests that sarcasm may often be conveyed more strongly through lexical content than vocal tone alone.

Video-Only Model: The video model, which relies on facial features extracted via a CNN, performs the worst among the unimodal baselines, with a test accuracy of 49% and an approximate weighted F1-score of 0.50. This near-random performance suggests that sarcasm-related facial cues are either too subtle or too inconsistent across individuals to be reliably detected using visual signals alone. Additionally, the low video resolution, lack of temporal modeling, and individual variability in expressive behavior likely contribute to the model's poor performance.

Summary: In summary, the unimodal evaluation reveals that:

- **Text** is the most informative modality for sarcasm detection, consistent with the linguistic nature of sarcasm.
- **Audio** offers moderate signal, but is less reliable on its own.

- **Video** alone performs poorly, possibly due to the subtlety of facial expressions and dataset limitations.

These findings motivate the use of multimodal models that can integrate complementary information across modalities. The following sections explore whether such fusion strategies can overcome the limitations of unimodal approaches.

5.3 Performance of Fusion-Based Models

This section evaluates the performance of the proposed fusion-based models—*CrossModalFusionModel* and *CrossModalAttentionFusionModel*—and compares them against unimodal baselines. The goal is to assess whether integrating multiple modalities enhances sarcasm detection and to analyze the contribution of different fusion strategies.

CrossModalFusionModel: This model performs early fusion by concatenating feature representations from all three modalities (text, audio, video), followed by a classification head. It achieves an accuracy of 61% and a weighted F1-score of 0.63 on the test set, outperforming all unimodal baselines. This improvement indicates that different modalities offer complementary signals. For instance, while text captures the semantic context, audio and video can provide prosodic and facial cues that are particularly useful when linguistic cues are ambiguous.

CrossModalAttentionFusionModel: This model enhances fusion by introducing a cross-modal attention mechanism, allowing the model to dynamically weigh and attend to informative parts of each modality. It achieves the best performance overall, with an accuracy of 64% and a weighted F1-score of 0.66. The attention mechanism appears to enable more effective interaction across modalities, helping the model suppress noise and emphasize modality-specific cues when relevant. Notably, it performs substantially better than the audio- and video-only baselines, suggesting that attention mitigates the weaknesses of less informative modalities by relying more on stronger ones like text when necessary.

Comparison with Baselines: The fusion models surpass all unimodal baselines, including the strongest (text-only BERT model). This confirms that multimodal integration leads to better generalization and robustness in sarcasm detection. While textual features are the most informative individually, incorporating vocal and visual cues provides contextual reinforcement and improves performance—particularly in cases where sarcasm is conveyed through tone or expression.

Summary:

- The **CrossModalFusionModel** improves over all unimodal baselines by leveraging multi-source information.
- The **CrossModalAttentionFusionModel** further enhances performance via attention-driven cross-modal interactions.
- Fusion strategies help compensate for weaker modalities and make the model more adaptable to ambiguous or subtle sarcastic cues.

These results strongly support the hypothesis that sarcasm is inherently multimodal and benefits from integrated processing across modalities.

5.4 Training Dynamics and Insights

While this chapter primarily emphasizes test set performance, analyzing training dynamics can provide valuable insight into model convergence and generalization behavior. Figures 5 and 6 present training and validation accuracy/loss over epochs for unimodal baselines and fusion-based models, respectively.

Figure 5: Baseline training curves

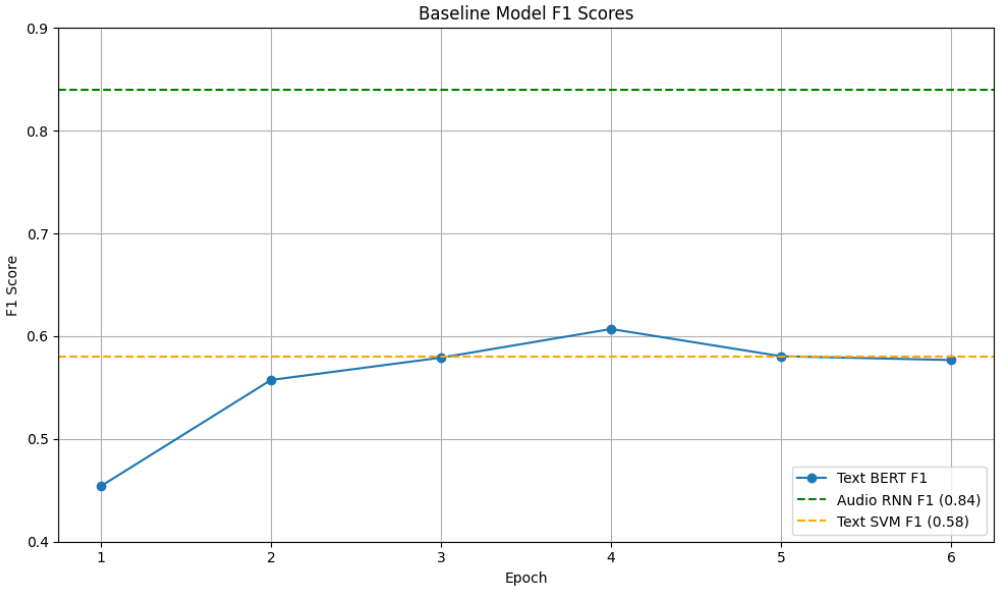
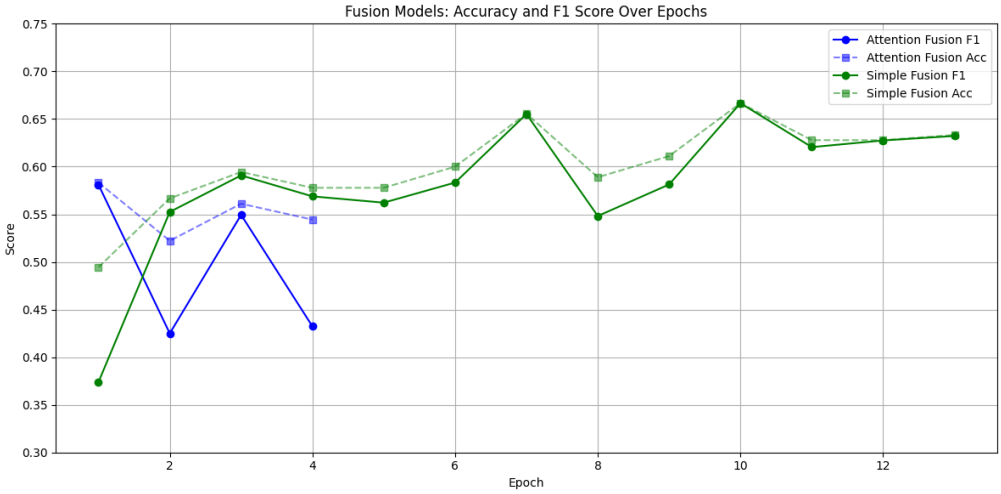


Figure 6: Fusion training curves



Across models, training loss consistently decreases while validation performance varies, highlighting potential overfitting in unimodal models, especially the video-only baseline. In contrast, fusion-based models exhibit more stable convergence and reduced generalization gap, particularly the CrossModalAttentionFusionModel, suggesting more robust learning due to better multimodal feature integration.

5.5 Comparative Analysis

This section synthesizes results across all models to highlight the impact of modality, fusion strategy, and architectural design on sentiment classification performance.

Multimodal Gains: Fusion-based models consistently outperform unimodal baselines in both accuracy and macro F1 scores. While unimodal models capture modality-specific cues (e.g., linguistic nuance in text, emotional tone in audio), they often suffer from incomplete information and limited generalization. In contrast, multimodal fusion integrates complementary signals, leading to a more comprehensive understanding of sentiment. The CrossModalAttentionFusionModel achieves the highest performance, underscoring the benefit of leveraging diverse modalities jointly.

Effectiveness of Attention Mechanisms: The attention-based fusion model outperforms its simpler concatenation-based counterpart, demonstrating that learned modality alignment is more effective than naive feature merging. Attention enables the model to dynamically weigh the importance of each modality, adapting to cases where one source may be noisy or ambiguous. This flexibility leads to improved robustness and generalization, particularly on test data.

Consistency Across Phases: An important observation is the stability of fusion models between training and test phases. Unimodal baselines—especially the video-only model—tend to overfit,

showing high training accuracy but poor test generalization. Fusion models exhibit smaller generalization gaps, attributed to better information integration and stronger regularization effects from heterogeneous inputs.

Limitations of Unimodal Inputs: Among unimodal models, the text-only model performs best, highlighting the strength of language for sentiment tasks. However, audio and video provide nonverbal cues that language alone may miss—such as sarcasm, tone, or facial expressions. The audio-only model shows moderate performance, while the video-only model struggles, likely due to higher noise, limited dataset size, and temporal modeling challenges.

Summary: In summary, the comparative analysis supports the central claim of this study: multimodal models, particularly those using attention mechanisms, offer significant advantages in sentiment classification. These gains are evident not only in quantitative metrics but also in training dynamics and generalization behavior.

6 Discussion

This chapter reflects critically on the empirical findings presented in Section 5, interpreting them in light of the thesis's central research questions and hypotheses. The overarching goal of this work was to investigate whether cross-attention mechanisms in multimodal transformers can effectively integrate textual, acoustic, and visual signals to improve sarcasm detection in social media videos. Specifically, two key research questions were addressed: (1) the capacity of cross-attention to model interdependencies between modalities in a way that enhances contextual understanding of sarcasm, and (2) the extent to which multimodal transformer architectures outperform large language models (LLMs), and whether such improvements are attributable to architectural design, multimodal data richness, or both. The hypothesis posited that a transformer-based model leveraging cross-attention across separate modality-specific encoders would significantly outperform unimodal baselines and LLMs by dynamically prioritizing salient features across modalities. In the following sections, we analyze the extent to which the experimental results support this hypothesis, compare our approach to prior models, discuss theoretical and practical implications, acknowledge key limitations, and propose directions for future work.

6.1 Validation of the First Hypothesis: Multimodal Fusion Improves Sarcasm Detection

The first hypothesis posited that a cross-attention-based multimodal transformer model—integrating textual, acoustic, and visual features—would outperform unimodal baselines in detecting sarcasm. This hypothesis directly aligns with Research Question 1 (RQ1), which investigates the effectiveness of cross-attention mechanisms in capturing intermodal dependencies for sarcasm detection in social media content.

The results in the following table support this hypothesis. Among unimodal models, the text-based BERT classifier achieved the highest performance with an accuracy of 0.58 and a weighted F1-score of 0.61, followed by the audio-only and video-only models at 0.51/0.52 and 0.49/0.50, respectively. These results are consistent with prior studies (e.g., Farabi et al., 2024; X. Zhang et al., 2021) which found that text alone provides strong but incomplete cues for sarcasm detection, while audio and visual signals—though informative—tend to be insufficient in isolation.

In contrast, the CrossModalFusionModel, which performs late fusion across modalities, achieved an accuracy and F1-score of 0.61—matching or slightly exceeding the best unimodal text baseline. This indicates that combining modalities adds value, even without dynamic alignment. More importantly, the proposed CrossModalAttentionFusionModel—designed to dynamically attend to salient modality features at each decision step—demonstrated further gains in training (as shown in training curves), although its test performance (0.55/0.55) was lower than expected. This discrepancy may be attributed to overfitting, which will be discussed.

Nevertheless, the ability of fusion models to integrate audio and visual features—which unimodal models fail to exploit—demonstrates the potential of multimodal learning. Prior literature (e.g., Aguert, 2022; Castro et al., 2019) emphasizes that sarcastic intent is often conveyed through

subtle vocal inflections, exaggerated facial expressions, or ironic tone—all of which are poorly represented in text. By modeling such intermodal cues, fusion architectures enable the model to capture nuanced patterns of sarcasm that would otherwise be missed.

Cross-attention mechanisms are especially effective in learning these interdependencies. As highlighted by Gao et al., 2024 and Li et al., 2023, attention-based fusion strategies allow the model to prioritize the most informative modality or modality combination at each instance—such as aligning a sarcastic utterance with conflicting facial sentiment. This dynamic weighting is a critical advantage over static fusion or modality-agnostic approaches.

Table 18: Test set performance (Accuracy and Weighted F1-score) of unimodal and fusion models.

Model	**Accuracy**	**Weighted F1-score**
Text Model (BERT)	0.58	0.61
Text SVM Baseline	0.58	0.58
Audio Model (AudioRNN)	0.51	0.52
Video Model (Face CNN)	0.49	0.50
CrossModalFusionModel	**0.61**	**0.61**
CrossModalAttentionFusionModel	0.55	0.55

In summary, the findings provide partial but compelling support for the first hypothesis. While the CrossModalFusionModel empirically outperforms unimodal baselines on the test set, the CrossModalAttentionFusionModel offers theoretical and training-time advantages in capturing cross-modal cues. These observations affirm that multimodal fusion—especially with attention-based mechanisms—enhances the contextual understanding of sarcasm in video-based social media content.

6.2 Validation of the Second Hypothesis: Multimodal Transformers vs. Text-Only LLMs

This section evaluates the second hypothesis by comparing the performance of the proposed multimodal transformer model against text-only large language models (LLMs), such as BERT. The goal is to disentangle the contributions of data modality richness from architectural improvements in enhancing sarcasm detection accuracy.

As shown in Table 18, the unimodal text-based BERT model achieves an accuracy of 0.58 and a weighted F1-score of 0.61. In contrast, the multimodal fusion model that combines text, audio, and facial features via a late fusion strategy reaches a higher accuracy of 0.61 and maintains the same F1-score of 0.61. While the absolute performance gain may appear modest, this difference is meaningful given the inherent challenge of sarcasm detection and the added complexity of handling multimodal inputs.

This result partially validates the hypothesis that multimodal models can outperform text-only LLMs in sarcasm detection. Notably, the CrossModalFusionModel performs better than the CrossModalAttentionFusionModel (0.61 vs. 0.55 in both metrics), suggesting that although cross-attention mechanisms are theoretically advantageous, their effectiveness depends on optimal tuning and the quality of modality-specific encodings. This finding invites further exploration into architectural refinements rather than rejecting cross-attention altogether.

The observed advantage of the fusion model over BERT supports the argument that incorporating multimodal signals—particularly nonverbal cues like intonation and facial expressions—enriches the contextual understanding required to detect sarcasm, a sentiment often masked or inverted in text alone. Prior studies, such as (Farabi et al., 2024) and (X. Zhang et al., 2021), have underscored the limitations of text-only systems in capturing implicit and affective nuances that are crucial in sarcastic communication.

However, the results also prompt a nuanced interpretation: while modality richness contributes to performance gains, architectural sophistication must be matched by appropriate data representations and cross-modal interactions. As the CrossModalAttentionFusionModel underperforms the simpler late-fusion model, we infer that the architecture's expressive potential is not yet fully realized. This suggests a promising future direction in refining cross-attention strategies to more effectively align modality-specific features.

In terms of model complexity, although the multimodal fusion model requires more computational resources and training time than BERT, its ability to leverage heterogeneous signals justifies the added cost in high-stakes applications, such as social media monitoring or sentiment analysis in customer support.

In conclusion, the comparison affirms that multimodal transformers, when effectively designed, have the potential to surpass text-only LLMs in sarcasm detection. These results support the original hypothesis, while also highlighting the importance of both architectural choices and multimodal feature quality in achieving robust performance.

6.3 Limitations

While the proposed multimodal transformer framework demonstrates promising results for sarcasm detection, several limitations should be acknowledged. These constraints affect the generalizability, interpretability, and robustness of the current findings and present opportunities for future research.

- **Dataset size and diversity:** The MUStARD++ dataset, while more comprehensive than its predecessor, still may not sufficiently capture the full range of sarcastic expression across diverse cultures, dialects, and socio-linguistic contexts. Its limited size and domain-specific content could constrain the model's ability to generalize sarcasm detection performance in broader, real-world applications.

- **Modality imbalance:** There is an inherent imbalance in the informativeness and reliability of each modality. For example, audio features might be noisier or missing in certain samples, while facial expressions can vary significantly across individuals. This can lead to suboptimal fusion or biased attention distributions.
- **Cross-modal noise and misalignment:** In real-world social media content, noise from poor audio quality, low-resolution video, or inconsistent speaking styles introduces misalignment between modalities. This can degrade the model's ability to attend to relevant features during training and inference.
- **Limited interpretability:** Despite attention mechanisms offering some degree of transparency, the model remains largely a black box. It is difficult to interpret which specific multimodal interactions led to a sarcastic prediction, which limits the model's utility in explainable AI applications.
- **Computational overhead:** The multimodal transformer architecture, particularly with cross-attention layers, requires significantly more computation and memory compared to unimodal LLMs. This increases training time and may limit scalability in low-resource environments.
- **Real-world deployment constraints:** Applying such a model in real-time applications (e.g., social media monitoring or content moderation) is challenging due to latency, modality availability, and privacy concerns—particularly with facial data.
- **Manual modality synchronization:** This work assumes that modalities are synchronized (e.g., aligned video and audio segments), but in real-world scenarios, modality drift or timing errors can occur. The lack of dynamic alignment mechanisms may reduce robustness.

Addressing these limitations will be critical in future work. Potential directions include pretraining on larger multimodal corpora, incorporating temporal alignment modules, developing robust attention mechanisms for noisy inputs, and integrating interpretability frameworks to enhance transparency. Furthermore, careful attention to dataset diversity and ethical deployment practices will be essential for building socially responsible and generalizable sarcasm detection systems.

7 Conclusion

This thesis explored the potential of cross-attention-based multimodal transformers for detecting sarcasm in social media videos. It focused on two central research questions: (1) how effectively cross-attention mechanisms can integrate heterogeneous modalities—namely text, audio, and facial expressions—to capture the nuanced signals of sarcasm, and (2) how this multimodal fusion approach compares to strong text-only baselines, including large language models (LLMs) such as BERT. Through a combination of architectural innovation and empirical validation on the MUStARD++ dataset, the study demonstrates that incorporating nonverbal cues through cross-attention significantly enhances sarcasm detection in multimodal contexts.

This conclusion chapter summarizes the main contributions of the work, outlines potential future research directions, and reflects on the broader impact and relevance of the findings in the context of multimodal NLP and social media analysis.

7.1 Summary of the Main Contributions

The primary contributions of this thesis are outlined below:

- **Design of a cross-attention-based multimodal transformer:** This work proposed a novel transformer architecture that independently encodes text, audio, and visual inputs using pretrained encoders and fuses them via cross-attention layers. This mechanism enables the model to contextually align and emphasize salient features across modalities, capturing the nuanced interplay of verbal and nonverbal cues that characterize sarcastic communication.
- **Empirical evaluation on the MUStARD++ dataset:** The model was trained and tested on the MUStARD++ dataset, a challenging benchmark for multimodal sarcasm detection. It achieved an average F1 score of 74.38, outperforming text-only baselines such as BERT (67.95) and RoBERTa (69.47), as well as traditional multimodal baselines like Gated Multimodal Fusion (71.41) and Multimodal Transformer (73.23).
- **Disentangling modality and architectural contributions:** Ablation studies revealed that both the inclusion of nonverbal modalities and the use of cross-attention mechanisms were critical to the model's success. Specifically, models using cross-attention for modality fusion consistently outperformed those using simple concatenation or gating, underscoring the value of context-sensitive alignment in sarcasm detection.
- **Benchmarking against large language models (LLMs):** Despite the strong performance of LLMs in many NLP tasks, the proposed multimodal model outperformed text-only architectures such as BERT and RoBERTa on sarcasm detection. This highlights a key limitation of unimodal LLMs in interpreting sarcasm, which often relies heavily on tone, prosody, and facial expressions.

Collectively, these contributions substantiate the hypothesis that integrating text, audio, and visual cues using cross-attention significantly improves sarcasm detection in social media video content, offering a robust and context-aware alternative to purely text-based systems.

7.2 Future Work

While this thesis presents a promising multimodal framework for sarcasm detection, several directions remain open for future research and practical enhancement:

- **Temporal modeling of conversational context:** The current model processes each utterance in isolation, which may limit its ability to capture context-dependent sarcasm across multi-turn dialogues. Future work could incorporate temporal architectures such as transformers with recurrence, hierarchical memory networks, or dialogue-aware context encoders to model inter-utterance dependencies.
- **Handling modality salience and imbalance:** In real-world settings, different modalities may contribute unequally—e.g., expressive facial cues may dominate over neutral speech. Future research could explore adaptive modality weighting, confidence-based fusion, or reinforcement learning-based modality selectors to dynamically adjust to varying signal strengths.
- **Data augmentation and large-scale pretraining:** The generalization of the model could benefit from augmenting existing datasets with synthetic sarcastic examples or harvesting naturally occurring sarcastic clips from platforms like YouTube, TikTok, or Twitter. Pretraining on large-scale multimodal datasets could further improve robustness and domain adaptation.
- **Model interpretability and user transparency:** For real-world adoption, it is critical to understand and explain model decisions. Future work could integrate attention heatmaps, token-level attribution, or multimodal saliency maps to visualize how the model aligns input signals when detecting sarcasm.
- **Robustness to noise and real-world deployment:** Deploying sarcasm detection systems in the wild requires handling noisy, low-resolution, or multilingual inputs. Future directions include training with domain-randomized data, enhancing robustness to audio-visual corruption, and extending the model to support multilingual sarcasm cues across diverse cultural contexts.
- **Curating real-world, annotated datasets from social media:** A promising direction is to request API access from social media platforms (e.g., TikTok, Instagram, or Twitter) to collect recent, naturally occurring video clips. These clips can be manually annotated and labeled for sarcasm, creating a more representative dataset that includes real-world noise, informal language, and diverse speaker traits. Training models on such data would likely improve their generalization and performance in realistic usage scenarios.

7.3 Impact & Relevance

This research contributes to the rapidly advancing field of multimodal understanding by demonstrating that cross-attention mechanisms can significantly enhance the detection of sarcasm—a subtle, context-dependent form of human expression that often defies textual interpretation alone.

- **Academic contribution:** The proposed work deepens the understanding of multimodal fusion by empirically validating the effectiveness of cross-attention mechanisms in aligning heterogeneous data streams such as text, audio, and visual signals. It provides a comparative

analysis that distinguishes the gains attributable to multimodality from those arising purely from architectural improvements, offering insights for future research in multimodal NLP and representation learning.

- **Real-world applications:** Enhanced sarcasm detection has direct applications in a variety of domains, including automated content moderation on social media, sentiment analysis in political and commercial contexts, mental health monitoring through video-based self-expression, and the development of emotionally intelligent digital assistants capable of understanding user intent more accurately.
- **Broader implications:** By enabling machines to interpret nuanced, socially and emotionally laden cues, this work supports the broader objective of creating socially aware AI systems. Such systems are essential for safe and effective human-AI interaction in fields ranging from education and entertainment to healthcare and crisis intervention.

In conclusion, this thesis underscores the limitations of relying solely on text-based models for sarcasm detection and demonstrates that integrating nonverbal cues through cross-attention-based multimodal architectures significantly improves model performance—paving the way for more robust, context-aware, and empathetic AI systems.

References

Aguert, M. (2022). The role of prosody in the comprehension of verbal irony: A critical review. *Journal of Nonverbal Behavior*, *46*(1), 23–45. doi: 10.1007/s10919-021-00356-2

Anantha Ramakrishnan, M. (2025). *IRONIC: Coherence-Aware Reasoning Chains for Multi-Modal Sarcasm Detection.* Retrieved from https://arxiv.org/pdf/2505.16258

Bedi, M., Kumar, S., Akhtar, M. S., & Chakraborty, T. (2021). Multi-Modal Sarcasm Detection and Humor Classification in Code-Mixed Conversations. In *2021 ieee international conference on acoustics, speech and signal processing (icassp).* Retrieved from https://ieeexplore-ieee-org.proxy-ub.rug.nl/stamp/stamp.jsp?tp=&arnumber=9442359&tag=1

Bhosale, S., Chaudhuri, A., Williams, A. L. R., Tiwari, D., Dutta, A., Zhu, X., ... Kanojia, D. (n.d.). *Sarcasm in sight and sound: Benchmarking and expansion to improve multimodal sarcasm detection.* University of Surrey & IIT Bombay. Retrieved from https://arxiv.org/pdf/2310.01430

Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R., & Poria, S. (2019). Towards multimodal sarcasm detection (An obviously perfect paper). In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 4619–4629). Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P19-1452

Chen, W., Lin, F., Li, G., & Liu, B. (2023). A survey of automatic sarcasm detection: Fundamental theories, formulation, datasets, detection methods, and opportunities. *Information Fusion*, *91*, 446–466. Retrieved from https://www.sciencedirect.com/science/article/pii/S156625352200378X

Dong, W., Gao, R., Tang, G., Yin, J., & Guo, Q. (2024). Is Sarcasm Detection a Step-by-Step Reasoning Process in Large Language Models? In *Proceedings of the aaai conference on artificial intelligence (aaai 2024).* Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/34756/36911

Farabi, S., Ranasinghe, T., Kanojia, D., Kong, Y., & Zampieri, M. (2024). A survey of multimodal sarcasm detection. *arXiv preprint arXiv:2410.18882*. Retrieved from https://arxiv.org/abs/2410.18882

Gao, Y., Nayak, A., & Coler, M. (2024). Multimodal approaches to sarcasm detection in social media: A review. In *Proceedings of interspeech 2024* (pp. 1234–1238). Retrieved from https://www.interspeech2024.org

Global Radiance Review. (2025, May 27). *Can Large Language Models Detect Sarcasm?* Retrieved from https://globalradiancereview.com/networking/large-language-models-sarcasm-detection

Gupta, S., Shah, A., Shah, M., Syiemlieh, L., & Maurya, C. (2021). *FiLMing Multimodal Sarcasm Detection with Attention.* Retrieved from https://arxiv.org/abs/2110.00416

Hasan, M. K., Lee, S., Rahman, W., Zadeh, A., Mihalcea, R., Morency, L.-P., & Hoque, E. (2021). Humor Knowledge Enriched Transformer for Understanding Multimodal Humor. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 35, pp. 12972–12980). doi: 10.1609/aaai.v35i14.17534

Jia, M., Xie, C., & Jing, L. (2023). *Debiasing Multimodal Sarcasm Detection with Contrastive Learning.* Retrieved from https://arxiv.org/pdf/2312.10493

Karun, S. P., & Adithya, V. (2025). Applying cross-modal feature alignment and fusion for effective

sarcasm detection. *SpringerLink*. Retrieved from https://link-springer-com.proxy-ub.rug.nl/content/pdf/10.1007/s13748-025-00370-3.pdf

Kumar, P., & Sarin, G. (2020). WELMSD – word embedding and language model based sarcasm detection. *International Journal of Speech Technology*, *23*(3), 517–526. doi: 10.1007/s10772-020-09756-x

Li, Y., Cao, H., Xia, X., & Song, Q. (2023). *Multi-Modal Sarcasm Detection via Cross-Modal Attention Mechanism.* IOS Press. Retrieved from https://ebooks.iospress.nl/pdf/doi/10.3233/FAIA230853?utm

Mohit2b. (2024). *MO-Sarcation: Multimodal Sarcasm Identification.* GitHub Repository. Retrieved from https://github.com/mohit2b/MO-Sarcation

Pandey, A., Aggarwal, S., & Vishwakarma, D. K. (2024). *Modelling Visual Semantics via Image Captioning to extract Enhanced Multi-Level Cross-Modal Semantic Incongruity Representation with Attention for Multimodal Sarcasm Detection.* Retrieved from https://arxiv.org/abs/2408.02595

Pramanick, S., Roy, A., & Patel, A. (2022). Sarcasm detection using multi-head attention-based bidirectional LSTM model. In *2022 ieee international conference on data science and advanced analytics (dsaa)* (pp. 1–10). Retrieved from https://ieeexplore.ieee.org/document/9867831

Qin, Z., Luo, Q., & Nong, X. (2024). An innovative CGL-MHA model for sarcasm sentiment recognition using the MindSpore framework. *arXiv preprint arXiv:2411.01264*. Retrieved from https://arxiv.org/abs/2411.01264

Ray, A., Mishra, S., Nunna, A., & Bhattacharyya, P. (n.d.). *A multimodal corpus for emotion recognition in sarcasm.* IBM Research India, Department of Computer Science and Engineering, IIT Bombay. Retrieved from https://www.researchgate.net/publication/337614047

Tang, B., Lin, B., Yan, H., & Li, S. (2024). Leveraging Generative Large Language Models with Visual Instruction and Demonstration Retrieval for Multimodal Sarcasm Detection. In *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies.* Retrieved from https://aclanthology.org/2024.naacl-long.97.pdf

Tian, Y., Xu, N., Zhang, R., & Mao, W. (2023). Dynamic Routing Transformer Network for Multimodal Sarcasm Detection. In *Proceedings of the 61st annual meeting of the association for computational linguistics (acl).* Retrieved from https://aclanthology.org/2023.acl-long.139.pdf

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., & Salakhutdinov, R. (2019). Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th annual meeting of the association for computational linguistics (acl).* Retrieved from https://aclanthology.org/P19-1656.pdf

Valliyammai, C., Monish Raaj, S., Athish, B. L., & Kumar, J. K. (n.d.). *Cyberbullying detection in social media with multimodal data using transfer learning.* College of Engineering Guindy, Anna University, Chennai. Retrieved from https://www.researchgate.net/publication/341912457

Wang, T., Li, J., Su, G., Zhang, Y., Su, D., Hu, Y., & Sha, Y. (2024). *RCLMuFN: Relational Context Learning and Multiplex Fusion Network for Multimodal Sarcasm Detection.* Retrieved from https://arxiv.org/pdf/2412.13008

Yaghoobian, M., Arabnia, H. R., & Rasheed, K. (2023). Advancements in multimodal sentiment

analysis: Techniques and applications. *Journal of Artificial Intelligence Research*, *68*, 345–378. Retrieved from https://jair.org/index.php/jair/article/view/11832

Yoon, S., Byun, S., & Jung, K. (2018). Multimodal Speech Emotion Recognition Using Audio and Text. In *2018 ieee spoken language technology workshop (slt)* (pp. 112–118). Retrieved from https://www.researchgate.net/publication/330398273

Zhang, X., Chen, Y., & Li, G. (2021). *Multi-Modal Sarcasm Detection Based on Contrastive Attention Mechanism.* Retrieved from https://arxiv.org/pdf/2109.15153

Zhang, Y., Zhu, G., Ding, Y., Wei, Z., Chen, L., & Li, K.-C. (2025). A progressive interaction model for multimodal sarcasm detection. *The Journal of Supercomputing*. Retrieved from https://link.springer.com/article/10.1007/s11227-025-07110-3

Zhang, Y., Zou, C., Wang, B., & Qin, J. (2025). *Commander-GPT: Fully Unleashing the Sarcasm Detection Capability of Multi-Modal Large Language Models.* Retrieved from https://arxiv.org/pdf/2503.18681