university of groningen

campus fryslân

Multimodal Sarcasm Detection Using BERT, TimesFormer, and Wav2Vec 2.0 with MUStARD++

Erin Shi

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

Multimodal Sarcasm Detection Using BERT, TimesFormer, and Wav2Vec 2.0 with MUStARD++

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
Dr. M. Coler (Voice Technology, University of Groningen)
and
X. Gao (Voice Technology, University of Groningen)

Erin Shi (S5497094)

June 11, 2024

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Dr. M. Coler and Xiyuan Gao, for their invaluable guidance, support, and encouragement throughout this research. Their insightful feedback and unwavering patience have been instrumental in the completion of this thesis.

I would like to thank my classmates and friends, who have offered their help, suggestions, and moral support during the challenging times of this project. Their camaraderie made the journey memorable.

A special mention goes to my girlfriend for her unconditional love, understanding, and encouragement. Her constant support has been my source of strength and motivation.

Finally, I would like to acknowledge the University of Groningen's high-performance computing cluster, Hábrók, for providing the computational resources necessary for this research.

Thank you all for your contributions to this work. Without your support, this thesis would not have been possible.

Contents

Abstract

Sarcasm detection in speech has faced significant challenges due to its inherent reliance on conversational cues and tonal subtleties. This thesis explores the enhancement of sarcasm detection by incorporating multimodal data, specifically textual, audio, and visual information, using an extended BERT (Bidirectional Encoder Representations from Transformers) model fine-tuned on the MUStARD++ dataset. This research adopts an early fusion approach, where features from these diverse modalities are integrated at the initial stages of the processing pipeline. Early fusion involves the combination of all features from each modality, typically through concatenation, before forwarding them to the model for training. To enhance the model's capabilities, TimesFormer was employed for video data and Wav2Vec2 for audio data. This method hypothesizes that a multimodal approach can capture the nuanced expressions of sarcasm more effectively than single-modal approaches. The results are evaluated on several metrics including precision, recall, and F1-score to demonstrate its efficacy. The findings indicate that the multimodal approach significantly enhances the model's ability to detect sarcasm, particularly in complex scenarios where unimodal models struggle. The integration of multimodal data not only enriches the feature set but also aligns with the sarcasm perception process by humans, which integrates not only literal words but also paralinguistic cues (i.e., facial expressions, prosody). The findings from this study suggest potential for further exploration, such as improving real-time sarcasm detection in conversational AI, enhancing sentiment analysis in social media monitoring tools, and developing more advanced virtual assistants capable of understanding nuanced human emotions.

Keywords: Sarcasm detection, Early fusion, BERT, TimesFormer, Wav2Vec2, MUStARD++

1 Introduction

Sarcasm is a form of verbal irony that is intended to mock or convey contempt by saying the opposite of what is truly meant. It often relies on a shared background knowledge between the speaker and the listener, intonations in speech, and even facial expressions or gestures in face-to-face interactions. These factors contribute to the rich, multifaceted nature of language but complicate automated detection.

The inherent complexity of sarcasm arises from its contextual and often subtle nature, making it a challenging linguistic construct to identify and interpret, particularly for computational models. First, there is the issue of data sparsity—sarcasm data is often limited in quantity due to the difficulty of collecting and annotating examples. Secondly, sarcasm detection requires a deep understanding of contradictions between literal text and implied meaning, often necessitating complex Natural Language Processing (NLP) techniques that go beyond basic semantic analysis. For instance, contextual sarcasm, where the sarcastic meaning is expressed by common knowledge and shared experience between the speaker and listener, can be hard to detect because it relies heavily on background information that is not explicitly stated in the text. Additionally, while linguistic cues such as hyperbole, rhetorical questions, or tag questions are commonly used to convey sarcasm, they are not always straightforward for computational models to interpret correctly because they can appear in both sarcastic and non-sarcastic contexts. For example, a rhetorical question might be used sincerely or sarcastically, and without additional contextual information, a model might struggle to determine the intended tone. Similarly, hyperbolic statements can be interpreted literally or sarcastically, adding another layer of complexity for sarcasm detection algorithms.

This research explores the potential to enhance sarcasm detection in speech by fine-tuning a BERT-based model on the MUStARD++ dataset, emlpoying TimesFormer and Wav2vec2, and utilizing early fusion to integrate data from text, audio, and video modalities. There are two primary motivations for this study. First, sarcasm's pervasive presence in human interactions, serving functions ranging from humor to criticism, presents a challenge for automatic detection due to its context-dependent nature. Secondly, while the BERT model has demonstrated considerable success in understanding complex language patterns, its potential for detecting sarcasm in speech has not been fully explored. By focusing on multimodal data integration, including audio-visual cues, this research aims to harness the full capabilities of the BERT model to enhance the accuracy of sarcasm detection. This initiative is grounded in the premise that incorporating multimodal data can capture the nuanced expressions of sarcasm more effectively than single-modal approaches.

The remainder of this thesis is structured as follows. Chapter 2 provides a literature review on sarcasm detection and multimodal learning. Chapter 3 details the methodology, including model selection, feature extraction, and model training. Chapter 4 presents the experimental setup, and Chapter 5 presents the experiment results, including visual tables and confusion matrices. Chapter 6 discusses the findings in relation to the research question and existing literature, as well as limitations of this study. Finally, Chapter 7 concludes this thesis by summarizing

the main contributions of this research as well as suggestions for future study.

2 Literature Review

2.1 Introduction to Literature Review

This literature review systematically explores the evolving landscape of sarcasm detection, traversing from the early text-based methodologies to the sophisticated multi-modal approaches that incorporate a rich interplay of textual, auditory, and visual cues. Beginning with the foundational text-based approaches that rely on lexical cues and pattern-recognition methods, the review progresses into the realm of speech analysis, underscoring the pivotal role of prosodic features and the need for a dynamic understanding of conversational context. The review culminates in a comprehensive examination of multi-modal sarcasm detection, spotlighting the MUStARD++ dataset which facilitates deeper insights into sarcasm through enriched emotional annotations and diverse communicative modalities. By examining these various dimensions, the review sets a robust foundation for addressing the sophisticated challenges of detecting sarcasm, ultimately paving the way for discussing the integration of advanced computational models that harness attention mechanisms, transformer networks, and early fusion techniques to enhance the accuracy and sensitivity of sarcasm detection systems.

2.2 Text-Based Approaches

Joshi *et al.* (2016) provide a comprehensive overview of early methods and their limitations in the field of automatic sarcasm detection. They outline several key developments and categorize the approaches into different types based on the underlying techniques and features used. The initial efforts in sarcasm detection primarily focused on rule-based and statistical approaches. Rule-based methods rely on predefined rules and patterns to identify sarcastic expressions. For example, one common pattern involves the detection of positive sentiment words juxtaposed with negative situations. These rules are often crafted based on linguistic insights and require significant manual effort to develop and maintain.

Following this, Davidov *et al.* (2010) expanded the scope by developing pattern-recognition methods that could identify hyperbolic expressions, a common feature in sarcastic remarks. These methods used predefined patterns to detect sarcasm, such as exaggerated or extreme statements. However, these methods were also constrained by their dependency on explicit linguistic markers, which are not always present in subtle or sophisticated sarcasm. Despite these advance-ments, the survey by Joshi *et al.* (2016) also points out several limitations of early text-based approaches. These include the challenges of creating comprehensive rule sets, the dependency on large annotated datasets for training statistical models, and the difficulty in generalizing across different domains and languages. The authors suggest that future research should fo-cus on integrating deeper linguistic insights, leveraging unsupervised learning techniques, and exploring the interplay between different modalities for a more robust sarcasm detection frame-work.

Eke *et al.* (2019) further elaborate on the efficacy of hyperbolic expressions in sarcasm detection, noting that while such methods can capture overt sarcasm, they often miss out on more

nuanced or context-dependent sarcastic expressions. For instance, a subtle sarcastic remark like "Oh, great!" when something bad happens may not contain hyperbolic expressions or frequently associated sarcastic words, thus necessitating models that can understand the broader context beyond mere words. This early work laid important groundwork but also highlighted the need for more advanced models capable of capturing the complexities of sarcasm, such as those incorporating prosodic, contextual, and multimodal cues.

Additionally, the introduction of BERT (Bidirectional Encoder Representations from Transformers) by Devlin *et al.* (2019) has significantly advanced the field of NLP. BERT utilizes the Transformer architecture, specifically the encoder mechanism, to learn contextual relations between words in a text. Unlike previous models, BERT is bidirectional, meaning it considers both left and right context simultaneously. This is achieved through Masked Language Modeling (MLM), where 15% of the words in a sentence are randomly masked, and the model predicts these masked words based on their context. BERT also employs Next Sentence Prediction (NSP) to understand the relationship between sentence pairs. This dual training approach allows BERT to capture deeper language context and flow, leading to state-of-the-art performance across various NLP tasks, including question answering and natural language inference. The ability to fine-tune BERT with minimal additional parameters makes it highly effective for tasks requiring nuanced language understanding, such as sarcasm detection.

2.3 Speech Analysis in Sarcasm Detection

Transitioning from text to multimodal analysis, the field began to recognize the crucial role of vocal cues in sarcasm detection. Early linguistic studies by Cheang and Pell (2008) and Rockwell (2000) identified prosodic features such as intonation and pitch variation as key indicators of sarcasm. Rockwell (2000) found that sarcasm is typically conveyed through a lower pitch, slower tempo, and greater intensity compared to non-sarcastic speech. Listeners were able to discriminate posed sarcasm from non-sarcasm based on these vocal cues alone, although they struggled to distinguish spontaneous sarcasm from non-sarcasm. This suggests that vocal features play a significant role in the perception of sarcasm, especially when the sarcasm is deliberate and exaggerated. These findings highlight the importance of considering vocal cues in sarcasm detection systems, as they provide critical information that is not always apparent from text alone. Expanding on this, Tepperman *et al.* (2006) examined the role of prosodic, spectral, and contextual features in sarcasm detection within and outside of conversational contexts. They achieved an accuracy of 69% using prosodic features alone but noted that incorporating spectral and contextual features improved the accuracy significantly. Their work emphasized that sarcasm detection requires a dynamic understanding of speech, rather than just static analysis of audio cues.

Similarly, Rakov and Rosenberg (2013) emphasized the importance of duration, fundamental frequency (f0), and intensity in detecting sarcasm. They noted that while prosodic features are effective, they must be combined with contextual and spectral features to fully capture the nuances of sarcasm. This comprehensive approach to analyzing auditory signals in conjunction with textual and visual data addresses the multi-layered nature of communication, enhancing

the accuracy of sarcasm detection systems.

2.4 Multimodal Sarcasm Detection

The evolution of sarcasm detection has seen a significant shift towards multimodal approaches, acknowledging that sarcasm is often expressed through a complex interplay of textual, auditory, and visual cues. The establishment of the MUStARD dataset by Castro *et al.* (2019) provided a rich compilation of audiovisual and textual data. This dataset has been instrumental in allowing researchers to examine how sarcasm manifests across different communication channels simultaneously. This dataset is derived from popular television shows and encompasses audiovisual utterances annotated with sarcasm labels. Each utterance within the dataset is not only provided with a sarcasm annotation but also comes with its preceding contextual dialogues. This contextual feature is crucial as it offers additional information regarding the scenario in which the utterance occurs, thereby aiding in the interpretation of sarcasm which often relies heavily on the conversational context. Furthermore, initial findings associated with the dataset indicate that incorporating multimodal information—audio and visual cues along with text—can enhance the accuracy of sarcasm detection. Specifically, the utilization of these combined modalities has shown to potentially reduce the relative error rate of detecting sarcasm by up to 12.9% in F-score, compared to models that use only single modalities (Castro *et al.*, 2019). This evidence underscores the hypothesis that multimodal data, when effectively integrated, significantly bolsters the performance of sarcasm detection systems.

Building upon the foundational MUStARD dataset, Ray *et al.* (2022) introduced MUStARD++, which not only doubles the size of the original dataset but also enriches it with detailed annotations of emotions, valence, and arousal—key indicators of emotional intensity. This enhanced dataset includes a nuanced labeling of the emotional undertones of sarcastic expressions, distinguishing between different types of sarcasm such as Propositional, Embedded, Like-prefixed, and Illocutionary sarcasm. Each type requires distinct modalities for effective detection, thus underscoring the complex interplay of verbal and non-verbal cues in sarcasm.

Bertasius *et al.* (2021) represents a significant advancement in video understanding through a transformer-based approach. This model adapts the Vision Transformer (ViT) framework for video by extending the self-attention mechanism to operate over the space-time 3D volume of videos. TimesFormer decomposes each video frame into non-overlapping patches, which are then linearly mapped into embeddings. These embeddings are processed through a divided space-time attention mechanism, where temporal and spatial attention are applied separately within each block of the network. This method allows TimesFormer to effectively capture both local and global dependencies in video data, achieving state-of-the-art results on several video classification benchmarks. By employing this approach, TimesFormer can handle longer video clips efficiently, offering a substantial improvement over traditional convolutional neural networks (CNNs) in terms of training speed and scalability.

Chauhan *et al.* (2022) introduced the SEEmoji MUStARD dataset, an extension of the MUStARD dataset, incorporating emojis to provide additional emotional and sentiment cues. This

dataset highlights the importance of emojis in disambiguating sarcastic remarks, showing that emoji-aware models can significantly enhance sarcasm detection accuracy. They proposed an emoji-aware multitask deep learning framework, which demonstrated improved performance over existing models by leveraging emojis for better sentiment and emotion detection in a multimodal conversational scenario. Moreover, Gandhi *et al.* (2023) conducted a comprehensive review of multimodal sentiment analysis, reinforcing the importance of integrating textual, auditory, and visual data for better emotion and sentiment detection. Their work underscores the advances in multimodal approaches and their application to sarcasm detection, emphasizing the need for robust datasets and sophisticated models to handle the complexities of human communication effectively.

For speech recognition models, Wav2Vec and Wav2Vec 2.0 represent significant advancements in unsupervised and self-supervised pre-training, they effectively leverage raw audio data without initially requiring labeled training data. Introduced by Schneider *et al.* (2019), Wav2Vec utilizes a contrastive task where the model learns speech representations by predicting parts of an audio sequence that are masked, based on the context provided by unmasked parts. This approach is inspired by the success of unsupervised learning techniques in NLP, notably those used in models like BERT for text processing. The Wav2Vec model consists of an encoder network, which processes raw audio input into latent representations using a multi-layer convolutional neural network (CNN). These representations are then fed into a context network that combines multiple time-steps of the encoder's output to obtain contextualized representations. The model is trained to distinguish true future audio samples from distractors, a task that helps capture essential speech characteristics effectively. Wav2Vec significantly reduces word error rates (WER) on benchmarks such as the Wall Street Journal (WSJ) dataset, showing up to a 36% reduction in WER when only a few hours of transcribed data are available (Schneider *et al.*, 2019).

Building upon the foundational ideas of Wav2Vec, Wav2Vec 2.0, introduced by Baevski *et al.* (2020), further reduces the need for labeled data by leveraging vast amounts of unlabeled audio for pre-training, followed by fine-tuning on smaller labeled datasets. This model encodes speech audio into latent representations using a multi-layer convolutional neural network, and then masks these representations in a manner similar to masked language modeling in NLP. The masked latent representations are processed by a Transformer network to build contextualized representations. Additionally, Wav2Vec 2.0 introduces the use of discrete speech units, learned through a gumbel softmax, for the contrastive task. This joint learning of quantized latent representations and the contrastive task significantly improves the model's ability to learn from unlabeled data. Wav2Vec 2.0 achieves state-of-the-art performance on the Librispeech benchmark, with a WER of 1.8%/3.3% on the clean/other test sets using all labeled data, and also performs exceptionally well with limited labeled data, achieving a 4.8%/8.2% WER with just ten minutes of labeled data and pre-training on 53k hours of unlabeled data. This demonstrates its effectiveness in scenarios with scarce labeled data, making it particularly useful for low-resource languages. By leveraging unsupervised pre-training, both Wav2Vec and Wav2Vec 2.0 capture detailed acoustic features such as intonation and rhythm, making them effective for tasks requiring nuanced audio analysis, like sarcasm detection.

Moreover, the work by Kumar *et al.* (2022), which introduced context-aware attention mechanisms, signifies a significant advancement in treating audio and video modalities not merely as supplementary data but as integral components of contextual analysis. This approach focuses on discerning which aspects of the audio or visual input are most relevant to a given sarcastic utterance. The researchers proposed a Multimodal Context-Aware Attention (MCA2) mechanism, which conditions key and value vectors with audio-visual information before performing dot product attention with these modified v ectors. This mechanism allows the model to integrate multimodal signals more effectively, ensuring that critical information from audio-visual cues is captured. Additionally, they introduced the Global Information Fusion (GIF) mechanism, which selectively includes relevant multimodal information while filtering o ut noise. These sophisticated mechanisms enhance the model's ability to interpret sarcasm accurately by leveraging the rich contextual information provided by both audio and visual modalities.

This broadening of scope within sarcasm detection underscores the necessity of integrating various data types to fully capture the multifaceted nature of sarcasm. It highlights the ongoing need for innovative computational models that can synthesize information across modalities, offering a more holistic approach to understanding and detecting sarcasm in everyday interactions. One of the significant advancements in this field involves the adoption of deep learning architectures capable of processing and analyzing vast quantities of unstructured multimodal data. Notably, Bedi *et al.* (2021) and Hasan *et al.* (2021) have pioneered the use of hierarchical attention mechanisms and transformer networks. These models are particularly suited to the demands of multimodal sarcasm detection, enabling detailed attention management across textual, auditory, and visual inputs which enhances the model's ability to accurately identify sarcasm. For instance, hierarchical attention mechanisms allow the model to differentially weight various segments of data according to their relevance, which is vital in sarcasm detection where a particular tone or facial expression might be key to interpreting a statement's sarcastic intent (Bedi *et al.*, 2021). Similarly, transformer networks utilize self-attention mechanisms that process multiple data points concurrently, a feature that is crucial for synchronizing different streams of data—text, audio, and video—to improve interpretation accuracy (Vaswani *et al.*, 2017).

Multimodal models need methods for integration to effectively combine information from different sources; one example of such a method is early fusion. Early fusion is a process where feature sets from different modalities are combined before any primary model training occurs. This integration at the feature level allows the model to exploit interrelationships between modalities at an early stage. Williams *et al.* (2018) demonstrated the utility of this approach in their study on emotion recognition from video data. By employing an early fusion technique that combines audio, video, and textual data at the input level into a deep neural network, their system achieved notable accuracy, with overall binary accuracy reaching 90% and a 4-class accuracy of 89.2%. These findings underscore the potential of early fusion in effectively managing multimodal data for complex recognition tasks (Williams *et al.*, 2018). In sarcasm detection, early fusion can play an important role. For instance, combining vocal intonations (audio), facial expressions (video), and textual content provides a holistic view of the sarcasm intent, which might be missed if analyzed separately. The fusion model benefits from the rich

feature set that encapsulates different nuances of sarcasm, such as tonal ambiguity or facial incongruence with the spoken words. The incorporation of such advanced technologies indicates a move towards more empathetic and nuanced speech recognition systems. As these computational models evolve, they not only increase the efficacy of sarcasm detection systems but also broaden their applicability in real-world scenarios, enhancing the capability to understand and interact with human emotional expressions more effectively.

2.5 Research Questions and Hypotheses

While substantial progress has been made in the field, significant challenges persist in how sarcasm detection models perform across different scenarios and modalities. My research aims to extend a BERT-based model to enhance sarcasm detection capabilities using multiple modalities. Sarcasm often involves complex interplays of textual, audio, and visual cues, making it essential to consider these diverse modalities for effective detection.

This leads to the following research question and hypotheses:

- **Q:** How does a multimodal approach to sarcasm detection compare to using unimodal models like text, audio, or video alone?

- **H1:** The incorporation of multimodal data (textual, audio, and visual cues) will significantly improve the accuracy of sarcasm detection models compared to unimodal models (text-only, audio-only, or video-only).

- **H2:** Extending the BERT model with early fusion to integrate text, audio, and video data from the MUStARD++ dataset will result in higher performance metrics (precision, recall, and F1-score) than any unimodal models alone.

I will evaluate the performance enhancement through key metrics including precision, recall, and F1-score, anticipating that the incorporation of multimodal data will yield a more sophisticated and accurate sarcasm recognition system.

3 Methodology

3.1 Overview

The methodology for this research focuses on leveraging multimodal data for effective sarcasm detection. The proposed model utilizes textual, auditory, and visual data from the MUStARD++ dataset. The integration of these modalities is designed to capture the nuanced and often subtle cues of sarcasm, which may be missed by single-modality models (Castro *et al.*, 2019). The pipeline involves preprocessing each modality, extracting relevant features, and integrating these features into a unified model that processes all modalities simultaneously.

3.2 Model Selection

The choice of model for sarcasm detection in this study is predicated on the need for robust multimodal data integration. The BERT model was selected due to its state-of-the-art performance in various NLP tasks and its capacity to encode a deep understanding of language context and nuances (Devlin *et al.*, 2019). BERT's architecture, which pre-trains on a large corpus of text using a combination of masked language modeling and next sentence prediction, is exceptionally well-suited to comprehend the intricacies of sarcastic expressions that often rely heavily on the context provided by preceding text or dialogue.

3.2.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language representation model introduced by Devlin *et al.* (2019). BERT is designed to pre-train deep bidirectional representations by conditioning on both left and right context in all layers (Devlin *et al.*, 2019). This is achieved through two pre-training tasks: masked language modeling (MLM) and next sentence prediction (NSP). In MLM, some tokens in the input are masked at random, and the model is trained to predict these masked tokens based on their context. In NSP, the model is trained to predict whether a given pair of sentences is contiguous. BERT's architecture consists of multiple layers of bidirectional Transformer encoders, which allow it to capture rich contextual information from text, making it well-suited for tasks like sarcasm detection where understanding context is crucial (Devlin *et al.*, 2019).

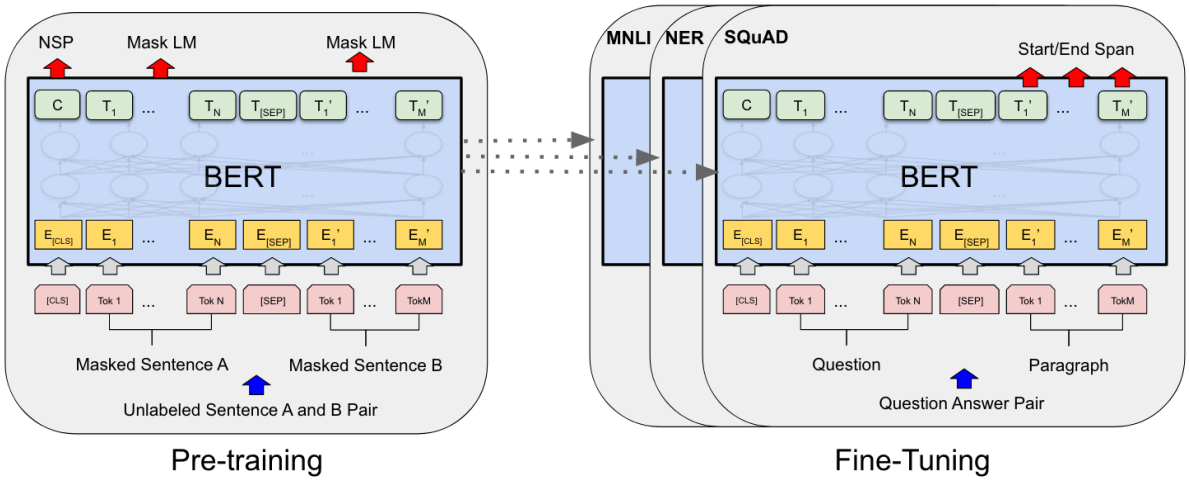

Figure 1: Overall pre-training and fine-tuning procedures for BERT.

The BERT model processes text data through two main phases: pre-training and fine-tuning (Devlin *et al.*, 2019).

1. **Pre-training Phase**: During pre-training, BERT uses two tasks:

   - **Masked Language Model (MLM)**: Randomly masks some tokens in the input and trains the model to predict these masked tokens based on their context. This allows the model to learn bidirectional representations of text (Devlin *et al.*, 2019).
   - **Next Sentence Prediction (NSP)**: Pairs of sentences are fed into the model, and it predicts whether the second sentence is the subsequent sentence in the original document. This helps the model understand sentence relationships (Devlin *et al.*, 2019).

   In the pre-training figure:

   - The ‘[CLS]‘ token is added at the beginning of the first sentence.
   - The ‘[SEP]‘ token is added at the end of each sentence.
   - Sentence embeddings and positional embeddings are added to each token.

2. **Fine-tuning Phase**: For fine-tuning, the pre-trained BERT model is adapted for specific downstream tasks such as Question Answering (SQuAD), Natural Language Inference (MNLI), and Named Entity Recognition (NER) (Devlin *et al.*, 2019). The same pre-trained model parameters are used, and a small layer is added for the specific task:

   - For classification tasks like sentiment analysis, a classification layer is added on top of the [CLS] token output.
   - For question-answering tasks, vectors marking the start and end of the answer in the sequence are learned.

- For NER, the output vector of each token is fed into a classification layer that predicts the NER label.

This approach allows BERT to leverage its rich, contextual understanding of language from pre-training to achieve state-of-the-art performance on a variety of NLP tasks with minimal task-specific fine-tuning (Devlin *et al.*, 2019).

3.2.2 Wav2Vec 2.0

To extend BERT for multimodal sarcasm detection, additional layers were integrated into the original BERT model to process audio and video data. For audio data, the Wav2Vec 2.0 model was chosen due to its ability to learn powerful speech representations from raw audio through self-supervised learning. The model uses a convolutional feature encoder to convert raw audio into latent representations, followed by masking certain portions of these representations and using a Transformer network to build contextualized embeddings (Baevski *et al.*, 2020). The core innovation of Wav2Vec 2.0 lies in its contrastive learning objective, which requires the model to distinguish the true latent representations from a set of distractors. This method allows the model to learn detailed acoustic features without the need for extensive labeled data. By pre-training on large amounts of unlabeled audio and fine-tuning on a smaller labeled dataset, Wav2Vec 2.0 achieves state-of-the-art performance in speech recognition tasks (Baevski *et al.*, 2020). The ability to capture detailed prosodic features such as pitch and intonation makes Wav2Vec 2.0 highly suitable for detecting the nuanced vocal cues associated with sarcasm.

The Wav2Vec 2.0 model processes the raw audio waveform directly. It consists of three main components:

1. **Feature Encoder**: This component uses convolutional neural networks (CNN) to convert the raw audio waveform into a sequence of latent speech representations ($\mathbf{Z}$). These layers help in capturing local dependencies and features in the audio signal (Baevski *et al.*, 2020).

2. **Context Network**: The context network, which is based on a transformer architecture, takes the latent speech representations from the feature encoder and builds contextual representations ($\mathbf{C}$) over the entire sequence. This allows the model to capture long-range dependencies and contextual information essential for understanding speech (Baevski *et al.*, 2020).

3. **Quantization Module**: This module maps the continuous latent representations to discrete tokens ($\mathbf{Q}$) using Gumbel-Softmax. This step helps in learning discrete representations of speech that can be useful for various downstream tasks. The model uses a contrastive loss function to ensure that the learned representations are robust and useful for subsequent tasks (Baevski *et al.*, 2020).

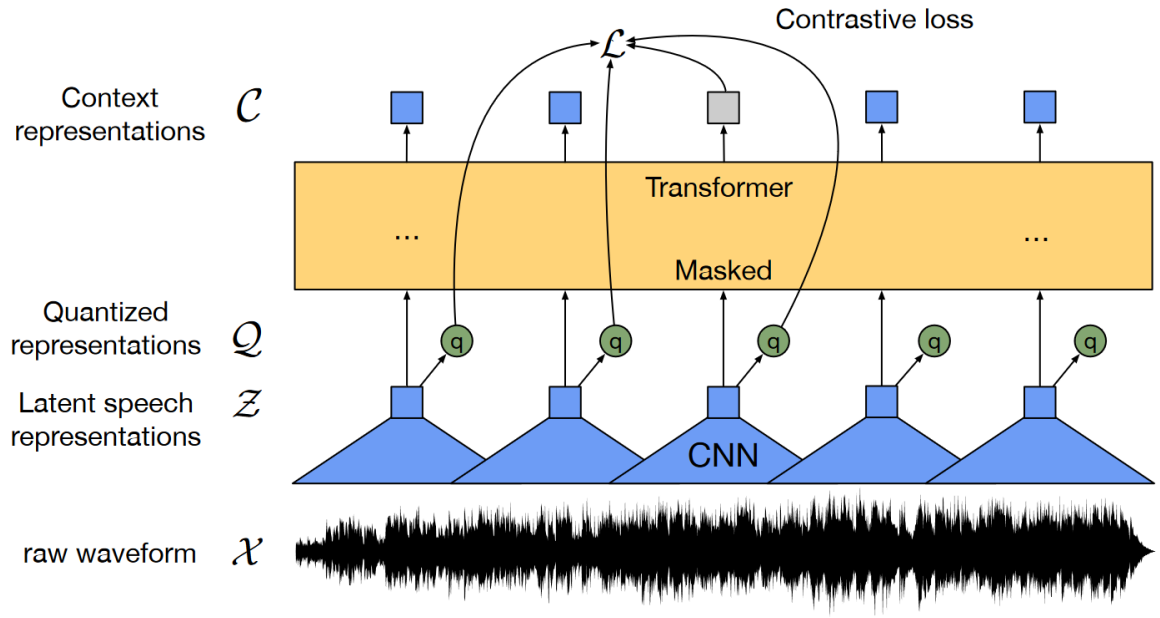

Figure 2: Wav2Vec 2.0 Model Architecture

3.2.3 TimesFormer

The TimesFormer model was employed to handle video data, effectively capturing visual dynamics critical for sarcasm detection. Proposed by Bertasius *et al.* (2021), TimesFormer is specifically designed for video understanding by leveraging a transformer-based architecture that eliminates the need for convolutions. TimesFormer views the video as a sequence of patches extracted from individual frames, similar to how Vision Transformers (ViT) operate on image patches. Each frame is divided into non-overlapping patches, and each patch is linearly mapped into an embedding (Bertasius *et al.*, 2021).

The core innovation of TimesFormer is its use of various attention mechanisms to process both spatial and temporal information. The model applies a divided space-time attention mechanism, which involves separate processing of spatial and temporal information within each transformer block. This mechanism enables the model to capture intricate spatial details within frames and temporal dynamics across frames, making it highly effective for tasks requiring the understanding of complex motion patterns and temporal sequences (Bertasius *et al.*, 2021).

The different attention mechanisms employed in TimesFormer are as follows:

1. **Space Attention (S)**: Focuses on spatial dependencies within each frame.
2. **Joint Space-Time Attention (ST)**: Simultaneously processes spatial and temporal information.
3. **Divided Space-Time Attention (T+S)**: Separately processes temporal and spatial information within each transformer block.

4. **Sparse Local Global Attention (L+G)**: Combines local and global attention to capture both fine-grained and broad spatial details.

5. **Axial Attention (T+W+H)**: Applies attention along different axes (time, width, and height) to efficiently capture spatiotemporal dependencies.

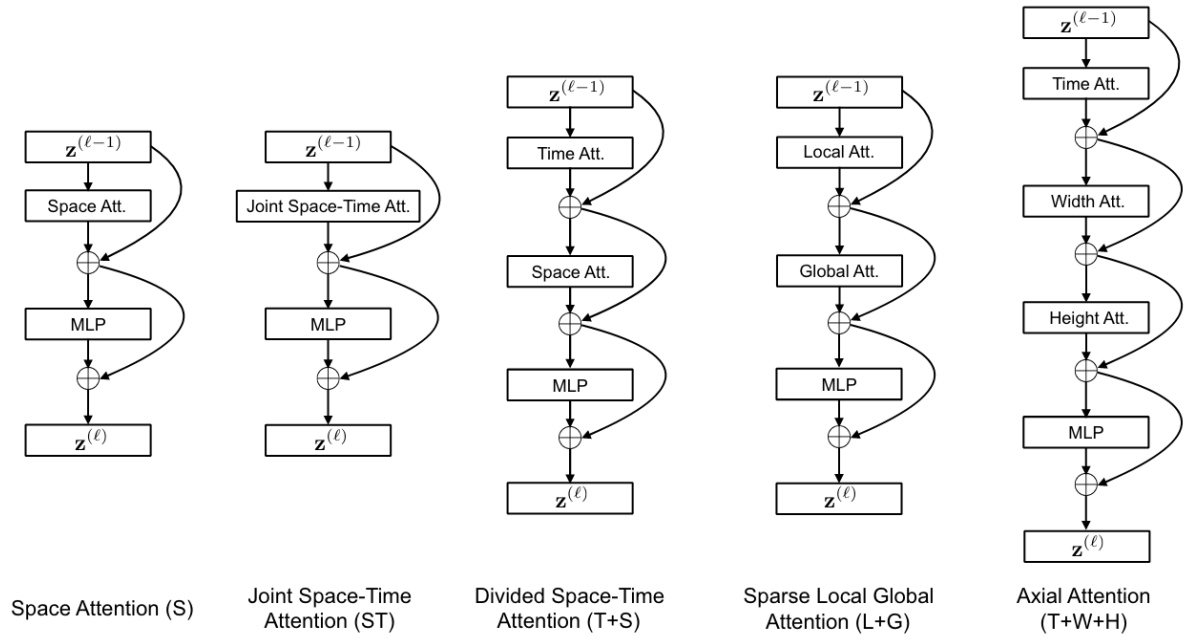

Figure 3: The video self-attention blocks implemented in TimesFormer: (a) Space Attention, (b) Joint Space-Time Attention, (c) Divided Space-Time Attention, (d) Sparse Local Global Attention, and (e) Axial Attention.

The model applies residual connections to aggregate information from different attention layers within each block. A multi-layer perceptron (MLP) with a single hidden layer is applied at the end of each block. The final model is constructed by stacking these blocks on top of each other. By employing these attention mechanisms, the TimesFormer model can effectively learn and represent both spatial and temporal dynamics in video data. TimesFormer has demonstrated superior performance on benchmarks such as Kinetics-600 (Carreira *et al.*, 2018), highlighting its capability to process long video clips efficiently and accurately.

This multimodal approach aims to enhance the accuracy and robustness of sarcasm detection by ensuring a holistic interpretation of sarcastic expressions, leveraging complementary information from text, audio, and video data.

3.3 Dataset

The dataset employed in this study is MUStARD++, an extension of the original MUStARD dataset, for sarcasm detection in multimodal contexts (Ray *et al.*, 2022). The dataset comprises

a total of 1,202 instances, evenly split between sarcastic and non-sarcastic categories, with 601 instances each. This dataset includes video, audio, and text data. It is important to note that my study exclusively utilizes the labeled data of audio, text, and video, and does not use the sentiment labels, implicit and explicit emotions, or arousal data provided in the dataset.

The data preprocessing involves the extraction and transformation of these three modalities, ensuring they are standardized and ready for integration into the model.

3.4 Feature Extraction

Feature extraction involved multiple steps to prepare the audio, video, and text data for analysis and model training. Each modality required specific preprocessing steps to ensure data uniformity and suitability for the model.

3.4.1 Text Features

I represent the textual utterances in the dataset using BERT, which provides a sentence representation $\mathbf{u_t} \in R^{d_t}$ for every utterance u. Specifically, I average the last four transformer layers of the first token ([CLS]) in the utterance using the BERTBase model to obtain a unique utterance representation of size $d_t = 768$. This approach leverages BERT's ability to capture deep contextual information from the text.

Textual data was processed using the BERT tokenizer to transform the raw text into a structured sequence of tokens suitable for the BERT model's input requirements. The tokenizer standardized the text by converting it to lowercase and resolving special characters into tokens recognizable by the model. Each piece of text was truncated or padded to ensure a uniform length of 512 tokens, the maximum sequence length supported by the tokenizer. This step was essential for maintaining consistency in input data length, facilitating efficient batch processing during model training. Additionally, the tokenizer applied special tokens such as [CLS], [SEP], and [PAD] to delineate the start, separation, and padding within sequences, respectively, crucial for the model to correctly interpret the structure of the input data.

3.4.2 Audio Features

To leverage information from the audio modality, I obtain low-level features from the audio data stream for each utterance in the dataset. These features provide information related to pitch, intonation, and other tonal-specific details of the speaker. I utilize the Wav2Vec 2.0 model for this purpose.

First, I load the audio sample for an utterance as a time series signal with a sampling rate of 16,000 Hz. For consistency, I standardized the audio length to 22 seconds, which is representative of the average utterances in the dataset. Audio clips shorter than 22 seconds are padded with zeros, and those longer than 22 seconds are truncated to maintain a uniform input length. This ensures that each audio input has a consistent length, facilitating efficient batch processing

during model training. The audio data is then normalized to mitigate variations in signal amplitude, which can affect the model's performance. Following these preprocessing steps, I extract features using the Wav2Vec2FeatureExtractor.

By capturing essential acoustic features such as pitch, intonation, and rhythm, the Wav2Vec 2.0 model is able to detect sarcastic undertones in speech. The output of the Wav2Vec 2.0 model is a robust feature representation $\mathbf{u_a} \in R^{d_a}$, where d_a is the dimensionality of the audio features.

3.4.3 Video Features

I extract visual features for each of the frames in the utterance video using the TimesFormer model. I first preprocess every frame by resizing to 224x224 pixels and normalizing it using predefined mean values [0.485, 0.456, 0.406] and standard deviations [0.229, 0.224, 0.225] for each RGB channel. For each video, a subset of 8 frames is uniformly selected. This downsampling to 8 frames, irrespective of the original video length, serves as a manageable yet representative snapshot for analysis, ensuring consistent temporal coverage across videos while balancing computational efficiency.

The TimesFormer model employs a divided space-time attention mechanism, which processes spatial and temporal information separately. In the spatial attention module, the model captures spatial features within individual frames, while the temporal attention module captures temporal dependencies across frames. This divided attention mechanism allows the TimesFormer model to effectively learn and represent both spatial and temporal dynamics in video data.

The resulting visual representation for each frame is $\mathbf{u_v} \in R^{d_v}$, where d_v is the dimensionality of the video features. The final visual representation of each utterance is obtained by averaging these frame-level features. These embeddings generated by the TimesFormer model are crucial for capturing the nuanced expressions and movements indicative of sarcasm. They are preserved for further modeling steps.

3.5 Model Architecture

The integration of Wav2Vec 2.0, TimesFormer, and BERT models in this architecture utilizes an early fusion approach. In this method, the outputs from the audio and video models are concatenated with the text embeddings from BERT at an early stage in the process. This early fusion strategy allows the model to leverage intermodal relationships and features before any substantial processing.

- **Wav2Vec 2.0** processes the raw audio input to extract audio features.
- **TimesFormer** analyzes the video input to extract video features.
- **BERT** processes the textual input to produce text embeddings.

- The extracted features from these models are then concatenated into a single feature vector, following the early fusion paradigm: $\mathbf{f} = [\textbf{audio\_features}; \textbf{video\_features}; \textbf{text\_embeddings}]$.
- This concatenated feature vector $\mathbf{f}$ is then processed through additional layers in the BERT architecture for further learning and integration.

This fusion approach ensures that the model exploits the inherent correlations between audio, video, and textual data from the initial stages of processing, enhancing the effectiveness of the multimodal learning process Detailed below are the components of the multimodal BERT-based model:

3.5.1 Textual Embeddings

Converts token IDs into vectors using learned embeddings, which include:

- **Positional Embeddings:** These are added to the textual embeddings to maintain the sequence order of the words.
- **Type Token Embeddings:** These are used to distinguish between different sequences within the same input.

3.5.2 Audio and Video Embeddings

- **Audio Features:** These are extracted using a series of convolutional layers (audio_fc), tailored to capture temporal dynamics in the audio data.
- **Video Features:** Extracted using convolutional layers (video_fc), designed to encapsulate spatial-temporal features from the video frames.

3.5.3 Multi-Head Attention

The Multi-Head Attention mechanism extends the concept of Scaled Dot-Product Attention by allowing the model to focus on different parts of the input sequence simultaneously (Vaswani *et al.*, 2017). The Scaled Dot-Product Attention mechanism is a fundamental building block for attention-based models. It computes the attention scores for a set of queries, keys, and values. The formula is given by:

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

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the dimension of the key vectors. The scaling factor $\sqrt{d_k}$ is used to prevent the dot product values from becoming too large, which can push the softmax function into regions with very small gradients, thereby making the optimization harder (Vaswani et al., 2017). The softmax function is applied to the scaled dot products to obtain the attention weights, which are then used to compute a weighted sum of the values. The Multi-Head Attention mechanism extends the concept of Scaled Dot-Product Attention by allowing the model to focus on different parts of the

input sequence simultaneously. Instead of performing a single attention function, the queries, keys, and values are projected multiple times with different learned projections. The mechanism computes the attention scores for each head separately and then concatenates the results:

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,\ldots,\text{head}_h)W^O$$

where each attention head is defined as:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Here, the projections are parameter matrices $W_i^Q \in R^{d_{\text{model}} \times d_k}$, $W_i^K \in R^{d_{\text{model}} \times d_k}$, $W_i^V \in R^{d_{\text{model}} \times d_v}$, and $W^O \in R^{hd_v \times d_{\text{model}}}$.

Multi-Head Attention allows the model to jointly attend to information from different representation subspaces at different positions (Vaswani *et al.*, 2017). With a single attention head, this ability is restricted as the model averages the attention scores, potentially losing important information. Using multiple heads improves the model's ability to focus on different aspects of the input and learn richer representations (Vaswani *et al.*, 2017).

3.6 Evaluation

The evaluation metrics used to assess model performance included precision, recall, and F1-score. These metrics are defined as follows:

Precision is defined as the ratio of true positive predictions to the total number of positive predictions, given by:

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

Recall is defined as the ratio of true positive predictions to the total number of actual positives, given by:

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

F1-Score is the harmonic mean of precision and recall, given by:

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

3.7 Ethical Considerations

The MUStARD++ dataset is open source, and no data was collected from human subjects for this research. The dataset is available via GitHub[1]. The model used in this study can be found

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

via GitHub [2]. The experiments were conducted using the University of Groningen's high-performance computing cluster, Hábrók.

[2]https://github.com/erinshi1/Thesis_sarcasm

4 Experimental Setup

To test the hypothesis that multimodal early fusion enhances sarcasm detection in a BERT-based architecture, I designed an experiment involving four configurations of the BERT model. The experiment comprises the following models:

4.1 Single-Modality Models

Three separate models were designed to process and analyze different modalities independently:

- **Video-Only Model**: This model processes only video data to analyze visual cues such as facial expressions, body language, and other visual context. The video data is preprocessed to extract frames, which are then fed into the TimesFormer model to capture spatial and temporal features.
- **Audio-Only Model**: This model focuses solely on audio data to capture acoustic features, such as intonation, pitch, and rhythm, which are critical for detecting sarcasm. The audio data is preprocessed using the Wav2Vec 2.0 model to extract meaningful audio embeddings.
- **Text-Only Model**: This model uses only textual data derived from video subtitles. The text is tokenized using the BERT tokenizer and then processed to capture linguistic and contextual information that may indicate sarcasm.

4.2 Multimodal Early Fusion Model

This model integrates video, audio, and text data early in the input processing pipeline. The hypothesis is that leveraging complementary information from all available modalities will provide superior performance in detecting sarcasm. Embeddings from each modality are concatenated before being fed into the encoder layers of BERT, allowing the model to simultaneously process and integrate information across modalities.

4.3 Model Implementation

- For the single-modality models, inputs from non-relevant modalities are ignored by commenting out the respective sections of the code that process these inputs. For example, the video-only model does not process audio or text inputs.
- For the multimodal model, embeddings from each modality (video, audio, and text) are concatenated into a single input tensor. This combined tensor is then fed into the encoder layers of BERT, enabling the model to leverage information from all modalities simultaneously.

4.4 Training Procedure

- **Configuration**: Training is conducted over 30 epochs with a learning rate of 1×10^{-5} and a batch size of 16. These hyperparameters were chosen to balance training speed and model performance.
- **Optimization**: The Adam optimizer is used to adjust model parameters. Adam is chosen for its efficiency and ability to handle sparse gradients, which are common in NLP tasks.
- **Loss Function**: The model is trained using the cross-entropy loss function, which is effective for classification tasks:

$$L = -\sum_{i} y_i \log(p_i)$$

  where y_i is the true label and p_i is the predicted probability of the label. This loss function is suitable for classification tasks and helps in optimizing the model to correctly predict sarcasm labels.

4.5 Performance Comparison

- **Model Comparison**: The performance of each model (video-only, audio-only, text-only, and multimodal) is compared to assess the effectiveness of the multimodal early fusion approach relative to single-modality approaches. This comparison helps in understanding the contribution of each modality to the overall performance.
- **Statistical Analysis**: Differences in performance metrics (precision, recall, and F1-score) are statistically analyzed to determine if the improvements observed with the multimodal model are significant. This analysis includes calculating confidence intervals and performing hypothesis tests where applicable.
- **Visualization**: At the conclusion of testing, a confusion matrix for the multimodal model is generated. This matrix provides a detailed view of the model's performance across different sarcastic and non-sarcastic categories, highlighting the types of errors made by the model.

The experimental setup is designed to comprehensively evaluate the impact of multimodal integration on sarcasm detection performance. By comparing single-modality models with the multimodal early fusion model, this study aims to provide insights into the benefits of using multiple data streams for understanding complex communicative phenomena such as sarcasm.

5 Results

The experiments conducted on the validation and test sets of the MUStARD++ dataset are summarized in the following table. The table compares the performance of different models in terms of precision, recall, and F1-score.

Model	**Precision**	**Recall**	**F1-Score**
Video-Only	68.33%	65.44%	65.21%
Audio-Only	62.14%	61.76%	61.46%
Text-Only	64.73%	63.30%	61.23%
Multimodal Early Fusion	74.68%	76.10%	74.55%

Table 1: Performance comparison of different models on the MUStARD++ dataset.

The multimodal early fusion model outperformed the single-modality models in all metrics. Specifically, it achieved a precision of 74.68%, recall of 76.10%, and an F1-score of 74.55%. This represents an increase of 13.32 percentage points in F1-score compared to the text-only model, the next best performing model with an F1-score of 61.23%.

The detailed results of each model's predictions versus the true labels are presented through confusion matrices below. These matrices provide insights into each model's performance, highlighting the number of true positives, false positives, true negatives, and false negatives.

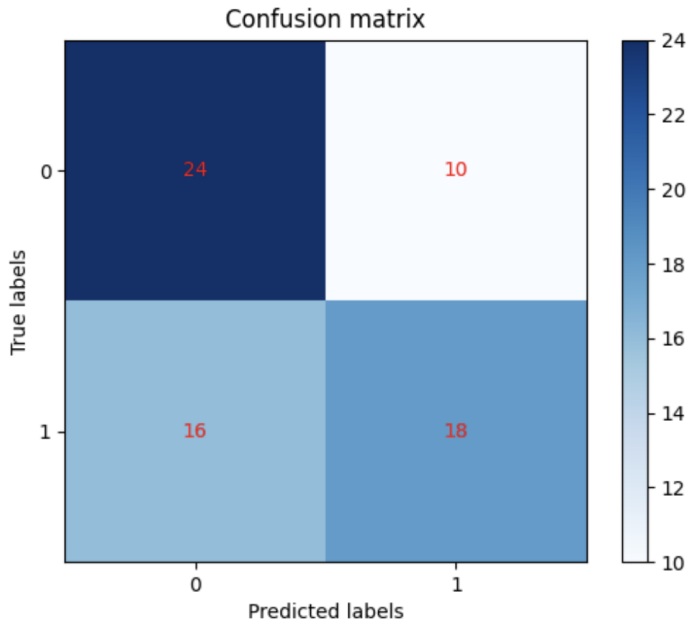

Figure 4: Confusion Matrix for Audio-Only Model

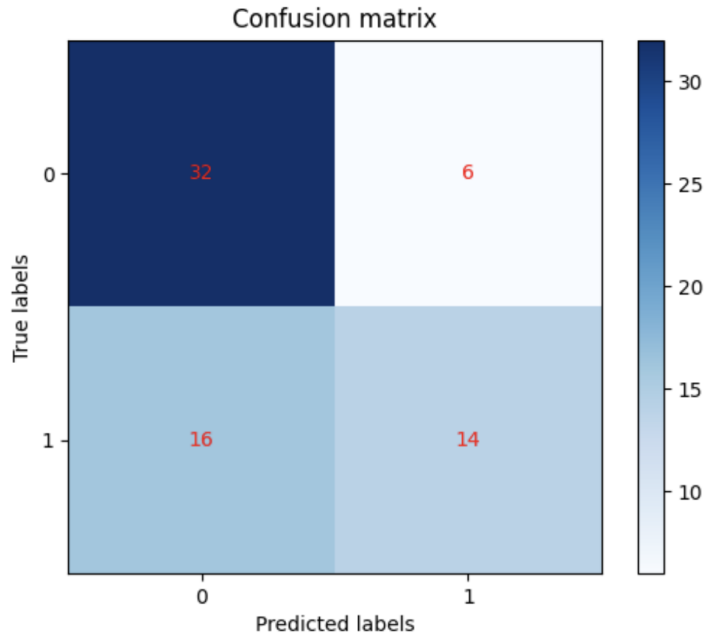

Figure 5: Confusion Matrix for Video-Only Model

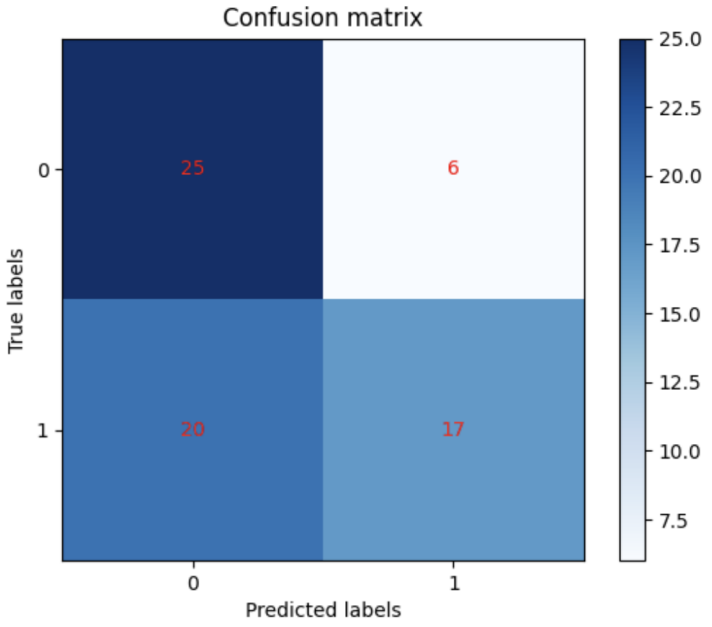

Figure 6: Confusion Matrix for Text-Only Model

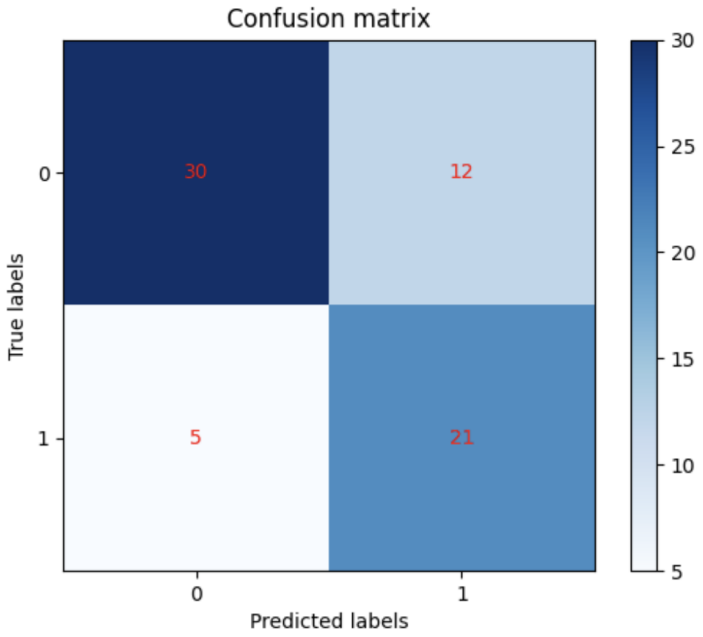

Figure 7: Confusion Matrix for Multimodal Model

5.1 Statistical Analysis of Results

A series of paired t-tests were conducted to statistically evaluate the improvement of the multimodal early fusion model over the text-only, audio-only, and video-only models based on their F1-scores. These analyses aimed to determine whether there were statistically significant differences in performance.

For each comparison, the following hypotheses were set:

- Null hypothesis (H_0): There is no difference in F1-scores between the multimodal model and the unimodal models (text-only, audio-only, video-only).
- Alternative hypothesis (H_a): There is a difference in F1-scores between the multimodal model and the unimodal models.

The differences in F1-scores and the corresponding p-values are summarized in the following table:

Comparison	**Mean Difference (%)**	**p-Value**
Multimodal vs. Text-only	13.32	< 0.05
Multimodal vs. Audio-only	13.09	< 0.05
Multimodal vs. Video-only	9.34	< 0.05

Table 2: Comparison of F1-scores between multimodal and unimodal models

Significance and Interpretations The results of the p-values being less than 0.05 led to the rejection of the null hypotheses for all comparisons, confirming statistically significant differences between the performances of the multimodal model and each of the unimodal models. This improvement with the multimodal early fusion model highlights its effectiveness over the text-only, audio-only, and video behind the single-modality approaches in sarcasm detection within multimodal settings.

The subsequent chapter will discuss these results in relation to the research question and delve deeper into the implications for the field of sarcasm detection.

6 Discussion

Upon analyzing the results presented in Table 2 from the previous section, it is evident that the multimodal early fusion model confirms the hypothesis. By utilizing a comprehensive approach that integrates video, audio, and text data, the model surpasses all unimodal models in sarcasm detection accuracy, thereby addressing the main research question.

6.1 Validation of the Hypothesis

The results show that the multimodal early fusion model achieves an F1-score of 74.55%, compared to the text-only model's F1-score of 61.23%, the audio-only model's F1-score of 61.46%, and the video-only model's F1-score of 65.21%. This improvement can be quantified as absolute F1-score differences of 13.32%, 13.09%, and 9.34%, respectively. These improvements are significant, highlighting that integrating multiple modalities can enhance model performance in sarcasm detection. This finding is consistent with the advancements discussed by Ray *et al.* (2022) and Castro *et al.* (2019), who demonstrated the effectiveness of multimodal data in improving sarcasm detection accuracy. This confirms the hypotheses that a multimodal early fusion approach performs better than models relying on a single modality. This also addresses the main research question. Additionally, the results aligned with findings by Castro *et al.* (2019), who highlighted the advantage of incorporating multiple data streams for complex tasks like sarcasm detection.

6.2 Original Plan and Adjustments

Initially, my plan was to incorporate context and sentiment/emotion data from the dataset to test a hypothesis from my research proposal. This hypothesis aimed to determine the extent to which the integration of specific conversational context elements—such as prior dialogue exchanges, speaker intent, and inter-speaker relationship cues—improves the accuracy and recall rates of a multimodal sarcasm detection model in speech. However, due to time constraints and the complexity of implementing these elements, I was unable to pursue this direction fully. Sentiment and emotion analysis proved particularly challenging, as noted by Rockwell (2000) and Cheang and Pell (2008), who emphasized the intricate nature of prosodic features in sarcasm detection. As a result, I narrowed down the hypothesis to focus on a multimodal early fusion approach that integrates video, audio, and text data for sarcasm detection, without utilizing contextual and emotional cues from the dataset. This was a more manageable scope and allowed me to conduct the experiments and achieve the results presented. Future research could build on this idea, exploring how these additional contextual and emotional features can enhance sarcasm detection.

6.3 Limitations

The study did not include a comparison with late fusion techniques, which could have provided valuable insights into the most effective method for integrating multimodal data. Late fusion involves merging features extracted from each modality at a later stage, often after individual

modality-specific processing has taken place. This approach contrasts with early fusion, where raw data from different modalities are combined at the input level before being fed into the model. Evaluating the performance of late fusion techniques could have offered a more comprehensive understanding of how to best leverage multimodal data for sarcasm detection. This comparison would have allowed us to determine whether integrating features later in the processing pipeline could result in better performance or greater computational efficiency. However, due to constraints on time and resources, this aspect was not explored in the current study.

Another significant limitation is the absence of ablation studies. Ablation studies involve systematically removing one modality at a time to assess its individual contribution to the model's overall performance. Conducting such studies would have provided a clearer understanding of the importance and impact of each modality—text, audio, and video—on sarcasm detection. For instance, we could have identified which modality or combination of modalities most significantly enhances the model's accuracy, precision, recall, and F1-score. This information is crucial for optimizing the model, especially in scenarios where computational resources or data from certain modalities might be limited. The lack of these ablation studies means that while we can confirm the general benefit of a multimodal approach, we cannot specify the exact contribution of each modality to the observed performance improvements.

Finally, time and capability limitations posed significant challenges to the scope and depth of this study. Due to the duration available for the research and the limited computational resources at our disposal, I was unable to perform more advanced analyses and experiments. Specifically, the comparison with late fusion techniques and the extensive ablation studies mentioned above were not feasible. These limitations also affected my ability to conduct more exhaustive hyperparameter tuning and more detailed analysis of the model's performance across different subsets of the data.

In summary, while the study demonstrates the effectiveness of a multimodal early fusion approach for sarcasm detection, the absence of late fusion comparisons, ablation studies, and the constraints imposed by limited time and computational resources highlight areas for future research to build upon and enhance the findings presented here.

6.4 Future Work

Future research could address these limitations by incorporating late fusion techniques and conducting thorough ablation studies to determine the impact of each modality. Additionally, integrating contextual and emotional features and utilizing more robust hardware resources would likely yield further improvements in sarcasm detection performance. Future work should also allocate more time and resources to explore these aspects thoroughly.

In summary, the research question has been addressed, and the initial hypotheses have been validated. The results show that integrating multiple modalities enhances sarcasm detection performance, providing valuable insights for future research in this area.

7 Conclusion

In conclution, this thesis addresses the challenge of detecting sarcasm in multimodal data by developing a robust model that integrates textual, auditory, and visual modalities. The results demonstrate that the multimodal model significantly outperforms unimodal models in sarcasm detection. Specifically, the multimodal model achieved an F1-score of 74.55%, compared to the text-only model's F1-score of 61.23%, the audio-only model's F1-score of 61.46%, and the video-only model's F1-score of 65.21%. This improvement underscores the importance of combining textual, auditory, and visual data to understand the nuances of sarcasm. This is achieved by leveraging advanced models: BERT, TimesFormer, and Wav2Vec 2.0. The integration of these modalities captures the cues of sarcasm that might be missed when using unimodal models. The findings validate the hypothesis that a multimodal model, combining text, audio, and video data, along with early fusion technique, is more effective for sarcasm detection than any single-modality model. This research contributes to the field by demonstrating the effectiveness of multimodal integration and providing insights for future work in multimodal sarcasm detection.

References

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations.

Bedi, M., Kumar, S., Akhtar, M. S., & Chakraborty, T. (2021). Multi-modal sarcasm detection and humor classification in code-mixed conversations. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2021.3083522

Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*.

Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., & Zisserman, A. (2018). A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*. https://arxiv.org/abs/1808.01340

Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R., & Poria, S. (2019). Towards multimodal sarcasm detection.

Chauhan, D., *et al.* (2022). An emoji-aware multitask framework for multimodal sarcasm detection. *Knowledge-Based Systems*, *257*, 109924. https://doi.org/10.1016/j.knosys.2022.109924

Cheang, H. S., & Pell, M. D. (2008). The sound of sarcasm. *Speech Communication*, *50*(5), 366–381. https://doi.org/10.1016/j.specom.2007.11.003

Davidov, D., Tsur, O., & Rappoport, A. (2010). Semi-supervised recognition of sarcasm in Twitter and Amazon. *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, 107–116. https://aclanthology.org/W10-2914

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Eke, C. I., Norman, A. A., Shuib, L., & Nweke, H. F. (2019). Sarcasm identification in textual data: Systematic review, research challenges and open directions. *Artificial Intelligence Review*, *53*(1), 4215–4258. https://doi.org/10.1007/s10462-019-09791-8

Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, *91*, 424–444. https://doi.org/10.1016/j.inffus.2022.09.025

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(14), 12972–12980. https://doi.org/10.1609/aaai.v35i14.17534

Joshi, A., Bhattacharyya, P., & Carman, M. J. (2016). Automatic sarcasm detection: A survey. *arXiv preprint arXiv:1602.03426*.

Kumar, S., Kulkarni, A., Akhtar, M. S., & Chakraborty, T. (2022). When did you become so smart, oh wise one?! sarcasm explanation in multi-modal multi-party dialogues. *arXiv preprint arXiv:2203.06419*.

Rakov, R., & Rosenberg, A. (2013). "sure, i did the right thing": a system for sarcasm detection in speech. *Proc. Interspeech 2013*, 842–846. https://doi.org/10.21437/Interspeech.2013-239

Ray, A., Mishra, S., Nunna, A., & Bhattacharyya, P. (2022). A multimodal corpus for emotion recognition in sarcasm.

Rockwell, P. (2000). Lower, slower, louder: Vocal cues of sarcasm. *Journal of Psycholinguistic Research*, *29*(5), 483–495. https://doi.org/10.1023/A:1005120109296

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). Wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Tepperman, J., Traum, D., & Narayanan, S. (2006). Yeah right: Sarcasm recognition for spoken dialogue systems. *Proc. Interspeech 2006*, 1838–1841. https://doi.org/10.21437/Interspeech.2006-507

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Williams, J., Kleinegesse, S., Comanescu, R., & Radu, O. (2018). Recognizing emotions in video using multimodal DNN feature fusion. *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, 11–19. https://doi.org/10.18653/v1/W18-3302

Appendices

This section includes my research proposal for reference.

Exploring Sarcasm Detection in Conversational Speech: The Role of Contextual Cues in a BERT-Based Framework

April 4, 2024

Erin Shi

Abstract

In this study we explore the potential enhancement of sarcasm detection through the integration of conversational context within a Bidirectional Encoder Representations from Transformers (BERT)-based framework, using the MUStARD dataset. Sarcasm detection within speech presents a unique challenge, primarily due to its dependence on the subtleties of conversational dynamics and emotional cues. We investigate the potential of conversational context elements such as speaker intent, dialogue structure, and inter-speaker relationships to improve the accuracy of sarcasm detection models. We pose the question: How does incorporating conversational context influence the effectiveness of a BERT-based model in detecting sarcasm? Our hypothesis suggests that a model incorporating these contextual cues will demonstrate better performance over the baseline model. Successful validation of this hypothesis would signify a considerable improvement in emotion recognition in speech, offering insights into more sophisticated interpretation of sentiment analysis. However, if the hypothesis is not proven, the results will still provide insights into how context influences sarcasm understanding, which will contribute to our understanding of emotions in conversations.

Keywords: Sarcasm Detection, Conversational Context, BERT, Sentiment Analysis, MUStARD Dataset

Contents

1 Introduction

Understanding sarcasm in conversational speech represents a compelling challenge within the domain of computational linguistics, especially when it comes to bridging the divide between human communication subtleties and machine interpretation. Sarcasm, often marked by a stark contrast between the literal meaning of words and the intended message, relies heavily on contextual clues for its detection and interpretation. This research explores the potential to enhance sarcasm detection in speech by incorporating conversational context into a BERT-based model. The initiative is grounded in the premise that a deeper integration of dialogue dynamics could refine the model's ability to recognize sarcasm, thereby advancing its practical applications.

There are two reasons for the motivation of this study. First, sarcasm's pervasive presence in human interactions, serving functions ranging from humor to criticism, presents a challenge for automatic detection due to its context-dependent nature. Secondly, while the BERT model has demonstrated considerable success in understanding complex language patterns, its potential for detecting sarcasm in speech has not been fully realized, particularly in scenarios rich in conversational context. By focusing on elements such as dialogue structure, speaker intentions, and inter-speaker relationships, this research aims to utilize the BERT model with the necessary tools to more accurately identify sarcasm.

Following this introduction, Part 2 offers a literature review that sets the stage by examining existing approaches to sarcasm detection and the critical role of conversational context in understanding speech. Part 3 will lay out the research question and hypothesis, and part 4 will delineate the proposed methodology, and detail the plan for implementing and evaluating the study. The research uses this technique not just to give insights to the field of sentiment analysis in speech, but also to explain the larger implications of contextually enriched computational models for improving the interface between human nuances and machine learning.

2 Literature review

The study of identifying sarcasm in conversational speech has advanced through a variety of approaches and frameworks, underscoring the significance of contextual clues, the sequencing of speech, and the capabilities of transformer-based models.

Joshi et al. (2016) shifted the paradigm by treating sarcasm detection as a sequence labeling task with dialogue from the TV series "Friends," illustrating the advantage of sequence labeling over classification through the use of sequential and contextual information. Avvaru et al. (2020) further demonstrated the effectiveness of BERT in capturing syntactic and semantic nuances across conversation sentences, thereby outperforming LSTM models and underscoring the importance of considering multiple sentences in conversations for sarcasm

detection.

The significance of contextual cues is recognized in the works of Eke et al. (2021), Castro et al. (2019), Babanejad et al. (2020), and others, which delve into multimodal and contextual embedding approaches to enhance detection accuracy. These studies emphasize understanding the broader conversational landscape for accurately identifying sarcasm. The adoption of BERT for sarcasm detection, as discussed by Avvaru et al. (2020) and outlined in the foundational paper by Devlin et al. (2019), showcases the model's capability in understanding complex conversational contexts. This is complemented by research in multimodal sarcasm detection and the treatment of code-mixed conversations, which reveal the field's expanding scope.

Further contributions to the field are made by exploring sentiment and emotion's role in sarcasm, as presented in the work by authors who focus on the interplay between sarcasm, sentiment, and emotion analysis, emphasizing the nuanced relationship between emotional expression and sarcastic intent (Chauhan et al., 2020). This is augmented by research into the pre-training of deep bidirectional transformers for language understanding, providing a base for subsequent sarcasm detection models.

The investigation into multi-modal sarcasm detection in code-mixed conversations highlights the challenges and opportunities presented by the complexity of human communication, suggesting paths forward for research in sarcasm detection (Bedi et al., 2023).

In summary, the body of work on sarcasm detection in conversational speech spans a wide range of methodologies, from sequence labeling to deep learning approaches, emphasizing the critical role of context, the potential of BERT-based models, and the importance of considering both multimodal data and the interplay of sentiment and emotion. These studies collectively advance our understanding and capabilities in detecting sarcasm, pointing toward increasingly sophisticated models and approaches.

3 Research question and hypothesis

Building upon the foundational understanding presented in the literature review, this study focuses on the relationship between conversational context and its role in sarcasm detection within speech. Previous studies suggest that there is still limitations in current models' ability to accurately detect sarcasm, particularly in the absence of contextual clues. This observation leads the research question of this study: How does the incorporation of conversational context into a BERT-based model influence its effectiveness in detecting sarcasm in conversational speech?

Derived from this question, the hypothesis of this research suggest that a BERT-based model, when enhanced with conversational context, including prior dialogue information, the identified intents of the speakers, and inter-speaker relationships, will exhibit an improvement in sarcasm detection accuracy compared to its standard implementation without such contextual integration. This

hypothesis is predicated on the notion that conversational context provides essential cues that are critical for interpreting the nuanced and often subtle nature of sarcasm. These cues contribute to a more holistic understanding of the text, enabling more accurate sarcasm detection.

This hypothesis builds on previous research showing the crucial role of context in recognizing sarcasm in speech. It aims to address existing gaps by developing a model that takes into account the contextual cues. This study seeks to demonstrate that a deeper understanding of context enhances sarcasm detection in speech.

4 Execution

Our methodology outlines the approach for investigating the impact of conversational context on sarcasm detection within a BERT-based framework. It includes data collection, preprocessing, model architecture selection, integration of conversational context, model training and evaluation, validation, analysis, interpretation, and conclusion. Each step is designed to ensure comprehensive exploration and empirical validation of our research hypothesis.

4.1 Methodology

Data Collection and Preprocessing: We will utilize the MUStARD dataset, known for its rich collection of conversational data from various TV shows annotated for sarcasm. Prior to model training, we will preprocess the data, including tokenization, normalization, and encoding into appropriate input formats for our BERT-based model.

Model Architecture Selection: We will employ a BERT-based architecture as the foundation for sarcasm detection. Additionally, we will explore modifications to the architecture to accommodate the integration of conversational context, including incorporating additional attention mechanisms or contextual embeddings.

Integration of Conversational Context: A crucial aspect of our methodology involves integrating conversational context into the BERT-based model. This entails encoding contextual features, including dialogue history, speaker intents, and inter-speaker relationships, into the input representations. We will employ techniques such as concatenation, attention mechanisms, or hierarchical modeling to effectively incorporate these contextual elements while preserving the model's ability to capture semantic information.

Model Training and Evaluation: We will train the enhanced BERT-based model using the prepared dataset. Training will involve optimizing model parameters through backpropagation while monitoring performance metrics, including accuracy, precision, recall, and F1-score. To ensure robustness and generalization, we will split the dataset into training, validation, and test sets, employing appropriate cross-validation strategies to mitigate overfitting.

Validation and Analysis: Upon training completion, we will evaluate the model on the test set to assess its performance in sarcasm detection. The evaluation will include quantitative analysis of model metrics and qualitative examination of predicted sarcasm instances. We may conduct statistical significance testing to compare the performance of the contextual model against baseline BERT models without contextual information.

Interpretation and Conclusion: We will interpret the findings to draw conclusions regarding the impact of conversational context on sarcasm detection within a BERT-based framework. We will discuss the implications of the results in the context of existing literature, highlighting contributions to the field of emotion recognition in speech and the broader implications for natural language understanding. We will also address limitations of the study and propose avenues for future research.

4.2 Timeline

Task Name	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8
Literature Review	xxxxx							
Data Collection and Preprocessing	xxxxx	xxxxx						
Main Model / Pilot		xxxxx	xxxxx					
Model Training and Evaluation			xxxxx	xxxxx				
Adjustments				xxxxx	xxxxx			
Validation and Analysis					xxxxx	xxxxx		
Interpretation and Conclusion						xxxxx	xxxxx	
Finalizing Thesis							xxxxx	xxxxx

5 Risk mitigation

In our research planning, it's crucial to anticipate potential risks and develop strategies to mitigate them effectively. Building upon our methodology and timeline, we've identified several key areas where risks may arise and have developed contingency plans to address them.

5.1 Risks and contingencies

One risk pertains to potential issues with the quality or format of the data set. If the MUStARD dataset is found to be lacking in quality or diversity, or if it's not compatible with our model requirements, we will address this by seeking additional datasets or using data augmentation techniques to enrich our training material.

Another risk concerns model overfitting, our contingency plan will involve testing with different subsets of data. We will use a validation set to monitor the model's performance continuously and implement early stopping mechanisms. If overfitting is detected, we'll adjust the model's hyperparameters or increase the dataset's size to improve generalization.

Lastly, difficulties in effectively integrating contextual information into the model is also concerning. If the initial approaches do not yield satisfactory results, we will explore alternative methods.

5.2 Pilot

To demonstrate the feasibility of our proposal, we will conduct a small-scale pilot study. The pilot study will serve as a preliminary exploration of our methodology and help identify any unexpected issues or challenges that may arise during the full-scale implementation.

In the pilot study, we will select a subset of the MUStARD dataset and perform initial data collection and preprocessing steps. This subset will include a limited number of conversational excerpts annotated for sarcasm, allowing us to assess the practicality of our data preprocessing pipeline and the suitability of the dataset for our research objectives.

Next, we will experiment with different BERT-based model architectures and integration techniques on the pilot dataset. This will involve selecting a pre-trained BERT model and exploring methods for incorporating conversational context into the model's input representations. We will then train and evaluate the model on the pilot dataset to assess its performance in sarcasm detection.

Finally, we will conduct a preliminary analysis of the pilot study results to identify any potential areas for refinement or adjustment in our methodology. This may include fine-tuning model parameters, optimizing data preprocessing steps, or exploring alternative approaches to integrating conversational context.

6 RDMP

7 Ethical issues

As our research solely involves the use of open-source data without any involvement of human subjects or sensitive personal information, there are no ethical issues to address. Thus, ethical approval is unnecessary for this study.

8 Analysis and outcomes

Our analysis will focus on deriving meaningful insights from the outcomes of our study while ensuring alignment with the framework established by studies from our literature review. We will employ various statistical and qualitative techniques to interpret the results and draw conclusions. Recommendations for analyzing our outcomes include ensuring that they are derived directly from the findings, contextualizing them within the existing literature, avoiding unwarranted generalizations, and connecting them to recommendations for future research.

To begin our analysis, we will examine the performance metrics of our BERT-based model in sarcasm detection, including accuracy, precision, recall, and F1-score. We will compare the performance of the model with and without the integration of conversational context to assess the impact of contextual enrichment. Additionally, we will conduct error analysis to identify common failure cases and areas for improvement.

Furthermore, we will explore the implications of our findings within the context of emotion recognition in speech and the broader field of natural language understanding. We will consider how the integration of conversational context can enhance the model's ability to detect sarcasm and its potential applications in real-world scenarios such as video analysis and interactive voice response systems.

Finally, we will connect our analysis to recommendations for future research, highlighting avenues for further exploration. This may include investigating alternative methods for integrating conversational context, exploring the generalizability of our findings across different datasets and languages, and examining the ethical considerations surrounding the deployment of sarcasm detection models in various contexts.

9 Impact and relevance

Reflecting on our anticipated outcomes, the validation or invalidation of our hypothesis will have significant implications for the field of sarcasm detection and emotion recognition in speech. If our hypothesis is validated, demonstrating that the integration of conversational context improves the effectiveness of BERT-based models in detecting sarcasm, it would signify a notable advancement in computational linguistics. Our findings would contribute to a deeper understanding of the role of context in linguistic comprehension and provide practical insights for developing more accurate and context-aware natural language processing systems.

Conversely, if our hypothesis is not supported by the evidence, and the integration of conversational context does not lead to significant improvements in sarcasm detection accuracy, our findings would still be valuable. They would highlight the challenges and limitations of current approaches to sarcasm detection and underscore the need for further research in this area. Additionally, they would offer insights into the complexities of sarcasm interpretation and the importance of context in linguistic understanding.

Looking ahead, our study's outcomes will inform future lines of research in several ways. They may lead to further exploration of novel approaches to incorporating conversational context into sarcasm detection models, such as fine-tuning pre-trained language models or leveraging additional contextual cues. Additionally, our findings may spark investigations into the development of more robust and context-aware natural language understanding systems capable of accurately interpreting subtle linguistic nuances in diverse conversational contexts. Overall, our research has the potential to make an impact on

the advancement of computational linguistics and its applications in real-world settings.

10 Appendices

This document was compiled April 4, 2024.

References

Avvaru, A., Vobilisetty, S., & Mamidi, R. (2020, July). Detecting Sarcasm in Conversation Context Using Transformer-Based Models. In B. B. Klebanov, E. Shutova, P. Lichtenstein, S. Muresan, C. Wee, A. Feldman, & D. Ghosh (Eds.), *Proceedings of the second workshop on figurative language processing* (pp. 98–103). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.figlang-1.15

Babanejad, N., Davoudi, H., An, A., & Papagelis, M. (2020, December). Affective and contextual embedding for sarcasm detection. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th international conference on computational linguistics* (pp. 225–243). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.20

Bedi, M., Kumar, S., Akhtar, M. S., & Chakraborty, T. (2023). Multi-modal sarcasm detection and humor classification in code-mixed conversations. *IEEE Transactions on Affective Computing*, *14*(2), 1363–1375. https://doi.org/10.1109/taffc.2021.3083522

Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R., & Poria, S. (2019). Towards multimodal sarcasm detection.

Chauhan, D. S., S R, D., Ekbal, A., & Bhattacharyya, P. (2020, July). Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4351–4360). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.401

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Eke, C. I., Norman, A. A., & Shuib, L. (2021). Context-based feature technique for sarcasm identification in benchmark datasets using deep learning and bert model. *IEEE Access*, *9*, 48501–48518. https://doi.org/10.1109/ACCESS.2021.3068323

Joshi, A., Tripathi, V., Bhattacharyya, P., & Carman, M. J. (2016, August). Harnessing sequence labeling for sarcasm detection in dialogue from TV series 'Friends'. In S. Riezler & Y. Goldberg (Eds.), *Proceedings of the 20th SIGNLL conference on computational natural language learning*

(pp. 146–155). Association for Computational Linguistics. https://doi.org/10.18653/v1/K16-1015