university of groningen

campus fryslân

A Lightweight Multimodal Framework for Context-Aware Punchline Detection

Yinzi Wang

university of groningen

campus fryslân

University of Groningen - Campus Fryslân

A Lightweight Multimodal Framework for Context-Aware Punchline Detection

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(Voice Technology, University of Groningen)
with the second reader being
Dr. Joshua Schäuble (Voice Technology, University of Groningen)

Yinzi Wang (S5933986)

June 11, 2025

Acknowledgements

First of all, my earnest appreciation comes to my supervisor, Xiyuan Gao, who has always given me clear guidance and valuable enlightenment in the process of thesis writing. It is Xiyuan's encouragement, helpful instruction and patience all through the tutoring that helps me to finish my paper with joy and achievements. I shall extend my thanks to Xiyuan Gao for her kindness and support.

I also acknowledge the Center for Information Technology of the University of Groningen for their technical support and for providing access to the Hábrók high-performance computing cluster.

Additionally, I would like to extend my heartfelt thanks to my fellow students for the engaging discussions and the shared sense of motivation. Their supports and collaborative spirit have been an important part of my academic experience. Special thanks go to Sixing Mi, whose encouragement sustained me during moments of doubt. Her unwavering support and help throughout this journey have been truly invaluable.

Lastly, I would like to express my gratitude to everyone who, in ways big or small, has contributed to my academic journey and supported the successful completion of this thesis.

Abstract

This thesis proposes a lightweight multimodal framework for punchline detection in spoken dialogue, aiming to enhance computational efficiency while maintaining classification accuracy. The architecture integrates three types of input features: (1) Textual representations from a pretrained ALBERT model, which incorporate both the punchline and its preceding conversational context; (2) Acoustic features derived from COVAREP, including pitch (F0), energy, harmonics-to-noise ratio, glottal parameters and so on; and (3) Humor-centric features (HCF), a handcrafted set of syntactic, semantic, and affective indicators empirically associated with humorous delivery. The model employs a cross-attention mechanism to align information across modalities, followed by max-pooling and a lightweight Multi-Layer Perceptron (MLP) classifier. Its design prioritizes low computational overhead, making it well-suited for deployment in latency-sensitive or resource-constrained environments.

Experiments conducted on the UR-FUNNY dataset demonstrate the effectiveness of the proposed model, which achieves an accuracy of 72.33% and an F1-score of 0.7231. To assess the relative contribution of each modality, we conduct ablation studies by removing one modality at a time. When acoustic features are excluded, the F1 score drops to 0.6504, indicating the importance of acoustic information in humor detection. Removing contextual input also results in a notable decline, with the F1 score decreasing to 0.6523. In comparison, the exclusion of HCF features causes a smaller reduction, with the F1 score falling to 0.6927. These results highlight the complementary nature of semantic, prosodic, and structurally-informed cues in spoken humor recognition. Overall, the proposed model offers a practical and interpretable approach to multimodal humor detection, contributing toward the development of more nuanced conversational AI systems.

Contents

1 Introduction

Humor is a core component in human communication. It can create a relaxed and pleasant atmosphere for dialogue, with both social and practical functions. Research has shown that humor can be used to grab attention, establish rapport, build trust, and boost persuasive power Choube and Soleymani (2020). In addition to its communicative value, humor contributes to social cohesion, reduces conversational tension, and can positively influence psychological well-being M. Xu, Chen, Lian, and Liu (2023). Given these benefits, the recognition and understanding of humor hold great research value in the field of artificial intelligence, especially for speech systems that aim to simulate social adaptability and emotionally intelligent interactions. In practice, the integration of humor detection and artificial intelligence presents promising applications in various fields: (1) It supports the development of robots, virtual assistants, and other human-computer interaction systems that can engage users in a more natural and compelling manner. By incorporating humor recognition, these systems are able to increase user satisfaction and reduce miscommunication, which in turn improves commercial value and economic benefits. (2) With more accuracy punchline detection, ASR pipelines can improve subtitle timing and alignment, aiding humor comprehension for non-native and hearing-impaired users, and potentially boosting audience retention and streaming platform revenue. (3) Humor recognition in assistive technology and companion robotics facilitates socially sensitive and emotionally acceptable reactions, especially in daily and long-term interaction settings. This can help build long-term trust and engagement in human-AI interaction.

In recent years, humor recognition has emerged as a popular research topic in the field of natural language processing. Research on humor recognition has experienced many stages, evolving from rule-based approaches with handcrafted features to more recent deep learning and multimodal systems. Early studies in humor detection mainly relied on handcrafted features derived from linguistic humor theories, which were used with shallow classifiers such as Naive Bayes and SVMs. These features include syntactic patterns (e.g., specific part-of-speech sequences), lexical ambiguity (e.g., pun-related or polysemous words), and incongruent semantic associations (e.g., surprising or contextually unexpected terms). Such handcrafted features aim to capture incongruity, ambiguity, and surprise—three key mechanisms in textual humor—and demonstrate the essential role of the text modality in humor recognition. Yang, Lavie, Dyer, and Hovy (2015)proposed a supervised classification framework using handcrafted features based on four linguistic humor theories. As neural methods gained traction, Chen and Soo (2018) were one of the first to apply deep learning architectures to humor detection, which inspired further developments using hybrid models like CNN-LSTM and highway networks to boost accuracy. Building on these developments, researchers later turned to Transformer-based models, which have great advantages in capturing long-range dependencies within humorous text (Weller & Seppi, 2020).

Although these earlier efforts primarily concentrated on textual features, recent advances have shifted toward multimodal approaches to better reflect how humor is conveyed in real-world communications. In real-world communication, humans convey meaning through multiple modalities. Humor does not only come from text information; speech prosody, facial expressions, and body movements are also involved. Therefore, integrating non-text modalities is crucial for effective humor recognition. The UR-FUNNY dataset proposed by Hasan et al. (2019) marked a turning point to multimodal-based research, which means that humor recognition research start to integrate textual, auditory, and visual modalities. Based on this dataset,Hasan et al. (2019) developed a Contextual Memory Fusion Network (C-MFN) that extends MFN by integrating sequential context through

LSTM-based unimodal encoders and Transformer self-attention, with context representations initializing MFN's memory components. The model encodes punchlines using MFN augmented with contextual memories, and classifies humor based on the combined state of LSTMs and gated memory, excelling in tasks requiring long-range context modeling. Choube and Soleymani (2020) proposed the HF (Hierarchical Fusion) model on the UR-FUNNY dataset, leveraging GRU-based contextual modeling and hierarchical weighted fusion of text, audio, and video to explicitly capture bimodal and trimodal interactions. More recently, Hasan et al. (2021) further advanced this direction by proposing a more semantically enriched framework—the Humor Knowledge Enriched Transformer (HKT). Their model fused Transformer-based representations from text, audio, and visual modalities using a Bimodal Cross-Attention mechanism, which performance strongly on UR-FUNNY dataset. Building on this line of research, M. Xu et al. (2023) developed the MuSE-Humour system, which leverages pseudo labeling and contextual modeling to improve performance on spontaneous speech.

As Zhou (2024) notes in a recent survey, although acoustic shifts, timing, and delivery are of great theoretical importance, they still remain underexplored in spoken humor research. Meanwhile, it is still a great challenge for most humor recognition models to deploy in real-time or resource-limited conditions for their complex structure. These methodological limitations highlight the need for lightweight, context-aware and multimodal models that can capture how punchlines are delivered in real-world speech.

To address these challenges, this study proposes a lightweight, context-aware, and multimodal framework that integrates textual, acoustic, and humor-centric features (HCF). To achieve this, the model is designed to be lightweight by removing the visual modality and using efficient transformer-based encoders with reduced parameter sizes. The model utilizes a pretrained ALBERT encoder to extract contextualized representations from the punchline and its preceding utterances. At the same time, acoustic features extracted from speech are encoded via a transformer-based acoustic encoder, and HCF are encoded using a separate transformer-based module. To model the interaction between modalities, we introduce a cross-attention mechanism in which the textual and HCF sequences jointly attend to the acoustic representation. This allows the model to capture fine-grained dependencies between language and prosodic signals. The resulting cross-modal representations are then combined with global summary vectors from each modality through a pooling-based fusion strategy, producing a unified representation for classification. This unified representation is passed to a fully connected layer to perform binary humor classification. The experiments are conducted on the UR-FUNNY dataset, leveraging aligned textual and acoustic inputs. Results demonstrate that the proposed model achieves a strong balance between accuracy and efficiency, making it well-suited for real-world deployment in resource-constrained settings.

Now that the motivation for this research has been presented, the structure of this thesis is as follows:

- Section 1.1 presents the research questions and hypotheses
- Section 2 reviews relevant literature and positions this work within current research
- Section 3 describes the methodological approach
- Section 4 details the experimental setup
- Section 5 presents and analyzes the results

- Section 6 discusses insights and the role of different components
- Section 7 concludes challenges, limitations, future directions and implications

1.1 Research Questions and Hypotheses

In light of the preceding discussion, this research addresses the following question:

> **How does integrating acoustic features (such as pitch (F0), energy, glottal parameters) into a lightweight, context-aware punchline detection model influence its binary classification performance, as evaluated by F1-score and Accuracy?**

From which the following subquestions are derived:

- How does incorporating contextual features influence the model's ability to detect humor?
- How does incorporating HCF influence the model's ability to detect humor?

[Hypothesis: Based on Choube and Soleymani (2020) findings that, in a context-aware hierarchical fusion architecture, integrating acoustic features with textual input improved punchline classification accuracy from 64.72% to 66.68% on the UR-FUNNY dataset—an absolute gain of 1.96 percentage points, we hypothesize that integrating acoustic features into a lightweight, context-aware punchline detection model will yield improvement in binary classification performance, as measured by F1 score and accuracy, compared to a text-only baseline.]

2 Literature Review

This section provides a comprehensive review of the existing research related to automatic punchline detection, with a particular focus on context-aware punchline detection using multimodal features. By conducting a thorough and critical analysis of the literature in this domain, this review attempts to offer valuable insights into the methods and effectiveness of applying multimodal modeling strategies to improve humor understanding in spoken dialogue. Special attention is given to approaches that incorporate acoustic and contextual inputs with textual input to enhance punchline classification accuracy.

This section is structured as follows. First, I outline the keywords and search strategies used in the literature review as well as the inclusion/exclusion criteria employed to select the most pertinent studies. Following this, I provide a thematic synthesis of the key contributions in the field of humor detection. The review begins by introducing commonly used humor-related corpora, followed by early text-based approaches and their limitations in modeling the complexity of humor. It then turns to state-of-the-art multimodal methods, with a particular focus on how acoustic and contextual signals are incorporated to enhance performance. Finally, the review highlights persisting limitations and open challenges in current research.

2.1 Search Strategy and Selection Criteria

To ensure comprehensive literature coverage, a multi-stage search procedure was conducted across major databases: Google Scholar, IEEE Xplore, ACL Anthology, EMNLP,INTERSPEECH, ICASSP, Scopus, and Web of Science. The search process followed a structured approach with detailed logging of keywords, filters, and screening criteria to support replicability.

- **Primary Search (Broad Field of Humor Detection)**

  Keywords: (”humor detection” OR ”humor recognition” OR ”punchline detection”)

  Purpose: Identify foundational work in automatic humor recognition

- **Secondary Search (UR-FUNNY and Multimodal Fusion)**

  Keywords: (”multimodal” OR ”acoustic features”) AND (”humor” OR ”punchline”)

  Focus: Studies involving audio or acoustic delivery features

- **Tertiary Search (Lightweight & Context Modeling)**

  Keywords: (”lightweight model” OR ”context-aware”) AND (”humor” OR ”punchline”)

  Goal: Identify recent studies on context-aware, lightweight humor detection

To streamline the paper selection process, the retrieved studies were organized by their relevance to these specific research topics. While numerous studies were retrieved during the literature search, not all were directly relevant to the goals of the current investigation. To ensure a focused and coherent review, specific selection criteria were established. First, studies that addressed only text-based humor detection—without incorporating acoustic or contextual elements—were excluded, as they fall outside the scope of this multimodal research. Second, in order to reflect recent technological developments, only peer-reviewed works published from 2010 onward were included. This time frame was chosen to capture contemporary advancements in speech-based humor recognition and multimodal modeling. By applying these filters, the resulting body of literature remains both current and thematically aligned, thereby offering a solid foundation for the present study, which seeks to improve spoken humor recognition through the integration of context and acoustic features within a lightweight neural framework.

Next, I outline the inclusion and exclusion criteria used to select the literature reviewed in this study. The inclusion criteria were as follows: (1) studies that focused on punchline or humor detection in spoken language, where delivery and timing play a significant role in humor perception; (2) empirical research incorporating acoustic features or applying acoustic analysis in humor classification; (3) studies that employed context-aware modeling approaches, such as GRUs or transformer-based architectures, in spoken humor detection tasks; (4) papers reporting quantitative evaluation metrics (e.g., accuracy, F1-score), especially those using multimodal datasets such as UR-FUNNY.

The exclusion criteria were as follows: (1) non-peer-reviewed materials, including blog posts, preprints, or informal reports without formal evaluation procedures or reproducibility; (2) studies that focused solely on clean text-based humor classification without consideration of prosody, speech delivery, or conversational context; (3) research limited to laughter detection or sentiment analysis lacking punchline-level annotation or interpretive modeling; (4) works relying exclusively on visual modalities (e.g., facial expressions) without incorporating linguistic or acoustic components; (5) studies based on non-English corpora that are not transferable to English-based ASR and acoustic modeling systems.

By applying these criteria, the selected literature is both timely and consistent with the objectives of this study. This targeted selection provides a solid foundation for understanding the current state of multimodal, context-aware punchline detection and provides a reference for the design of a lightweight framework that integrates acoustic and contextual signals to improve spoken humor recognition.

Based on these criteria, the selected literature is grouped thematically to reflect the main trends and methodologies in spoken humor detection. Each of the following subsections (2.2-2.3) focuses on a different research perspective.

2.2 Humor-Related Corpora

In natural language processing (NLP) research, corpora are important resources for model training and evaluation. Especially for the complex task of humor recognition, a suitable corpus is of paramount importance. A suitable corpus should not only cover various forms of humor, but also provide appropriate annotations to accurately identify humorous content. However, the subjectivity and diversity of humor make it particularly difficult to collect high-quality corpora. Nevertheless, with the rise of humor recognition research, a number of humor corpora have emerged in recent years, providing valuable resources for research in this field.

Yang et al. (2015) constructed the ”Pun of the Day” dataset, which collected more than 2,000 positive samples from the pun website ”Pun of the Day”, while negative samples came from the Associated Press, the New York Times, Yahoo Answers, and proverbs. It is clear that there are distinct differences between the positive and negative datasets and all selected negative samples were limited to vocabulary present in the positive samples. To address the limitations of humor datasets in terms of type and size, Weller and Seppi (2020) collected more than 550,000 jokes from the r/Jokes section on Reddit. The humor level of each joke was quantified based on user feedback from the r/Jokes community. In contrast to purely text-based humor datasets, Hasan et al. (2019) introduced the UR-FUNNY dataset, a large-scale multimodal corpus based on TED-style monologues. It consists of 1,866 videos from over 1,700 speakers and covers a wide range of topics. Each utterance is annotated for humor using transcript-based laughter tags. The dataset provides tri-modal alignment (text, audio, video) at the word level and includes preceding conversational context for each humorous instance. These features make UR-FUNNY highly suitable for punchline detection and context-aware modeling in spoken language. Following this, Patro et al. (2021) developed the Multimodal Humor Dataset (MHD), using dialogue scenes from the sitcom The Big Bang Theory. Humor labels were derived from laugh tracks, and multiple lines of dialogue were grouped into segments to reflect the importance of contextual buildup. Additional attributes, such as speaker identity and scene timing, were also included to support more fine-grained modeling.

2.3 Approach for Humor Detection

2.3.1 Text-based Humor Detection Approaches

Early research on humor recognition primarily focused on textual input, using linguistic features and contextual cues to identify humor. In traditional machine learning approaches, researchers usually preprocess the text, extract handcrafted features, and train classifiers such as Support Vector Machines (SVMs) or Random Forests for humor binary classification. For early foundational work, Yang et al. (2015) proposed a supervised classification framework using handcrafted features based on four linguistic humor theories: incongruity, denoting semantic contradiction or unexpected contrast; ambiguity, involving multiple plausible interpretations of words or phrases; interpersonal effect, reflecting social or emotional intent conveyed through sentiment-laden language; and phonetic style, characterized by sound-based devices such as rhyme or alliteration. Their systems have achieved initial success in distinguishing humorous and non-humorous, but the system lack temporal modeling and multimodal input, which limits their applicability in natural dialogue systems. Although these methods are relatively efficient and easy to interpret, especially on small datasets, they often have difficulties in scalability and generalization. Since they rely heavily on hand-crafted features, they may capture superficial patterns in language rather than the underlying humorous intent.

To address the shortcomings of traditional models, recent researches have paid more attention to deep learning and pre-trained language models. Unlike traditional methods, deep learning models do not depend on manually crafted humor features. Deep learning models can automatically extract high-level semantic features from large-scale data through end-to-end learning, which helps to reduce the need for manual feature extraction and enabling more accurate capture of the deeper semantics of humor. Bertero and Fung (2016) proposed a pioneering model to detect the setup–punchline in conversational humor analysis. They trained an LSTM-based classifier using dialogue transcripts from The Big Bang Theory. The architecture combined convolutional neural networks (CNNs) for encoding individual utterances with a long short-term memory (LSTM) network to capture contextual dependencies across dialogue turns. In addition, the model incorporated several high-level linguistic features such as sentence length, part-of-speech ratio, antonym occurrence, sentiment score, and speaker identity. Their CNN-LSTM framework achieved an F1 score of 62.9% on the test set, an 8% improvement over the conditional random field (CRF) baseline model. Compared to traditional n-gram-based methods, the model improved recall and reduced false positives. While their work demonstrated the effectiveness of sequential modeling in identifying humor in dialogue, the model was still limited in terms of input modality and generalizability. Subsequent studies attempted to address these limitations by incorporating more diverse features and deeper network structures.

Building on this line of research, Chen and Soo (2018) introduced a deep CNN model that combined filter-size variation with Highway Networks for humor classification. They tested the model on both English datasets—including Pun of the Day, 16000 One-Liners, and Short Jokes—as well as a Chinese joke corpus (PTT Jokes). The model achieved strong results in F1 scores, reaching 0.943 on the PTT dataset and 0.903 on One-Liners. These results showed that CNNs are effective in extracting surface-level and lexical humor features across different languages and joke styles. However, their model did not include contextual or acoustic inputs, 1 which limited its ability in dealing with spoken or dialogue-based humor. In later work, researchers started to introduce large-scale pretrained language models. Weller and Seppi (2020) developed a humor detection model based on BERT(Devlin, Chang, Lee, & Toutanova, 2019). By using self-attention, their system captured contextual semantics in short jokes. Evaluated on Short Jokes, Pun of the Day, and Reddit, the BERT-based model achieved an F1 score of 0.986 on Short Jokes and outperformed a CNN + Highway Layer baseline (F1 = 0.951). It even exceeded crowd-sourced human performance on Reddit data. That said, the model did not account for structural relationships between joke setup and punchline—an important element in humor delivery—and also lacked awareness of broader conversational context, such as previous Reddit comments.

Despite the advances in textual humor detection, most existing approaches remain restricted to linguistic input and overlook the role of acoustic signals—such as pitch variation, pausing, and intonation, that are often critical for recognizing humor in spoken language. Prior linguistic research (Schuller, Batliner, Steidl, & Seppi, 2011) has demonstrated that prosodic cues contribute significantly to the perception of irony, sarcasm, and emotional salience, which are frequently employed in humor. However, few computational models incorporate these signals effectively. For instance, Mao and Liu (2019) proposed a BERT-based framework for humor detection in Spanish tweets, participating in the IberLEF 2019 HAHA shared task. Their system achieved strong performance (F1 = 0.784) by fine-tuning a multilingual BERT model and applying a score-based reclassification strategy. While effective in textual domains, their method, much like that of Weller and Seppi (2019), focused solely on textual input and did not consider nonverbal or prosodic cues. This limitation reduces the model's capacity to capture the nuances of humor in spoken contexts, where delivery features such as exaggerated stress or rhythmic phrasing are often key to triggering laughter.

This review follows the development path of automatic humor recognition, starting from early rule-based and feature-engineered methods, and moving toward deep learning models that can learn contextualized semantic representations. Although Transformer-based architectures like BERT have brought notable improvements in detecting humor from text, most of these models still focus only on linguistic input. Suprasegmental cues, such as prosody, rhythm, and timing, are rarely included, even though they play a key role in understanding humor in spoken interactions. In addition, many existing models pay little attention to dialogic context. They are often trained on scripted or short-form texts, which makes it difficult to apply them effectively to spontaneous and real-life conversations. Due to these shortcomings, it becomes clear that there is a strong need for humor recognition models that are not only lightweight and context-aware, but also capable of handling multiple modalities. Such models are expected to better match how humor is actually expressed and interpreted in natural speech.

2.3.2 Multimodal Humor Detection Approaches

While Transformer-based architectures have shown great promise in capturing textual humor, they remain limited in their ability to model delivery, prosody, and timing—elements that are critical for understanding humor in spoken communication. As a result, recent research has shifted toward multimodal approaches that incorporate audio and visual information alongside text.

A major leap in multimodal modeling came with the work of Hasan et al. (2019), who introduced the Contextual Memory Fusion Network (C-MFN) on the UR-FUNNY dataset. Their architecture extends MFN by integrating sequential context through LSTM-based unimodal encoders and Transformer self-attention, with context representations initializing MFN's memory components. The model encodes punchlines using MFN augmented with contextual memories, and classifies humor based on the combined state of LSTMs and gated memory, excelling in tasks requiring long-range context modeling. C-MFN achieved 65.23% binary classification accuracy with three modalities and was one of the first systems to explicitly incorporate dialog-level memory for punchline prediction. However, the model's computational complexity and lack of interpretability may hinder its scalability in real-world applications. Choube and Soleymani (2020) introduced the HF (Hierarchical Fusion) model on the UR-FUNNY dataset. HF integrates text, audio, and video via GRU-based context modeling and hierarchical modality fusion, achieving 67.84% binary accuracy. Unlike prior work, HF explicitly models bimodal and trimodal interactions through weighted linear combinations, though it uses pre-extracted acoustic features (e.g., prosodic cues from COVAREP) without separate prosody isolation. The model's hierarchical design enhances humor detection but may face scalability challenges in resource-constrained scenarios.Hasan et al. (2021) proposed a more semantically enriched framework—the Humor Knowledge Enriched Transformer (HKT). Their model fused Transformer-based representations from text, audio, and visual modalities using a Bimodal Cross-Attention mechanism, while enriching the text with knowledge-based embeddings from ConceptNet and sentiment information from NRC-VAD lexicons. HKT achieved state-of-the-art accuracy of 77.36% on UR-FUNNY and 79.41% on MUStARD(Castro et al., 2019), outperforming competitive models such as MAG-XLNet and MISA. Nevertheless, the model showed sensitivity to noisy visual data and did not provide explicit modeling of acoustic contributions within the acoustic modality.

H. Xu et al. (2022) introduced a Hybrid Multimodal Fusion Model (HMF-MD) in the MuSe 2022 Humor Sub-Challenge. Their two-stage pipeline first extracted unimodal features using BiLSTMs and Transformers, then performed multimodal fusion across speech, text, and video using attention-based alignment. While their model achieved a strong AUC of 0.8945 on the Passau-SFCH dataset (Amiriparian et al., 2022), its reliance on a relatively small dataset and implicit prosody modeling raised concerns about generalizability and robustness.

To improve efficiency, Pramanick, Roy, and Patel (2022) leveraged optimal transport theory to align modality-specific embeddings for multimodal sarcasm and humor detection. Their lightweight framework achieved competitive performance with reduced computational cost. However, their focus remained on text-image and video-text pairs, lacking audio or ASR input—key components for humor detection in spoken settings.

More recently, M. Xu et al. (2023) participated in the MuSe 2023 Cross-Cultural Humor Sub-Challenge by introducing a high-capacity multimodal pipeline using Whisper for audio, Eva02 for video, and mBERT for text. Their model utilized pseudo-labeling and post-smoothing to mitigate temporal misalignment and modality imbalance, achieving an AUC of 0.9112 on Passau-SFCH. Despite its performance, the system's reliance on visually-heavy features and lack of explicit prosody modeling limits its transparency and adaptability to voice-only contexts.

Taken together, these studies demonstrate the progress and diversity of multimodal humor recognition systems. However, few models explicitly focus on acoustic features—such as pitch, rhythm, and speech rate—which are vital for interpreting timing-sensitive humor. Furthermore, the reliance on high-capacity fusion networks often results in computational burdens, making them computationally inefficient. These gaps highlight the need for compact, context-aware multimodal architectures with targeted acoustic modeling, particularly for speech-driven humor applications.

2.4 Summary and Observed Limitations

Early humor detection models based on artificial linguistic features are interpretable, but they lack the ability to capture temporal dynamics and context. With the rise of deep learning, Transformer-based models (such as BERT) have enhanced text representation learning. However, most of these methods are still limited to textual input and ignore the prosodic and contextual cues that are crucial for spoken humor. Multimodal methods have addressed some of these issues by integrating audio and visual modalities and have achieved remarkable results on datasets such as UR-FUNNY and MUStARD. Nevertheless, most models have the following drawbacks: First, they incur high computational costs, which restricts real-time applications and makes deployment difficult; In addition, they rely heavily on visual inputs that are susceptible to noise, while paying less attention to acoustic features and contextual features independently. These limitations point to a clear need for a lightweight, context-acoustic multimodal humor detection system that are better suited for punchline detection in natural spoken dialogue.

3 Methodology

In this section, I will outline the methodology used to address the research question and validate the hypothesis on a high level. First, in subsection 3.1, I will discuss the datasets utilized for training and testing the models. Next, subsection 3.2 will focus on the introduction of the model framework. Subsection 3.3 will then elaborate on the evaluation method and metric employed. Finally, in subsection 3.4 I will discuss the Ethics and Research Integrity.

3.1 Dataset Description

For this study, I employed the UR-FUNNY dataset (Hasan et al., 2019), which is a multimodal corpus specifically made for computational punchline recognition in spoken language. UR-FUNNY consists of 1,866 TED-style talk videos, delivered by 1,741 distinct speakers across 417 diverse topics, ensuring a wide range of speaker identities, discourse types, and speaking styles. Because of this, the dataset has a wide range of speaker identities, speaking styles, and topic types, which makes it useful for studying humor in a real-world setting, not only scripted jokes.

Each video is segmented into utterances and annotated for humor using laughter cues present in the transcripts. If an utterance is directly followed by a [LAUGHTER] tag, it is labeled as humorous. Conversely, non-humorous instances are selected from utterances that are not followed by laughter, ensuring they are structurally similar but lack humorous markers. The dataset includes 8,257 humorous and 8,257 non-humorous utterances, providing a balanced binary classification setup. This makes it a balanced dataset for binary classification. UR-FUNNY also provides preceding dialogue context for each labeled instance, which is defined as the segment of speech between the current utterance and the most recent preceding humorous utterance (or from the beginning of the video if no earlier humorous utterance exists). Including this context allows models to better capture how humor is constructed across multiple utterances, rather than treating each line as independent. This makes the dataset especially useful for punchline detection systems that require discourse-level understanding.

A significant strength of UR-FUNNY lies in its tri-modal structure. It combines text, audio, and video features in a word-aligned format, which supports more detailed and synchronized multimodal analysis. (1)Textual features include the raw utterance transcripts and associated word-level embeddings. (2)Acoustic features are extracted using the COVAREP toolkit at a sampling rate of 30 Hz, capturing low-level descriptors such as pitch (F0), energy, harmonics-to-noise ratio, glottal parameters, and spectral slope—several of which have been shown to correlate with acoustic markers of humor, such as timing, exaggeration, and vocal dynamics. (3)The video part is extracted with OpenFace(Baltrušaitis, Robinson, & Morency, 2016). It includes facial Action Units, gaze, head movement, and mouth shape, which help detect non-verbal cues.

The average humorous utterance in the dataset lasts approximately 5.2 seconds, with the preceding context averaging around 15.4 seconds, giving ample temporal depth for both utterance-level and dialogue-level modeling. To sum up, UR-FUNNY is a robust and versatile dataset for exploring punchline recognition in spoken language. Its multimodal design, real-speech sources, balanced labels, and detailed alignment make it suitable for training models that focus on both context and multiple modalities, especially in tasks related to humor detection.

Description	**Quantity**
Total number of video clips	16,514
Total video duration (hours)	90.23
Number of speakers	1,741
Total number of utterances	63,727
Average number of utterances per clip	3.86
Total number of words	965,573
Number of unique words	32,995
Average utterance length (words)	15.15
Average utterance duration (seconds)	4.64

Table 1: UR-FUNNY Dataset Statistics

3.2 Model Framework

This section describes the architecture of our proposed lightweight multimodal model for punchline detection. The model is structured into three stages: (1) Unimodal Representation, where linguistic, prosodic, and humor-centric features are individually encoded; (2) Bimodal Cross-Attention Layer, where linguistic and acoustic features are fused via cross-modal attention; and (3) Multimodal Fusion, where pooled representations are concatenated and passed to a classification layer.

3.2.1 Feature Extraction

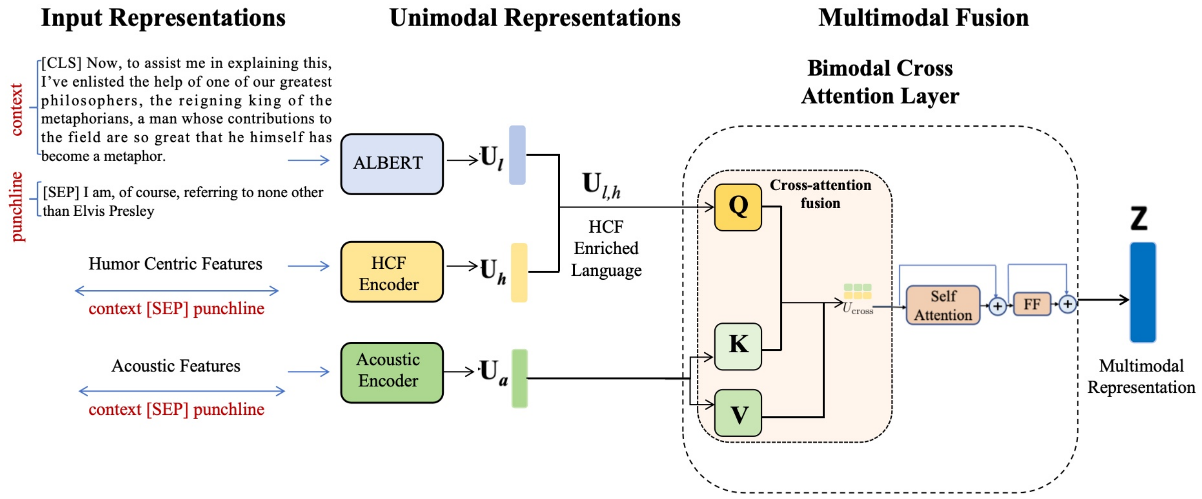
Input Representations
Unimodal Representations
Multimodal Fusion
context
punchline
[CLS] Now, to assist me in explaining this, I've enlisted the help of one of our greatest philosophers, the reigning king of the metaphorians, a man whose contributions to the field are so great that he himself has become a metaphor.
[SEP] I am, of course, referring to none other than Elvis Presley
Humor Centric Features
context [SEP] punchline
Acoustic Features
context [SEP] punchline
ALBERT
HCF Encoder
Acoustic Encoder
U_l
U_h
U_a
$U_{l,h}$
HCF Enriched Language
Bimodal Cross Attention Layer
Cross-attention fusion
Q
K
V
U_{cross}
Self Attention
FF
Z
Multimodal Representation

Figure 1: Overview of the proposed multimodal humor recognition architecture.

As shown in Figure1, the input to our model consists of three modalities: language (l), acoustic (a), and humor-centric features (HCF) (h). For the language input, we use a pretrained ALBERT model (Lan et al., 2019) to obtain contextualized embeddings. Each instance includes a punchline and up to five preceding utterances as context. These are concatenated and tokenized in the format:

$$X_l = [\text{CLS}], C_l, [\text{SEP}], P_l \tag{1}$$

where C_l and P_l represent the context and punchline respectively. The sequence is passed through ALBERT to obtain $U_l \in R^{\tau \times d_l}$, with $d_l = 768$ and τ being the sequence length. The first token $U_l[0]$ (i.e., u_{cls}) is used for global textual representation.

For acoustic information, we use 81-dimensional features extracted via the COVAREP toolkit, aligned at the word level. These features are processed by a custom Transformer encoder, consisting of one self-attention layer. The resulting embedding is denoted as $U_a \in R^{\tau \times d_a}$, where $d_a = 81$.

Humor-Centric Features (HCF) consist of four affective and semantic attributes: valence, arousal, dominance (from NRC-VAD), and ambiguity (computed via ConceptNet and GloVe). These are structured as $X_h \in R^{\tau \times d_h}$ with $d_h = 4$, and encoded using a single-layer Transformer encoder to produce U_h.

The language and HCF outputs are concatenated token-wise to form an enriched sequence:

$$U_{l,h} = [U_l | U_h] \in R^{\tau \times (d_l + d_h)} \tag{2}$$

This serves as the query input to the subsequent cross-modal fusion layer.

3.2.2 Bimodal Cross-Attention Layer

To model intermodal dependencies between linguistic and prosodic features, we introduce a Bimodal Cross-Attention Layer. The textual-humor enriched sequence $U_{l,h} \in R^{\tau \times (d_l + d_h)}$ is used as the query input, and the acoustic representation $U_a \in R^{\tau \times d_a}$ serves as the key and value. The standard scaled dot-product attention is employed:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V \tag{3}$$

where $Q = W_q U_{l,h}$, $K = W_k U_a$, and $V = W_v U_a$. To enable bidirectional alignment, we also compute the reverse direction, applying attention from U_a to $U_{l,h}$ symmetrically. This allows each modality to capture the most informative dimensions of the other, enriching their respective representations.

Following bidirectional cross-attention, the outputs from both directions are concatenated to form a fused sequence that captures intermodal dependencies at each timestep. This sequence is then passed through a single-head self-attention layer, allowing further contextual integration across the sequence. Finally, a feed-forward (FF) sublayer is applied to transform the fused features into the final joint representation $Z \in R^{\tau \times (d_l + d_h + d_a)}$.

This Bimodal Cross-Attention Layer introduces only one layer of cross-attention and one layer of self-attention, making the design shallow and computationally efficient, yet effective at capturing key multimodal interactions between textual and prosodic information. Residual connections and layer normalization are applied after each sub-layer to ensure stable optimization.

[3] https://github.com/Yinnnz/Context-Aware-Punchline-Detection.git

3.2.3 Multimodal Fusion

After obtaining the enriched linguistic representation $U_{l,h}$, the prosodic encoding U_a, and the cross-attention output Z, we proceed to generate a compact yet informative vector for final classification through a multistage fusion strategy. The goal of this stage is to combine multiple sources of information in a way that preserves their respective semantic contributions while minimizing computational overhead. We begin by extracting three key representations:

- $u_{\text{cls}} \in R^{d_l}$: the output of the [CLS] token from ALBERT, which is widely used in sentence-level tasks due to its global contextual representation of the input sequence.
- $a_{\text{max}} \in R^{d_a}$: he max-pooled vector over the time dimension of the acoustic encoder output U_a. This operation captures the most salient prosodic cues present in the input utterance, such as pitch peaks, energy shifts, and glottal dynamics.
- $z_{\text{max}} \in R^{d_l+d_h+d_a}$: the max-pooled vector over the cross-modal attention output Z, which represents fine-grained linguistic–acoustic interactions.

These three vectors are concatenated into a single unified representation:

$$o = [u_{\text{cls}} | a_{\text{max}} | z_{\text{max}}] \in R^{d_o}, \quad d_o = d_l + d_a + d_l + d_h + d_a \tag{4}$$

This results in a dense vector that captures global semantics (u_{cls}), raw prosody (a_{max}), and deep cross-modal dependencies (z_{max}).

The concatenated feature vector o is then passed through a two-layer fully connected network for final classification. The first layer applies a ReLU activation to introduce non-linearity, followed by a dropout regularization (rate = 0.2366) to mitigate overfitting. The second layer projects the output to a single scalar, which is interpreted as the probability of the punchline being humorous via the sigmoid function:

$$p = \sigma(W_o \cdot \text{ReLU}(W_f o + b_f) + b_o) \tag{5}$$

where W_f, b_f, W_o, and b_o are learnable parameters, and σ denotes the sigmoid activation function.

This architecture is intentionally designed to be shallow and computationally efficient, with only one attention layer per modality and one cross-attention layer. Despite its simplicity, our experimental results in Section 5 demonstrate its effectiveness in capturing multimodal dependencies for punchline detection.

3.3 Evaluation Methodology

In this study, the evaluation of the proposed punchline detection model is conducted by usingtwo standard metrics in binary classification: Accuracy and F1 score. These two metrics canreflect the overall prediction ability of the model well and they also work well when the dataset is not perfectly balanced — which is often the case in humor-related tasks. These two metrics are calculated using the following classification outcomes:True Positives (TP) refer to the number of humorous utterances correctly classified as humorous, and True Negatives (TN) denote the number of non-humorous utterances correctly classified as non-humorous. False Positives (FP) represent non-humorous utterances incorrectly labeled as humorous and False Negatives (FN) refer to humorous utterances that are wrongly labeled as non-humorous. Accuracy measures the proportion of correctly predicted instances—both humorous andnon-humorous—over the total number of predictions. It is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

F1-score is the harmonic mean of precision and recall. This is helpful when dealing withclass imbalance, which often appears in humor detection. F1-score offers a more balancedevaluation of the model's ability to correctly identify humorous punchlines. Therefore, F1-score is used as the primary metric for early stopping and model checkpointing during training. The F1-score is calculated as:

$$\text{F1-score} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{7}$$

3.4 Ethics and Research Integrity

This study adheres to the ethical guidelines of responsible research practices and reflects a sustained commitment to integrity throughout the research process. During the development of the proposed lightweight multimodal humor detection model, we have consistently considered the ethical implications related to dataset usage, algorithmic fairness, model transparency, and computational sustainability. Particular attention was given to ensuring that data sources were appropriate and that model behavior would not reinforce social biases or exclude underrepresented voices. The following subsections will elaborate in more detail on how each of these dimensions has been addressed in the design, implementation, and evaluation of the model.

3.4.1 Data Ethics and Privacy

This study utilizes the UR-FUNNY dataset, a public multimodal humor dataset that contains video clips of English stand-up comedy performances. All the data used has been anonymized and is only for academic purposes. This dataset does not contain any Personally Identifiable Information (PII). The participants in the original recordings are all public figures in performance scenarios, which minimizes privacy concerns. Additionally, this study has neither collected nor processed any other user data.

3.4.2 FAIR Principles Implementation

The dataset and code resources used in this study comply with the FAIR principles: Findable: All resources have been indexed and can be accessed through public code repositories (such as HuggingFace and GitHub). Accessible: The UR-FUNNY dataset and pre-trained models (such as BERT and COVAREP) are publicly available under academic licenses. Interoperable: The data processing pipeline is implemented using standard Python libraries and open frameworks. Reusable: All experimental codes are modular and have been documented to ensure reproducibility. Hyperparameters, training configurations, and model weights can be reused in future research.

3.4.3 Open Science Practices

This project actively contributes to the advancement of open science by utilizing a suite of widely adopted open-source tools and frameworks, including HuggingFace Transformers, PyTorch, and the pre-trained ALBERT model. Throughout the research process, we maintained a commitment to transparency and reproducibility by thoroughly documenting the entire modeling pipeline—from dataset preprocessing and feature extraction to model architecture, training configuration, and evaluation procedures. All relevant source code, including data loading scripts, model definitions, and training routines, will be made publicly available via GitHub upon acceptance of the thesis. In addition, key metrics and training dynamics have been continuously logged and visualized using Weights & Biases (wandb), ensuring that experimental workflows are fully traceable. This approach facilitates future replication, adaptation, and community-driven improvement of our proposed framework.

3.4.4 Bias and Fairness

Although the UR-FUNNY dataset offers a relatively diverse collection of speech samples in terms of speaker identities, topics, and delivery styles, it is predominantly constructed from English-language stand-up comedy performances originating in Western cultural contexts, particularly from the United States and Canada. This cultural homogeneity may introduce both linguistic and socio-cultural biases into the model. As a result, the patterns learned by the model may reflect humor norms and delivery mechanisms that are specific to Western English-speaking audiences, limiting the model's ability to generalize to other cultures, languages, or humor genres.

Furthermore, humor is inherently subjective, and the process of annotating punchlines is susceptible to annotator bias. For instance, what one annotator perceives as humorous may be interpreted as neutral or even offensive by another, depending on their individual background, cultural exposure, or sense of humor. Since the UR-FUNNY dataset relies on binary humor annotations (humorous vs. non-humorous) typically derived from audience laughter cues, it may oversimplify the nuanced nature of humor perception and overlook subtle or culturally specific forms of humor.

These limitations highlight important ethical and methodological concerns regarding fairness and inclusivity in humor classification research. To mitigate these issues, future work should aim to incorporate multilingual and cross-cultural humor datasets, including performances from non-Western cultures and non-English speakers. Additionally, implementing annotation strategies that involve multiple annotators with diverse backgrounds could help capture a broader spectrum of humor perceptions. Techniques such as consensus-based labeling, weighted voting, or subjective scoring distributions could further address inter-annotator variability and improve the fairness and robustness of humor recognition systems.

3.4.5 Reproducibility and Replicability

To enhance reproducibility, all model codes, configuration files, and data preprocessing scripts will be publicly released. The model is based on open-source pre-trained components and is trained on the publicly available UR-FUNNY dataset. Although fixed random seeds have not been applied to all components, the core processes are deterministic and replicable. This enables other researchers to reproduce the reported performance and further validate the research findings.

In summary, this study has tried to take ethical factors into consideration from data selection, model design to training process. We strive to use public data to avoid privacy issues, and adopt efficient model structures to reduce unnecessary consumption of computing resources. We also make the code and experimental settings public to facilitate other researchers to reproduce and verify the results. As humor recognition technology is promoted in more languages and cultures, future research also needs to continue to pay attention to fairness, inclusiveness and the boundaries of technology use to ensure that these systems are used responsibly.

4 Experimental Setup

This chapter presents the technical and practical settings of the conducted experiments, including-data preprocessing, input construction, dataset partitioning, model variants, and the computational environment. All experiments are carried out on the UR-FUNNY dataset with a focus onevaluating a context-aware, multimodal approach to spoken punchline detection using text andacoustic features.

4.1 Data Preparation

The input to the proposed model is constructed from three sources of information: textual features, acoustic features, and HCF. These features are extracted from the UR-FUNNY dataset, a multimodal humor corpus. All features are aligned at the word level and serialized into .pkl format prior to training, using the official script provided with the dataset.

Language Input: Each data point consists of a punchline and up to five preceding context utterances. All utterances are tokenized using the AlbertTokenizer from HuggingFace Transformers. The tokenized sequence is formatted as [CLS] context [SEP] punchline, where [CLS] is a classification token and [SEP] separates the context and punchline segments. The combined token sequence is then passed through a pre-trained ALBERT model (albert-base-v2) to extract contextual embeddings. For each token, the corresponding hidden state is obtained from the final transformer layer. The output embedding dimension is 768, and sentence-level representations are subsequently aligned with acoustic and HCF features for joint modeling.

Acoustic Input: Acoustic features are extracted using the COVAREP toolkit (Degottex, Kane, Drugman, Raitio, & Scherer, 2014). These include a rich set of low-level descriptors such as fundamental frequency (F0), harmonic-to-noise ratio (HNR), mel-cepstral coefficients, glottal source parameters, and other prosodic indicators. For each word, the start and end time are determined using forced alignment with P2FA, and the corresponding acoustic frames are sliced and averaged across time to obtain a fixed-length vector. Each utterance is padded or truncated to a predefined number of words (e.g., 20), ensuring temporal consistency across samples.

Humor-Centric Features (HCF): HCF are designed to capture affective and semantic properties relevant to humor, following the theoretical basis of the ambiguity and superiority theories of humor. Each word is assigned a 4-dimensional HCF vector comprising: 1. Valence (positive/negative sentiment), 2. Arousal (calm/excited), 3. Dominance (submissive/dominant), 4. Ambiguity (degree of semantic uncertainty). The first three dimensions—valence, arousal, and dominance—are extracted from the NRC VAD lexicon (Mohammad, 2018), which assigns each English word a score between 0 and 1 for each emotional dimension. The ambiguity score is computed based on ConceptNet(Liu & Singh, 2004) and GloVe(Pennington, Socher, & Manning, 2014) embeddings. Specifically, for each word, its top-N related concepts are retrieved from ConceptNet, and their 300-dimensional GloVe embeddings are obtained. Pairwise cosine distances between the sense vectors are calculated and averaged to quantify ambiguity: the greater the distance, the higher the ambiguity score. This method reflects the number and diversity of plausible meanings a word can convey in context. For each utterance, the HCF vectors of individual words are averaged to obtain a fixed-length 4-dimensional representation aligned with the ALBERT and acoustic features. These representations are then padded or truncated as needed to match the token sequence length.

All modalities are synchronized at the word level using alignment indices provided by the dataset. The resulting features are stored in a .pkl file and loaded dynamically during training and evaluation.

4.2 Data Splitting

The dataset used for this study is the UR-FUNNY and data is divided using the official `data_folds.pkl` split file. To ensure a balanced evaluation, the dataset was split into three subsets:

• Train Dataset: Consisting of 70% of the total data, used for model optimization.

• Dev Dataset: The development dataset, used for validation during training, makes up 15%of the total data.

• Test Dataset: The test dataset, also comprising 15% of the total data, is used to evaluate the final model performance.

To ensure reproducibility, we set a fixed random seed (seed=100) for all stochastic processes, including Python, NumPy, and PyTorch. Additionally, we disabled non-deterministic CuDNN behavior by setting torch.backends.cudnn.deterministic=True and benchmark=False.

4.3 Experimental Environment

4.3.1 Hardware

All experiments were conducted on the Hábrók high-performance computing (HPC) cluster provided by the University of Groningen. Each job was executed on a compute node equipped with an NVIDIA V100 GPU (32 GB VRAM), a 32-core Intel Xeon CPU, and 128 GB of RAM. This configuration provided sufficient computational resources for training and evaluating the proposed multimodal model with minimal latency or memory constraints.

4.3.2 Software Environment

The implementation was developed in Python 3.10. Model training and inference were performed using PyTorch 2.0.1. The HuggingFace Transformers library (version 4.36.2) was used to load the pre-trained ALBERT model and process token embeddings. Acoustic features were handled using torchaudio 2.1.0 and librosa 0.10.1. Additional utility libraries, including NumPy (1.24.3), pandas (1.5.3), and scikit-learn (1.3.0), were employed for data manipulation and metric computation. All experiments were executed within a Conda-managed environment to ensure reproducibility and consistent dependency versions across runs.

4.3.3 Training and Evaluation

The model was trained using the AdamW optimizer. To account for differences in modality complexity and feature scale, distinct learning rates were applied to each encoder: 5×10^{-5} for the ALBERT text encoder (to fine-tune pre-trained weights conservatively), 3×10^{-3} for the acoustic encoder (to accelerate learning of low-level features), and 3×10^{-4} for the HCF encoder. A warmup ratio of 0.07178 was used with a linear learning rate scheduler to prevent unstable weight updates at the early stages of training.

The loss function was binary cross-entropy, which is suitable for the binary nature of the humor classification task. To reduce overfitting, a dropout rate of 0.2366 was applied to the transformer encoder layers and the fusion module, selected empirically based on preliminary validation performance.

Training was conducted for 15 epochs with a batch size of 16. The maximum sequence length was fixed at 85 tokens per example, covering both punchline and context. All transformer-based encoders employed a single attention layer with 1 head, and the cross-modal attention module included 1 layer with 4 heads. The fusion module projected the concatenated features to a 172-dimensional latent space before classification.

Model performance was monitored on the validation set during training, using both Accuracy and F1-score as evaluation metrics. The model checkpoint yielding the highest F1-score on the validation set was retained for final testing. No early stopping mechanism was applied. All test results reported in this study are based on the model selected by validation performance.

4.4 Baseline models

To assess the performance of our proposed model, we compare it against several representative baselines widely used in multimodal sentiment and humor classification. The selected models include both early fusion and hierarchical fusion strategies. Below is a brief description of each:

- **SVM(Support Vector Machine)**: A classical early-fusion baseline that concatenates features from multiple modalities—typically text and acoustic vectors—into a single feature vector, which is then classified using a linear or kernel-based Support Vector Machine. In humor recognition tasks, this model is typically applied only at the punchline level, ignoring contextual information from preceding utterances. Due to the lack of temporal modeling and modality-specific attention, SVM fails to capture inter-modal dependencies or sequential humor cues. Although simple and computationally efficient, it is limited in its ability to model subtle or temporally grounded humor structures, especially in conversational data.

- **CNN (Convolutional Neural Network)**: An early-fusion approach where multimodal feature vectors are concatenated and treated as sequences (e.g., word-level inputs) before being passed through convolutional layers. The CNN applies local filters to extract spatial or short-range dependencies, making it suitable for identifying low-level patterns in the input. However, it lacks mechanisms for long-range modeling or hierarchical context integration. As such, CNN-based models are often unable to capture the nuanced interplay between modalities or exploit the sequential context that is critical for detecting humor in multi-utterance dialogues.

- **TFN (Tensor Fusion Network)**: A neural fusion architecture introduced by Zadeh, Chen, Poria, Cambria, and Morency (2017) and designed for multimodal data analysis. It first leverages LSTM to extract sequential features from the textual modality, capturing the temporal dynamics of language. Then, it employs an outer - product operation to fuse features across three modalities (e.g., text, audio, visual). Through this fusion, TFN models intricate interactions at unimodal, bimodal, and trimodal levels, enabling it to capture both simple single - modality patterns and complex cross - modality relationships for tasks like sentiment analysis.

- **C-MFN (Contextual Memory Fusion Network)**: An extension of the Memory Fusion Network (MFN), this model is specifically designed to incorporate sequential context in multimodal interactions(Hasan et al., 2019). It uses separate LSTM encoders to capture unimodal context representations, and applies a Transformer-based self-attention mechanism to model intra- and inter-modal interactions hierarchically. The outputs of these unimodal and multimodal context modules are used to initialize the memory components of the MFN. For each prediction, the punchline is encoded using MFN, and the final classification is based on the last state of the memory. C-MFN is particularly effective in dialogue-based tasks like humor recognition, where long-range dependencies and temporal context play a crucial role.

- **bc-LSTM (Bidirectional Contextual LSTM)**: An architecture that employs bidirectional LSTM layers to model contextual dependencies among utterances in multimodal sequences (Poria et al., 2017). Multimodal features are first extracted separately, then concatenated and fed into the bidirectional LSTM to capture sequential information from both preceding and succeeding utterances. This structure enables the model to integrate cross-utterance contextual cues while leveraging fused multimodal features, making it suitable for spoken dialogue analysis where temporal order and inter-utterance relations are critical.

- **HF (Hierarchical Fusion)**: A context-aware multimodal architecture for punchline detection(Choube & Soleymani, 2020). It uses GRU to model sequential dependencies among utterances, capturing the contextual flow crucial to humor understanding. Before fusion, multimodal feature vectors are dimensionally aligned to ensure consistency across modalities. This design allows HF to effectively integrate context and modality-specific cues, improving performance in spoken dialogue scenarios.

5 Results

This section presents the experimental results and evaluates the proposed multimodal model for punchline detection on the UR-FUNNY dataset. The model's performance is assessed using Accuracy and F1-score, which provide complementary insights into its classification effectiveness. We compare our approach with several baseline models from prior work to highlight its relative advantages. In addition, an ablation study is conducted to investigate the contribution of individual modalities. These analyses provide a comprehensive understanding of the model's performance, both in its full configuration and under reduced settings.

5.1 Results

To evaluate the effectiveness of the proposed lightweight multimodal framework, we conducted a series of experiments on the UR-FUNNY dataset. The full model incorporates three types of input: textual features (including both the punchline and up to five preceding utterances as context), acoustic features extracted from speech signals using the COVAREP toolkit, and HCF designed to capture structured indicators such as incongruity, exaggeration, or polarity shift. Under this configuration, the model achieves an accuracy of 72.33% and an F1-score of 0.7231, representing the highest performance among all evaluated settings. As shown in Table 2, our model consistently outperforms all baseline systems, including classical early-fusion methods such as SVM and CNN, as well as more advanced hierarchical architectures like HF, TFN, C-MFN and bc-LSTM. While many prior models offer high accuracy at the cost of computational complexity, our proposed approach achieves a favorable trade-off between accuracy and efficiency. The performance gain, combined with its compact architecture, demonstrates that incorporating prosodic and contextual features into a lightweight design can significantly enhance punchline classification in spoken dialogue settings.

Method	**Accuracy (%)**	**F1 Score**
SVM	61.22	0.6036
CNN	63.98	0.6631
TFN	65.83	–
C-MFN	65.23	–
be-LSTM	66.99	0.6565
HF	67.84	0.6885
Our Model	**72.33**	**0.7231**

Table 2: Comparison of Accuracy and F1-score between different models

5.2 Ablation Studies

To investigate the contribution of each modality to the model's predictive performance, we conducted a series of ablation studies, each involving the removal of one input type from the full multimodal configuration. The goal was to assess the relative importance of contextual, acoustic, and HCF by observing the resulting impact on classification metrics.

The removal of contextual information—operationalized as using only the punchline and excluding preceding utterances, yielded an accuracy of 65.39% and an F1-score of 0.6523. This highlights the importance of conversational history in humor understanding, as contextual build-up often plays a critical role in comedic effect.

When acoustic features were excluded, the model's performance dropped to an accuracy of 65.29% and an F1-score of 0.6504, resulting in the largest performance decline among all conditions. This indicates that acoustic features—such as pitch, energy, and speech rhythm—contribute meaningfully to punchline detection, likely by capturing delivery-based cues that go beyond textual content.

Finally, when the HCF were removed, the model achieved an accuracy of 69.82% and an F1-score of 0.6927. Although the drop is relatively smaller, it still reflects a meaningful loss in performance, especially considering the low dimensionality of HCF. These features appear to provide structured, interpretable signals—such as incongruity, polarity shifts, and semantic exaggeration—that are not easily inferred from raw text or audio streams alone.

Taken together, these findings confirm that all three modalities contribute in complementary ways. Context offers high-level discourse structure, acoustics convey delivery-related nuances, and HCF introduces domain-aware semantic signals. Their combination yields the best results, validating the effectiveness of the proposed multimodal design.

Modality Combination	**Accuracy (%)**	**F1 Score**
no context	65.39	0.6523
no acoustic	65.29	0.6504
no hcf	69.82	0.6927
full multimodal	**72.33**	**0.7231**

Table 3: Results of Ablation Study across different modality combinations

6 Discussion

This chapter discusses the experimental findings in detail and reflects on the effectiveness of the proposed lightweight multimodal model. It begins by comparing the model's performance with a range of established baselines, highlighting its ability to achieve a favorable balance between accuracy and computational efficiency. Furthermore, it provides an in-depth examination of how different input modalities—namely acoustic features, conversational context, and humor-centric cues—contribute to the task of humor recognition. Through a series of ablation studies, we explore the specific role and value of each component. The discussion also considers broader implications for multimodal modeling, emphasizing how the proposed architecture supports both interpretability and practical deployment.

6.1 Overall Performance and Comparison with Baselines

Compared to several established baseline models, our proposed lightweight multimodal framework demonstrates clear and consistent advantages in both accuracy and F1-score for spoken humor classification. Traditional early fusion methods, such as SVM and CNN, operate by directly concatenating features from different modalities and feeding them into a classifier. While these approaches are easy to implement, they lack the capacity to model modality-specific dependencies or contextual relationships—both of which are critical in humor recognition tasks that rely heavily on nuanced delivery, timing, and semantic subtleties.

In our experiments, these early fusion baselines achieved relatively poor performance, with F1-scores of only 0.6036 (SVM) and 0.6631 (CNN), respectively. These results are significantly lower than the 0.7231 F1-score achieved by our proposed model. The stark performance gap highlights the inability of these simple models to capture deeper inter-modality interactions or leverage temporal discourse context effectively, both of which are essential for understanding punchlines in spoken dialogue.

In contrast, more sophisticated multimodal fusion strategies, including TFN , C-MFN, bc-LSTM, and HF, incorporate temporal or hierarchical structures to better model cross-modal dependencies. These models are theoretically more expressive and capable of capturing fine-grained interactions across text, audio, and context. However, this enhanced expressiveness comes at the cost of increased model complexity, heavier computational overhead, and lower interpretability. Their reliance on deep stacking, tensor operations, or multiple sequential modules makes them less practical for real-time or resource-constrained applications.

Among these advanced models, the HF architecture achieved the best result with an F1-score of 0.6885. Despite its relatively strong performance, it still lags behind our model by 3.46 percentage points. This margin is non-trivial and indicates that our lightweight model not only competes with, but surpasses, more elaborate systems while maintaining efficiency and interpretability. The observed performance gap underscores the potential of our design to retain essential multimodal information without incurring excessive computational burden.

One of the key strengths of our model lies in its modular architecture. Each modality—context, acoustic features, and HCF—is processed independently by a dedicated encoder. The resulting embeddings are then integrated via a cross-attention module, which explicitly learns the interaction dynamics between modalities. This fusion mechanism is computationally light yet effective, avoiding the redundancy of deep fusion blocks or static tensor operations. It also promotes a more transparent understanding of how modality contributions affect final predictions.

In conclusion, our model achieves an effective balance between performance and efficiency. It consistently outperforms both simple and complex baselines, demonstrating superior capability in capturing the multimodal nature of humor while retaining a lightweight and interpretable structure. This makes it particularly well-suited for real-world applications where computational resources are limited or fast inference is required.

6.2 Role of Different Components

To examine the contribution of individual modalities to humor detection, we conducted a series of ablation experiments, each targeting a specific input modality. In each case, the rest of the model architecture and training setup were kept constant. The results are summarized in Table 5.2, and each sub-section below addresses one of the research questions.

To evaluate the role of acoustic input in the humor recognition process, we removed the acoustic features from the model while retaining both the textual inputs and HCF. Under this configuration, the model achieved an accuracy of 65.29% and an F1-score of 0.6504, representing the largest performance drop of 7.27 points in F1 compared to the full model. This outcome underscores the importance of acoustic information in spoken humor. Acoustic features—such as pitch inflection, emphasis, hesitation, rhythm, or elongated pauses—often signal irony or serve as cues that lead up to a punchline. For example, elongated pauses or rising pitch patterns often precede punchlines, helping to build anticipation or signal a shift in tone (Naz, Farooq, & Jabeen, 2023). Similarly, ironic or sarcastic utterances are frequently marked by slower speech rate, greater pitch variability, and exaggerated intonation (Bryant, 2010), indicating that acoustic cues can serve as powerful signals for humorous intent. These findings underscore the importance of modeling prosody in humor recognition systems, particularly in multimodal settings. The drop in performance following the removal of acoustic input suggests that this modality contributes valuable and non-redundant information to the classification process, improving both interpretability and the model's sensitivity to delivery style.

To further investigate the influence of conversational context, we removed all context and retained the punchline, acoustic and HCF as input. This variant achieved an F1-score of 0.6523, reflecting a drop of 7.08 points relative to the full model. Although slightly smaller than the acoustic ablation, the degradation still underscores that context plays a major role in humor comprehension. Such a decline highlights the critical role of discourse-level information in understanding humor. In many stand-up comedy scenarios, humor arises not from the punchline in isolation but from a buildup of expectations, narrative progression, or thematic contrast created in the preceding context. Without this background, the model is deprived of key semantic signals needed to detect incongruity, resolve ambiguity, or appreciate comedic timing. In contrast, the full model leverages up to five prior utterances, which allows it to learn speaker intent, track dialogue flow, and capture longer-range dependencies. The sharp performance drop confirms that humor detection is not a standalone utterance-level classification problem but instead requires a more holistic understanding of dialogue context.

To assess the value of humor-centric features, we removed the HCF input and preserved only the textual and acoustic modalities. This configuration resulted in an F1-score of 0.6927, a relatively modest drop of 3.04 points compared to the complete system. Although the magnitude of the drop is smaller than that observed when removing context or acoustic features, it remains a meaningful indicator of the utility of HCF. These features are specifically designed to encode high-level humor constructs, including exaggeration, incongruity, polarity shifts, and repetition—all of which are difficult to extract from raw text or audio alone. The absence of HCF leads to reduced sensitivity to subtle structural humor patterns, particularly in cases where punchlines are ambiguous or contextually nuanced. By introducing these targeted signals, the HCF module provides inductive bias that enhances the model's interpretability and improves its ability to make fine-grained decisions.

Taken together, the ablation studies highlight that each modality contributes uniquely and meaningfully to the task of humor recognition. Acoustic features plays the most dominant role, followed by contextual features, with HCF offering an additional performance boost. The relative stability of the model when HCF is excluded suggests that these features may serve as an optional but effective enhancement in resource-constrained scenarios. More broadly, these results validate the model's design principles, showing that our lightweight architecture effectively integrates complementary signals from multiple modalities to support nuanced, context-aware humor classification.

6.3 Summary

In summary, this chapter highlights the superior performance and design efficiency of our lightweight multimodal model. Through comparative evaluation and ablation studies, we demonstrated that each modality—acoustic, contextual, and humor-centric—plays a distinct and complementary role in spoken humor recognition. Unlike traditional models that compromise interpretability for complexity, our framework achieves robust performance through modular design and effective cross-modal interaction. These insights confirm the practical value of integrating structured, task-relevant features into lightweight architectures, making our approach a promising solution for real-world humor detection applications.

7 Conclusion

This chapter provides a comprehensive conclusion to the study by summarizing its main findings, reflecting on technical and methodological challenges, and outlining key limitations and opportunities for future work. The research aimed to develop a lightweight multimodal framework for spoken humor recognition, striking a balance between performance, interpretability, and efficiency. Through a series of experiments and ablation studies, the model demonstrated strong performance compared to both simple and complex baselines. This chapter also highlights unresolved issues, such as limited modality coverage and task generalization, and proposes future directions to improve model robustness, expand functionality, and explore real-world deployment potential.

7.1 Challenges

Throughout the course of this study, several significant challenges were encountered, primarily related to technical capabilities and data handling. These challenges are outlined as follows:

1) Dataset preprocessing and data format issues

The UR-FUNNY SDK-format dataset, while preprocessed to an extent, still required significant additional handling. One of the challenges lies in the nested structure of data entries, where textual and acoustic features are stored under different keys (e.g., $punchline_{f}eatures$, $context_{f}eatures$). Some entries unexpectedly contain string representations instead of numerical vectors, resulting in error during tensor conversion. I had to manually inspect the data set, perform safety checks and write secure conversion functions to ensure that the model could process all samples without interruption. Although these steps are not part of the model design itself, they are crucial for establishing a stable training process.

2) Cross-modal Shape Alignment and Fusion Design

One of the main difficulties in this project was aligning features from different modalities, which often vary in both length and dimensionality. The text and audio encoders produced outputs with different shapes, so extra steps were needed to ensure that the punchline and context representations could be compared and fused correctly. Getting this alignment right was important for allowing the cross-attention module to work effectively. We also spent time experimenting with different ways of combining the two types of input. The goal was to let each modality contribute its own information without losing meaning in the process. Though it might seem simple in theory, building this fusion mechanism involved a lot of adjustment and testing to get it to function reliably.

3) Model Adaptation for Ablation Studies

Conducting ablation studies—such as removing the acoustic modality or the humor-centric features (HCF)—required careful reconfiguration of both the model architecture and the training pipeline. Since the original model integrates modalities through tightly coupled cross-attention and concatenation operations, each modality is embedded at a different stage of the architecture. As a result, removing a single modality was not a matter of simply disabling an input channel; it necessitated structural rewrites. Specifically, modifications had to be made to the model.py file to define variant architectures that exclude one of the modalities, as well as to main.py to handle different data preprocessing and model instantiation logic. Additionally, the inference script (test.py) required adaptation to accommodate changes in input dimensionality and feature flow. Each ablation variant—text-only, text + HCF, and text + acoustic—demanded distinct structural considerations, particularly because the affected components were often embedded within layers such as the cross-attention block or the fusion layer. Maintaining consistency across these multiple model variants without breaking compatibility or duplicating code proved to be a considerable challenge. This task was further complicated by the interdependencies among modality-specific encoders and fusion mechanisms, underscoring the complexity of conducting rigorous ablation experiments in multimodal learning frameworks.

4) Uncertainties in Experimental Design Decisions

A key challenge throughout this study was managing various uncertainties in experimental design—especially when it came to selecting hyperparameters and determining suitable training strategies. As is common in postgraduate research, many important decisions had to be made without established guidelines or known optimal configurations. Among the parameters that most affected performance were the dropout rate, the dimensionality of the fusion layer, the number of attention heads in the cross-modal alignment module, and the total number of training epochs. Each of these choices had a direct influence on how stable the model's training process was and how well it ultimately performed.

Early experiments revealed that the model exhibited strong sensitivity to certain hyperparameter combinations. This sensitivity was especially pronounced in multimodal settings, where nonlinear interactions between input streams exacerbated instability. For instance, setting the dropout rate too low led to overfitting, while overly high values caused underfitting or even training collapse. Similarly, increasing the number of attention heads or the width of the fusion layer sometimes resulted in abnormal fluctuations in training loss, gradient vanishing, or non-convergence. To address these issues, I adopted a hybrid approach combining manual tuning with automated search. A systematic grid search was conducted on key parameters to strike a balance between model capacity and generalization ability. In addition, gradient clipping and learning rate warm-up strategies were employed to prevent gradient explosion and stabilize parameter updates during early training.

Overall, resolving these uncertainties in experimental design was essential for developing a reproducible and stable model. It also provided valuable insights into the nuanced relationship between architectural complexity, modality fusion strategies, and training dynamics. Through extensive experimentation and fine-tuning, I was able to construct a lightweight multimodal system that performed robustly across different configurations, including ablation scenarios.

7.2 Limitations

This study proposes a lightweight multimodal model that shows promise in balancing accuracy, computational efficiency, and practical applications. Rather than pursuing optimal performance at the expense of increased complexity, the model is designed to intentionally prioritize a compact structure suitable for practical applications. Despite these advantages, there are still some limitations that are worth exploring.

First of all, although the model has shown competitive results on the UR-FUNNY dataset, with the highest test accuracy of 72.33% and an F1 score of 0.7231 in the full multimodal setting, there is still room for improvement.Given the design priority on efficiency, the model uses only one self-attention layer per modality and a single-layer cross-attention module for fusion, which may limit its ability to capture long-range dependencies and subtle inter-modal interactions. As a result, certain complex humor patterns—such as jokes that rely on long dialogue history, nuanced speaker intent, or subtle cultural references—may not be fully represented or correctly interpreted. In addition, This study focuses on the textual and acoustic modalities and constructs a humor model based on acoustic features and semantic content. Visual cues such as facial expressions and gestures play an important role in human humor perception. Although the visual modality was deliberately ignored to maintain the simplicity of the model, the absence of the visual modality may limit the model's ability to recognize multimodal humor, as non-verbal signals are crucial in multimodal humor.

Meanwhile, this study focuses on the binary classification recognition of humor, that is, determining whether a segment is humorous. This task definition is concise and clear and is suitable for building a basic humor recognition system. However, as a complex linguistic phenomenon, humor has diverse expression styles and emotional hierarchies. Simply classifying it as ”humorous” or ”non-humorous” is still rather crude. The model cannot distinguish between different types of humor (such as sarcasm, exaggeration, irony, black humor, etc.), nor can it identify dimensions such as humor intensity, style, or the emotional response of the audience. This singularity in the modeling hierarchy limits the application scope of the model in more complex scenarios.

Last but not least, the training process in this study adopted a fixed setup: a predefined learning rate, the AdamW optimizer and the standard binary classification cross-entropy loss. Although the model is suitable for baseline evaluation, future research can explore whether enhancement methods (such as cyclic learning rates, advanced data augmentation (such as acoustic noise simulation), or curriculum learning) can further improve the performance without deviating from the lightweight concept.

In conclusion, this study proposed a lightweight multimodal architecture that achieves a balance between deployability and reasonable accuracy in humor recognition tasks. Although the model shows robust performance under limited computing resources, future research should still focus on improving model performance and humor level understanding capabilities, exploring richer modal feature inputs and more efficient training strategies, while always adhering to the core design concept of lightweight and efficient.

7.3 Future Work

Future research can further expand and refine the lightweight multimodal humor recognition model proposed in this study to enhance its practicality, generalizability, and overall system performance.

First of all, we can introduce more expressive yet low-computational-overhead feature enhancement mechanisms. For example, lightweight attention modules, adaptive feature recalibration methods, or inter-modal contrastive learning strategies are all expected to improve the collaborative modeling effect between modalities without significantly increasing the number of parameters. Additionally, it is also a direction worth exploring to further explore the deep connections between acoustic information and semantic information on the current basis, or to conduct more meticulous dynamic fusion of features in different time segments.

Secondly, considering that non-verbal cues such as facial expressions, eye gaze, and body movements play a crucial role in humor perception, in the future, we can consider introducing compressed visual features (such as using lightweight visual encoders like MobileNet) to construct a three-modal version of the lightweight network. Such an expansion can significantly enhance the model's ability to perceive and understand multimodal humor while maintaining computational efficiency

At the same time, Although the binary classification setup is concise and clear and suitable for establishing a basic humor recognition system, humor essentially has rich types and semantic hierarchies, and simply modeling it through binary classification is too simplistic. Future research can try to expand the task definition. For instance, we can construct multi-class humor recognition(such as distinguishing sarcasm, exaggeration, black humor, etc.) or introduce regression modeling of humor intensity, thus enhancing the system's ability to distinguish humor styles, expression intensities, and emotional responses, and making the model closer to the humor dissemination and perception mechanisms in real contexts.

In addition, more flexible optimization techniques can be introduced, such as cyclic learning rates, data augmentation methods with acoustic perturbations, curriculum learning, etc., to improve the model's adaptability to complex training samples and the convergence speed, thereby enhancing the model's robustness and generalization ability.

It is worth noting that currently, the model is only trained and evaluated on the UR-FUNNY dataset. Although it has a certain degree of diversity, it still mainly reflects the characteristics of English TED speech corpora. Humor shows significant differences among different languages, cultures, and audience groups. For future improvement, we can explore cross-language andcross-cultural adaptability, and try to migrate the model to multiple language datasets or spontaneous spoken dialogues to evaluate its robustness in multi-language and multi-style humor recognition scenarios. At the same time, strategies such as unsupervised pre-training and cross-modal alignment can also be introduced to reduce dependence on annotated corpora and improve the practicality of the model in low-resource scenarios.

Finally, It is expected that the model's lightweight nature will enable its use in real-world systems with limited resources. Future research can further evaluate its deployment feasibility and operational efficiency in mobile devices, embedded systems, and interactive voice platforms. For example: integrating humor recognition capabilities in voice assistants to enhance the naturalness and personalized experience of human-computer interaction; or combining the humor perception module with the speech recognition system (ASR pipeline) to label and understand the humorous content in comedies and talk shows in real time, thereby enhancing the audience's immersive experience in auditory media.

In conclusion, the lightweight multimodal humor recognition model provides new ideas for constructing intelligent systems with real-time performance and deployability. Future work should continue to deepen feature modeling, expand the ability to understand context, and strengthen practical application value while maintaining an efficient structure, so as to promote the application of humor recognition technology to a wider range of practical scenarios.

7.4 Impact and Relevance

This study focuses on humor recognition in spoken dialogue and contributes to the field of multimodal affective computing. Compared to traditional sentiment or emotion classification, humor is more complex because it often depends on timing, delivery, and context rather than just word meaning. The lightweight multimodal model proposed in this work addresses both practical needs (e.g., real-time performance) and theoretical questions about how spoken humor can be captured by machines.

Experimental results show that adding acoustic features—such as pitch, energy, glottal parameters—improves the model's ability to detect humor, especially in spoken settings. These nonverbal signals help highlight moments of exaggeration or punchline delivery. Contextual utterances also helps the model to understand when something is funny based on previous dialogue turns, rather than judging each sentence in isolation.

Another important aspect is the use of HCF, which provide structural information based on common humor patterns, such as repetition or contrast. These features make the model's behavior a bit easier to interpret and support the overall learning process. This aspect also relates to recent work in explainable AI, though in a more focused and practical way.

The proposed framework is not only compact and efficient but also modular and extensible. Its design makes it suitable for real-world deployment in applications such as intelligent agents, virtual companions, and interactive learning environments. By enabling real-time humor understanding on resource-limited devices, this model meets the growing need for responsive and socially aware AI systems.

Academically, the work offers a solid foundation for further research in multimodal humor detection. It encourages future studies to explore richer fusion strategies, cultural variation in humor, and the inclusion of additional modalities such as facial expressions or laughter. Overall, the study provides practical insights into how multimodal cues can be leveraged for nuanced language understanding, contributing to more natural and engaging human-computer interactions.

References

Amiriparian, S., Christ, L., König, A., Meßner, E.-M., Cowen, A., Cambria, E., & Schuller, B. W. (2022). Muse 2022 challenge: Multimodal humour, emotional reactions, and stress. In *Proceedings of the 30th acm international conference on multimedia* (pp. 7389–7391).

Baltrušaitis, T., Robinson, P., & Morency, L.-P. (2016). Openface: an open source facial behavior analysis toolkit. In *2016 ieee winter conference on applications of computer vision (wacv)* (pp. 1–10).

Bertero, D., & Fung, P. (2016). A long short-term memory framework for predicting humor in dialogues. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 130–135).

Bryant, G. A. (2010). Prosodic contrasts in ironic speech. *Discourse Processes*, *47*(7), 545–566.

Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R., & Poria, S. (2019). Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv preprint arXiv:1906.01815*.

Chen, P.-Y., & Soo, V.-W. (2018). Humor recognition using deep learning. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 2 (short papers)* (pp. 113–117).

Choube, A., & Soleymani, M. (2020). Punchline detection using context-aware hierarchical multimodal fusion. In *Proceedings of the 2020 international conference on multimodal interaction* (pp. 675–679).

Degottex, G., Kane, J., Drugman, T., Raitio, T., & Scherer, S. (2014). Covarep—a collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 960–964).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171–4186).

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).

Hasan, M. K., Rahman, W., Zadeh, A., Zhong, J., Tanveer, M. I., Morency, L.-P., et al. (2019). Ur-funny: A multimodal language dataset for understanding humor. *arXiv preprint arXiv:1904.06618*.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Liu, H., & Singh, P. (2004). Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, *22*(4), 211–226.

Mao, J., & Liu, W. (2019). A bert-based approach for automatic humor detection and scoring. *IberLEF@ SEPLN*, *2421*, 197–202.

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 174–184).

Naz, A., Farooq, M. U., & Jabeen, S. (2023). Prosodic analysis of humor in stand-up comedy. *Journal of English Language, Literature and Education*, *5*(3), 1–25.

Patro, B. N., Lunayach, M., Srivastava, D., Singh, H., Namboodiri, V. P., et al. (2021). Multimodal humor dataset: Predicting laughter tracks for sitcoms. In *Proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 576–585).

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., & Morency, L.-P. (2017). Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 873–883).

Pramanick, S., Roy, A., & Patel, V. M. (2022). Multimodal learning using optimal transport for sarcasm and humor detection. In *Proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 3930–3940).

Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech communication*, *53*(9-10), 1062–1087.

Weller, O., & Seppi, K. (2019). Humor detection: A transformer gets the last laugh. *arXiv preprint arXiv:1909.00252*.

Weller, O., & Seppi, K. (2020). The rjokes dataset: a large scale humor collection. In *Proceedings of the twelfth language resources and evaluation conference* (pp. 6136–6141).

Xu, H., Liu, W., Liu, J., Li, M., Feng, Y., Peng, Y., … Wang, M. (2022). Hybrid multimodal fusion for humor detection. In *Proceedings of the 3rd international on multimodal sentiment analysis workshop and challenge* (pp. 15–21).

Xu, M., Chen, S., Lian, Z., & Liu, B. (2023). Humor detection system for muse 2023: contextual modeling, pesudo labelling, and post-smoothing. In *Proceedings of the 4th on multimodal sentiment analysis challenge and workshop: Mimicked emotions, humour and personalisation* (pp. 35–41).

Yang, D., Lavie, A., Dyer, C., & Hovy, E. (2015). Humor recognition and humor anchor extraction. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 2367–2376).

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Zhou, X. (2024). A review of humor recognition research. *Applied and Computational Engineering*, *109*, 50–56.

Appendix

A AI Declaration

I hereby declare that this Master's thesis was written independently by me and that the work herein is my own, except where explicitly stated otherwise in the text. This thesis has not been submitted for any other degree or professional qualification, nor has it been published. Where the work of others has been used (from any source: printed, internet or other), it has been duly acknowledged and referenced.

During the preparation of this thesis, I used **ChatGPT** for the following purposes:

- In Section 2.3 (Approach for Humor Detection), I used the tool to help organize the structure of the literature review, and check for logical consistency. It also assisted with improving sentence fluency and correcting grammar where needed.;
- In Section 5.2 (Ablation Results), the tool helped format table templates and assisted in restructuring the result summaries to improve clarity, coherence, and conciseness;
- In Sections 7.1 (Challenges) and 7.4 (Impact and Relevance), I used the tool to refine language flow, improve grammar, and enhance the logical progression of the arguments. No conceptual content was generated by AI; the ideas and insights were fully developed by me.
- In other parts of the thesis, I occasionally used ChatGPT for minor grammar corrections and to suggest smoother sentence structures. All outputs were reviewed and substantially revised before inclusion.

All content generated with the aid of this tool was subsequently reviewed, verified, and substantially modified by me. I take full responsibility for the contents and conclusions of this thesis.

Yinzi Wang
June 2025