university of groningen

campus fryslân

An Exploration of Cross-Lingual Model Transfer in Multimodal Sarcasm Detection

Meiling Zhang

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

An Exploration of Cross-Lingual Model Transfer in Multimodal Sarcasm 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
Dr. Shekhar Nayak (Voice Technology, University of Groningen)
with the second reader being
Xiyuan Gao (Voice Technology, University of Groningen)

Meiling Zhang (S5511038)

June 11, 2025

Acknowledgements

First and foremost, I would like to express my heartfelt gratitude to my supervisors, Dr. Shekhar Nayak and Xiyuan Gao. It has been a true honor to participate in the development of the MCSD database. It is thanks to the use of MCSD that this thesis became possible. I am deeply thankful for the generous time, patience, and thoughtful guidance they devoted to me throughout this process. Their mentorship has been like a lighthouse, illuminating my path and guiding me forward, especially during times when I felt lost or made little progress. Their excellent qualities—rigorous academic attitude, enthusiasm for research, and unwavering perseverance—have always served as examples for me to follow. Their passion for research has inspired me deeply and has inspired me to look forward to the possibility of pursuing an academic career in the future.

I would also like to express my sincere appreciation to Dr. Matt Coler. His Research Design course had a profound influence on my way of thinking about academic research. His keen attention to detail and ability to pinpoint the essence of a problem have benefited me tremendously, teaching me to approach research with greater depth.

Finally, I want to thank all those who have supported and accompanied me throughout my academic journey. Whether classmates, friends, or family, your encouragement and support have been a constant source of motivation for me. It is precisely these moments of warmth that enabled me to complete this thesis and made my journey of growth more solid and meaningful.

Abstract

Sarcasm detection poses unique challenges due to the contrast between literal expressions and intended meaning, especially in spoken and multimodal communication. Misinterpretation of sarcasm can negatively impact sentiment analysis, human-computer interaction, and online content moderation. While significant progress has been made for English text, robust and generalizable approaches for other languages—especially tonal languages such as Mandarin Chinese—remain underexplored, as does the integration of multimodal cues.

This study presents a transfer learning framework for multimodal sarcasm detection across English and Mandarin Chinese. Models are constructed and evaluated using three complementary modalities: text, audio, and visual information. For text features, BERT-based sentence embeddings capture deep semantic and contextual nuances. High-level audio features are extracted using VGGish, a deep audio representation model pre-trained on large-scale audio datasets; these features may implicitly reflect intonation, emotion, and other paralinguistic cues. ResNet-152 is employed to extract high-level visual representations from video segments, capturing facial expressions and gestures relevant to sarcasm. These modality-specific features are integrated through a fusion mechanism to provide complementary information, enabling comprehensive multimodal modeling.

Experiments are conducted on the public English MUStARD dataset and MCSD, a Mandarin Chinese sarcasm dataset, both containing aligned multimodal data. The task is formulated as binary classification using support vector machines as the baseline classifier. To address the scarcity of labeled Mandarin data, we design cross-lingual transfer learning experiments in both zero-shot and few-shot settings. In the zero-shot scenario, models trained exclusively on English data are directly evaluated on the Chinese test set; in the few-shot scenario, a small number of labeled Chinese samples adapt the model, simulating low-resource transfer conditions.

In cross-lingual few-shot transfer from English (MUStARD) to Mandarin (MCSD), providing 40 labeled target samples increases the macro F1 of multimodal fusion (text+audio+video) from 46.8% (zero-shot) to 61.2%. Conversely, transferring from MCSD to MUStARD under the same setting, the macro F1 of multimodal fusion improves from 47.9% to 63.6%. Similar improvements are observed for text and audio modalities.

These findings highlight the effectiveness and generalizability of few-shot multimodal transfer learning in both directions. BERT-based text embeddings and VGGish-based audio features contribute most to cross-lingual generalization, while ResNet-based visual features provide complementary cues. The greater performance gains for audio models when transferring to Mandarin suggest that paralinguistic cues—such as tone and prosody—may be more salient for sarcasm detection in Mandarin than in English.

This work systematically investigates multimodal sarcasm detection across English and Mandarin, demonstrating that integrating text, audio, and visual cues—together with transfer learning—enables robust performance in both high- and low-resource settings. Our results show that zero-shot and few-shot cross-lingual adaptation can effectively extend sarcasm detection to underexplored languages and modalities.

Keywords: Sarcasm Detection, Multimodal Learning, Transfer Learning, Cross-Lingual, Mandarin Chinese

Contents

1 Introduction

Sarcasm, in which the intended meaning contrasts with the literal expression Eke, Norman, Liyana Shuib, and Nweke (2020), is a linguistically complex phenomenon that poses challenges for both human understanding and machine interpretation. In natural communication, sarcasm often relies on subtle shifts in tone, exaggerated expressions, and contextual cues that go beyond surface-level semantics. Early computational approaches to sarcasm detection were largely rule-based or relied on sentiment inversion techniques applied to textual inputs Joshi, Sharma, and Bhattacharyya (2015). These methods, however, often failed to capture the pragmatic and multimodal nature of sarcasm, especially in dynamic, informal settings such as social media.

From a social perspective, accurate sarcasm detection holds practical value for content moderation, sentiment analysis, and engagement modeling. On platforms like YouTube and Twitter, the misclassification of sarcasm can distort sentiment measurements Schifanella, de Juan, Tetreault, and Cao (2016), compromise moderation decisions, and reduce the accuracy of content recommendations. Sarcastic but benign expressions may be incorrectly flagged as toxic, while negative sarcasm may go undetected. Given these risks, the ability to accurately detect sarcasm is increasingly important.

Recent advancements in artificial intelligence and affective computing have demonstrated that incorporating multimodal features—such as vocal tone, facial expression, and body language—significantly improves the detection of sarcasm Farabi, Ranasinghe, Kanojia, Kong, and Zampieri (2024). Text-only models often misinterpret sarcastic cues, particularly when contextual or prosodic indicators are absent. To address this, researchers have shifted toward multimodal architectures that integrate textual, acoustic, and visual signals. Gao, Nayak, and Coler (2024) highlight a clear evolution in sarcasm detection, from early rule-based and unimodal approaches to deep learning classifiers, and more recently to sophisticated multimodal models that incorporate complex communicative cues through attention mechanisms. Numerous studies have attempted to improve sarcasm detection using multimodal fusion techniques. For instance, Cai, Cai, and Wan (2019) combined textual inputs with image semantics and visual attributes, showing that richer visual signals can significantly improve sarcasm classification. Similarly, Castro et al. (2019) explored sarcasm detection in videos using handcrafted features and support vector machines.

Despite progress in multimodal sarcasm detection, two critical research gaps remain. First, cross-lingual multimodal sarcasm detection—particularly transfer from English to Mandarin Chinese—has yet to be systematically investigated, limiting the general applicability of current approaches An et al. (2024). Second, there is a lack of in-depth quantitative analysis regarding the effectiveness and practical value of transfer learning strategies (such as zero-shot and few-shot adaptation) in cases where annotated sarcasm data in the target language is extremely limited Basabain, Cambria, Alomar, and Hussain (2023). To address these issues, this work systematically studies cross-lingual multimodal sarcasm detection with a focus on transfer learning. We utilize the MUStARD English sarcasm dataset and the MCSD Mandarin Chinese sarcasm dataset, extracting features from text, audio, and video modalities using BERT, VGGish, and ResNet-152, respectively. The performance and generalization capabilities of the models are systematically evaluated in both zero-shot and few-shot transfer settings, simulating realistic scenarios with limited annotated target-language data.

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 implications and insights
- Section 7 concludes with key findings and future directions

1.1 Research Questions and Hypotheses

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

> **How effective are transfer learning strategies (zero-shot and few-shot) for multimodal sarcasm detection from English to Mandarin Chinese, and what are the respective contributions of textual, audio, and visual modalities in cross-lingual settings?**

This main question can be broken down into the following sub-questions:

- **RQ1**: To what extent does integrating text, audio, and video features improve sarcasm detection accuracy in cross-lingual transfer scenarios compared to unimodal approaches?
- **RQ2**: How much does few-shot transfer learning (with a limited number of annotated Mandarin samples) enhance model performance relative to zero-shot transfer?
- **RQ3**: What are the respective impacts of text, audio, and video modalities on overall performance in cross-lingual sarcasm detection between English and Mandarin?

Based on these research questions, the following hypotheses are proposed:

- **H1**: Multimodal transfer learning—integrating text, audio, and video features—will significantly outperform unimodal approaches (such as text-only or audio-only) in cross-lingual sarcasm detection Yue, Shi, Mao, Hu, and Cambria (2024). This is because multimodal signals provide complementary contextual and affective information that single modalities cannot fully capture Castro et al. (2019).
- **H2**: Few-shot transfer learning, which leverages a small number of annotated Mandarin samples for adaptation, will yield substantial improvements in sarcasm detection performance compared to zero-shot transfer Gao, Nayak, and Coler (2022). This demonstrates the data efficiency and practical value of few-shot adaptation in low-resource target domains.
- **H3**: Among the three modalities, text and audio features (extracted via BERT and VGGish, respectively) will contribute more to cross-lingual sarcasm detection than video features (extracted via ResNet-152), due to the dominant roles of semantics and prosody in the expression and interpretation of sarcasm Tomar, Tiwari, Saha, and Saha (2023).

2 Literature Review

This section provides a comprehensive review of research on multimodal sarcasm detection, with a particular emphasis on the development of transfer learning strategies from English to Mandarin Chinese. While significant progress has been made in sarcasm detection—especially through the use of multimodal approaches and deep learning—major challenges persist regarding cross-lingual transfer and the scarcity of annotated sarcasm corpora in non-English languages. This review identifies these persistent gaps and highlights the necessity for robust cross-lingual models.

2.1 Search Strategy and Selection Criteria

To ensure a comprehensive overview of recent developments in multimodal sarcasm detection and transfer learning, a systematic literature review was conducted in May 2025 using three major academic databases: Google Scholar, ISCA Archive, and IEEE Xplore. The search strategies and criteria were as follows.

For Google Scholar, the query ”multimodal sarcasm detection” and ”transfer learning” were used, restricted to publications from 2015 to 2025, resulting in approximately 244 results.

In the ISCA Archive, the keyword ”sarcasm” was used, with the publication years limited to 2015–2025, yielding 8 relevant papers.

In IEEE Xplore, the search term ”multimodal sarcasm detection” was applied, and the results were filtered to include only conference papers published between 2019 and 2025, resulting in 57 papers.

The initial screening process involved reviewing the titles and abstracts, followed by a full-text assessment. Inclusion criteria required studies to be peer-reviewed and published in English, report original experimental work with clear methodological descriptions, include at least two modalities (such as text, audio, or visual data), use publicly available datasets or provide sufficient documentation for reproducibility, and report standard evaluation metrics such as accuracy, precision, recall, or F1-score. Exclusion criteria ruled out studies published before 2015, those using only unimodal input, lacking quantitative evaluation, relying on inaccessible datasets, or lacking sufficient experimental detail.

After applying these criteria, a total of 40 studies were retained for detailed review and synthesis in this research.

2.2 Progress in Multimodal Sarcasm Detection

Early research on sarcasm detection primarily relied on textual data, using rule-based systems Veale Tony and Hao Yanfen (2010) or sentiment-inversion techniques Riloff et al. (2013)to identify incongruity between literal and intended meaning Joshi et al. (2015). While these approaches offered basic insights, their performance was limited in real-world settings, especially on informal platforms like social media, where sarcasm often relies on subtle cues that are not explicitly present in text Joshi, Bhattacharyya, and Carman (2016). With the advent of deep learning, models began to capture more complex linguistic features, improving accuracy in text-based sarcasm classification Poria, Cambria, Hazarika, and Vij (2017). However, these systems still struggled with ambiguous or context-dependent expressions. This limitation prompted a shift toward multimodal approaches Schifanella et al. (2016), which incorporate acoustic features such as intonation and

prosody, as well as visual cues like facial expressions and gestures Castro et al. (2019). Multimodal sarcasm detection leverages the complementary strengths of each modality to resolve ambiguities that text alone cannot capture Sangwan, Akhtar, Behera, and Ekbal (2020). Datasets such as MUSTARD and its extensions enabled the development of models that simultaneously process text, speech, and video. Researchers experimented with various fusion strategies. Cai et al. (2019); Schifanella et al. (2016)—ranging from early feature concatenation to late decision-level integration—to improve performance. These studies demonstrated that combining modalities significantly enhances sarcasm recognition, particularly in conversational or emotionally expressive contexts. However, most studies to date have focused on English-language datasets, leaving open questions regarding cross-lingual generalizability and cultural variation in sarcasm expression.

2.3 Cross-Lingual and Transfer Learning for Sarcasm Detection

The evolution of transfer learning and few-shot paradigms in speech and language understanding has followed the broader advances in machine learning. Early research on transfer learning was inspired by the human ability to generalize prior knowledge from one context to another, and initially focused on homogeneous tasks within similar feature spaces, such as text classification and sentiment analysis Zhuang et al. (2020). In these early stages, typical approaches relied on feature-based transfer or domain adaptation strategies, requiring substantial annotated data in both source and target domains.

With the emergence of deep learning, especially large-scale pre-trained models like word embeddings and later BERT, transfer learning methods became increasingly effective and prevalent, allowing robust adaptation from resource-rich domains (such as English) to low-resource languages Song, Wang, Mondal, and Sahoo (2022). These developments enabled the adaptation of models pre-trained on massive English corpora to new languages or domains with limited annotated data, and also facilitated the use of multimodal signals.

As the limitations of fully supervised learning became more apparent, few-shot and zero-shot learning paradigms were introduced. Originally popularized in computer vision, few-shot learning approaches, such as metric learning and meta-learning, rapidly extended to NLP and speech, providing solutions for scenarios where only a handful of labeled target samples are available Feng and Chaspari (2023); Song et al. (2022). Metric learning methods, particularly those based on Siamese neural networks, have proven effective in audio classification, speaker verification, and emotion or facial expression recognition, learning distance-based embeddings that allow effective knowledge transfer even with minimal supervision Feng and Chaspari (2023).

In the context of speech-based affective computing, traditional transfer learning methods—such as adaptive SVMs, neural network fine-tuning, progressive neural networks, and adversarial learning—have achieved promising results for cross-corpus or cross-domain emotion recognition, but typically require a relatively large number of labeled samples from the target domain Feng and Chaspari (2023). Few-shot learning, by contrast, provides a compelling alternative, especially when the target domain is low-resource or exhibits significant distributional differences from the source domain.

Early sarcasm detection studies were predominantly focused on English-language data, leveraging large-scale annotated corpora sourced from social media and television show dialogues Castro et al. (2019). These approaches typically relied on supervised learning frameworks, which require a significant volume of labeled examples to achieve robust performance Wang et al. (2022). With the

rise of deep learning, multimodal sarcasm detection methods—incorporating text, audio, and visual modalities—have demonstrated substantial gains in accuracy on English benchmarks Ray, Mishra, Nunna, and Bhattacharyya (2022).

However, extending sarcasm detection to other languages, particularly Mandarin Chinese, presents fundamental challenges. A key barrier is the lack of large, high-quality annotated sarcasm corpora for Mandarin; most existing resources are limited in both scale and linguistic diversity Yue et al. (2024).This data scarcity makes supervised learning approaches largely infeasible for sarcasm detection in low-resource languages. Furthermore, sarcasm is inherently context-dependent and is often conveyed through language-specific cues—such as intonation, syntax, or cultural references—that do not easily transfer across languages Peters, Wilson, Boiteau, Gelormini-Lezama, and Almor (2016).

To overcome these limitations, research in natural language processing has extensively explored transfer learning and domain adaptation Zhuang et al. (2020). The prevailing paradigm involves pre-training models on large English datasets and adapting the learned representations to target languages with limited labeled data. In related tasks such as sentiment analysis and machine translation, advances in multilingual embeddings and joint semantic spaces have enabled robust cross-lingual transfer and zero-shot learning, where models can generalize to new languages without requiring additional annotation Artetxe and Schwenk (2019). However, the effectiveness of such methods for sarcasm detection is constrained by the unique challenges of multimodal and cultural transfer, as well as the lack of parallel multimodal sarcasm datasets in most non-English Song et al. (2022).

In response, few-shot and zero-shot learning strategies have gained attention as promising solutions, enabling models to leverage knowledge from resource-rich domains and adapt to new languages with minimal supervision Eriguchi, Johnson, Firat, Kazawa, and Macherey (2018).

2.4 Summary and Research Gap

Although transfer learning and few-shot learning have shown strong results in many areas of natural language and speech processing, their use in multimodal sarcasm detection—especially across different languages—is still rare Huang et al. (2021). Most research so far has focused on English data and has tested models only in single-language settings.

There are still few studies that test how well zero-shot and few-shot methods work for sarcasm detection in other languages, such as Mandarin Chinese. Several challenges remain: First, there are not enough large, well-annotated sarcasm datasets in languages other than English. Second, it is difficult to transfer cues that are specific to a language or a data type, such as voice tone or facial expression, between languages. Third, there are no standard benchmarks for testing cross-lingual and multimodal sarcasm detection.

Because of these gaps, there is a need for more studies that use both zero-shot and few-shot learning and combine text, audio, and video. This work aims to fill this gap. We test how well multimodal sarcasm detection models trained on English can be transferred to Mandarin, using datasets like MUStARD and MCSD. Both zero-shot and few-shot setups are studied, and analyze how helpful each modality is. The results will help in building better sarcasm detection systems that work across languages.

Table 1: Summary of Key Literature

Reference	Key Findings	Theme
An et al. (2024)	Propose prompt learning framework with data augmentation and contrastive learning, significantly improving cross-lingual sarcasm detection.	Sarcasm Detection
Artetxe & Schwenk (2019)	Develop massively multilingual sentence embeddings, enabling zero-shot cross-lingual transfer across 93 languages without target data.	Transfer Learning
Basabain et al. (2023)	Introduce label-semantic augmentation for few/zero-shot Arabic text classification, improving feature extraction and performance.	Transfer Learning
Cai et al. (2019)	Propose hierarchical fusion model for multimodal sarcasm detection in Twitter, integrating text and image features for better results.	Sarcasm Detection
Castro et al. (2019)	Present multimodal sarcasm detection using text, audio, and video; emphasize importance of visual and acoustic cues.	Sarcasm Detection
Eke et al. (2020)	Systematic review of sarcasm identification, highlighting major challenges and open research directions.	Sarcasm Detection
Eriguchi et al. (2018)	Use multilingual NMT for zero-shot cross-lingual classification, enabling direct transfer without target-language training data.	Transfer Learning
Farabi et al. (2024)	Survey multimodal sarcasm detection, summarizing fusion techniques, datasets, and evaluation strategies.	Sarcasm Detection
Feng & Chaspari (2023)	Apply few-shot Siamese neural networks to emotion recognition from speech, showing effectiveness in low-resource scenarios.	Transfer Learning
Gao et al. (2024)	Improve sarcasm detection by attention-based fusion of speech and text, leveraging emotional and sentiment cues.	Sarcasm Detection
Huang et al. (2021)	Multilingual multimodal pre-training enables zero-shot cross-lingual transfer of vision-language models.	Transfer Learning
Joshi et al. (2015)	Highlight the role of context incongruity in sarcasm detection, proposing methods to leverage it.	Sarcasm Detection
Joshi et al. (2016)	Comprehensive survey on automatic sarcasm detection, reviewing models from rule-based to deep learning.	Sarcasm Detection
Peters et al. (2016)	Find native language speakers process sarcastic cues in speech more efficiently than non-natives.	Sarcasm Detection
Poria et al. (2017)	Deep convolutional networks for sarcasm detection in tweets, outperforming traditional feature-based approaches.	Sarcasm Detection

Reference	Key Findings	Theme
Ray et al. (2022)	Introduce a multimodal corpus for emotion recognition in sarcasm, supporting affective computing research.	Sarcasm Detection
Riloff et al. (2013)	Model sarcasm as contrast between positive sentiment and negative situations, proposing detection frameworks.	Sarcasm Detection
Sangwan et al. (2020)	Explore multimodality and conversational context for sarcasm detection, showing non-textual cues are valuable.	Sarcasm Detection
Schifanella et al. (2016)	Pioneer multimodal sarcasm detection by combining text and images from social media posts.	Sarcasm Detection
Song et al. (2022)	Survey few-shot learning, its evolution, applications, challenges, and opportunities in NLP.	Transfer Learning
Tomar et al. (2023)	Show that modality order matters in multi-modal sarcasm detection, with tone having greater impact than visual cues.	Sarcasm Detection
Veale & Hao (2010)	Early system to detect ironic intent in creative comparisons, foundational for sarcasm research.	Sarcasm Detection
Wang et al. (2022)	Propose unsupervised masking and generation methods for sarcasm detection, avoiding labeled data dependency.	Sarcasm Detection
Yue et al. (2024)	Introduce SarcNet, a large-scale multilingual multimodal sarcasm detection dataset in English and Chinese.	Sarcasm Detection
Zhuang et al. (2020)	Comprehensive survey of transfer learning theory and applications in NLP, including cross-lingual transfer.	Transfer Learning

3 Methodology

This section outlines the overall methodology employed to investigate multimodal sarcasm detection across English and Chinese, with a particular focus on transfer learning strategies. We first describe the datasets and their relevance to our research questions, followed by the core modeling approaches for fusing textual, audio, and visual modalities. Theoretical motivations and the rationale for each methodological choice are presented, emphasizing how our approach enables a systematic evaluation of both in-domain and cross-lingual sarcasm detection. Detailed implementation and experimental procedures are deferred to Section 4. By clearly separating conceptual design from experimental specifics, this section aims to provide a coherent framework for understanding how our methods address research objectives.

3.1 Dataset Description

This research utilizes two primary datasets for multimodal sarcasm detection: the MUStARD dataset for English and the MCSD dataset for Mandarin Chinese.

MUStARD is a publicly available multimodal sarcasm dataset compiled from popular TV shows, including Friends, The Big Bang Theory, and The Golden Girls. The dataset contains audiovisual utterances manually annotated for sarcasm, with each utterance accompanied by its conversational context—namely, the preceding turns in the dialogue. The balanced version of MUStARD comprises 690 utterances (345 sarcastic, 345 non-sarcastic), each with three modalities: text, audio, and video. On average, utterances are about 5.2 seconds long, and the context spans approximately 14 seconds. Each utterance in our dataset is coupled with its context utterances, which are preceding turns by the speakers participating in the dialogueCastro et al. (2019).

The MCSD (Mandarin Chinese Sarcasm Dataset) is a large-scale Mandarin Chinese dataset specifically designed for cross-lingual sarcasm detection. The dataset is constructed from episodes of the popular Chinese stand-up comedy show Rock & Roast. It contains 2,705 annotated clips, with a total duration exceeding 10 hours, covering both sarcastic and non-sarcastic examples. Each sample provides aligned text, audio, and video modalities, enabling comprehensive multimodal analysis. The MCSD dataset features natural, spontaneous conversational speech with authentic multimodal cues, making it a valuable resource for evaluating the generalizability of sarcasm detection models in a cross-lingual context.

Both datasets are highly relevant to the research questions of this study, as they enable systematic exploration of multimodal and cross-lingual sarcasm detection. MUStARD allows for benchmarking on widely-used English content, while MCSD provides a unique testbed for transfer learning and domain adaptation between English and Mandarin. Notable features include balanced class distributions in MUStARD and the scale and naturalness of MCSD. Limitations include potential domain biases because the dataset is drawn mainly from TV shows and stand-up comedy.

The combination of these datasets justifies their choice: together, they support comprehensive investigation into the effectiveness and transferability of multimodal approaches to sarcasm detection across languages and domains.

3.2 Core Methods and Models

The core methodological framework for this research centers on modular, supervised learning for multimodal sarcasm detection and cross-lingual transfer. Each utterance is represented by three distinct modalities: text, audio, and video.

Text features are obtained using a pretrained multilingual BERT model Devlin, Chang, Lee, and Toutanova (2019). For each utterance, the mean pooled [CLS] embeddings from the top four hidden layers are used to produce a contextualized sentence vector. Audio features are extracted with VGGish Simonyan and Zisserman (2015), a convolutional network pre-trained on large-scale audio data, resulting in 128-dimensional embeddings that capture prosodic and paralinguistic information. Video features are extracted by applying a pre-trained ResNet-152 network He, Zhang, Ren, and Sun (2016) to each video frame, collecting the 2048-dimensional activations from the final average pooling (pool5) layer.

For each sample, all available modality-specific features are concatenated at the feature level (early fusion) to form a single input vector. This fused feature is then passed to a support vector machine (SVM) classifier with a radial basis function (RBF) kernel, which performs binary sarcasm prediction. The methodology is designed to flexibly enable unimodal, bimodal, or trimodal settings, as well as the inclusion or exclusion of conversational context, depending on experimental configuration.

In addition to independent training and evaluation on each language, we further perform joint training by combining source and target data. Under this joint training regime, the model is trained on a mixture of labeled samples from both languages to encourage more effective cross-lingual feature alignment. Following joint training, two transfer learning scenarios are systematically evaluated: (1) zero-shot transfer, where the model is trained only on source language data and directly tested on the target language without access to target labels Pourpanah et al. (2023); and (2) few-shot transfer, where a limited number of labeled target-language examples are incorporated into training to facilitate adaptation Song et al. (2022). This comprehensive experimental design enables us to compare independent, joint, zero-shot, and few-shot learning paradigms for cross-lingual multimodal sarcasm detection.

3.3 Technical Framework

The technical framework of this study is based on modular supervised learning for multimodal sarcasm detection and cross-lingual transfer. All experiments are conducted in Python, utilizing open-source libraries including HuggingFace Transformers for BERT-based text encoding, `torchvggish` for audio feature extraction, and `torchvision` for video feature processing.

For each utterance or conversational context, we extract modality-specific features as follows. Let $\mathbf{x}_{\text{text}} \in \mathcal{R}^{d_t}$ denote the 768-dimensional BERT embedding, $\mathbf{x}_{\text{audio}} \in \mathcal{R}^{d_a}$ the 128-dimensional VGGish embedding, and $\mathbf{x}_{\text{video}} \in \mathcal{R}^{d_v}$ the 2048-dimensional ResNet-152 pool5 embedding. For the MUStARD dataset, both utterance-level and context-level features are evaluated, while for the MCSD and joint datasets, only context-level (conversational context) features are available due to the data format Castro et al. (2019).

All available features are concatenated to form a single input vector using early fusion:

$$\mathbf{x}_{\text{fusion}} = [\mathbf{x}_{\text{text}}; \mathbf{x}_{\text{audio}}; \mathbf{x}_{\text{video}}]. \tag{1}$$

This fused representation is used as input to a support vector machine (SVM) classifier with a radial basis function (RBF) kernel. The SVM seeks a hyperplane defined as:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right), \tag{2}$$

where $K(\mathbf{x}_i, \mathbf{x}) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}\|^2)$ is the RBF kernel, α_i are the learned weights, $y_i \in \{-1, 1\}$ are class labels, and b is the bias term.

For cross-lingual transfer, both zero-shot and few-shot paradigms are adopted. In the zero-shot setting, the classifier is trained solely on source-language data and evaluated directly on target-language data without further adaptation. In the few-shot setting, a small number of labeled target samples are integrated into the training set to facilitate transfer.

Specifically, the MUStARD dataset is evaluated under both utterance-level and context-level configurations, enabling detailed analysis of contextual effects. The MCSD and joint datasets only provide context-level video segments, and all experiments on these datasets are therefore performed with context-aware models.

Model evaluation is based on weighted precision, recall, and F1 score, and all feature extraction and model training are performed on Linux servers equipped with NVIDIA GPUs.

3.4 Evaluation Methodology

Model evaluation in this study follows robust and widely-accepted classification protocols. For all in-domain experiments, stratified 5-fold cross-validation is employed to ensure balanced class distributions and to provide a reliable estimate of model performance. In the joint training and transfer learning experiments (including both zero-shot and few-shot settings), predefined train/test splits are used to reflect realistic deployment scenarios and guarantee non-overlapping partitions between training and evaluation sets.

For each experimental run, several metrics are computed and reported. These include overall accuracy as well as precision, recall, and F1-score, both on a per-class basis and as macro- and weighted-averaged scores. The confusion matrix is also provided to enable qualitative error analysis and to highlight common misclassifications.

Model performance is systematically compared across different modality combinations, including unimodal, bimodal, and trimodal settings, as well as with and without conversational context.

3.5 Ethics and Research Integrity

This research was conducted in accordance with standard institutional and academic ethical guidelines. The study did not involve any human subjects, new data collection, or experimental intervention on individuals. All datasets used, including MUStARD and MCSD, are derived from publicly available television shows and stand-up comedy performances, containing no personally identifiable or sensitive information.

Throughout this project, copyright, privacy, and data protection regulations were strictly observed. All analyses were performed at the aggregate, anonymized level, and no attempt was made to infer or track the identities of any individuals.

The research outcomes and models are intended solely for scientific and academic purposes, with no use in surveillance, profiling, or automated decision-making regarding individuals or groups. By adhering to these principles, the project maintains a high standard of research integrity, transparency, and social responsibility.

3.5.1 Data Ethics and Privacy

This research does not involve any original data collection, human participants, or sensitive personal information. All experiments are conducted exclusively on publicly available datasets (MUStARD and MCSD), which are compiled from broadcast television shows and stand-up comedy programs. As such, there is no requirement for participant consent, privacy protection measures, or anonymization procedures beyond those already in place in the original data releases. Data usage complies with the terms of use and licensing associated with the original datasets.

3.5.2 FAIR Principles Implementation

In accordance with the FAIR (Findable, Accessible, Interoperable, Reusable) principles, this research relies solely on publicly available datasets, which are already documented and indexed by the research community. The datasets can be found through persistent identifiers and standard repositories. No additional data was created or shared by this project. All experimental procedures and methodology are fully described in this manuscript to ensure transparency and support future reuse and reproducibility. The project code and relevant documentation are publicly available at `https://github.com/MeilingZhang0107/enchtransfer` to facilitate independent verification, future reuse, and reproducibility.

3.5.3 Open Science Practices

This study follows open science principles by employing publicly available benchmark datasets (MUStARD and MCSD), and by providing detailed methodological descriptions in the manuscript. All experimental procedures, feature extraction methods, and evaluation protocols are described comprehensively to facilitate independent verification and reproducibility.

To further support open science and transparent research, the codebase and relevant resources are made available at: `https://github.com/MeilingZhang0107/enchtransfer`. The repository includes documentation feature extraction, and model training, as well as guidelines for citation and contribution.

The released code is licensed under the MIT License[1], allowing broad reuse for academic and research purposes. Users are encouraged to cite this repository if the code or methods are used in derivative works. Detailed citation instructions and contribution policies are included in the repository's `README`.

3.5.4 Bias and Fairness

Potential sources of bias and fairness concerns are acknowledged in this research. The datasets used, while diverse, are drawn from English-language television series and Chinese stand-up com-

[1] `https://github.com/MeilingZhang0107/enchtransfer/blob/main/LICENSE`

edy, which may not fully represent the broader demographic, linguistic, or cultural contexts. This may lead to biases in both training and evaluation, and model performance could vary across underrepresented groups or scenarios Pagano et al. (2023). To mitigate such risks, experiments include checks for class balance and error analysis, and all findings are reported transparently with an emphasis on their limitations. The models and results are intended solely for academic research and should not be directly deployed in sensitive or decision-making applications without further fairness assessments.

3.5.5 Environmental Impact

All computational experiments for this study were conducted on the Habrok GPU cluster at the University of Groningen (RUG). The scale of model training and inference was moderate, focusing on neural network-based feature extraction and SVM classification without large-scale hyperparameter tuning or training of deep models. As such, the energy consumption and carbon footprint are relatively modest compared to resource-intensive deep learning research. Nonetheless, environmental sustainability remains a consideration for future work, and resource usage will be monitored and optimized as the project scales.

3.5.6 Reproducibility and Replicability

To support reproducibility and replicability, all methodological steps—including data preprocessing, feature extraction, experimental design, and evaluation metrics—are described in detail within this manuscript. The software environment, hardware (Habrok GPU cluster), and random seed settings are documented. Although the code and models are not yet publicly released, the provided descriptions, parameter settings, and configuration options are sufficient for knowledgeable researchers to replicate the experiments. Known limitations include potential variations due to differences in computational environments and any inherent dataset biases. Further improvements in reproducibility will be pursued in future work, possibly including code and data releases as project policies allow.

In summary, this methodology establishes a robust framework for multimodal sarcasm detection and cross-lingual transfer, leveraging state-of-the-art feature extraction, early fusion, and SVM classification. The approach is designed to investigate the contributions of different modalities, evaluate both in-domain and cross-domain scenarios, and ensure research transparency, fairness, and reproducibility. The subsequent section provides a detailed account of the experimental setup, implementation details, and evaluation protocols used to empirically validate the proposed methodology.

4 Experimental Setup

This section documents the experimental setup in comprehensive detail. Complete descriptions are provided for data preparation, feature extraction, training and evaluation protocols, and all relevant parameter settings. This enables other researchers to independently replicate the results and verify the conclusions.

All experimental code builds on or extends established open-source toolkits, notably the original MUStARD baseline implementation and the Google Research BERT repository. Any additional scripts, configuration files, and environment specifications used in this study are described below. Section 4.1 details the procedures for dataset preparation and feature extraction, while subsequent sections cover data splitting, model training, and evaluation.

By following the practices described here, this research aims to meet the highest standards of scientific reproducibility.

4.1 Data Preparation

All experiments in this study are conducted on two multimodal sarcasm detection datasets. The first is MUStARD, an English-language dataset that provides annotated MP4 video clips together with JSON files containing utterance-level transcriptions and sarcasm labels. The second is MCSD, a Mandarin Chinese dataset, where the video data is distributed as MP4 files and the corresponding transcriptions and annotations are provided in CSV format. The CSV file contains, for each sample, the sentence identifier, transcript and sarcasm label.

For textual feature extraction, utterance transcripts were parsed from the JSON annotation file in MUStARD and from the CSV file in MCSD. We used the HuggingFace Transformers library (version 4.49.0) with the `bert-base-multilingual-cased` model to encode each utterance and its context. For each sentence, the mean of the [CLS] embeddings from the top four hidden layers was used as the 768-dimensional textual representation. The extracted BERT features were stored in line-delimited JSON format for subsequent modeling. Text feature extraction was performed in a dedicated environment with `torch` 2.6.0, `numpy` 2.0.2, and `scikit-learn` 1.6.1. Random seeds were fixed in all stages of tokenization and model inference to ensure reproducibility.

Audio features were extracted by first converting each MP4 video into a 16,kHz mono-channel WAV file using FFmpeg. For each utterance, a 128-dimensional embedding was computed using the VGGish model via the `torchvggish` package (version 0.1/0.2). Audio segments shorter than 0.96 seconds were zero-padded to meet the input length requirement of VGGish. The audio feature extraction was performed in a dedicated environment with `torch` 2.2.2, `numpy` 2.1.3, `torchaudio` 2.5.1, `librosa` 0.11.0, `scikit-learn` 1.6.1, `h5py` 3.13.0, and `jsonlines` 4.0.0. All audio features were saved in pickle format.

Visual features were extracted by first decomposing each utterance-level video into JPEG frames using FFmpeg. Each frame was then processed using a pre-trained ResNet-152 model (torchvision 0.16.2, PyTorch 2.1.2) to obtain a 2048-dimensional pool5 feature vector. Visual feature extraction was performed in an environment with `torch` 2.1.2, `torchvision` 0.16.2, `numpy` 1.26.4, and `h5py` 3.12.1. The mean-pooled vector across all frames for each utterance was used as its visual representation and stored in HDF5 format.

All feature extraction was performed on a server equipped with NVIDIA GPUs (CUDA 12.1) for deep model inference and CPUs for frame extraction. Random seeds were fixed throughout all pre-

processing steps, and all scripts used for data processing are available to ensure full reproducibility. The final extracted features are stored as follows: BERT features in `.jsonl` files, VGGish features in `.p` files, and ResNet features in HDF5 files.

We extracted modality-specific features using publicly available code and official pretrained models to ensure reproducibility and consistency with prior work.

4.2 Data Splitting

For all supervised in-domain experiments, we employ five-fold stratified cross-validation (`StratifiedKFold` with $n_splits = 5$) to ensure robust and balanced evaluation. In each fold, four splits are used for training and one split is reserved for testing, with class distributions preserved in both sets. All split indices are fixed by random seed for reproducibility, and the same splits are consistently used across unimodal and multimodal experiments.

For cross-lingual transfer experiments, we evaluate both directions: English to Mandarin (MUStARD to MCSD) and Mandarin to English (MCSD to MUStARD). In the zero-shot setting, the model is trained on all labeled samples from the source-domain dataset and directly evaluated on the entire target-domain dataset without access to any labeled target-domain examples during training Kodirov, Xiang, Fu, and Gong (2015). To mitigate domain shift, feature normalization parameters are fitted exclusively on the target-domain data at inference time, following standard zero-shot transfer protocols. This procedure is applied symmetrically in both transfer directions.

In the few-shot setting, a small number of labeled target-domain samples per class—specifically 10, 20, 30, 40, or 50—are randomly sampled and incorporated into the source-domain training set. The remaining target-domain samples are reserved for evaluation and are not used during training. This sampling range simulates real-world low-resource scenarios, where only a limited amount of labeled data is available in the target domain. The 10-to-50 samples per class setting balances practical feasibility with research challenge Xu, Zeng, Lian, and Ding (2022). Few-shot sampling is performed independently for each transfer direction and repeated with multiple random seeds to account for variance; final results are averaged over several runs. Stratification by sarcasm label is maintained to ensure class balance in both the few-shot training and test splits.

4.2.1 Experiment 1: In-Domain Supervised Baseline on MUStARD

The primary objective of this experiment was to establish a supervised in-domain baseline for sarcasm detection on the MUStARD dataset. Both unimodal and multimodal models were compared, and the impact of including conversational context was assessed.

Text features were extracted using the `bert-base-multilingual-cased` model (Transformers 4.49.0, PyTorch 2.6.0), yielding 768-dimensional embeddings for each utterance. In the context setting, preceding utterances were concatenated with the target utterance and jointly encoded by BERT. Audio features were extracted using VGGish (torchvggish 0.1, PyTorch 2.2.2, librosa 0.11.0), resulting in 128-dimensional embeddings. Video features were obtained via ResNet-152 (torchvision 0.16.2, PyTorch 2.1.2), producing 2,048-dimensional vectors for each utterance.

All features were standardized and used as input to a linear Support Vector Machine (SVM) classifier (scikit-learn 1.6.1, $C = 1.0$). Both unimodal (text, audio, video) and multimodal (T+A, T+A+V) configurations were evaluated in both utterance-only and context-based settings.

Performance was evaluated using weighted precision, weighted recall, and weighted F1-score, following the standard evaluation metrics in sarcasm detection. All reported results were averaged across five folds using stratified K-fold cross-validation to ensure robustness and fair comparison. For each fold, the model was trained and tested on non-overlapping splits, and performance metrics were computed on the held-out test set. Feature extraction was performed using NVIDIA V100 GPUs. This experiment serves as a reference for subsequent cross-lingual transfer experiments and quantifies the contribution of both conversational context and multimodal features in supervised sarcasm detection on English dialogue.

4.2.2 Experiment 2: In-Domain Supervised Baseline on MCSD

The goal of this experiment is to establish the in-domain baseline for sarcasm detection on the Mandarin MCSD dataset, using the same pipeline as for MUStARD to enable direct comparison.

Text features were extracted using `bert-base-multilingual-cased` (Transformers 4.49.0), resulting in 768-dimensional embeddings for each utterance or context segment. Audio features were extracted with VGGish, producing 128-dimensional embeddings. Video features were obtained from ResNet-152 (2,048 dimensions).

All features were standardized and used as input to a linear SVM classifier (scikit-learn 1.6.1, $C = 1.0$). Both unimodal and multimodal (T+A, T+A+V) experiments were conducted.

Evaluation was based on weighted precision, weighted recall, weighted F1, and macro F1, with results averaged over five runs for robustness. The same software stack and hardware as in the previous experiment were used for full consistency.

This MCSD baseline provides a reference point for Mandarin sarcasm detection and supports fair comparison with English in-domain results and cross-lingual transfer experiments.

4.2.3 Experiment 3: Joint Cross-Lingual Supervised Training (MUStARD + MCSD)

The objective of this experiment is to assess whether jointly training on both English (MUStARD) and Mandarin (MCSD) data enhances the generalizability of sarcasm detection models across languages and domains.

For both MUStARD and MCSD, text features were extracted with `bert-base-multilingual-cased` (Transformers 4.49.0, PyTorch 2.6.0). Audio features were extracted using VGGish (torchvggish 0.1, librosa 0.11.0), and video features with ResNet-152 (torchvision 0.16.2). All features were standardized and concatenated for multimodal experiments (T+A, T+A+V).

The joint training set was formed by concatenating the full MUStARD and MCSD datasets, covering both languages. Models were trained using linear SVMs (scikit-learn 1.6.1, $C = 1.0$) under the same settings as previous in-domain experiments. Evaluation was performed on a mixed test set as well as individually for each language to assess both overall and cross-lingual performance.

Weighted precision, weighted recall, and weighted F1 were reported for all modalities and combinations. Hardware and software configurations were identical to the in-domain baselines. This joint training experiment allows for the examination of feature transferability in a multilingual setting.

4.2.4 Experiment 4: Zero-Shot Cross-Lingual Transfer

The objective of this experiment is to evaluate the ability of sarcasm detection models to generalize across languages without any target-domain supervision. Specifically, the model is trained on the entire source-domain dataset (either English MUStARD or Mandarin MCSD) and directly tested on the target-domain data, simulating a true zero-shot transfer scenario.

For all zero-shot experiments, features are extracted consistently from both domains to ensure comparability. Text features are derived from `bert-base-multilingual-cased` using Transformers 4.49.0, resulting in 768-dimensional utterance or context embeddings. Audio features are obtained via VGGish (128 dimensions), and video features via ResNet-152 (2,048 dimensions). All features are standardized within the target-domain set before inference, following standard cross-domain evaluation protocol.

A linear SVM classifier (scikit-learn 1.6.1, $C = 1.0$) is trained using the full labeled source-domain data for each modality combination: unimodal (text, audio, video), bimodal (text+audio), and trimodal (text+audio+video). No target-domain labels are used during training or model selection.

The trained model is applied directly to all target-domain samples. To prevent information leakage, feature normalization (mean and variance scaling) is performed by fitting the scaler only on the target-domain features. For each modality, macro F1, accuracy, class-specific F1 scores, and confusion matrices are reported.

All experiments are conducted using the same Python software stack and hardware (NVIDIA V100 GPUs for feature extraction) as in previous experiments. Model file names and configuration keys correspond to the specific modality settings to ensure reproducibility.

This experiment provides a direct assessment of cross-lingual transferability under realistic low-resource conditions and serves as a baseline for evaluating the added value of few-shot adaptation in subsequent experiments.

4.2.5 Experiment 5: Few-Shot Cross-Lingual Transfer

This experiment investigates the effectiveness of few-shot learning for cross-lingual sarcasm detection. Specifically, we evaluate how incorporating a small number of labeled target-domain samples can improve model transfer performance from a source language (MUStARD or MCSD) to a target language. For each experimental run, we first extract all features for both the source and target domains using `bert-base-multilingual-cased` for text, VGGish for audio, and ResNet-152 for video. From the target domain, we randomly sample $k \in 10, 20, 30, 40, 50$ labeled examples per class to construct a small labeled set; the remaining target-domain samples are reserved as the test set. The entire source domain is always included in the training set. To ensure fairness and robustness, each k-shot configuration is repeated over five random seeds, and results are averaged.

All extracted features are standardized using `StandardScaler`, which is fit only on the training set (i.e., source data plus few-shot target samples) and then applied to both training and test sets to prevent information leakage. For classification, we use a support vector machine with an RBF kernel (scikit-learn 1.6.1, $C = 1.0$). Experiments are conducted for unimodal settings (text, audio, video) and multimodal combinations (text+audio, text+audio+video). For each modality and shot setting, macro F1, weighted precision, weighted recall, weighted F1, and confusion matrices are computed on the target-domain test set. All experiments are implemented in Python, ensuring comparability

across experimental conditions. The experimental procedure directly reflects a realistic cross-lingual few-shot transfer pipeline, enabling systematic analysis of data efficiency and modality contributions in low-resource sarcasm detection.

5 Results

This section presents the experimental results for sarcasm detection across multiple settings, including in-domain supervised baselines, joint training, and cross-lingual transfer via zero-shot and few-shot learning. The performance is evaluated for five modality settings: text, audio, video, text and audio (t+a), and text, audio, and video (t+a+v). All main results are summarized in Tables 2, 3, 4, 5, 6 and 7.

For each experimental condition, we report standard metrics such as accuracy and macro F1 score to facilitate comparison. Results are presented in a logical order, beginning with in-domain baselines, followed by joint training, zero-shot transfer, and few-shot adaptation in both transfer directions.

Table 2: Weighted precision, recall, and F1 (%) for SVM models on MUStARD with different modalities and input types.

Modality	**Input Type**	**Precision (%)**	**Recall (%)**	**F1 (%)**
Text	Utterance	66.1	65.7	65.7
Audio	Utterance	70.6	70.0	69.9
Video	Utterance	68.1	67.4	67.4
T+A	Utterance	68.2	67.5	67.6
T+A+V	Utterance	70.9	70.4	70.5
Text	Context	69.0	69.0	69.0
Audio	Context	66.2	65.8	65.8
Video	Context	72.5	71.9	71.9
T+A	Context	69.9	69.6	69.5
T+A+V	Context	72.7	72.3	72.3

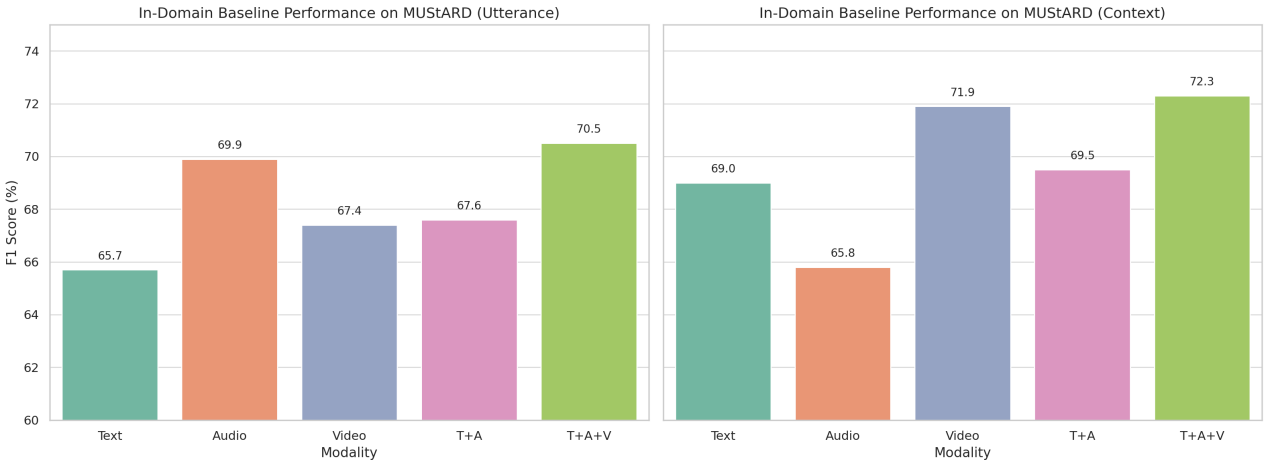

Figure 1: In-Domain Baseline Performance on MUStARD

5.1 In-Domain Baseline Performance on MUStARD

From the results, several clear trends emerge: First, contextual information markedly enhances overall performance for most modalities. For example, when only text features are used, incorporating context raises all metrics from 65.7% (utterance-level F1) to 69.0% (context-level F1), reflecting a consistent 3.3 percentage point improvement. This trend is also observed in the multimodal T+A+V setting, where F1 increases from 70.5% (utterance) to 72.3% (context). Such improvements underline the value of integrating conversational context in sarcasm detection.

Second, among unimodal models, the audio-only modality performs best at the utterance level (F1 = 69.9%), slightly surpassing both text (65.7%) and video (67.4%). This result suggests that prosodic and paralinguistic audio cues carry significant discriminative power for sarcasm detection, possibly capturing subtle tone variations or emphasis not present in text or visual streams.

Third, multimodal fusion consistently outperforms unimodal inputs. The T+A+V (utterance) model achieves a 70.5% F1, while adding context further boosts this to 72.3%, representing the highest overall performance. This demonstrates that textual, acoustic, and visual signals provide complementary information, and their integration is crucial for robust sarcasm identification Sun, Zhang, Yang, and Wang (2022).

It is also notable that the effect of context varies by modality. While context offers substantial gains for text and video, its impact on audio is mixed. For audio, the F1 actually decreases from 69.9% (utterance) to 65.8% (context). This drop may be due to noise introduced when aggregating audio from multiple utterances, which could blur prosodic cues essential for detecting sarcasm.

Furthermore, the video modality benefits the most from context, with F1 rising from 67.4% (utterance) to 71.9% (context)—an impressive gain of 4.5 percentage points. This suggests that sequential visual information, such as facial expressions and gestures over several dialogue turns, provides important context for sarcasm understanding.

Finally, the performance of bimodal models (T+A) is consistently between the best unimodal and trimodal models, confirming the incremental value added by each additional modality.

In summary, the experimental results on MUStARD confirm that (1) contextual information significantly improves sarcasm detection, especially for text and video; (2) audio cues are highly effective on their own at the utterance level; and (3) multimodal fusion, especially when context is available, delivers the most robust and accurate performance.

Table 3: In-domain SVM performance on MCSD dataset using different modalities.

Modality	**Precision (%)**	**Recall (%)**	**F1 (%)**
Text	62.8	61.7	61.5
Audio	67.4	66.8	66.9
Video	57.4	56.5	56.5
T+A	67.2	66.7	66.7
T+A+V	59.4	58.8	58.9

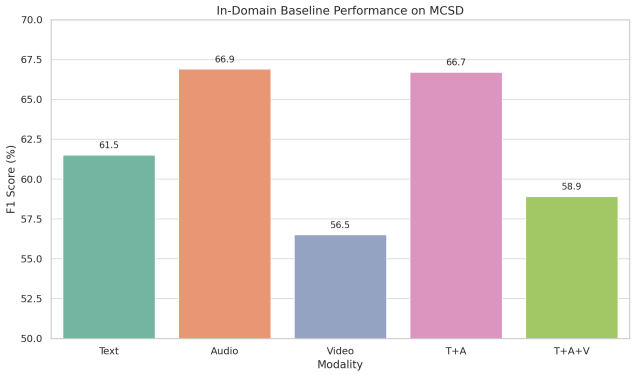

Figure 2: In-Domain Baseline Performance on MCSD

5.2 In-Domain Baseline Performance on MCSD

Table 3 shows the weighted precision, recall, and F1 scores for sarcasm detection on the MCSD dataset using different modalities and their combinations.

The best single-modality performance is achieved by audio, with a weighted F1 of 0.669, precision of 0.674, and recall of 0.668. This indicates that acoustic cues play a key role in Mandarin sarcasm detection, and the audio-based model outperforms the text-only model, which achieves an F1 of 0.615. The text model, with a precision of 0.628 and recall of 0.617, still demonstrates substantial effectiveness, highlighting the contribution of linguistic features.

The video modality alone results in a considerably lower F1 score of 0.565, suggesting that visual information by itself is less predictive for sarcasm in this dataset. One possible explanation lies in the annotation granularity of MCSD: sarcasm labels are often assigned at the context level, encompassing multiple utterances, rather than being tied to isolated statements Ghosh, Fabbri, and Muresan (2018). This labeling scheme may reduce the ability of visual models to capture specific sarcasm-related expressions within each segment.When sarcasm labels are assigned at the context level—covering multiple utterances instead of specific sentences—visual modalities such as facial expressions or gestures may struggle to align with the target sarcastic intent, thus reducing their utility in model training.

Additionally, cultural and communicative factors may affect the expressiveness of sarcastic cues Techentin, Cann, Lupton, and Phung (2021). In Mandarin conversational settings, facial expressions and gestures signaling sarcasm are frequently more subtle and less exaggerated than those typically found in English-language media, limiting the discriminative power of visual features Yue et al. (2024). Furthermore, the ambiguity of sarcasm labels across extended context may weaken the correlation between visual signals and the intended meaning. Because a sarcastic label may cover an entire dialogue or scene, individual visual frames do not necessarily correspond to overt sarcastic behavior, further reducing the utility of video-based features.

When combining modalities, the text+audio (T+A) model achieves a weighted F1 of 0.667, nearly matching the audio-only performance, while the precision and recall remain very close (0.672 and 0.667, respectively). However, adding video to form the text+audio+video (T+A+V) model leads to a noticeable decrease in performance, with the F1 dropping to 0.589. This pattern suggests that the addition of video features may introduce noise or redundancy, diluting the benefits of audio and text.

In summary, these results highlight the importance of audio and text modalities for sarcasm detection in Mandarin, while the video modality does not provide additional value and may even hinder performance when fused with other modalities.

Table 4: Performance of SVM models trained jointly on the MUStARD and MCSD datasets with context.

Modality with context	**Precision (%)**	**Recall (%)**	**F1 score (%)**
Text	62.9	62.0	62.0
Audio	66.2	65.8	65.8
Video	65.7	65.1	65.0
T+A	65.7	65.2	65.3
T+A+V	67.3	66.8	66.8

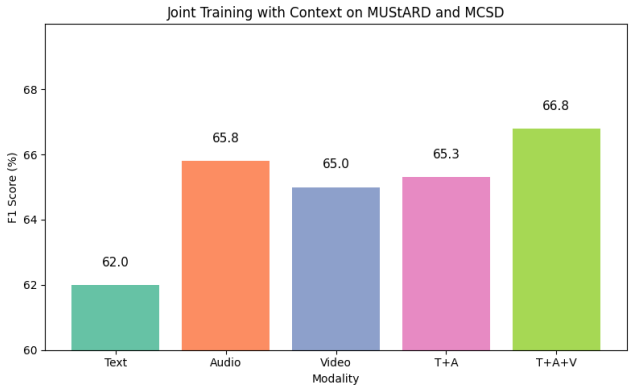

Figure 3: Joint Training with Context on MUStARD and MCSD

5.3 Joint Training with Context on MUStARD and MCSD

Table 4 presents the weighted precision, recall, and F1 scores for SVM models jointly trained on the combined MUStARD and MCSD datasets, using contextual input for all modalities. Comparing these results with in-domain baselines reveals several important trends.

First, the trimodal (text+audio+video) model with context achieves the best overall F1 score under joint training, reaching 66.8%. This result is substantially higher than the trimodal F1 for MCSD alone (58.9%), but still lower than the corresponding result for MUStARD alone (72.3%). This pattern indicates that joint training can improve performance on the more challenging MCSD data, but does not fully close the gap with the best-case results obtainable on MUStARD.

The audio-only model achieves a weighted F1 of 65.8% under joint training, which is close to its MCSD baseline (66.9%) and somewhat below its performance on MUStARD (69.9%). The text-only and video-only models obtain F1 scores of 62.0% and 65.0%, respectively, both of which fall between their respective in-domain baselines: for text, MUStARD achieves 69.0% and MCSD 61.5%; for video, MUStARD achieves 71.9% and MCSD 56.5%.

It is noteworthy that in joint training, the addition of video features helps the trimodal model regain the best overall result (F1 = 66.8%), whereas in the MCSD dataset alone, adding video had previously led to a reduction in F1. This suggests that the increased diversity and size of the training data in the joint setting can help mitigate noise or redundancy from visual features, and better exploit complementary information.

Despite these advantages, joint training does not achieve the highest scores observed on the MUStARD dataset alone. This relative performance decline can be attributed to domain shift and annotation differences between the two datasets Elsahar and Gallé (2019). Differences in conversational style, annotation granularity, and modality-specific expressions of sarcasm introduce additional variation, which increases the learning difficulty for a single unified model Helal, Hassan, Badr, and Afify (2024).

Overall, these findings demonstrate that joint training with context-aware, multimodal inputs improves generalizability and robustness, particularly benefiting performance on MCSD. However, the highest scores achieved through joint training remain lower than the best unimodal or trimodal results on MUStARD alone, reflecting inherent domain differences and annotation challenges between the datasets. The results underscore the value of fusing textual, acoustic, and visual signals, and indicate that training on multilingual, multi-source datasets is a promising direction for building more universal sarcasm detection systems.

Table 5: Zero-shot SVM performance (%) for sarcasm detection across languages.

Trained on MUStARD, Tested on MCSD				
Modality	**Accuracy (%)**	**F1 (Class 0) (%)**	**F1 (Class 1) (%)**	**Macro F1 (%)**
Text	51.8	52.7	50.8	51.8
Audio	52.0	49.0	54.7	51.8
Video	49.2	55.1	41.4	48.3
T+A	55.6	55.0	56.1	55.6
T+A+V	51.5	54.1	48.5	51.3
Trained on MCSD, Tested on MUStARD				
Text	48.8	51.6	45.8	48.7
Audio	54.2	51.2	56.8	54.0
Video	51.0	54.0	47.7	50.8
T+A	54.8	55.2	54.4	54.8
T+A+V	48.6	53.6	42.3	47.9

5.4 Zero-shot Cross-lingual Performance

Table 5 presents the results of zero-shot cross-lingual experiments. In these settings, models are trained on one language (MUStARD or MCSD) and directly evaluated on the other, without access to any labeled samples from the target language.

When trained on MUStARD and tested on MCSD, the best macro F1 score (55.6%) is obtained by the bimodal text+audio (T+A) model. The text-only (51.8%) and audio-only (51.8%) models

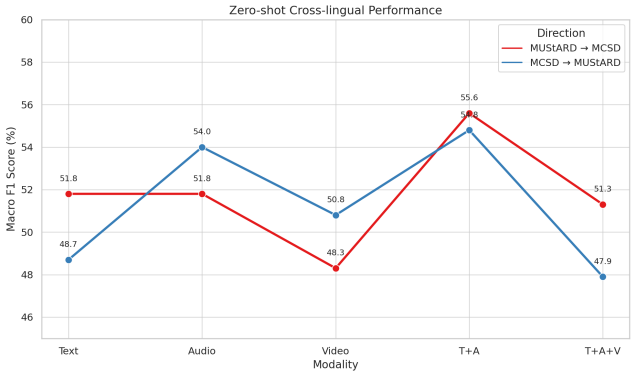

Figure 4: Zero-shot Cross-lingual Performance

achieve nearly identical macro F1, while the video-only model is lower (48.3%). Notably, adding video to the multimodal system (T+A+V) does not further improve performance (macro F1 drops to 51.3%), suggesting that the inclusion of visual features may introduce domain-specific noise in this transfer direction.

In the reverse transfer—training on MCSD and testing on MUStARD—the T+A model again achieves the best macro F1 (54.8%). Here, the audio-only model (macro F1 = 54.0%) outperforms the text-only (48.7%) and video-only (50.8%) models, further confirming the strong cross-lingual generalizability of acoustic features. Similar to the previous setting, adding video to the fusion does not enhance performance; the macro F1 for T+A+V drops to 47.9%.

Across both directions, bimodal text+audio models consistently yield the best zero-shot transfer results, outperforming unimodal and trimodal systems. This pattern indicates that textual and acoustic cues are more transferable across languages than visual features under domain shift, and that naive multimodal fusion may be suboptimal without explicit domain adaptation.

Overall, these results underscore the challenges of zero-shot sarcasm detection across languages. While leveraging both text and audio provides clear advantages, tri-modal fusion can sometimes degrade performance due to mismatched visual context or annotation schemes in the source and target datasets. Future work should explore more robust fusion strategies and domain-invariant modeling to further improve cross-lingual transfer.

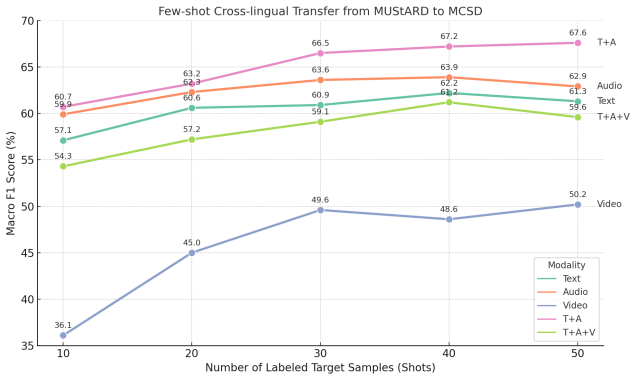

Figure 5: Few-shot Cross-lingual Transfer from MUStARD to MCSD

5.5 Few-shot Cross-lingual Transfer from MUStARD to MCSD

For the MUStARD to MCSD direction, the performance of each modality can be quantitatively distinguished as follows:

Text-only models start at 57.6% accuracy and a macro F1 of 57.1% with 10 shots. As the number of labeled target samples increases, accuracy climbs to 62.2% at 40 shots and then slightly drops to 61.3% at 50 shots. The macro F1 follows a similar trend, peaking at 62.2% (40 shots) before a slight decrease. F1 (0) and F1 (1) are initially imbalanced (52.4% vs 61.7%), but this gap narrows at higher shots (62.7% vs 59.8% at 50 shots), showing that with more supervision, the model becomes more balanced in classifying sarcastic and non-sarcastic utterances.

Audio-only models begin at a higher starting point, with 60.0% accuracy and a macro F1 of 59.9% at 10 shots. These metrics steadily rise with more data, reaching 63.9% accuracy and macro F1 at 40 shots, and slightly decreasing to 63.1%/62.9% at 50 shots. F1 for the sarcastic class (Class 1) consistently outperforms Class 0 at high shots (65.5% vs 60.3% at 50 shots), indicating that audio features may capture prosodic cues more associated with sarcasm.

The video-only model is notable for its severe class imbalance at low shots: with 10 labeled samples, F1 (0) is as high as 66.1%, but F1 (1) is only 6.2%. As the shot number increases, F1 (1) improves to 42.4% at 50 shots, but macro F1 (36.1% to 50.2%) and accuracy (50.2% to 51.4%) remain far behind the other modalities, confirming the difficulty of cross-lingual transfer for visual features in this direction.

Multimodal fusion models show clear quantitative advantages. T+A starts at 61.8% accuracy and a macro F1 of 60.7%, already surpassing all unimodal baselines at 10 shots. It improves rapidly to 67.6% accuracy and macro F1 at 50 shots. Importantly, the F1 for both classes is nearly equal at high shots (67.2% and 67.9%), reflecting not only superior but also more balanced detection. The T+A+V model, however, does not show as clear an advantage; its macro F1 only increases from 54.3% (10 shots) to 61.2% (40 shots) and even declines at 50 shots (59.6%), suggesting the addition of video can sometimes dilute the effectiveness of the fusion in this scenario.

To illustrate the effect of few-shot adaptation, we primarily report results at 40 or 50 shots depending on where each model achieves its optimal or most stable performance. In many cases, macro F1 plateaus or reaches its peak at 40 shots, after which further increases in labeled data yield marginal or inconsistent gains. However, for some modalities, the best performance is observed at 50 shots. Therefore, our analysis references the most representative or highest-performing shot setting for each configuration to provide an accurate account of few-shot improvements.

In the MUStARD to MCSD direction, text, audio, and T+A models all exhibit steady and balanced improvements in both F1 (0) and F1 (1) as the number of labeled target samples increases. For instance, the T+A model's F1 (0) rises from 54.0% to 67.2% and F1 (1) from 67.4% to 68.2% when increasing shots from 10 to 40. This indicates that few-shot learning enhances the model's ability to identify both sarcastic and non-sarcastic utterances, mitigating class bias. However, for the video-only model, F1 (0) remains substantially higher than F1 (1) (e.g., 66.1% vs. 6.2% at 10 shots), revealing a pronounced bias toward the majority class and limited utility of visual cues alone for cross-lingual sarcasm detection. The trimodal T+A+V system offers moderate improvements for both classes, but the gains are less pronounced and fluctuate with increasing shots, reflecting the potential introduction of modality-specific noise.

In summary, text and audio modalities both benefit significantly from few-shot adaptation, with audio offering a slight advantage for sarcastic utterances and text achieving better balance at higher

shots. Video lags far behind, especially in classifying sarcasm. Fusion of text and audio provides the strongest and most stable gains, while the further addition of video does not guarantee improvement and can even be detrimental when transferring from MUStARD to MCSD.

Figure 6: Few-shot Cross-lingual Transfer from MCSD to MUStARD

5.6 Few-shot Cross-lingual Transfer from MCSD to MUStARD

When transferring from MCSD to MUStARD, the quantitative trends of each modality are as follows:

The text-only model begins with 59.4% accuracy and a macro F1 of 59.0% at 10 shots, steadily rising to 62.9% for both metrics at 50 shots. The class-wise F1 scores for text converge with more shots, reaching 63.4% for non-sarcastic and 62.3% for sarcastic at 50 shots, indicating a well-balanced performance as in the previous direction.

Audio-only performance is more modest. It starts at 50.3% accuracy and macro F1 at 10 shots, rising to 55.4% for both at 50 shots. The class-wise F1 scores also become balanced at higher shots, but remain lower than those for text and fusion models.

The video-only model stands out for its marked improvement and strong results in this direction. It starts at 54.8% accuracy and macro F1 of 52.7% at 10 shots, and climbs to 60.2% accuracy and 59.6% macro F1 at 50 shots. F1 (0) for video reaches 64.6% and F1 (1) 54.5% at 50 shots, outperforming audio and even approaching text in macro F1, which is in stark contrast to the MUStARD to MCSD results. This suggests that visual sarcasm cues are more transferable and relevant in the English TV-style MUStARD dataset.

Fusion models T+A and T+A+V both exhibit strong and stable improvements. T+A starts at 56.6% accuracy and 55.8 macro F1, growing to 63.9% accuracy and 63.7% macro F1 at 50 shots. T+A+V is the best overall, reaching 64.1% accuracy and macro F1 at both 30 and 50 shots. The class-wise F1 difference is minimized at high shots for both fusion models: for T+A+V, F1 (0) and F1 (1) reach 68.0% and 59.1% respectively at 50 shots. For the MCSD to MUStARD transfer direction, the text, audio, and T+A models all achieve relatively balanced F1 (0) and F1 (1) scores, with overall improvements as the number of shots increases. For example, the T+A model improves from 61.6% (F1 (0)) and 50.1% (F1 (1)) at 10 shots to 66.4% and 61.1% at 50 shots, respectively. This trend suggests that few-shot learning helps to mitigate class imbalance, allowing the model to become more sensitive to both sarcastic and non-sarcastic utterances.

The text-only and audio-only models show similar upward trends, although the improvement for the audio modality is slower. Notably, F1 (1) for audio remains consistently lower than F1 (0), indicating that audio-based sarcasm recognition is more challenging to transfer across languages.

For the video-only model, F1 (0) is always higher than F1 (1)—for instance, 62.6% versus 42.7% at 10 shots and 64.6% versus 54.5% at 50 shots. However, compared to the MUStARD to MCSD direction, the class imbalance is less severe, and F1 (1) increases more rapidly as the number of labeled samples grows. This suggests that the transfer barrier for visual cues is somewhat lower in this direction.

For the T+A+V model, both F1 (0) and F1 (1) increase together (e.g., reaching 68.0% and 59.1% at 50 shots), but the gains are less pronounced than those achieved by the T+A model alone. This indicates that, while combining visual features can still boost overall performance, the improvement for the sarcastic class remains limited by visual noise, even as the bias diminishes. To summarize, for MCSD to MUStARD, all modalities benefit from more labeled data, but video features contribute much more than in the reverse direction. Text and fusion models still provide the highest and most balanced performance, but video is competitive and valuable in this setting. The data suggests that sarcasm is expressed in a more visually consistent manner in English TV data than in Mandarin conversational data, and that fusion models leveraging all modalities achieve the best and most robust adaptation in the few-shot regime.

Table 6: Few-shot cross-lingual sarcasm detection from MUStARD to MCSD (%).

Shots	**Modality**	**Accuracy (%)**	**F1 (0) (%)**	**F1 (1) (%)**	**Macro F1 (%)**
10	Text	57.6	52.4	61.7	57.1
20	Text	61.1	56.3	64.9	60.6
30	Text	61.1	58.0	63.8	60.9
40	Text	62.2	61.5	62.9	62.2
50	Text	61.3	62.7	59.8	61.3
10	Audio	60.0	61.7	58.1	59.9
20	Audio	62.4	60.9	63.7	62.3
30	Audio	63.7	61.8	65.3	63.6
40	Audio	63.9	63.1	64.6	63.9
50	Audio	63.1	60.3	65.5	62.9
10	Video	50.2	66.1	6.2	36.1
20	Video	50.7	62.7	27.3	45.0
30	Video	52.0	60.6	38.6	49.6
40	Video	51.4	60.4	36.8	48.6
50	Video	51.4	58.0	42.4	50.2
10	T+A	61.8	54.0	67.4	60.7
20	T+A	63.8	58.7	67.7	63.2
30	T+A	66.5	65.4	67.6	66.5
40	T+A	67.2	66.1	68.2	67.2
50	T+A	67.6	67.2	67.9	67.6
10	T+A+V	55.2	60.8	47.7	54.3
20	T+A+V	57.2	57.4	57.0	57.2
30	T+A+V	59.1	59.8	58.4	59.1
40	T+A+V	61.3	63.7	58.7	61.2
50	T+A+V	59.7	61.4	57.8	59.6

Table 7: Few-shot cross-lingual sarcasm detection from MCSD to MUStARD (%).

Shots	**Modality**	**Accuracy (%)**	**F1 (0) (%)**	**F1 (1) (%)**	**Macro F1 (%)**
10	Text	59.4	54.8	63.1	59.0
20	Text	59.9	58.1	61.5	59.8
30	Text	61.1	57.1	64.4	60.8
40	Text	62.0	60.8	63.1	61.9
50	Text	62.9	63.4	62.3	62.9
10	Audio	50.3	53.0	47.2	50.1
20	Audio	52.8	55.2	50.1	52.6
30	Audio	54.3	53.8	54.7	54.3
40	Audio	53.8	51.2	56.1	53.6
50	Audio	55.4	55.2	55.7	55.4
10	Video	54.8	62.6	42.7	52.7
20	Video	59.9	62.2	57.1	59.7
30	Video	59.8	63.8	54.9	59.4
40	Video	61.2	65.7	55.2	60.4
50	Video	60.2	64.6	54.5	59.6
10	T+A	56.6	61.6	50.1	55.8
20	T+A	59.1	62.5	54.9	58.7
30	T+A	58.7	59.1	58.3	58.7
40	T+A	61.8	63.2	60.3	61.8
50	T+A	63.9	66.4	61.1	63.7
10	T+A+V	58.1	59.2	56.8	58.0
20	T+A+V	63.9	63.3	64.3	63.8
30	T+A+V	64.1	66.0	62.1	64.1
40	T+A+V	63.0	65.8	59.6	62.7
50	T+A+V	64.1	68.0	59.1	63.6

6 Discussion

The results reported in Section 5 clearly demonstrate that both multimodal integration and few-shot learning significantly enhance the effectiveness of cross-lingual sarcasm detection. These findings directly address the primary research questions of this study. In this section, we critically evaluate each hypothesis in light of empirical evidence, relate our results to prior research, and discuss broader implications.

6.1 Validation of the First Hypothesis

Our first hypothesis proposed that multimodal transfer learning would significantly outperform unimodal baselines in cross-lingual sarcasm detection. The results largely confirm this, but the magnitude and consistency of improvement depend strongly on the dataset and modality.

In the in-domain MUStARD results (Table 2), we observe that adding modalities nearly always improves performance. For instance, at the utterance level, the T+A+V model achieves an F1 of 70.5%, higher than any unimodal configuration (text: 65.7%, audio: 69.9%, video: 67.4%). This effect is even more pronounced when context is included, with T+A+V reaching 72.3%. The same pattern holds for the joint training setting (F1 = 66.8% for T+A+V), further supporting the hypothesis that integrating complementary cues from text, audio, and video enhances robustness and generalizability.

However, the benefits of multimodal fusion are not uniform across datasets. On MCSD, although audio-only models perform best among unimodal options (F1 = 66.9%), fusing all three modalities (T+A+V) actually leads to a lower F1 (58.9%) than text+audio (66.7%) or audio alone. This result reveals that adding video can sometimes introduce noise and degrade overall performance, rather than improve it Yue et al. (2024). Possible explanations include the less expressive visual behaviors in Mandarin conversational settings which makes it harder for the model to exploit visual cues effectively.

Notably, in joint training across MUStARD and MCSD, the addition of video once again improves trimodal performance, suggesting that with more diverse data, the model can better learn when to use or ignore video cues. This highlights that multimodal fusion is most effective when each modality provides reliable, complementary information, and its value depends on both the properties of the data and the annotation granularity. In the transfer from MUStARD to MCSD, models fusing text and audio (T+A) consistently achieved the best macro F1 scores, surpassing either modality alone. For example, with 50 few-shot samples per class, the T+A model reached a macro F1 of 67.6%, representing a notable improvement over text-only (61.3%) and audio-only (62.9%) systems.

By contrast, in the reverse direction (MCSD to MUStARD), trimodal fusion (T+A+V) generally performed slightly better than bimodal models. For instance, with 40 shots, the T+A+V model achieved a macro F1 of 62.7%, and both outperforming unimodal baselines. With 50 shots, T+A+V further improves to 63.5%.

In summary, the evidence robustly supports the first hypothesis in most cases, but also demonstrates that blindly fusing all available modalities does not always guarantee performance gains. The value contributed by each modality is highly dependent on the characteristics of both the source and target domains. Indiscriminate fusion of all available modalities does not necessarily guarantee optimal performance.

6.2 Validation of the Second Hypothesis

H2 posited that few-shot adaptation would yield substantial improvements over zero-shot transfer.

The experimental results strongly support this hypothesis. In both transfer directions, most models—including text, audio, and fusion configurations—demonstrate clear and consistent gains in macro F1 as the number of labeled target samples increases from 10 to 40 or 50 per class. For example, in the MUStARD to MCSD direction, the T+A model's macro F1 rises from 60.7% at 10 shots to 67.6% at 50 shots, a gain of nearly 7 percentage points. Similarly, in the MCSD to MUStARD direction, the T+A+V model increases from 58.0% (10 shots) to 63.6% (50 shots). Even unimodal models, such as text-only, show marked improvement, with macro F1 increasing from 57.6% to 61.3% in MUStARD to MCSD and from 59.4% to 62.9% in MCSD to MUStARD. These findings confirm that few-shot adaptation is highly effective and practical for low-resource sarcasm detection tasks, in line with prior work on data efficiency in cross-lingual paralinguistic transfer.

However, it is important to note that the improvement trend is not strictly monotonic or uniform across all modalities. In particular, the video-only models display substantial variability. For instance, in the MUStARD to MCSD direction, the macro F1 for video-only rises from 36.1% (10 shots) to 49.6% (30 shots), but then plateaus and fluctuates, reaching only 50.2% at 50 shots. Similarly, the T+A+V fusion model in this direction achieves its highest macro F1 at 40 shots (61.2%) before dropping slightly at 50 shots (59.6%). This non-monotonic pattern suggests that simply increasing the number of labeled target samples does not always guarantee improved performance, especially for modalities such as video that may introduce domain-specific noise or less informative features.

These observations highlight that the effectiveness of few-shot adaptation depends not only on the number of labeled samples but also on the quality and cross-lingual transferability of the input features. While textual and acoustic cues generally lead to robust improvements, visual features show more variability and can even degrade performance in some settings. This aligns with prior research suggesting that modality mismatch, domain shift, and annotation inconsistency can limit the utility of certain modalities in cross-lingual transfer. In practical terms, adaptive modality selection or feature weighting may be necessary to maximize the benefits of few-shot adaptation, particularly in heterogeneous or noisy data scenarios.

Overall, our results demonstrate that few-shot cross-lingual learning substantially improves performance over zero-shot transfer, but that careful consideration of modality-specific factors is critical for optimal adaptation.

6.3 Validation of the Third Hypothesis

H3 hypothesized that text and audio features would contribute more to cross-lingual sarcasm detection than video features. The experimental results provide strong and nuanced support for this hypothesis. In both zero-shot and few-shot experiments, models based on text and audio consistently outperform those that rely solely on video information. For example, in the MUStARD to MCSD zero-shot transfer, the macro F1 for the video-only model is 48.3%, while both text and audio models achieve 51.8%. As more labeled target samples are added, this performance gap becomes even more apparent: at 50 shots, the macro F1 for video-only models peaks at just 50.2%, compared to 61.3% for text and 62.9% for audio. Similar trends are seen in the MCSD to MUStARD transfer, further illustrating the limited generalizability of visual cues.

Fusion models that include video (T+A+V) do not consistently outperform text+audio (T+A) models, and sometimes even show reduced performance. For example, in the MUStARD to MCSD setting, T+A+V reaches a macro F1 of 59.6% at 50 shots, significantly lower than the 67.6% achieved by T+A. This suggests that the addition of video can introduce more noise than useful information in cross-lingual settings, depending on the dataset.

This pattern can be attributed to several key factors: First, annotation granularity plays a critical role. In the MUStARD dataset, sarcasm is annotated at the utterance level, which allows visual and acoustic cues to be precisely aligned with the text and the intended meaning of each spoken line. In contrast, the MCSD dataset sometimes employs coarser annotations, labeling sarcasm at the dialogue or scene level. This coarse-grained labeling can blur the association between visual signals and actual sarcastic content, causing the model to extract less relevant or even misleading visual features.

Second, cultural norms of sarcasm expression differ across languages and domains. In English-language scripted television (MUStARD), sarcasm is often performed with overt and exaggerated facial expressions and gestures, which can be reliably captured by video features. However, in spontaneous Mandarin conversations (MCSD), sarcastic intent is frequently conveyed through subtler cues and less expressive body language, diminishing the value of visual modality for automatic detection.

Taken together, these findings demonstrate that while visual features can provide auxiliary information, their contribution is highly dependent on the alignment between data annotation and cultural context. In contrast, text and audio consistently deliver strong and transferable cues for cross-lingual sarcasm detection, regardless of the underlying dataset. This underscores the need for adaptive, context-aware multimodal fusion strategies, rather than assuming uniform utility of all modalities across domains.

6.4 Limitations

While this study provides a comprehensive analysis of cross-lingual multimodal sarcasm detection, several limitations should be acknowledged to ensure an objective interpretation of the results and to inform future research directions.

The current work relies on traditional supervised machine learning, specifically support vector machines (SVMs) with early fusion for multimodal integration. While SVMs offer interpretability and strong baselines, they may underperform compared to recent deep learning architectures that can model more complex interactions between modalities and adaptively fuse information. Future work should investigate transformer-based or contrastive multimodal frameworks to fully exploit cross-modal dependencies and large-scale pretraining.

All experiments are implemented in Python and executed on a fixed Linux server infrastructure. Hardware limitations—such as GPU memory, storage, and computational throughput—may restrict the scale of experiments, particularly for more computationally intensive neural architectures or larger datasets. In addition, the preprocessing pipelines for audio and video may lose fine-grained temporal cues due to mean pooling or dimensionality reduction.

Both the MUStARD and MCSD datasets, while multimodal and cross-lingual, exhibit specific domain characteristics. MUStARD consists of scripted English TV dialogue, whereas MCSD features Mandarin stand-up comedy. These genre and language differences, along with variations in

annotation granularity (e.g., utterance-level vs. context-level sarcasm labels), introduce inherent distribution shifts. Moreover, the MCSD dataset, though substantial, is still limited in size compared to large-scale NLP benchmarks, potentially constraining model generalizability and robustness.

Class imbalance, annotator subjectivity, and cultural variation in the expression of sarcasm all pose potential sources of bias Abercrombie and Hovy (2016); Joshi, Bhattacharyya, Carman, Saraswati, and Shukla (2016). For example, sarcasm in Mandarin may be more subtle or expressed differently than in English, which can affect the transferability of features and the consistency of human annotations Liu et al. (2014). While cross-validation and few-shot adaptation help mitigate some biases, further work is needed to systematically control for these factors.

The feature extraction process depends on specific pre-trained models and software libraries, such as HuggingFace Transformers, `torchvggish`, and `torchvision`. Library version inconsistencies, dependency issues, or updates may affect reproducibility. Additionally, fixed hardware configurations may preclude exploration of more advanced or computationally demanding architectures.

Due to practical time constraints, this study focuses primarily on binary sarcasm detection in a cross-lingual, multimodal context. The scope does not include fine-grained sarcasm types, multilingual data beyond English and Mandarin. Future research could expand the analysis to more languages, domains, or sarcasm subtypes, as well as investigate temporal modeling for sequential sarcasm detection.

Collectively, these limitations may restrict the generalizability of the findings to other languages, domains, or more complex conversational scenarios. Nevertheless, by highlighting both the strengths and weaknesses of current multimodal and cross-lingual sarcasm detection approaches, this work provides a valuable empirical foundation for future improvements and broader application.

7 Conclusion

This thesis systematically investigated the effectiveness of zero-shot and few-shot transfer learning strategies for multimodal sarcasm detection across English and Mandarin. The study focused on evaluating the respective contributions of text, audio, and video modalities, and analyzed the performance of different fusion using the MUStARD and MCSD datasets.

7.1 Summary of the Main Contributions

The main findings can be summarized as follows. First, multimodal fusion models, particularly those combining text and audio, generally outperformed unimodal models in both in-domain and cross-lingual settings. However, the benefits of adding video were found to be highly dependent on dataset characteristics, with video features sometimes introducing noise rather than useful information—especially when annotation granularity was coarse or visual expressions of sarcasm were subtle. Second, few-shot adaptation consistently led to substantial improvements over zero-shot transfer, demonstrating the practical value of leveraging a small number of labeled target-domain samples to boost model performance. Third, the empirical results confirmed that text and audio features are more robust and transferable cues for sarcasm detection than visual features, whose utility is limited by annotation quality and cultural or domain differences.

Despite these advances, several limitations remain. The current study was restricted to two datasets and a traditional SVM-based modeling framework. Dataset size, domain specificity, class imbalance, annotation subjectivity, and pre-trained feature extractor limitations all place constraints on generalizability. The analysis was limited to binary sarcasm classification and did not address more granular distinctions or multi-turn dialog modeling.

7.2 Future Work

Future research should consider expanding to additional languages, domains, and larger-scale datasets, as well as exploring more sophisticated neural models for adaptive and dynamic multimodal fusion. Investigating methods for mitigating bias, improving annotation quality, and incorporating sequential or contextual modeling could further enhance robustness. Finally, the development of more universal and generalizable sarcasm detection systems will require ongoing efforts in both data collection and methodological innovation.

7.3 Impact & Relevance

The findings of this study provide empirical support for the design of cross-lingual, multimodal affective computing systems. They demonstrate the promise—and practical limitations—of data-efficient transfer learning for social signal processing across languages and cultures. By clarifying the conditions under which different modalities are informative or redundant, this work helps guide future research toward more robust, adaptive, and generalizable approaches to multimodal sarcasm detection.

In summary, this thesis advances the understanding of multimodal and cross-lingual sarcasm detection by highlighting both the strengths and boundaries of current transfer learning strategies, and by identifying key avenues for future improvement.

References

Abercrombie, G., & Hovy, D. (2016, August). Putting sarcasm detection into context: The effects of class imbalance and manual labelling on supervised machine classification of twitter conversations. In H. He, T. Lei, & W. Roberts (Eds.), *Proceedings of the ACL 2016 Student Research Workshop* (pp. 107–113). Berlin, Germany: Association for Computational Linguistics. doi: 10.18653/v1/P16-3016

An, T., Yan, P., Zuo, J., Jin, X., Liu, M., & Wang, J. (2024, June). Enhancing cross-lingual sarcasm detection by a prompt learning framework with data augmentation and contrastive learning. *Electronics*, *13*(11), 2163. doi: 10.3390/electronics13112163

Artetxe, M., & Schwenk, H. (2019, November). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, *7*, 597–610. doi: 10.1162/tacl_a_00288

Basabain, S., Cambria, E., Alomar, K., & Hussain, A. (2023). Enhancing arabic-text feature extraction utilizing label-semantic augmentation in few/zero-shot learning. *Expert Systems*, *40*(8), e13329. doi: 10.1111/exsy.13329

Cai, Y., Cai, H., & Wan, X. (2019, July). Multi-modal sarcasm detection in twitter with hierarchical fusion model. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2506–2515). Florence, Italy: Association for Computational Linguistics. doi: 10.18653/v1/P19-1239

Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R., & Poria, S. (2019, June). *Towards multimodal sarcasm detection (an _obviously_ perfect paper)* (No. arXiv:1906.01815). arXiv.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May). *BERT: Pre-training of deep bidirectional transformers for language understanding* (No. arXiv:1810.04805). arXiv. doi: 10.48550/arXiv.1810.04805

Eke, C. I., Norman, A. A., Liyana Shuib, & Nweke, H. F. (2020, August). Sarcasm identification in textual data: Systematic review, research challenges and open directions. *Artificial Intelligence Review*, *53*(6), 4215–4258. doi: 10.1007/s10462-019-09791-8

Elsahar, H., & Gallé, M. (2019, November). To annotate or not? Predicting performance drop under domain shift. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 2163–2173). Hong Kong, China: Association for Computational Linguistics. doi: 10.18653/v1/D19-1222

Eriguchi, A., Johnson, M., Firat, O., Kazawa, H., & Macherey, W. (2018, September). *Zero-shot cross-lingual classification using multilingual neural machine translation* (No. arXiv:1809.04686). arXiv. doi: 10.48550/arXiv.1809.04686

Farabi, S., Ranasinghe, T., Kanojia, D., Kong, Y., & Zampieri, M. (2024, August). A survey of multimodal sarcasm detection. In *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence* (pp. 8020–8028). Jeju, South Korea: International Joint Conferences on Artificial Intelligence Organization. doi: 10.24963/ijcai.2024/887

Feng, K., & Chaspari, T. (2023, April). Few-shot learning in emotion recognition of spontaneous speech using a siamese neural network with adaptive sample pair formation. *IEEE Transactions on Affective Computing*, *14*(2), 1627–1633. doi: 10.1109/TAFFC.2021.3109485

Gao, X., Nayak, S., & Coler, M. (2022, September). Deep CNN-based inductive transfer learning

for sarcasm detection in speech. In *Interspeech 2022* (pp. 2323–2327). ISCA. doi: 10.21437/Interspeech.2022-11323

Gao, X., Nayak, S., & Coler, M. (2024). Improving sarcasm detection from speech and text through attention-based fusion exploiting the interplay of emotions and sentiments. In *186th Meeting of the Acoustical Society of America and the Canadian Acoustical Association* (p. 060002). Ottawa, Ontario, Canada. doi: 10.1121/2.0001918

Ghosh, D., Fabbri, A. R., & Muresan, S. (2018, December). Sarcasm analysis using conversation context. *Computational Linguistics*, *44*(4), 755–792. doi: 10.1162/coli_a_00336

He, K., Zhang, X., Ren, S., & Sun, J. (2016, June). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). Las Vegas, NV, USA: IEEE. doi: 10.1109/CVPR.2016.90

Helal, N. A., Hassan, A., Badr, N. L., & Afify, Y. M. (2024, July). A contextual-based approach for sarcasm detection. *Scientific Reports*, *14*(1), 15415. doi: 10.1038/s41598-024-65217-8

Huang, P.-Y., Patrick, M., Hu, J., Neubig, G., Metze, F., & Hauptmann, A. (2021, April). *Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models* (No. arXiv:2103.08849). arXiv. doi: 10.48550/arXiv.2103.08849

Joshi, A., Bhattacharyya, P., Carman, M., Saraswati, J., & Shukla, R. (2016, August). How do cultural differences impact the quality of sarcasm annotation?: A case study of Indian annotators and American text. In N. Reiter, B. Alex, & K. A. Zervanou (Eds.), *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 95–99). Berlin, Germany: Association for Computational Linguistics. doi: 10.18653/v1/W16-2111

Joshi, A., Bhattacharyya, P., & Carman, M. J. (2016, September). *Automatic sarcasm detection: A survey* (No. arXiv:1602.03426). arXiv. doi: 10.48550/arXiv.1602.03426

Joshi, A., Sharma, V., & Bhattacharyya, P. (2015, July). Harnessing context incongruity for sarcasm detection. In C. Zong & M. Strube (Eds.), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 757–762). Beijing, China: Association for Computational Linguistics. doi: 10.3115/v1/P15-2124

Kodirov, E., Xiang, T., Fu, Z., & Gong, S. (2015, December). Unsupervised domain adaptation for zero-shot learning. In *2015 IEEE International Conference on Computer Vision (ICCV)* (pp. 2452–2460). Santiago, Chile: IEEE. doi: 10.1109/ICCV.2015.282

Liu, P., Chen, W., Ou, G., Wang, T., Yang, D., & Lei, K. (2014). Sarcasm detection in social media based on imbalanced classification. In D. Hutchison et al. (Eds.), *Web-Age Information Management* (Vol. 8485, pp. 459–471). Cham: Springer International Publishing. doi: 10.1007/978-3-319-08010-9_49

Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Peixoto, R. M., Guimarães, G. A. S., Cruz, G. O. R., ... Nascimento, E. G. S. (2023, January). Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing*, *7*(1), 15. doi: 10.3390/bdcc7010015

Peters, S., Wilson, K., Boiteau, T. W., Gelormini-Lezama, C., & Almor, A. (2016, March). Do you hear it *now* ? A native advantage for sarcasm processing. *Bilingualism: Language and Cognition*, *19*(2), 400–414. doi: 10.1017/S1366728915000048

Poria, S., Cambria, E., Hazarika, D., & Vij, P. (2017, July). *A deeper look into sarcastic tweets using deep convolutional neural networks* (No. arXiv:1610.08815). arXiv. doi: 10.48550/

arXiv.1610.08815

Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C. P., ... Wu, Q. M. J. (2023, April). A review of generalized zero-shot learning methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(4), 4051–4070. doi: 10.1109/TPAMI.2022.3191696

Ray, A., Mishra, S., Nunna, A., & Bhattacharyya, P. (2022, June). A multimodal corpus for emotion recognition in sarcasm. In N. Calzolari et al. (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 6992–7003). Marseille, France: European Language Resources Association.

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., & Huang, R. (2013, October). Sarcasm as contrast between a positive sentiment and negative situation. In D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu, & S. Bethard (Eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 704–714). Seattle, Washington, USA: Association for Computational Linguistics.

Sangwan, S., Akhtar, M. S., Behera, P., & Ekbal, A. (2020, July). I didn't mean what I wrote! Exploring multimodality for sarcasm detection. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). Glasgow, United Kingdom: IEEE. doi: 10.1109/IJCNN48605.2020.9206905

Schifanella, R., de Juan, P., Tetreault, J., & Cao, L. (2016, October). Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM international conference on Multimedia* (pp. 1136–1145). doi: 10.1145/2964284.2964321

Simonyan, K., & Zisserman, A. (2015, April). *Very deep convolutional networks for large-scale image recognition* (No. arXiv:1409.1556). arXiv. doi: 10.48550/arXiv.1409.1556

Song, Y., Wang, T., Mondal, S. K., & Sahoo, J. P. (2022, May). *A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities* (No. arXiv:2205.06743). arXiv. doi: 10.48550/arXiv.2205.06743

Sun, Y., Zhang, H., Yang, S., & Wang, J. (2022, November). EFAFN: An efficient feature adaptive fusion network with facial feature for multimodal sarcasm detection. *Applied Sciences*, *12*(21), 11235. doi: 10.3390/app122111235

Techentin, C., Cann, D. R., Lupton, M., & Phung, D. (2021, June). Sarcasm detection in native English and english as a second language speakers. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale*, *75*(2), 133–138. doi: 10.1037/cep0000241

Tomar, M., Tiwari, A., Saha, T., & Saha, S. (2023, October). Your tone speaks louder than your face! Modality order infused multi-modal sarcasm detection. In *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 3926–3933). Ottawa ON Canada: ACM. doi: 10.1145/3581783.3612528

Veale Tony, & Hao Yanfen. (2010). Detecting ironic intent in creative comparisons. In *Frontiers in Artificial Intelligence and Applications.* IOS Press. doi: 10.3233/978-1-60750-606-5-765

Wang, R., Wang, Q., Liang, B., Chen, Y., Wen, Z., Qin, B., & Xu, R. (2022, July). Masking and generation: An unsupervised method for sarcasm detection. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2172–2177). Madrid Spain: ACM. doi: 10.1145/3477495.3531825

Xu, B., Zeng, Z., Lian, C., & Ding, Z. (2022). Few-shot domain adaptation via mixup optimal transport. *IEEE Transactions on Image Processing*, *31*, 2518–2528. doi: 10.1109/TIP.2022.3157139

Yue, T., Shi, X., Mao, R., Hu, Z., & Cambria, E. (2024, May). SarcNet: A multilingual multimodal sarcasm detection dataset. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 14325–14335). Torino, Italia: ELRA and ICCL.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., … He, Q. (2020, June). *A comprehensive survey on transfer learning* (No. arXiv:1911.02685). arXiv. doi: 10.48550/arXiv.1911.02685

Appendices

A Declaration

I hereby affirm that this Master thesis was composed by myself, and that the work herein is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification except as specified, nor has it been published. Where other people's work has been used (from any source: printed, internet, or other), this has been carefully acknowledged and referenced.

During the preparation of this thesis, I used OpenAI ChatGPT 4.1 for the following purposes: sentence restructuring in Sections 2.2 and 2.3, generating alternative explanations for technical concepts in Chapter 3, creating initial code documentation templates, and providing translation and refinement of my own writing. I used it to help me quickly understand and summarize academic papers relevant to my topic. When I encountered complex or densely written sections in technical papers, I provided selected excerpts to ChatGPT and asked it to explain the content in simpler terms.

All content produced with the assistance of ChatGPT was subsequently reviewed, verified, and substantially modified by me to ensure accuracy, originality, and alignment with academic standards.

Meiling Zhang / 11 June 2025