



A Multimodal Chinese Dataset for Cross-lingual Sarcasm Detection

Xiyuan Gao¹, Bruce Xiao Wang², Meiling Zhang¹, Shuming Huang², Zhu Li¹, Shekhar Nayak¹, Matt Coler¹

¹Speech Technology Lab, University of Groningen, the Netherlands

²English and Communication, The Hong Kong Polytechnic University, Hong Kong

xiyuan.gao@rug.nl, brucex.wang@polyu.edu.hk, m.zhang.43@student.rug.nl, shuming.huang@polyu.edu.hk, zhu.li@rug.nl, s.nayak@rug.nl, m.coler@rug.nl

Abstract

Sarcasm is expressed through subtle cues like pitch, speech rate, and facial expressions, with patterns varying across languages, e.g., English speakers lower the pitch while Cantonese speakers raise it. While humans readily interpret these signals, computational models struggle, creating challenges for Human-Machine Interaction. Most multimodal sarcasm recognition research focuses on English and the lack of high-quality datasets for other languages hinders cross-lingual and cross-cultural studies. We introduce the Multimodal Chinese Sarcasm Dataset (MCSD), containing 10.57 hours of video. We propose a standardized annotation framework that captures annotator certainty to reflect the subjectivity of sarcasm, achieving a Fleiss' kappa of 0.74 (unweighted) and 0.79 (certainty-weighted). Validation of our dataset using SVM achieves a 76.64% F1-score in sarcasm detection. MCSD lays the foundation for robust cross-lingual sarcasm detection, contributing to advanced, human-centric systems.

Index Terms: multimodal sarcasm dataset, human-machine interaction, computational paralinguistics

1. Introduction

Sarcasm relies not just on *what* is said, but on *how* it is said, integrating verbal content with non-verbal cues like pitch, speaking rate, and facial expressions. For humans, these multimodal signals are essential for processing non-literal language [1]. More specifically, Jacob *et al.*[2] suggested that the incongruity between verbal and non-verbal cues plays a key role in facilitating sarcasm comprehension. While modern AI systems like voice assistants and chatbots are increasingly prevalent in daily life, they still struggle to understand non-literal language use like sarcasm.

Current AI systems in Human-Machine Interaction (HMI) struggle to detect sarcasm reliably. However, researchers have found that a multimodal approach offers a promising solution to interpret sarcasm with greater precision [3, 4, 5, 6, 7, 8]. These multimodal approaches have been proven more effective than unimodal methods. However, most existing methods are based on English data, leaving cross-lingual and cross-cultural studies unexplored. Additionally, existing multimodal sarcasm datasets are limited in size (e.g., MUSTARD [3] contains 690 utterances) compared to those of related fields, which may hinder the generalizability and robustness of models trained on them. Furthermore, the inherent ambiguity of sarcasm poses challenges in standardizing the annotation process, leading to low Inter-Annotator Agreement (IAA) and weak reliability. For example, the MUSTARD [3] dataset achieved an IAA score of 0.59, indicating weak annotator consensus.

These limitations restrict the development of multilingual

systems. To tackle them, we introduce a novel, well-curated Multimodal Chinese Sarcasm Dataset (MCSD). The dataset consists of 10.57 hours of video from stand-up comedy shows, with a balanced distribution of sarcastic and non-sarcastic instances. We propose a standardized and replicable annotation protocol, enhancing reproducibility within the community. Our protocol includes a linguistics-based conceptual framework of sarcasm definition, and a streamlined manual annotation process. We achieve an unweighted Fleiss' kappa of 0.74 and a certainty-weighted value of 0.79, indicating substantial consensus, and validating the efficiency and effectiveness of our annotation strategy. Finally, we apply a Support Vector Machine (SVM) to assess the dataset's feasibility for multimodal sarcasm detection. The main contributions of our paper are as follows:

- We introduce a novel multimodal sarcasm dataset in Mandarin Chinese, laying the groundwork for cross-lingual studies of sarcasm.
- We incorporate linguistic insights in constructing the annotation guideline to balance sarcasm's inherent ambiguity with annotator variability in interpretation.
- We optimize efficiency while maintaining data integrity by proposing a sarcasm-first strategy for annotation.
- We standardize the data collection and annotation pipeline to enhance reproducibility and foster the development of reliable cross-lingual datasets. We ensure ethical compliance in data collection, prioritizing privacy concerns and maintaining research transparency.

This paper is organized as follows: Section 2 reviews related work. Section 3 introduces the dataset and curation method. Section 4 details the experiment and results. Section 5 discusses the findings. Section 6 concludes the paper and explains future research directions.

2. Related works

2.1. Multimodal sarcasm detection

Castro *et al.*[3] introduced MUSTARD, an English multimodal sarcasm dataset from American sitcoms, achieving 71.5% F1-score with multimodal SVM. Building on this, Zhang *et al.*[9] developed a contrastive-attention architecture that improved performance to 72.26% F1-score by capturing incongruities between verbal and non-verbal cues. Ray *et al.*[4] expanded the MUSTARD dataset to MUSTARD++, doubling its size and enriching the annotations with emotions, sentiments, and refined subtype sarcasm labels. They employed a collaborative gating architecture and achieved a 70.3% F1-score using multimodal data. Sequentially, Devraj *et al.*[7] employed a Graph Attention Network to capture the intra-modal dependencies alongside a cross-modal Contrastive Attention Mechanism designed to cap-

ture emotional incongruities between modalities. Their multimodal approach achieved a 74.96% F1-score.

Previous research suggests that integrating multimodal data is an effective way to enhance computational models’ ability to recognize sarcasm. Not only addressing data scarcity by incorporating more sarcasm-related content, the multimodal approach also captures the subtle nuances of sarcastic expressions by leveraging sophisticated architectures.

2.2. Limitations of current multimodal sarcasm datasets

Overall, current multimodal sarcasm datasets face several limitations: (a) lack of linguistic diversity, with most datasets focused on English; (b) insufficient data volume, with the largest dataset (MUSStARD++) containing only 7.36 hours of video. Meanwhile, traditional annotation strategies exhaustively label all utterances as sarcastic or non-sarcastic, assuming class balance and relying on binary decisions [3, 4, 10, 11]. This overlooks the subjective and context-dependent nature of sarcasm, leading to inefficiencies and low IAA scores (0.59-0.68) [3, 6, 12, 13, 5]. To generate a high-quality dataset, it is essential to take into account the inherent ambiguity of sarcasm and the variability in how annotators perceive and interpret sarcasm.

3. MCS D dataset

We introduce MCS D, a well-curated collection of annotated stand-up comedy videos. We propose a framework that implements a sarcasm-first approach and subset re-annotation to streamline the process. Moreover, we standardize the collection and annotation pipeline to ensure reproducibility. The detailed annotation process and dataset are available on our github¹.

3.1. Data collection

We aim to capture sarcasm as it appears in everyday conversations, reflecting both individual stylistic differences and demographic diversity to ensure a representative collection of sarcastic expressions. We selected the Chinese stand-up comedy show *Talk Show Gala* for its rich sarcastic content and diverse range of speakers. The data consists of six seasons in MP4 format², with each season containing ten to twelve episodes. These videos were then manually annotated for sarcasm.

3.2. Annotation process

3.2.1. Annotation guidelines

Defining sarcasm is far more nuanced than classifying images as cats or dogs. The interpretation of sarcasm can vary significantly based on annotators’ cultural backgrounds and personal understanding, leading to inconsistencies. Rather than suppressing the discrepancies, we capture the natural diversity in sarcastic expression. To respond to it, we developed annotation guidelines comprising: (a) sarcasm definition and characteristics, (b) examples of sarcastic and non-sarcastic expressions, and (c) self-testing cases. Our conceptual framework draws from linguistic theories to define sarcasm as expressions where the intended meaning differs from the literal statement. These expressions typically serve purposes such as self-deprecation, critique, mocking, or humor. Sarcasm can be expressed unimodally or multimodally through words, tone, facial expres-

sions, and body languages. Sarcastic remarks can also be conveyed through rhetorical devices like hyperbole, understatement and mockery imitation. Our guideline provides annotators comprehensive key features of sarcasm while allowing the flexibility to interpret. To maintain quality and consistency, we employed three annotators with linguistics background and provided them with training, held weekly meetings to review examples and address ambiguous cases through discussion, progressively refining alignment.

3.2.2. Annotation strategy

To create a balanced dataset and avoid inessential effort annotating non-sarcastic instances, which dominate in natural discourse (>95%), we adopted a novel sarcasm-first strategy. Unlike prior work that performs segmentation and annotation across the entire dataset, our sarcasm-first strategy prioritizes the identification and annotation of sarcastic utterances. Each is paired with a nearby non-sarcastic segment to maintain balance. Annotators used LosslessCut³ to select and extract the target instances directly. This strategy not only improved annotation efficiency, but also preserved the natural context for sarcasm interpretation.

3.2.3. IAA calculation

To assess annotation reliability, annotators blindly re-annotating 10-20% of each other’s samples that are randomly sampled [14]. Sarcasm was treated as a spectrum and annotators marked the certainty (very certain = 2, somewhat certain = 1). We calculated Fleiss’ Kappa using the statsmodels package⁴. The unweighted agreement was 0.74, and when weighted by certainty, it increased to 0.79, indicating substantial agreement.

3.3. Data statistics

We reviewed all videos, removing corrupted files and segments with incomplete sentences. The final dataset contains 1,350 sarcastic and 1,355 non-sarcastic instances, totaling 2,705 videos and 10.57 hours of data. We used Whisper-base⁵ to transcribe videos, followed by thorough manual post-editing to ensure accurate alignment with the video content. Detailed statistics are provided in Table 1.

Table 1: Statistical summary of the dataset

Metric	Value
Avg. length (tokens)	41
Avg. duration (sec)	14
Total duration (hour)	10.57
Total video	2705
No. of sarcasm	1350
No. of non-sarcasm	1355

The dataset exhibits three main characteristics:

- **Multimodal complexity:** The data combines verbal cues with nonverbal ones, such as prosody, timing, facial expressions, and body gestures, with sarcasm often conveyed through their interplay. Figure 1 illustrates how incongruity among these cues conveys sarcasm.
- **Contextual dependency:** Sarcastic remarks usually occur at the end of a discourse, while the preceding context is essential for accurate interpretation. In the dataset, each video includes

¹<https://github.com/x-y-g/MCS D/wiki>

²We utilized the platform’s official API (https://developers.google.com/youtube/v3/getting-started/?target=_blank) and ensured compliance with fair use guidelines in research.

³<https://github.com/mifi/lossless-cut?tab=readme-ov-file>

⁴<https://www.statsmodels.org/stable/index.html>

⁵<https://github.com/openai/whisper>

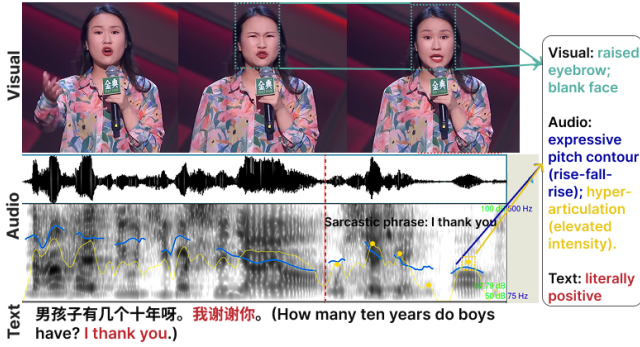


Figure 1: An example of sarcasm expressed through multimodal cues. The blue lines superimposed on spectrogram indicate pitch; yellow lines indicate intensity (dB).

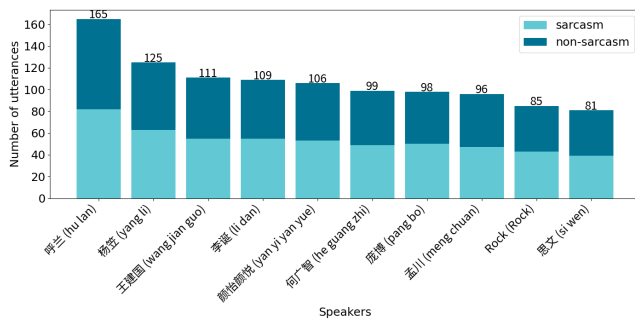


Figure 2: Utterance distribution for the top 10 speakers. Overall, 139 speakers are included, each contributing a balanced number of sarcastic and non-sarcastic instances.

necessary context.

- **Diverse speakers:** The dataset features 139 speakers (37 females, 102 males), each with distinct speaking styles. This diversity enhances the generalizability of the dataset. Figure 2 shows the utterance distribution of the top 10 speakers, with Hu Lan contributing the most (165 utterances).

Our data collection and annotation process captures the pragmatic nature of sarcasm, yielding a high-quality dataset that enhances reliability and reproducibility.

4. Experiments

In this section, we conduct experiments to validate the collected dataset, particularly, we examine the contribution of each modality and their combinations. Given that SVM has been effectively used to benchmark sarcasm detection in English, we adopt it to assess the effectiveness of our dataset.

4.1. Feature extraction

We obtain features from the text (t), audio (a) and visual (v) of our dataset:

Text features: We use BERT-Chinese⁶ to extract text embeddings. Each input is tokenized into τ tokens via WordPiece tokenization⁷, producing $d_t = 768$ -dimensional contextual em-

⁶<https://huggingface.co/google-bert/bert-base-chinese>

⁷<https://research.google/blog/a-fast-wordpiece-tokenization-system/>

beddings for each token. The entire utterance’s embedding is represented as $U_t \in \mathbb{R}^{\tau \times d_t}$, with the $[CLS]$ token providing a fixed-size global representation.

Audio features: We apply Wav2Vec 2.0 fine-tuned on Chinese⁸ to extract audio embeddings. Input WAV files are first converted to mono and resampled at 16 kHz. The model applies a multi-layer CNN encoder to extract low-level acoustic features. These are then passed through a transformer network to capture dependencies and contextual relationships in speech. The resulting frame-level embeddings are denoted as: $U_a \in \mathbb{R}^{w \times d_a}$ where w is the number of extracted audio frames, and $d_t = 1024$ is the dimension of each frame embedding. Last, we aggregate the embeddings by averaging all frame-level embeddings.

Visual features: We employ TimeSformer⁹ [15] to extract visual embeddings. Input videos are resized to 224x224 and normalized for consistency. We extract $f = 8$ frames from video, dividing each into patches and projecting them into $d_v = 768$ -dimensional embeddings. These embeddings pass through spatiotemporal transformer layers to produce frame-level representations, $U_v \in \mathbb{R}^{f \times d_v}$. To obtain a fixed-size video representation, we compute the mean of all frame embeddings.

To generate a unified representation from all modalities, we concatenate embeddings generated from each modality.

4.2. Experimental setup

We perform five-fold cross-validation to evaluate model performance, using Stratified K-folds to ensure balanced class distribution between sarcastic and non-sarcastic samples. One fold is used as the test set, while the remaining four are split 85-15% for training and hyperparameter tuning. Hyperparameters are tuned via GridSearchCV, exploring combinations of $C = [0.01, 0.1, 1, 10]$, $\text{kernel} = [\text{linear}, \text{rbf}]$, and $\text{gamma} = [0.0001, 0.001, 0.01, 0.1]$. The best parameters ($C = 10$, $\text{gamma} = 0.0001$, $\text{kernel} = \text{rbf}$) are used to train the SVM model with sklearn. We evaluate model performance using precision (P), recall (R), and F1-score (F1) to ensure consistency with related research in English.

5. Results

Table 2 presents a comparison of performance across different modalities in our dataset. The highest performance is achieved when data from all three modalities are used, reaching 76.64%, surpassing all unimodal scores. Audio achieves a high score (74.94%), while text and visual scores are lower unimodally. These results validate that a multimodal approach enhances sarcasm recognition performance in Chinese.

Table 2: Comparative performances (macro) on MCSD. T indicates text modality, A indicates audio modality, and V indicates visual modality.

Models	Modality	P (%)	R (%)	F1 (%)
SVM	T	69.83	69.85	69.78
	A	75.02	74.99	74.94
	V	56.29	56.09	55.47
	T+A	74.63	74.63	74.55
	T+V	70.47	70.45	70.38
	A+V	73.78	73.78	73.76
	T+A+V	76.69	76.72	76.64

⁸<https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-chinese-zh-cn>

⁹<https://github.com/facebookresearch/TimeSformer>

Table 3: Examples of mis-recognizing sarcasm with contextual nuance. S means Sarcasm, NS means Non-sarcasm.

Index	Utterance	True	Predicted
319	<p>但是我还是很吃惊，一个节目拒绝一个人的理由居然可以是不够好看。这放在今天绝对是不可思议的。你比如说在这个脱口秀大会的舞台上。</p> <p>Translation: I'm still shocked that a show would reject someone just for not being good-looking. That's almost unthinkable today, especially on a stage like this one.</p>	S	NS
325	<p>那帮人也不会放过我了。我豁出去了，自从我说了脱口秀以后，我冬天出门再也不觉得冷了，因为我头上扣的全是帽子。</p> <p>Translation: Those people won't let me off either. I've decided to go all in. Ever since I started doing stand-up, I never feel cold when I go out in winter, because I'm always wearing hats.</p>	S	NS

We performed a qualitative analysis of misclassified examples to identify error patterns. We found that speaker style poses a challenge for the model. Analyzing the top five speakers responsible for false positives (FP) and false negatives (FN), we observed that four of these speakers contribute to both errors when using visual data alone. While adding text and audio modalities reduces the overall error rate, these speakers remain problematic. Their reliance on text, tonal changes, stress, and tempo, rather than visual cues, is challenging for sarcasm recognition, particularly with visual data alone.

Additionally, the model struggles with recognizing sarcasm when it relies on subtle contextual cues. For example, utterance 319 in Table 3, the phrase “on the stage of this stand-up comedy show” is interpreted as sarcastic only when the audience captures that the speaker is implying the lack of attractive comedians in the show, which is repeatedly emphasized in the show, though not explicitly stated here. Similarly, in utterance 325, the speaker says, the sarcastic meaning is conveyed when the audience understands the Chinese idiom 扣帽子 (translated to: putting hats on someone), refers to unfairly labeling or criticizing someone.

6. Discussion

Our analysis reveals important differences in how sarcasm manifests across languages. While we confirm that multimodal analysis improves sarcasm detection in Chinese, as it does in English, the relative importance of different modalities varies between languages. In contrast to English [3], visual modality performed the weakest in Chinese. This can be attributed to technical limitation, data source, and culture differences: (a) We experimented with ViT¹⁰, which allowed an increased frame rate of 16, resulting in a decreased performance of 45.27%. This highlights the importance of dynamic context in the visual modality and calls for further technical exploration. (b) Our dataset, from stand-up comedy, values scripting that emphasizes textual and speech for sarcasm, whereas the MUsTARD dataset, based on TV performances, relies more on visual cues. (c) Also, cross-cultural differences explain the varying role of visual modality

¹⁰https://github.com/google-research/vision_transformer

in sarcasm detection. Following Hall’s [16] framework, Chinese, as a high-context culture, relies more on situational and contextual cues, while English-speaking cultures favor explicit visual signals like facial expressions. This cultural distinction is reflected in the research focus, with extensive studies on visual cues in English [17, 18, 19] but limited exploration in Chinese.

The audio modality is significant for sarcasm detection in Mandarin Chinese due to its tonal nature. Research by Li *et al.*[20] demonstrates that emotional expressions correlate with specific tone patterns, suggesting a complex interaction between lexical tones and sarcastic expression. More research is needed to reveal the specific tonal change related to sarcasm expression in Mandarin Chinese.

7. Conclusion and future work

To advance cross-lingual and cross-cultural sarcasm recognition in HMI, we introduce:

- **A novel, multimodal Chinese sarcasm dataset:** It comprises 10.57 hours of stand-up comedy videos, balanced for sarcasm and non-sarcasm labels.
- **Linguistic insights:** The annotation is based on a conceptual framework derived from linguistic research, mitigating sarcasm’s inherent ambiguity while preserving interpretative diversity. Unlike prior work that treats sarcasm as a binary label, we incorporate certainty levels in sarcasm labeling to capture varying intensities of sarcasm.
- **Annotation efficiency:** We strategically optimize the annotation process by a sarcasm-first approach, reducing manual effort while maintaining high data quality, in contrast to fully exhaustive annotation pipelines used in earlier datasets.
- **Standardization:** We establish a data collection and annotation pipeline to enhance reproducibility and support reliable sarcasm dataset development. Ethical compliance, privacy concerns, and research transparency are considered and prioritized during our data collection.

However, we acknowledge the limitations: (a) Sarcasm, as noted by Camp [21], includes subtypes like *propositional*, *embedded*, *like-prefixed*, and *illocutionary sarcasm*. Incorporating detailed subtypes into our dataset would provide a more nuanced perspective. (b) Our dataset captures diversity in geographic regions, accents, and speaking styles; however, 73% of the speakers are male. Since sarcasm varies by gender [22], balancing gender representation ensures robust model performance. Additionally, our dataset currently focuses on stand-up comedy. This domain was intentionally selected due to its rich use of diverse sarcasm subtypes and high-quality visual cues. Nonetheless, we acknowledge the limitation in domain scope and will extend future data collection to broader contexts to improve generalizability.

Looking ahead, we plan to expand the dataset by refining sarcasm subtypes, enriching emotion labels, and increasing gender diversity. Future work will benchmark MCS D to assess cross-lingual generalizability. Our study highlights the need for further research in Chinese sarcasm, particularly the role of visual cues and the interaction between lexical tone and sarcastic intent. These efforts will support cross-lingual studies and contribute to the development of multilingual systems for more human-centric HMI.

8. Acknowledgements

This research was supported by the EU COST Action CA19102 “Language In The Human-Machine Era” (LITHME) under the Short Term Scientific Missions (STSMs) framework. We are grateful for the support of Dr. Bruce Xiao Wang and the Department of English and Communication at The Hong Kong Polytechnic University for hosting the research visit associated with this project. We also extend our sincere thanks to Devraj Raghuvanshi from the Department of Data Science at Brown University for his valuable contributions to this work.

9. References

- [1] G. Deliens, K. Antoniou, E. Clin, E. Ostashchenko, and M. Kissine, “Context, facial expression and prosody in irony processing,” *J. Mem. Lang.*, vol. 99, pp. 35–48, 2018.
- [2] H. Jacob, B. Kreifelts, S. Nizielski, A. Schütz, and D. Wildgruber, “Effects of emotional intelligence on the impression of irony created by the mismatch between verbal and nonverbal cues,” *PLOS ONE*, vol. 11, no. 10, pp. 1–17, 2016.
- [3] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, “Towards multimodal sarcasm detection (an _obviously_ perfect paper),” in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Florence, Italy, Aug. 2019, pp. 4619–4629.
- [4] A. Ray, S. Mishra, A. Nunna, and P. Bhattacharyya, “A multimodal corpus for emotion recognition in sarcasm,” in *Proc. 13th Lang. Resour. Eval. Conf. (LREC)*, Marseille, France, June 2022, pp. 6992–7003.
- [5] D. S. Chauhan, D. S. R. A. Ekbal, and P. Bhattacharyya, “Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment, and emotion analysis,” in *Proc. 58th Annu. Meet. Assoc. Comput. Linguistics (ACL)*, Seattle, WA, USA, Jul. 2020, pp. 4351–4360.
- [6] D. S. Chauhan, G. Singh, A. Arora, A. Ekbal, and P. Bhattacharyya, “An emoji-aware multitask framework for multimodal sarcasm detection,” *Knowl.-Based Syst.*, vol. 257, 2022, doi: 10.1016/j.knosys.2022.109924.
- [7] D. Raghuvanshi, X. Gao, Z. Li, S. Bansal, M. Sharma, and S. Nayak, “Intra-modal relation and emotional incongruity learning using graph attention networks for multimodal sarcasm detection,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Hyderabad, India, 2025, pp. 1–5.
- [8] X. Gao, S. Nayak, and M. Coler, “Improving sarcasm detection from speech and text through attention-based fusion exploiting the interplay of emotions and sentiments,” in *Proc. Meet. Acoust.*, vol. 54, no. 1, Jan. 2024, p. 060002, doi:10.1121/2.0001918.
- [9] X. Zhang, Y. Chen, and G. Li, “Multi-modal sarcasm detection based on contrastive attention mechanism,” in *Proc. Nat. Lang. Process. Chin. Comput.*, vol. 13028, 2021, pp. 822–833.
- [10] K. Alnajjar and M. Hämmäläinen, “¡Qué maravilla! multimodal sarcasm detection in spanish: A dataset and a baseline,” in *Proc. 3rd Workshop Multimodal Artif. Intell.*, Mexico City, Mexico, Jun. 2021, pp. 63–68.
- [11] H. Gent, C. Adams, C. Shih, and Y. Tang, “Deep learning for acoustic irony classification in spontaneous speech,” in *Proc. Interspeech*, Incheon, South Korea, Sep. 2022, pp. 3993–3997.
- [12] K. Alnajjar and M. Hämmäläinen, “¡Qué maravilla! multimodal sarcasm detection in spanish: A dataset and a baseline,” in *Proc. 3rd Workshop Multimodal Artif. Intell.*, Mexico City, Mexico, Jun. 2021, pp. 63–68.
- [13] M. Bedi, S. Kumar, M. S. Akhtar, and T. Chakraborty, “Multimodal sarcasm detection and humor classification in code-mixed conversations,” *IEEE Trans. Affective Comput.*, vol. 14, no. 2, pp. 1363–1375, 2023.
- [14] K. A. Neuendorf, *The Content Analysis Guidebook*, 2nd ed. SAGE Publications, Inc, 2017.
- [15] G. Bertasius, H. Wang, and L. Torresani, “Is space-time attention all you need for video understanding?” in *Proc. Int. Conf. Mach. Learn.*, July 2021.
- [16] E. T. Hall, *Beyond Culture*, 1976, garden City, New York: Anchor Press/Doubleday.
- [17] D. Muecke, “Irony markers,” *Poetics*, vol. 7, no. 4, pp. 363–375, 1978.
- [18] S. Attardo, J. Eisterhold, J. Hay, and I. Poggi, “Multimodal markers of irony and sarcasm,” *Humor - Int. J. Humor Res.*, vol. 16, no. 2, pp. 243–260, Jan. 2003.
- [19] S. Tabacaru and M. Lemmens, “Raised eyebrows as gestural triggers in humour: the case of sarcasm and hyper-understanding,” *Eur. J. Humor Res.*, vol. 2, no. 2, pp. 11–31, 2014.
- [20] A. Li, Q. Fang, and J. Dang, “Emotional intonation in a tone language: Experimental evidence from chinese,” in *Proc. 17th Int. Congr. Phonetic Sci. (ICPhS)*, Hong Kong, China, Aug. 2011, pp. 1198–1201.
- [21] E. Camp, “Sarcasm, pretense, and the semantics/pragmatics distinction,” *Noûs*, vol. 46, no. 4, pp. 587–634, 2012.
- [22] H. S. Cheang and M. D. Pell, “The sound of sarcasm,” *Speech Commun.*, vol. 50, no. 5, pp. 366–381, May 2008.