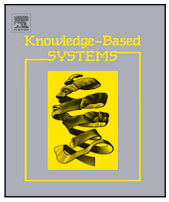
Knowledge-Based Systems 287 (2024) 111457

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



An attention-based, context-aware multimodal fusion method for sarcasm detection using inter-modality inconsistency

Yangyang Li [a,*], Yuelin Li [a], Shihuai Zhang [a], Guangyuan Liu [a], Yanqiao Chen [b], Ronghua Shang [a], Licheng Jiao [a]

[a] Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Joint International Research Laboratory of Intelligent Perception and Computation, International Research Center for Intelligent Perception and Computation, Collaborative Innovation Center of Quantum Information of Shaanxi Province, School of Artificial Intelligence, Xidian University, No. 2 Taibai South Road, Yanta District, Xi'an, 710071, Shaanxi, China
[b] The 54th Research Institute of China Electronics Technology Group Corporation, No. 589 Zhongshan West Road, Qiaoxi District, Shijiazhuang, 050081, Hebei, China

ARTICLE INFO

Keywords:
Multimodal sarcasm
Multimodal sentiment
Multimodal emotion
Attention mechanism
Deep learning

ABSTRACT

Sarcasm, a subtle and complex form of expression, presents significant challenges in detection, especially in the context of social media and meta universe applications where communication extends beyond text to include videos, images, and audio. Traditional sarcasm detection methods relying solely on text data often fail to capture the emotional incongruities and subtleties inherent in sarcasm. To address these challenges, this paper introduces a novel multimodal sarcasm detection method that not only processes multimodal data but also focuses on modeling the emotional mismatch between different modalities, a crucial aspect often overlooked by conventional approaches. Our method employs an intermodal emotional inconsistency detection mechanism, a contextual scenario inconsistency detection mechanism, and a cross-modal and segmented attention mechanism. These innovations enable a richer and more nuanced feature representation, capturing the essence of sarcasm more effectively. Experimental results on the dataset MUStARD Extended confirm the superiority of our approach, establishing it as the new state-of-the-art in sarcasm detection compared to existing models.

1. Introduction

Sarcasm, which means saying something opposite of what the speaker really wants to say, is a subtle way to express frustration, ridicule, contempt, or humor and wit. It is characterized by a tone of voice, facial expressions, or choice of words that convey a sense of irony or contempt. The implicit nature of sarcasm lies in its ability to convey a hidden meaning or subtext, requiring the listener to understand the intended irony behind the words. For instance, when someone says 'Oh, great! Another meeting. Just what I needed!', they are using sarcasm to express their dissatisfaction or annoyance with the situation, while the words themselves may seem positive on the surface.

Past studies in sentiment classification [1,2] have shown that while machine learning methods like SGD or SVM may yield satisfactory results in text data, their effectiveness diminishes in sarcasm detection due to sarcasm's implicit and contrastive nature. Deep learning approaches [3], such as CNNs or LSTMs, although more sophisticated, struggle with the computational demands and are limited in handling multimodal data, which is increasingly prevalent in the self-media era. To overcome these challenges, our model employs multimodal sentiment analysis, utilizing diverse types of information more effectively, akin to human perception. To effectively tackle the intricacies of sarcasm detection, our model integrates multimodal sentiment analysis, tapping into a diverse array of information types, which is pivotal in mirroring the human capacity to interpret sarcasm. This approach extends beyond mere text analysis, leveraging the rich context provided by multiple data types. A critical element of our methodology is the attention mechanism, which is adept at selectively focusing on the most pertinent information. This mechanism not only allows for a dynamic weighting of different input elements but also overcomes the drawbacks associated with fixed-length encoding vectors, making it adaptable to various data formats. The fusion of multimodal analysis with the attention mechanism represents a significant advancement in our model, offering a more nuanced and human-like understanding of sarcasm. It effectively addresses the challenges posed by sarcasm's implicit and contrastive nature, ensuring a sophisticated, context-rich analysis that aligns closely with human cognitive processes.

* Corresponding author.
E-mail address: yyli@xidian.edu.cn (Y. Li).

https://doi.org/10.1016/j.knosys.2024.111457
Received 2 November 2023; Received in revised form 13 January 2024; Accepted 27 January 2024
Available online 30 January 2024

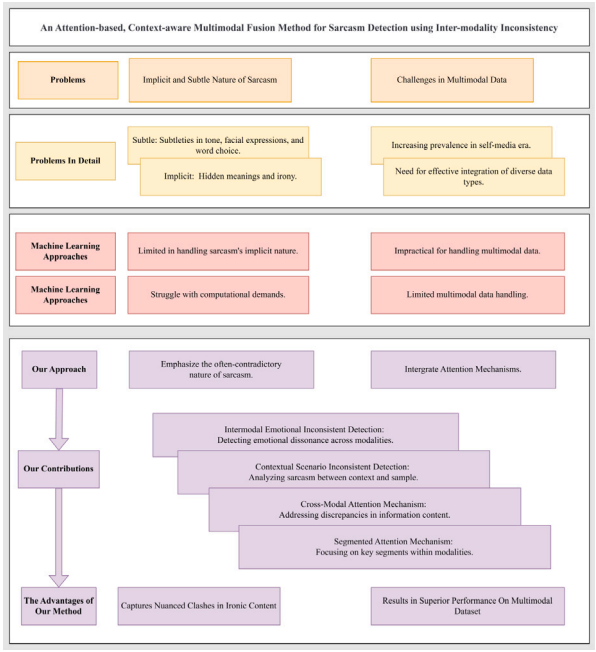


Fig. 1. A visual summary of the sarcasm detection problem and our approach.

Detecting sarcasm presents a formidable challenge, and there is no one-size-fits-all solution for this issue. We are confident that a substantial improvement in the accuracy of sarcasm detection can be attained by formulating our approach with a profound comprehension of sarcasm's intricacies. We focus on two key aspects in our approach: first, we emphasize the often-contradictory nature of sarcasm, and second, we integrate an attention mechanism to enhance the subtler and more effective identification of sarcastic elements. To overcome these challenges, our model employs multimodal sentiment analysis, utilizing diverse types of information more effectively, akin to human perception. A key component of our approach is the attention mechanism, which selectively focuses on the most relevant information, overcoming the drawbacks of fixed-length encoding vectors and adapting to various data formats. This integration of multimodal analysis and attention mechanism in our model marks a significant advancement in accurately detecting sarcasm.

For a visual summary of the problem addressed in this paper and our contributions, refer to Fig. 1. Additionally, part of our code and datasets used in this study are available on our GitHub repository at https://github.com/XDU-AI-LYYLab/MultimodalSarcasmDetection, facilitating further research and development in this field.

The main contributions and attributes of our work are as follows:

- We proposed an intermodal emotional inconsistent detection mechanism for sarcasm emotions with emotional dissonance properties.
- We proposed a contextual scenario inconsistent detection mechanism for sarcasm between context and sample.
- We introduced a cross-modal attention mechanism to address the huge discrepancy in information content between modalities.
- We introduced a segmented attention mechanism to address the problem that key segments within modalities are ignored.

The remainder of this paper is organized as follows: In Section 2, we review the literature on sarcasm sentiment, attention mechanism, and model evaluation. In Section 3, we explain the datasets and notations used in our study. In Section 4, we present the overall architecture of the model and its components. In Section 5, we report the results of our experiments. In Section 6, we summarize the paper and discuss the future work.

2. Related works

2.1. Sarcasm sentiment

According to a survey on sarcasm [4], sentiment can be categorized into five types: sarcasm as a disparity of sentiments, sarcasm as a means of conveying emotion, sarcasm as a form of written expression, sarcasm as a function of expertise, and behavior-based sarcasm. Given the inherent ambiguity of this sentiment, Chaudhari et al. [4] underscore several challenges scenarios in detecting sarcasm, including hyperbole, the expression of negative sentiment through positive words, and the use of brief text.

2.2. Sarcasm sentiment detection

Numerous researchers have focused on the task of sarcasm detection, mostly driven by the growing desire to identify negative intentions in user statements on social media platforms. Justo et al. [5] explore different sets of features based on various criteria and evaluates them using two classifiers. They find that sarcasm detection requires linguistic and semantic features, while nastiness detection can be done with surface patterns or indicators. Del et al. [6] propose a method to classify satirical and non-satirical tweets using a variety of psycholinguistic features. They find that the psycholinguistic features are effective in detecting satire.

Deep learning method has been widely applied in various fields, such as autonomous driving [7], news aggregation and fraud detection [8], natural language processing [9], and medical image processing [10–12]. Recently, with the development of deep learning, there has been a surge in research utilizing neural networks like transformers [13]. Ortega-Bueno et al. [14] propose a novel deep learning model, MvAttLSTM, for the detection of irony and satire in tweets written in various Spanish variants. The system achieves state-of-the-art performance in irony and satire detection in Spanish variants and competitive results in humor recognition. Potamias et al. [15] develop a Transformer-based method called Recurrent CNN-RoBERTA model for detecting sarcastic statements in a given dataset. Their method is proved effective on different datasets, including SemEval, Reddit's Political Sarcastic statements, and Rillof's sarcastic dataset. Aljedaani et al. [16] conduct a study on sentiment analysis of Twitter data using TextBlob and deep learning models, focusing on the US airline industry. The authors used a Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) to extract features automatically for analyzing sentiments and classification of reviews or opinions labeled into two polarity as positive or negative. The proposed model has better performances on benchmark datasets. Mazroui et al. [17] propose a system for multi-dialectal Arabic sentiment analysis that can detect the implied sarcastic features in the text. The system uses a hybrid approach that combines rule-based and machine learning techniques to identify the sarcasm indicators and their polarity. The system also uses a lexicon-based method to assign sentiment scores to the text segments. The results show that the system achieves high accuracy and F-measure scores for both datasets, and outperforms the baseline methods.

As described, the previous studies have primarily focused on analyzing the textual characteristics of sarcasm sentiment. However, this kind of approach may not be sufficient to effectively detect sarcasm, especially considering the abundance of multimedia data on social networks. In order to harness the potential of multimedia data more efficiently, we have incorporated multimodal detection into our model, resulting in a significant improvement in performance on social media datasets. Additionally, we have specifically designed our model to capture the fundamental characteristics of sarcasm sentiment.

2.3. Multimodal sentiment analysis

Considering the massive amount of multimedia content generated on social media, multimodal sentiment analysis is gaining more and more attention.

Liu et al. [18] propose a novel approach to multimodal sentiment analysis. The paper focuses on the development of a model that can analyze sentiment from multiple communication channels such as text, voice, and facial expressions. The paper provides a comprehensive overview of the state-of-the-art methods in multimodal sentiment analysis. Huang et al. [19] propose a text-centered fusion network with crossmodal attention (TeFNA) for multimodal sentiment analysis (MSA), which aims to leverage the text modality as the primary modality to improve the fusion of sentiment information from multiple modalities. Chauhan et al. [20] propose a novel approach to detect sentiment in multimodal conversational scenarios, using text, speech, and facial features along with emoji information. The paper presents an emoji-aware multitask deep learning framework that leverages the emoji information to improve the performance of sarcasm detection, as well as sentiment and emotion detection as auxiliary tasks.

2.4. Definition of multimodal sarcasm detection task

Sarcasm refers to the use of irony to mock or express contempt, according to definitions from Oxford Languages. While some researchers [21] have categorized sarcasm into more nuanced labels like political, humorous, and satirical, most view sarcasm detection as a binary classification task. To better focus on the intricacies of sarcasm, we adopt the latter perspective and define the sarcasm detection task as follows:

Definition 1 (*Sarcasm Detection on Multimedia Data*)**.** Given a set of multimedia input data consisting of audio A, text T, and video V segments, along with ground-truth sarcasm labels L, where the data segments have been aligned to a common timeline I and speakers P, the objective is to train a model M to perform sarcasm detection.

Specifically, the model takes aligned input segments for a given time i and speaker p:

$$M(A(i,p),T(i,p),V(i,p)) = l' \tag{1}$$

Where $i \in I$ and $p \in P$, and l' stands for the label predicted by the model. The model is optimized to match the ground-truth labels L.

2.5. Attention mechanism

The Attention Mechanism represents a great breakthrough in artificial intelligence, fundamentally transforming model capabilities across various domains. This mechanism, akin to human cognitive processes, enables AI models to dynamically concentrate on relevant information while filtering out extraneous inputs. Its application spans from computer vision [22] to natural language processing [23], demonstrating its versatility and effectiveness.

At its core, the Attention Mechanism is vital in numerous neural network architectures [24]. It empowers these models to identify and focus on the most pertinent segments of extensive input sequences, thereby enhancing predictive accuracy. The mechanism functions by calculating a weighted sum of the input, with the weights derived from a similarity measure between each input component and the model's current state, facilitating selective attention to crucial inputs.

The fundamental operation of the Attention Mechanism [13] is encapsulated in the formula:

$$\text{Attention}(Q,K,W) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)W \tag{2}$$

Fig. 2. Partial video frames of the extended MUStARD dataset.

Here, Q represents the queries, K the keys, and V the values. The softmax function is employed to normalize the weights, ensuring they sum to 1.

With the Attention Mechanism's ability to highlight subtle aspects, it holds the potential to significantly improve sarcasm detection accuracy, particularly in scenarios characterized by subtlety.

3. Datasets and notations

3.1. Datasets

3.1.1. MUStARD extended dataset

Castro et al. [25] proposed the MUStARD dataset, a pivotal resource for multimodal sarcasm detection, featuring a collection of 6365 videos, with 345 labeled as sarcastic and 6020 as non-sarcastic. Some video frames of the expanded MUStARD dataset are shown in Fig. 2. Predominantly composed of clips from TV shows like "Friends", "The Golden Girls", and "The Big Bang Theory", it is designed to aid research in detecting sarcasm through visual, auditory, and textual analysis. The dataset's annotation process utilized a custom web interface, enabling accurate sarcasm assessment through context clips. To ensure a balanced analysis, an equal number of sarcastic and non-sarcastic videos were selected, resulting in 690 videos. Crucially, MUStARD includes the context for each video, focusing on preceding dialogue turns to fully capture conversational dynamics. This feature is essential in sarcasm detection, underscoring the importance of context alongside content.

Chauhan et al. [26] expanded the MUStARD dataset to create the MUStARD Extended dataset, adding sentiment tags to its 690 audiovisual conversational utterances. This dataset is evenly split between 345 ironic and 345 non-ironic samples from four American TV series. Each sample includes text, video, and audio, contextualized by preceding content. The addition of sentiment tags provides a nuanced approach to analyzing emotional layers in sarcastic expressions, enhancing its use in sarcasm detection research. The dataset's label distribution is detailed in Table 1, which shows the ratio of implicit and explicit sentiment labels, and in Table 2, which outlines the ratio of implicit and explicit emotion labels. These tables offer a concise overview of the sentiment and emotion annotations, underscoring the dataset's depth in capturing emotional and sarcastic nuances.

In the experimental section, we evaluate our method on this dataset using two experimental settings. The first setting is speaker-independent. In this setting, the utterances from the Friends TV show are used as test data, and the remaining samples are used as training data. The second setting is speaker-related. In this setting, the dataset is divided into five parts, so that in five iterations, the ith part is selected as the test set each time, where $i \in 1,2,\ldots,5$, and the rest are used for training, thus producing five test sets.

3.2. Notations

All notations present in the paper are shown in Table 3.

Table 1
Sentiment distribution of dataset MUStARD Extended.

Implicit sentiment			Explicit sentiment		
Neg	Neu	Pos	Neg	Neu	Pos
391	89	210	246	119	325

Table 2
Emotion distribution of dataset MUStARD Extended.

Implicit Emotion									
Sentiment	An	Ex	Fr	Sd	Sp	Fs	Hp	Neu	Dg
Count	97	18	14	121	29	57	143	198	39
Explicit Emotion									
Sentiment	An	Ex	Fr	Sd	Sp	Fs	Hp	Neu	Dg
Count	54	30	6	118	35	23	206	228	10

Table 3
Notations.

	Notation	Description
Input	T_c^r	Contextual text information
	SP_c	Speaker information
	T_u^r	Text information of the sample
	SP_u	Speaker information corresponding to the sample
	V_c^r	Contextual video frame information
	V_u^r	Video frame information of the sample
	A_u^r	Audio information of the sample
Features	v_u	visual attributes at the discourse level
	t_u	textual attributes at the discourse level
	a_u	acoustic attributes at the discourse level

4. Methodology

In this section, the structure of the proposed multimodal sarcasm detection will be introduced and is shown as Fig. 3. The model consists of two main parts: feature extraction and sarcasm detection. Feature extraction involves using deep learning models to extract features from image, text, and audio data. The sarcasm detection module consists of four components that identify signals of sarcasm from different aspects. A $softmax$ function is used to obtain the final prediction of whether sarcasm is present or not.

4.1. The overall structure of the model

The overall structure of the model is as shown as Fig. 3. Our model takes in multimodal input consisting of visual, textual, and auditory data. Different data modalities are processed by separate deep learning feature extraction modules to obtain informative representations.

The input data including contextual contents and speaker details are specifically made. In text modality, T_c^r denotes the text information of the context, SP_c indicates the speaker information. Since there might be multiple speakers in the context, an input method that follows each speaker's utterance with their speaker information is adopted. T_u^r and SP_u are the text information and the corresponding speaker information of this sample, respectively. In visual modality, because the dataset does not provide video frames at sentence level, different speakers in the context cannot be segmented. Therefore, speaker information is not considered and only V_c^r as the video frame information of the context is used. V_u^r and SP_u are the video frame information and speaker information of this utterance respectively. In auditory modality, due to the difficulty of eliminating overlapping audio from multiple speakers and laughter from comedy (non-dialogue content), any auditory content from context is not used in our method. Since then, we only use audio A_u^r and speaker SP_u information in auditory modality.

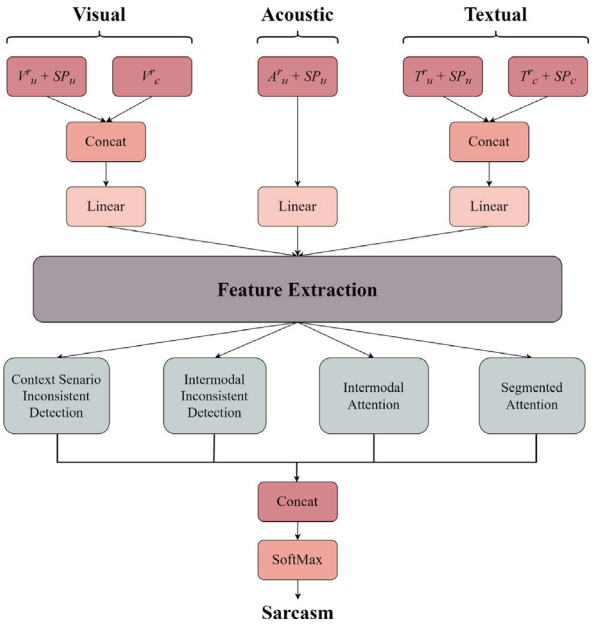


Fig. 3. The overall structure of the proposed model.

The features from each modality are then concatenated together into a joint multimodal representation, capturing both low-level lexical information as well as high-level contextual cues. This multimodal feature vector summarizes the semantic, discourse, acoustic, and visual aspects of the sarcasm presented in the input. It serves as the input to subsequent sarcasm classification modules in the architecture of our model.

4.2. Feature extraction

Inspired by Castroet et al. [25], two types of features, discourse-level and word-level multimodal features to detect sarcasm are used. In the realm of acoustic characteristics, OpenSmile is utilized to derive speech-level acoustic attributes that offer a range of complex features in an acoustic context. Ultimately, visual attributes, textual attributes, and acoustic attributes at the discourse level as v_u, t_u, and a_u are represented respectively, where $v_u \in \mathbb{R}^{d_v}$, $t_u \in \mathbb{R}^{d_t}$, $a_u \in \mathbb{R}^{d_a}$. Because discourse-level multimodal feature extraction targets whole sentences, the features only contain global information and can easily overlook subtle changes in actions and acoustics. To counter this challenge, word-level multimodal attributes that supply local information are introduced. Based on these attributes, the level of discordance between modalities to enhance the precision of sarcasm emotion detection tasks are determined. Our feature extraction network model structure is shown as Fig. 4.

4.2.1. Alignment in words

To acquire word-level multimodal features, it is necessary to align the tri-modal data at a word level prior to feature extraction. Given an input discourse u that includes words $w_1, w_2, \ldots w_l$, (where l is determined by the quantity of words forming the sentence), The Python library GENTLE for forced alignment is initially employed. This aligns the audio with each individual word. Based on these initial alignment results, we segment video frames to obtain audio segments and video frames corresponding to each word. It is important to note that the alignment is not required when obtaining discourse-level features.

4.2.2. Extract textual features

BERT is used to extract text utterances [27] from the dataset and convert each utterance u into a sentence representation t_u. Assuming there are n words in an utterance $w_1, w_2, \ldots w_n$, where $w_j \in \mathbb{R}^{d_t}$ and

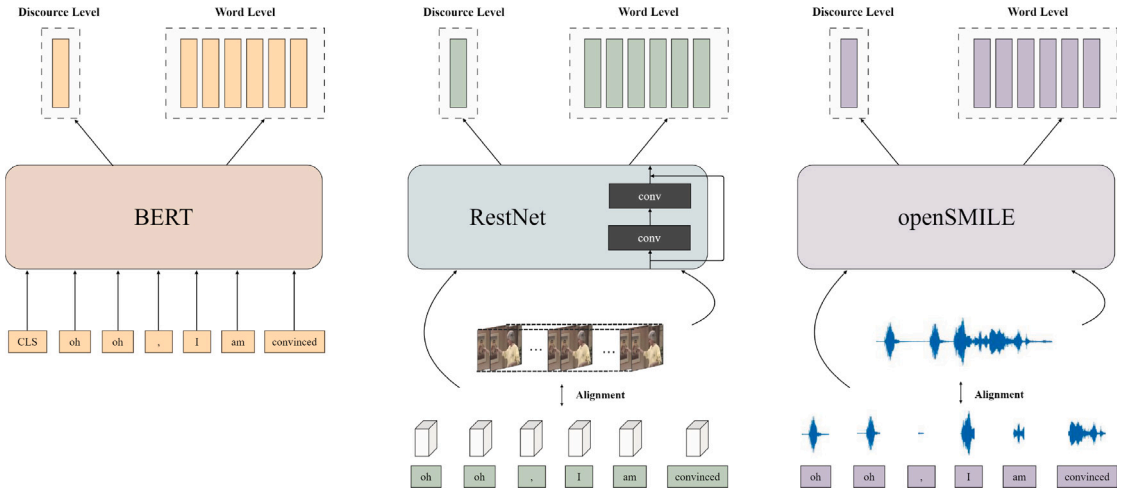


Fig. 4. Feature Extraction Network Structure.

w_j is a vector obtained using fastText word embeddings. In our model, the strategy of averaging the first token (Token)-[CLS] in the last four layers of Transformer is adopted to obtain a unique representation of 768 dimensions for the utterance $t_{w_1}, t_{w_2}, \ldots, t_{w_n}$, where $t_{w_i} \in \mathbb{R}^{d_t^w}$. When extracting utterance-level features, vector w_u is also obtained by using fastText word embeddings and inputting it into BERT [28] to obtain a feature representation of 768 dimensions.

4.2.3. Extract visual features

ResNet-152 [29] image classification model, which is pre-trained on ImageNet [30], is implemented to extract visual features $v_{w_1}, v_{w_2}, v_{w_3}, \ldots, v_{w_n}$ from the pool5 layer [31] of the video due to the richness of ironic cues in the visual modality. All features adhere to the condition $v_{w_i} \in \mathbb{R}^{d_v}$. Before features are extracted, the video frames are first scaled to 224 × 224 pixels and center cropping and normalization are performed. A $d_v = 2048$ dimensional feature vector v_{wi}^j is then computed for each frame $j \in \{1, \ldots, f\}$, and finally the average feature vector $v_u = \frac{1}{f}(\sum_j v_{wi}^j) \in \mathbb{R}^{d_v^w}$ for each word or utterance is obtained. Note that when extracting utterance-level visual features, f is the product of the utterance duration and the frame rate per second. When extracting word-level visual features [32], f is the product of the word duration and the frame rate per second.

4.2.4. Extract acoustic features

In order to obtain information from the auditory modality, the module of acoustic features extraction extracts low-level acoustic features from each utterance's audio file and uses them to infer details related to the speaker's intonation, pitch and tone. The speech processing tool OpenSmile [33] is employed for extracting acoustic features. It first loads the utterance's audio samples as a time series signal with a sampling rate of 22050 Hz [34], then removes the background noise from the signal using the Sox method, and finally segments the audio signal into d_w non-overlapping windows. From each window, it extracts local features such as Mel-scale Frequency Cepstral Coefficients (MFCC), Mel-spectrum, zero-crossing rate, spectral centroid and their time derivatives. The segmentation aims to obtain a fixed-length representation of the audio signal, otherwise the dimensionality of acoustic features would vary with different utterance lengths. Since multimodal features at word level are required, the segmentation length for word-level acoustic features depends on the duration of each word. Conversely, for utterance-level features it depends on the whole audio duration. All extracted acoustic features are concatenated to form a joint representation $d_a = 1000$ of window with dimensionality $\{a_{ui}^j\}_{i=1}^{d_w}$, where $j \in 1, \ldots, l$. The final acoustic representation of each utterance

Fig. 5. An example of a sarcastic utterance paired with a contradictory facial expression.

is obtained by computing the mean of all windows $\{a_{w1}, a_{w2}, \ldots, a_{w_l}\}$, where $a_{w_j} \in \mathbb{R}^{d_a^w}$. The formula [34] is as follows:

$$a_{w_j} = \frac{1}{d_w}(\sum^{i} a_{ui}^j) \tag{3}$$

4.3. Intermodal emotional inconsistent detection

Sarcasm detection hinges on recognizing its inherent implicitness and inconsistency, where the latter refers to a mismatch between expressed emotions and the actual sentiment. The provided image Fig. 5 illustrates this concept vividly; the man, Chandler, exhibits an emotional incongruity between his verbal sarcasm and his facial expressions. His words seem to praise, yet his expressions and tone suggest critique, an archetypal sarcastic delivery. Acknowledging this, we introduce the 'Intermodal Emotional Inconsistency' detection mechanism in our paper, a crucial component for sarcasm identification. This mechanism is designed to detect discrepancies between verbal and non-verbal cues, such as a sarcastic remark paired with a contradictory facial expression, as exemplified in the image. By harnessing this intermodal emotional inconsistency, our model effectively captures the essence of sarcasm, improving detection accuracy.

The inter-modal emotional inconsistency module, illustrated in Fig. 6, is anchored by an inconsistency scoring mechanism that detects emotional disparities across modalities. This mechanism discerns contrasts like negative verbal expressions paired with positive facial cues or a positive vocal tone accompanying negative textual content.

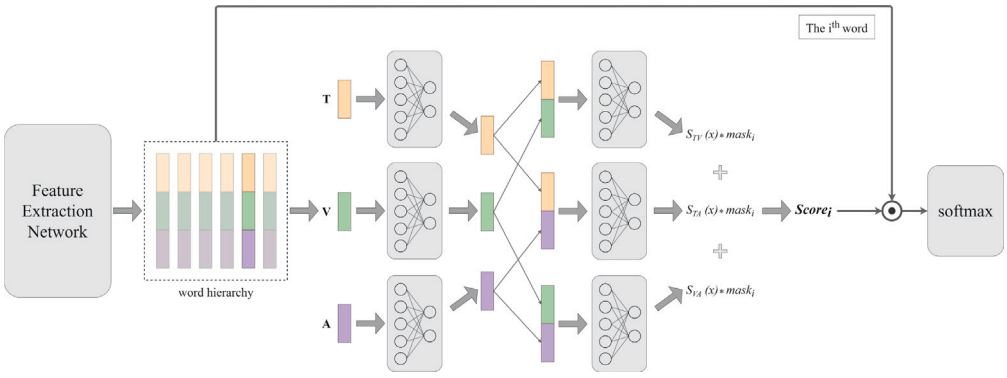


Fig. 6. Intermodal Emotional inconsistent Sarcasm Detection Network Structure.

Such discrepancies are telltale signs of sarcasm, serving as linguistic markers of inter-modal emotional inconsistency. Building on the previously discussed extraction of word-level multimodal features, we have developed scoring functions—S_{TV}, S_{TA}, and S_{VA}—to quantify the inconsistency levels between the text, visual, and auditory modalities. These functions leverage emotional scores from lexicons to adjust the inconsistency ratings for each word. For lexicon-absent words, their inherent incongruity scores remain unaltered. Further exposition on these scoring functions is provided below.

First, a non-linear layer is used to transform the word-level features and learn feature representations, and the calculation formulas are as follows:

$$\begin{aligned} r^t_{w_i} &= \tanh(W_t t_{w_i} + b_t) \\ r^a_{w_i} &= \tanh(W_a a_{w_i} + b_a) \\ r^v_{w_i} &= \tanh(W_v v_{w_i} + b_v) \end{aligned} \tag{4}$$

Where $W_t \in \mathbb{R}^{d\times d_t^w}, W_a \in \mathbb{R}^{d\times d_a^w}, W_v \in \mathbb{R}^{d\times d_v^w}, b_t \in \mathbb{R}^d, b_a \in \mathbb{R}^d, b_v \in \mathbb{R}^d$ are trainable weights, and d is the number of units in the hidden layer.

As shown in Fig. 6, the text modality, visual modality, and auditory modality are cross-concatenated to obtain the inter-modal interaction features. This inter-modal fusion can make full use of the correlation in heterogeneous data and reduce data waste. Next, the inter-modal scoring functions S_{TV}, S_{TA}, S_{VA} are used to measure the degree of incongruity between two modalities, which leverages the inter-modal incongruity and enhances the theoretical foundation of the model from a linguistic perspective. Its score is positively correlated with the degree of incongruity of words, and then the incongruity score is adjusted according to the emotional score of words and different levels of attention are applied to words based on the word incongruity score. The specific calculation formulas are as follows:

$$\begin{aligned} S_{TV}(r^t_{w_i}, r^v_{w_i}) &= ReLU(W_p[r^t_{w_i} \oplus r^v_{w_i}] + b_p) \\ S_{TA}(r^t_{w_i}, r^a_{w_i}) &= ReLU(W_q[r^t_{w_i} \oplus r^a_{w_i}] + b_q) \\ S_{VA}(r^v_{w_i}, r^a_{w_i}) &= ReLU(W_o[r^v_{w_i} \oplus r^a_{w_i}] + b_o) \end{aligned} \tag{5}$$

Where $W_p, W_q, W_o \in \mathbb{R}^{1\times 2d}$ are the weights of the fully connected network, $b_p, b_q, b_o \in \mathbb{R}^1$ are the network biases, and $\oplus$ denotes the one-dimensional concatenation between vectors. Here, a fully-connected neural network is used to model the mapping relationship between the degree of incongruity and the inter-modal interaction features. The non-linearity of the network can extract more rich information and fully capture the incongruity between modalities.

In Section 2.1, the basic knowledge of sarcasm was introduced, revealing that there are three types of irony. It was learned that two of these types (irony that reflects emotional differences, and sarcasm as a medium for conveying emotions) are related to the degree of emotion. Whether it is the emotional contrast between words and phrases and the overall context, or expressing emotions with opposite words, there are many words with strong emotions in the speech.

Table 4
Examples of emotional scores of words in emotional dictionary.

Word	Type	Emotional score
nice	adjective	0.708
quality	noun	0.352
fairly	adverb	−0.034
quiet	adjective	−0.218
looking	verb	0.012
big	adjective	0.103
bought	verb	0.083

Therefore, it can be inferred that the stronger the emotion degree of the words in the speech, the higher the probability of irony tendency. By using SenticNet [35] emotion dictionary, the emotion score of each word is obtained before text features are extracted. An emotion score corresponds to each word, and Table 4 shows some specific examples. The value range of emotion score is −1 to 1, and the emotion intensity of the word is indicated by the score. The stronger the positive emotion intensity of the word, the closer the score is to 1. The stronger the negative emotion intensity of the word, the closer the score is to −1. The closer the emotion polarity is to neutral, the closer the score is to 0. The final discordance score is obtained by adjusting the discordance degree based on the emotion score of the word. The specific calculation formula is as follows:

$$\begin{aligned} s_{w_i} &= W_x * [S_{TV}(r^t_{w_i}, r^v_{w_i}) \\ &\quad + S_{TA}(r^t_{w_i}, r^a_{w_i}) + S_{VA}(r^v_{w_i}, r^a_{w_i})] \\ W_x &= f(s^e_{w_i}) \\ f(x) &= x^2 \end{aligned} \tag{6}$$

Where s_{w_i} is the discordance score of the ith word in the speech, W_x is the discordance degree adjustment coefficient based on the emotion dictionary. $s^e_{w_i}$ is the emotion score of the ith word. The function $f(\cdot)$ affects how much the word emotion score adjusts the discordance score.

After the incongruity score of words has been obtained, the softmax function is applied to convert the score s_{w_i} into the attention weight p_{w_i}. The calculation formula is as follows:

$$p_{w_i} = \frac{e^{s_{w_i}}}{\sum_{i=1}^{l} e^{s_{w_i}}} \tag{7}$$

Where p_{w_i} represents the attention weight of the ith word in the speech.

To obtain the features of the entire speech, the word-level features of three modalities are concatenated and weighted according to the attention weight of words. The calculation formula is as follows:

$$\begin{aligned} r_{w_i} &= t_{w_i} \oplus a_{w_i} \oplus v_{w_i} \\ r_{w_i} &= p_{w_i} r_{w_i} \\ r_w &= \sum_{i=1}^{l} r_{w_i} \end{aligned} \tag{8}$$

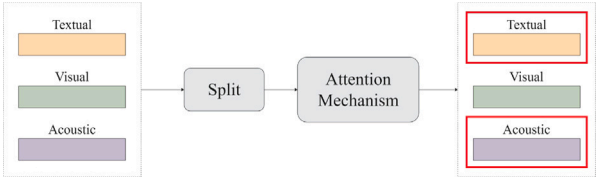


Fig. 7. A Schematic Diagram of the Cross-Modal Attention Computation Process.

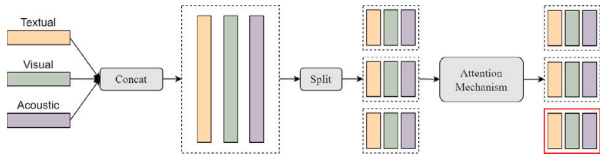


Fig. 8. A Schematic Diagram of the Segmented Attention Computation Process.

Where r_{w_i} is the word-level multimodal feature, which assigns different levels of attention based on the degree of incongruity and aggregates the features to obtain the word-level feature r_w of the whole speech. The word-level multimodal feature of the complete speech is input into the non-linear activation function tanh to perform the transformation, and the final representation r_c is obtained. The calculation formula is as follows:

$$r_c = \tanh(r_w) \tag{9}$$

Given that the irony emotion detection task is a binary classification task, the classic softmax classifier is utilized for irony prediction, and cross entropy is chosen as the loss function. The formula is as follows:

$$p = softmax(W_c r_c + b_c) \tag{10}$$

Where $W_c \in \mathbb{R}^{(d_t+d_v+d_a)}$, and $b_c \in \mathbb{R}^2$ are the parameters of the softmax classifier.

4.4. Attention mechanism

4.4.1. Cross-modal attention mechanism

The overall computation process is shown in Fig. 7. The modal attention module aims to predict irony using the multimodal information of a single speech. Given a speech u, the attention feature map of u needs to be computed, which represents the self-attention of the multimodal information of a single speech. Due to the different dimensions of the modal features, zero-padding is used to align the dimensions based on the modal feature with the highest dimension $d'_t = d'_v = d'_a = d_m$. After concatenating the three modalities, the resulting feature matrix is $X_u \in \mathbb{R}^{3\times d_m}$. Initially, the matching matrix $M_u \in \mathbb{R}^{3\times 3}$ of the feature matrix X_u is computed. Following this, the softmax function is utilized to compute the probability distribution score $N_u \in \mathbb{R}^{3\times 3}$ of the matching matrix M_u. This method essentially computes the attention weight between modalities, and applies soft attention to the speech-level multimodal features to compute the attention representation of different modalities $O_u \in \mathbb{R}^{3\times d_m}$. The computation formula is as follows:

$$\begin{aligned} M_u &= X_u \cdot X_u^T \\ N_u(i,j) &= \frac{e^{M_u(i,j)}}{\sum_{k=1}^{3} e^{M_u(i,k)}}, \quad for\ i,j = 1,2,3 \\ O_u &= N_u \cdot X_u \end{aligned} \tag{11}$$

Finally, the attention weight matrix O_u is applied to the speech-level multimodal features X_u, completing the attention allocation of different modalities to obtain balanced multimodal features. The formula is as follows:

$$A_u = O_u \odot X_u \tag{12}$$

Where $A_u \in \mathbb{R}^{3\times d_m}$.

4.4.2. Segmented attention mechanism

This module focuses on extracting the key segments from the speech, using the same dimension alignment method as described in Section 4.4.1. The speech-level multimodal feature X_u is obtained by concatenating the feature vectors of the three modalities t_u, v_u, a_u for each speech. The feature vector is then divided into k equal segments, and each segment feature $X_{u_i} \in \mathbb{R}^{3\times \frac{d_m}{k}}$, where $i \in 1,\ldots,k$. The segmented representation of the multimodal features is obtained at this point. This module aims to establish the influence relationship among different segment feature vectors. The overall structure of the segment attention module is shown in Fig. 8.

The calculation of discourse segment attention consists of two main steps:

(1) Extract multi-modal features at the discourse level and evenly divide them into k segments, concatenate the k segment features to obtain the segment feature matrix S_r.

(2) Use the segment feature matrix S_r as input to calculate the matching matrix, then apply the softmax function to compute the probability distribution scores $P(i,j)$ for each segment of the segment feature matrix, and calculate the attention representation O between segments via soft attention mechanism, finally multiply the attention weight matrix and the feature matrix element-wise to allocate more attention to key segments.

After the computation of the inter-modal attention module and segment attention module, the fused feature representations $A_u, B_u \in \mathbb{R}^{3\times d_m}$ are obtained, respectively, which are then concatenated and fed into the softmax classifier to get the classification probability p_m. The calculation formula is:

$$p_m = softmax(W_m(A_u \oplus B_u) + b_m) \tag{13}$$

Where $W_m \in \mathbb{R}^{(6\times d_m)}, b_m \in \mathbb{R}^2$ are parameters of the $softmax$ classifier.

4.5. Contextual emotional inconsistent detection

The aim of this module is to capture and model the emotional differences between the context and the speech. As shown in Fig. 9, the context consists of several sentences preceding the sample speech, and there may be different emotions among different speeches.

This module divides the context features and builds an emotional contrast structure between the context and the sample speech. It captures the emotional contrasts between the speech features and the context segment features, computes the emotional contrast scores based on them, and assigns attention to the context segments according to their level of difference.

The multimodal speech features and multimodal context features are obtained by using the feature extraction network. The multimodal features are simply concatenated to get the multimodal context features and multimodal speech features: $R_c, R_u \in \mathbb{R}^{(d_t+d_v+d_a)}$. The multimodal information is transformed by applying a non-linear function:

$$\begin{aligned} r_c &= \tanh(W_c R_c + b_c) \\ r_u &= \tanh(W_u R_u + b_u) \end{aligned} \tag{14}$$

Where $r_c, r_u \in \mathbb{R}^{(d_t+d_v+d_a)}$ are the feature representations of the context and speech multimodal information, $W_c, W_u \in \mathbb{R}^{d\times(d_t+d_v+d_a)}, b_c, b_u \in \mathbb{R}^d$ are trainable weight.

Contextual features are evenly divided into M segments $c_1, c_2, \ldots, c_M$ before identifying emotional variances. Each segment is allowed to merge with dialogue features. An emotional variance scoring function S_{ciu} is introduced by this method. The emotional variances between different segments and dialogues can be discerned based on the results of this scoring function. A larger emotional variance between a contextual segment and dialogue is indicated by a higher score, suggesting a higher

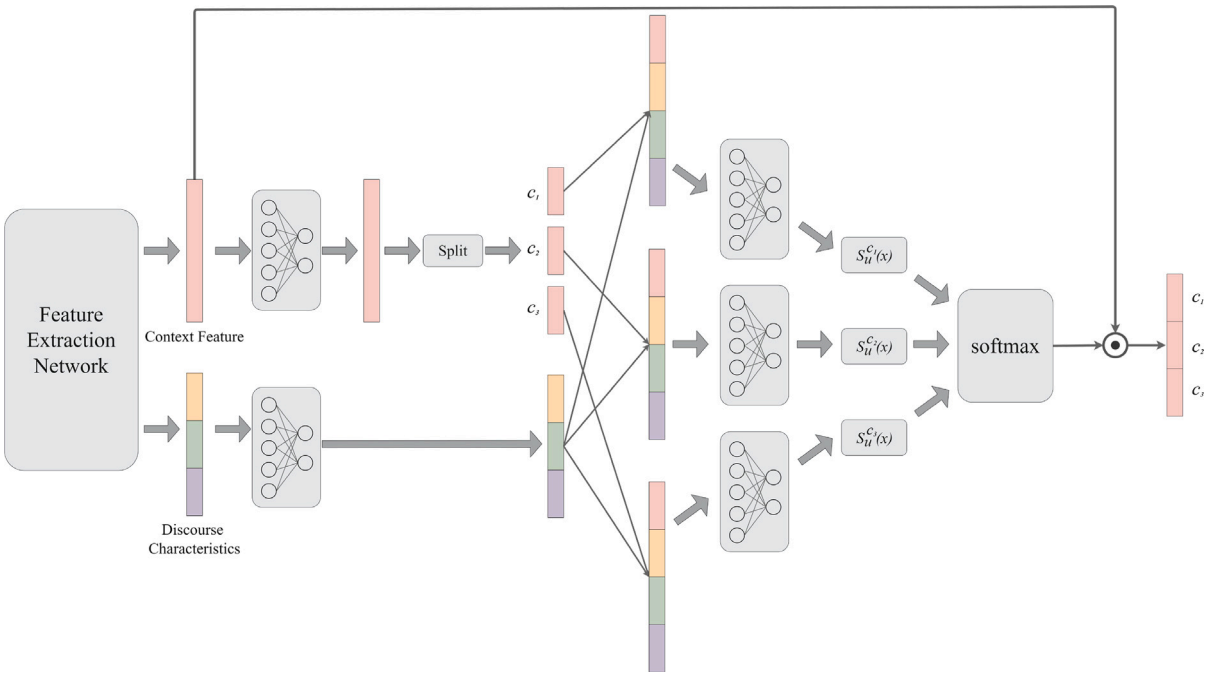


Fig. 9. Contextual Emotional inconsistent Sarcasm Detection Network Structure.

probability of sarcasm within the dialogue. The specific calculation formula is as follows:

$$S_u^{c_i}(r_{c_i}, r_u) = ReLU(W_i[r_{c_i} \oplus r_u] + b_i) \tag{15}$$

Where $c_i \in \mathbb{R}^{dt+dv+da}/M$ represents the ith segment of contextual features (where i ranges from 1 to M), $b_i \in \mathbb{R}^1$, $W_i \in \mathbb{R}^{1\times 2d}$. Merged features and emotional variance scores are mapped by using a fully connected neural network, effectively capturing the emotional variances between context and dialogue.

According to the emotional difference scores of different segments, the emotional difference attention distribution of the multimodal features of the context is obtained by the softmax function. The context features are weighted by this distribution, and the multimodal feature representation of the context is obtained $R_f \in \mathbb{R}^{dt+dv+da}$. The calculation formula is as follows:

$$p_{ci} = \frac{e^{S_u^{ci}}}{\sum_{i=1}^{M} e^{S_u^{ci}}}$$
$$R_f = p_{c_1} c_1 \oplus p_{c_2} c_2 \oplus \cdots \oplus p_{c_i} c_i \tag{16}$$

The final feature representation is obtained by combining the modal emotional incongruity module, the modal attention module, the segment attention module, and the context emotional contrast module. The irony probability p_f is obtained by using softmax. The calculation formula is as follows:

$$p_f = softmax(W_f(r_c \oplus A_u \oplus B_u \oplus R_f) + b_f) \tag{17}$$

where $W_f \in \mathbb{R}^{2\times(d_t+d_v+d_a)+6\times d_m}$, and $b_f \in \mathbb{R}^2$ are the classifier parameters.

5. Experiments and results

This section outlines our experimental methodology and the corresponding results.

5.1. Experimental setup

In our study, the PyTorch library was utilized to implement the model. We adopted Precision (P), Recall (R), and F_1-score (F1) as our key metrics for evaluating sarcasm detection performance.

Table 5
Network configuration.

Parameter	Value
Optimizer	Adam
Learning rate	1×10^{-4}
Batch size	32
Maximum number of iterations	50
Loss function	Categorical cross-entropy
Number of neurons in fully connected network	200
Feature dimension of text modality	768
Feature dimension of visual modality	2048
Feature dimension of auditory modality	1000

Details of the network configuration are provided in Table 5.

To mitigate the risk of overfitting, a dropout layer with a 50% probability was integrated into the experiment. Notably, the extended MUStARD dataset, used in our experiments, consists of only 690 samples, maintaining the same scale as the original dataset. Considering the limited size of the dataset, we opted for a Support Vector Machine (SVM) for testing purposes instead of the standard softmax classifier. The penalty term C of the SVM was set to 1 for the speaker-dependent setting and 1000 for the speaker-independent setting.

5.2. Results and analysis

Our extensive experiments demonstrate the effectiveness of our proposed multimodal approach in irony detection. Analyzing the extended MUStARD dataset, we compared models based on different combinations of modalities: text T, video V, and audio A. The findings, summarized in Table 6 and Table 7, reveal a significant improvement when utilizing all three modalities.

Specifically, the model incorporating T, V, and A achieved the highest performance metrics in both Speaker-Dependent and Speaker-Independent settings. In the Speaker-Dependent condition, this model outperformed its counterparts with precision, recall, and F_1-score values of 76.2%, 74.2%, and 0.752, respectively. Similarly, in the Speaker-Independent scenario, it showed superior precision (71.9%), recall (69.9%), and F_1-score (0.709).

Table 6
Performance comparison of different modality combinations on the extended MUStARD dataset under Speaker-Dependent condition.

Modality			Performance measures					
T	V	A	Sensitivity	Specificity	Accuracy	Precision	Recall	F_1-score
✓	✓	×	70.9%	72.97%	71.94%	72.4%	70.9%	0.716
✓	×	✓	70.4%	72.89%	71.65%	72.2%	70.4%	0.713
✓	✓	✓	**74.2%**	**76.82%**	**75.51%**	**76.2%**	**74.2%**	**0.752**

Table 7
Performance comparison of different modality combinations on the extended MUStARD dataset under Speaker-Independent condition.

Modality			Performance measures					
T	V	A	Sensitivity	Specificity	Accuracy	Precision	Recall	F_1-score
✓	✓	×	66.3%	68.22%	67.26%	67.6%	66.3%	0.669
✓	×	✓	67.7%	68.29%	67.99%	68.1%	67.7%	0.679
✓	✓	✓	**69.9%**	**72.68%**	**71.29%**	**71.9%**	**69.9%**	**0.709**

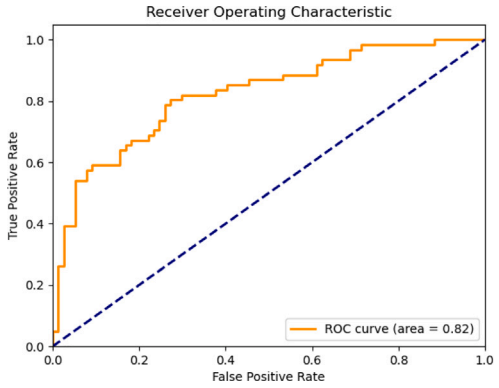


Fig. 10. ROC curve of our method on the extended MUStARD dataset under Speaker-Dependent condition.

The effectiveness of our tri-modal model is attributed to its ability to capture the intricate nature of sarcasm through a comprehensive analysis of textual, visual, and auditory cues. The fusion of these modalities allows for a deeper understanding of the context and the subtle nuances often present in sarcastic expressions. This is particularly evident in cases where sarcasm is conveyed through a combination of verbal irony and non-verbal cues, such as facial expressions or tone of voice, which our model adeptly identifies and interprets.

In addition to the numerical performance metrics, we further evaluated our models using Receiver Operating Characteristic (ROC) curves for both the independent and dependent settings. ROC curves are instrumental in assessing the balance between true positive and false positive rates. These curves for the independent and dependent experiments are depicted in Fig. 11 and Fig. 10, respectively. They provide a visual measure of our models' capability to distinguish between classes effectively.

Moreover, to gain a deeper insight into the classification accuracy and error patterns of our algorithms, we examined the confusion matrices for both experiments. These matrices, illustrated in Fig. 13 and Fig. 12, detail the instances of correct and incorrect classifications. The analysis of these matrices is vital for identifying specific error types made by the algorithms and for a comprehensive understanding of their classification accuracy and misclassification trends.

5.3. Comparative analysis

In this section, we analyze the performance of various sarcasm detection algorithms on the extended MUStARD dataset, with a specific focus on comparing our method's efficacy in both speaker-dependent and speaker-independent settings. The results are detailed in Tables 8 and Table 9, where the top performance achieved by our method is highlighted in bold.

Our analysis reveals that our method consistently outperforms others in both modes of the dataset. Notably, in the speaker-independent mode, it surpasses the state-of-the-art (SOTA) method by 0.032 in the F_1-score metric, indicating a significant advancement in sarcasm detection. In the speaker-dependent mode, there is a slight yet meaningful increase in recall by 0.2 percentage points. This improvement

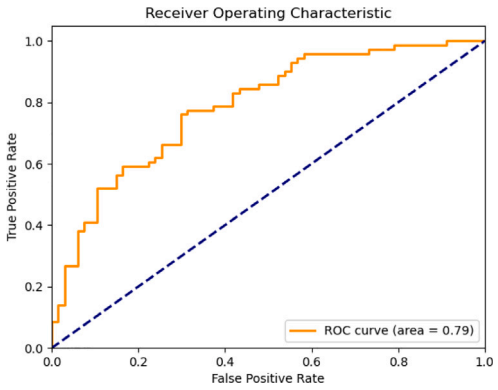


Fig. 11. ROC curve of our method on the extended MUStARD dataset under Speaker-Independent condition.

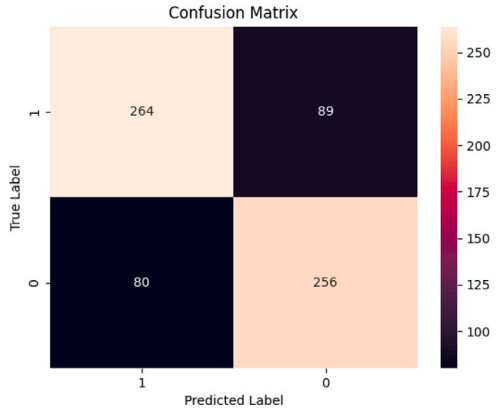


Fig. 12. Confusion matrix of our method on the extended MUStARD dataset under Speaker-Dependent condition.

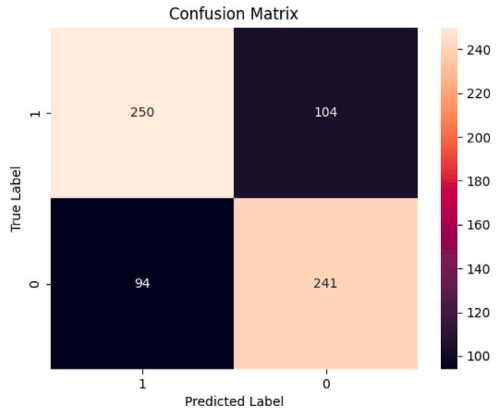


Fig. 13. Confusion matrix of our method on the extended MUStARD dataset under Speaker-Independent condition.

Table 8
Performance comparison of different methods on the extended MUStARD dataset under Speaker-Dependent condition.

Modality	Method	Performance measures					
		Sensitivity	*Specificity*	*Accuracy*	*Precision*	*Recall*	F_1-score
T	BERT [28]	66.9%	67.79%	67.34%	67.5%	66.9%	0.668
T	SMSD [36]	61.0%	61.97%	61.49%	61.6%	61.0%	0.611
T	MIARN [37]	64.0%	65.08%	64.54%	64.7%	64.0%	0.639
TV	FiLM [38]	66.2%	67.83%	67.02%	67.3%	66.2%	0.667
TAV	ConAttSD [39]	70.3%	70.44%	70.37%	70.4%	70.3%	0.703
TAV	GRU-based [40]	70.9%	72.02%	71.46%	71.7%	70.9%	0.708
TAV	EF-Concate [25]	70.8%	71.36%	71.08%	71.2%	70.8%	0.710
TAV	IAIE [26]	71.6%	72.29%	71.95%	72.1%	71.6%	0.719
TAV	MCER [41]	74.1%	74.1%	74.1%	74.1%	74.1%	0.741
TAV	Our method	**74.2%**	**76.7%**	**75.5%**	**76.2%**	**74.2%**	**0.752**

Table 9
Performance comparison of different methods on the extended MUStARD dataset under Speaker-Independent condition.

Modality	Method	Performance measures					
		Sensitivity	*Specificity*	*Accuracy*	*Precision*	*Recall*	F_1-score
T	BERT [28]	56.7%	59.3%	58.0%	58.2%	56.7%	0.574
T	SMSD [36]	48.2%	54.9%	51.6%	51.7%	48.2%	0.499
T	MIARN [37]	55.2%	63.8%	59.5%	60.4%	55.2%	0.577
TV	FiLM [38]	59.4%	61.7%	60.6%	60.8%	59.4%	0.601
TAV	ConAttSD [39]	63.8%	64.6%	64.2%	64.3%	63.8%	0.640
TAV	GRU-based [40]	62.6%	63.8%	63.3%	63.4%	62.7%	0.629
TAV	EF-Concate [25]	63.1%	65.0%	64.0%	64.3%	63.1%	0.637
TAV	IAIE [26]	65.5%	66.3%	65.9%	66.0%	65.5%	0.658
TAV	MCER [41]	66.0%	71.0%	68.5%	69.5%	66.0%	0.677
TAV	Our method	**69.9%**	**72.7%**	**71.3%**	**71.9%**	**69.9%**	**0.709**

Table 10
Performance comparison of the methods in this chapter on the extended MUStARD dataset under Speaker-Dependent condition.

Modality	Method			Performance measures					
	Intermodal	Attention	Contextual	*Sensitivity*	*Specificity*	*Accuracy*	*Precision*	*Recall*	F_1-score
TAV	✓	✓	✓	74.1%	74.77%	74.44%	74.6%	74.1%	0.743
TAV	✓	×	✓	74.7%	75.50%	75.10%	75.3%	**74.7%**	0.750
TAV	✓	✓	×	73.2%	76.37%	74.79%	75.6%	73.2%	0.744
TAV	✓	×	×	72.9%	75.96%	74.43%	75.2%	72.9%	0.740
TAV	✓	✓	✓	73.4%	76.44%	74.92%	75.7%	73.4%	0.745
TAV	✓	✓	✓	74.2%	76.82%	75.51%	**76.2%**	74.2%	**0.752**

suggests that our method can effectively discern irony by analyzing the emotional contrasts in context and utterance, irrespective of speaker-specific traits.

5.4. Ablation study

Our ablation studies, focusing on different modality combinations, shed light on the individual and collective contributions of each modality in our sarcasm detection method. Given that our approach hinges on modality interaction, we employed combinations of at least two modalities. We excluded the auditory–visual combination due to its lack of key contextual and speaker-specific information.

Further, we dissected the impact of each component on model performance, using the feature extraction module as a baseline for multimodal feature acquisition. Additional ablation experiments were performed on other components, with results detailed in Tables 10 and 11. Notably, the combination of the attention mechanism with contextual emotion differences yielded the most effective results in speaker-dependent settings. Incorporating the attention mechanism, modality-emotion inconsistency mechanism, and contextual emotion difference mechanism enhanced precision and recall by 0.4 and 0.9 percentage points, respectively, compared to using only two methods.

The attention mechanism, focusing on modality differences, assigns adaptive weights to multimodal features, resulting in a higher recall rate. The methods leveraging emotion inconsistency and contextual emotion differences capitalize on the distinct characteristics of sarcasm for more precise identification. Combining all three methods achieves the highest F_1-score, balancing precision and recall effectively.

5.5. Impact of hyperparameter

We investigated the impact of the hyperparameter M on our model's performance. In our approach, contextual features are segmented into M parts and then integrated with utterance features, making M a crucial factor in determining performance. This study was carried out using the Speaker-Dependent setting of the extended MUStARD dataset, with a focus on the F_1-score as the primary performance metric.

Table 12 presents the results of varying M. The data clearly indicates that the model's optimal performance is achieved when M is set to 10. This finding underscores the significance of selecting an appropriate value for M to maximize the efficiency of feature fusion in sarcasm detection".

This revision succinctly explains the role of the hyperparameter M in the model's performance, emphasizing its impact on feature segmentation and fusion. It highlights the key findings of the experiments in a clear and concise manner, focusing on the optimal setting for M.

6. Conclusion and future work

In the field of multimodal irony detection, we have identified and addressed critical challenges such as gaps in modal information and the overlooking of key segments. Our approach has yielded significant advancements:

- Our inter-modal emotion inconsistency mechanism effectively captures the nuanced clashes in ironic content, providing a detailed analysis by fusing modalities and modeling inconsistencies.

Table 11
Performance comparison of the methods in this chapter on the extended MUStARD dataset under Speaker-Independent condition.

Modality	Method			Speaker-Independent					
	Intermodal	Attention	Contextual	*Sensitivity*	*Specificity*	*Accuracy*	*Precision*	*Recall*	F_1-score
TAV	✓	✓	✓	69.7%	70.55%	70.13%	70.3%	69.7%	0.700
TAV	✓	×	✓	68.3%	70.45%	69.37%	69.8%	68.3%	0.690
TAV	✓	✓	×	70.1%	71.64%	70.87%	71.2%	70.1%	0.706
TAV	✓	×	×	69.7%	71.53%	70.62%	71.0%	69.7%	0.703
TAV	✓	✓	✓	68.9%	71.45%	70.17%	70.7%	68.9%	0.698
TAV	✓	✓	✓	69.9%	72.68%	71.29%	**71.9%**	**69.9%**	**0.709**

Table 12
Performance comparison of the methods with different hyperparameter M.

Parameter	$M = 4$	$M = 6$	$M = 8$	$M = 10$	$M = 12$	$M = 14$
F_1-score	0.749	0.751	0.750	0.752	0.747	0.748

This results in our network's superior performance in detecting irony compared to classical models.

- The introduction of inter-modal and segment-level attention mechanisms in our model enhances its ability to discern and prioritize crucial information across different modalities and within utterances. This leads to a more focused analysis, further improving performance.
- By incorporating context-utterance emotion differences, our model effectively recognizes and utilizes the subtleties of irony, making it more adept at detecting sarcasm.

However, our model has its limitations. Firstly, it may not perform efficiently in real-time computing scenarios, posing a challenge for applications requiring immediate analysis. Secondly, our current approach treats sarcasm as binary data, which might not fully capture its complex nature. Finally, the model's dependency on large quantities of training data could limit its adaptability and scalability.

Looking ahead, our future work will focus on two main areas. Firstly, we plan to explore sarcasm detection as a multi-label classification problem, considering the presence of other emotions like anger or grief that often accompany sarcasm in social media texts. Our goal is to develop a model capable of effective multi-class emotion recognition. Secondly, we aim to create a faster and more efficient sarcasm detection system suitable for real-time applications, such as sentiment analysis and hazardous speech detection. This will involve designing a lightweight network and employing techniques like pruning and compression to reduce model complexity while maintaining high accuracy. These developments will significantly enhance our model's practicality and effectiveness in various real-world scenarios.

CRediT authorship contribution statement

Yangyang Li: Writing – review & editing, Validation, Supervision, Methodology, Conceptualization. **Yuelin Li:** Writing – review & editing, Writing – original draft, Data curation. **Shihuai Zhang:** Writing – review & editing, Data curation. **Guangyuan Liu:** Writing – review & editing, Investigation. **Yanqiao Chen:** Writing – review & editing, Investigation. **Ronghua Shang:** Investigation. **Licheng Jiao:** Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work was supported in part by the Research Project of Song-Shan Laboratory, China under Grant YYJC052022004, in part by the National Natural Science Foundation of China under Grant 62101517 and 62176200, in part by the Natural Science Basic Research Program of Shaanxi, China under Grant No. 2022JC-45, in part by the Shaanxi Province Innovation Capability Support Plan, China under Grant 2023-CX-TD-09, and in part by the Fund for Foreign Scholars in University Research and Teaching Programs, China (the 111 Project).

References

[1] R. Prabowo, M. Thelwall, Sentiment analysis: A combined approach, J. Informetr. 3 (2) (2009) 143–157.
[2] B. Pang, L. Lee, et al., Opinion mining and sentiment analysis, Found. Trends Inf. Retr. 2 (1–2) (2008) 1–135.
[3] I. Chaturvedi, E. Cambria, R.E. Welsch, F. Herrera, Distinguishing between facts and opinions for sentiment analysis: Survey and challenges, Inf. Fusion 44 (2018) 65–77.
[4] P. Chaudhari, C. Chandankhede, Literature survey of sarcasm detection, in: 2017 International Conference on Wireless Communications, Signal Processing and Networking, WiSPNET, IEEE, 2017, pp. 2041–2046, http://dx.doi.org/10.1109/wispnet.2017.8300120.
[5] R. Justo, T. Corcoran, S.M. Lukin, M. Walker, M.I. Torres, Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web, Knowl.-Based Syst. 69 (2014) 124–133, http://dx.doi.org/10.1016/j.knosys.2014.05.021, URL https://www.sciencedirect.com/science/article/pii/S0950705114002226.
[6] M. del Pilar Salas-Zárate, M.A. Paredes-Valverde, M.Á. Rodriguez-García, R. Valencia-García, G. Alor-Hernández, Automatic detection of satire in Twitter: A psycholinguistic-based approach, Knowl.-Based Syst. 128 (2017) 20–33, http://dx.doi.org/10.1016/j.knosys.2017.04.009, URL https://www.sciencedirect.com/science/article/pii/S0950705117301855.
[7] S. Chen, Y. Leng, S. Labi, A deep learning algorithm for simulating autonomous driving considering prior knowledge and temporal information, Comput.-Aided Civ. Infrastruct. Eng. 35 (4) (2020) 305–321.
[8] S. Shehnepoor, R. Togneri, W. Liu, M. Bennamoun, HIN-RNN: A graph representation learning neural network for fraudster group detection with no handcrafted features, IEEE Trans. Neural Netw. Learn. Syst. (2021).
[9] I. Lauriola, A. Lavelli, F. Aiolli, An introduction to deep learning in natural language processing: Models, techniques, and tools, Neurocomputing 470 (2022) 443–456.
[10] S.M. Naguib, H.M. Hamza, K.M. Hosny, M.K. Saleh, M.A. Kassem, Classification of cervical spine fracture and dislocation using refined pre-trained deep model and saliency map, Diagnostics 13 (7) (2023) 1273.
[11] M.A. Kassem, S.M. Naguib, H.M. Hamza, M.M. Fouda, M.K. Saleh, K.M. Hosny, et al., Explainable transfer learning-based deep learning model for pelvis fracture detection, Int. J. Intell. Syst. 2023 (2023).
[12] Y.S. Alsahafi, M.A. Kassem, K.M. Hosny, Skin-net: a novel deep residual network for skin lesions classification using multilevel feature extraction and cross-channel correlation with detection of outlier, J. Big Data 10 (1) (2023) 105.
[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 30 (2017) URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
[14] R. Ortega-Bueno, P. Rosso, J.E. Medina Pagola, Multi-view informed attention-based model for Irony and Satire detection in Spanish variants, Knowl.-Based Syst. 235 (2022) 107597, http://dx.doi.org/10.1016/j.knosys.2021.107597, URL https://www.sciencedirect.com/science/article/pii/S0950705121008595.
[15] R.A. Potamias, G. Siolas, A.-G. Stafylopatis, A transformer-based approach to irony and sarcasm detection, Neural Comput. Appl. 32 (2020) 17309–17320, http://dx.doi.org/10.1007/s00521-020-05102-3.

[16] W. Aljedaani, F. Rustam, M.W. Mkaouer, A. Ghallab, V. Rupapara, P.B. Washington, E. Lee, I. Ashraf, Sentiment analysis on Twitter data integrating TextBlob and deep learning models: The case of US airline industry, Knowl.-Based Syst. 255 (2022) 109780, http://dx.doi.org/10.1016/j.knosys.2022.109780, URL https://www.sciencedirect.com/science/article/pii/S0950705122009017.

[17] I. Touahri, A. Mazroui, Enhancement of a multi-dialectal sentiment analysis system by the detection of the implied sarcastic features, Knowl.-Based Syst. 227 (2021) 107232, http://dx.doi.org/10.1016/j.knosys.2021.107232, URL https://www.sciencedirect.com/science/article/pii/S0950705121004949.

[18] Y. Liu, Z. Li, K. Zhou, L. Zhang, L. Li, P. Tian, S. Shen, Scanning, attention, and reasoning multimodal content for sentiment analysis, Knowl.-Based Syst. 268 (2023) 110467, http://dx.doi.org/10.1016/j.knosys.2023.110467, URL https://www.sciencedirect.com/science/article/pii/S0950705123002174.

[19] C. Huang, J. Zhang, X. Wu, Y. Wang, M. Li, X. Huang, TeFNA: Text-centered fusion network with crossmodal attention for multimodal sentiment analysis, Knowl.-Based Syst. 269 (2023) 110502, http://dx.doi.org/10.1016/j.knosys.2023.110502, URL https://www.sciencedirect.com/science/article/pii/S0950705123002526.

[20] D.S. Chauhan, G.V. Singh, A. Arora, A. Ekbal, P. Bhattacharyya, An emoji-aware multitask framework for multimodal sarcasm detection, Knowl.-Based Syst. 257 (2022) 109924, http://dx.doi.org/10.1016/j.knosys.2022.109924, URL https://www.sciencedirect.com/science/article/pii/S0950705122010176.

[21] A. Reyes, P. Rosso, T. Veale, A multidimensional approach for detecting irony in twitter, Lang. Res. Eval. 47 (2013) 239–268, http://dx.doi.org/10.1007/s10579-012-9196-x.

[22] J. Sun, J. Jiang, Y. Liu, An introductory survey on attention mechanisms in computer vision problems, in: 2020 6th International Conference on Big Data and Information Analytics, BigDIA, 2020, pp. 295–300, http://dx.doi.org/10.1109/BigDIA51454.2020.00054.

[23] G. Liu, J. Guo, Bidirectional LSTM with attention mechanism and convolutional layer for text classification, Neurocomputing 337 (2019) 325–338, http://dx.doi.org/10.1016/j.neucom.2019.01.078, URL https://www.sciencedirect.com/science/article/pii/S0925231219301067.

[24] G. Brauwers, F. Frasincar, A general survey on attention mechanisms in deep learning, IEEE Trans. Knowl. Data Eng. 35 (4) (2023) 3279–3298, http://dx.doi.org/10.1109/tkde.2021.3126456.

[25] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, S. Poria, Towards multimodal sarcasm detection (an _Obviously_Perfect paper), in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4619–4629, http://dx.doi.org/10.18653/v1/P19-1455, URL https://aclanthology.org/P19-1455.

[26] D.S. Chauhan, D. S R, A. Ekbal, P. Bhattacharyya, Sentiment and emotion help sarcasm? A multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4351–4360, http://dx.doi.org/10.18653/v1/2020.acl-main.401, URL https://aclanthology.org/2020.acl-main.401.

[27] A. Baruah, K. Das, F. Barbhuiya, K. Dey, Context-aware sarcasm detection using BERT, in: B.B. Klebanov, E. Shutova, P. Lichtenstein, S. Muresan, C. Wee, A. Feldman, D. Ghosh (Eds.), Proceedings of the Second Workshop on Figurative Language Processing, Association for Computational Linguistics, Online, 2020, pp. 83–87, http://dx.doi.org/10.18653/v1/2020.figlang-1.12, URL https://aclanthology.org/2020.figlang-1.12.

[28] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/N19-1423, URL https://aclanthology.org/N19-1423.

[29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778, http://dx.doi.org/10.1109/cvpr.2016.90.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255, http://dx.doi.org/10.1109/cvpr.2009.5206848.

[31] Q. Zhang, J. Xu, L. Xu, H. Guo, Deep convolutional neural networks for forest fire detection, in: 2016 International Forum on Management, Education and Information Technology Application, Atlantis Press, 2016, pp. 568–575, http://dx.doi.org/10.2991/ifmeita-16.2016.105.

[32] M. Rohanian, J. Hough, M. Purver, Detecting depression with word-level multimodal fusion, in: Proc. Interspeech 2019, 2019, pp. 1443–1447, http://dx.doi.org/10.21437/Interspeech.2019-2283.

[33] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: The munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM International Conference on Multimedia, MM '10, Association for Computing Machinery, New York, NY, USA, 2010, pp. 1459–1462, http://dx.doi.org/10.1145/1873951.1874246.

[34] D. Kumar, P.K.V. Patil, A. Agarwal, S.M. Prasanna, Fake speech detection using opensmile features, in: International Conference on Speech and Computer, Springer, 2022, pp. 404–415, http://dx.doi.org/10.1007/978-3-031-20980-2_35.

[35] E. Cambria, D. Olsher, D. Rajagopal, SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 28, No. 1, 2014, http://dx.doi.org/10.1609/aaai.v28i1.8928.

[36] Sarcasm detection with self-matching networks and low-rank bilinear pooling, in: The World Wide Web Conference, WWW '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 2115–2124, http://dx.doi.org/10.1145/3308558.3313735.

[37] Y. Tay, A.T. Luu, S.C. Hui, J. Su, Reasoning with sarcasm by reading in-between, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 1010–1020, http://dx.doi.org/10.18653/v1/P18-1093, URL https://aclanthology.org/P18-1093.

[38] S. Gupta, A. Shah, M. Shah, L. Syiemlieh, C. Maurya, FiLMing multimodal sarcasm detection with attention, in: T. Mantoro, M. Lee, M.A. Ayu, K.W. Wong, A.N. Hidayanto (Eds.), Neural Information Processing, Springer International Publishing, Cham, 2021, pp. 178–186, http://dx.doi.org/10.1007/978-3-030-92307-5_21.

[39] X. Zhang, Y. Chen, G. Li, Multi-modal sarcasm detection based on contrastive attention mechanism, in: L. Wang, Y. Feng, Y. Hong, R. He (Eds.), Natural Language Processing and Chinese Computing, Springer International Publishing, Cham, 2021, pp. 822–833, http://dx.doi.org/10.1007/978-3-030-88480-2_66.

[40] M. Firdaus, H. Chauhan, A. Ekbal, P. Bhattacharyya, MEISD: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations, in: D. Scott, N. Bel, C. Zong (Eds.), Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 4441–4453, http://dx.doi.org/10.18653/v1/2020.coling-main.393, URL https://aclanthology.org/2020.coling-main.393.

[41] A. Ray, S. Mishra, A. Nunna, P. Bhattacharyya, A multimodal corpus for emotion recognition in sarcasm, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Lang. Res. Eval. Conference, European Language Resources Association, Marseille, France, 2022, pp. 6992–7003, URL https://aclanthology.org/2022.lrec-1.756.