Breakthrough from Nuance and Inconsistency: Enhancing Multimodal Sarcasm Detection with Context-Aware Self-Attention Fusion and Word Weight Calculation

Hongfei Xue, Linyan Xu, Yu Tong, Rui Li, Jiali Lin and Dazhi Jiang*
Department of Computer Science, Shantou University, China
{21hfxue, linyanxu, tongyu, ruili, jllin, dzjiang}@stu.edu.cn

Abstract

Multimodal sarcasm detection has received considerable attention due to its unique role in social networks. Existing methods often rely on feature concatenation to fuse different modalities or model the inconsistencies among modalities. However, sarcasm is often embodied in local and momentary nuances in a subtle way, which causes difficulty for sarcasm detection. To effectively incorporate these nuances, this paper presents Context-Aware Self-Attention Fusion (CAAF) to integrate local and momentary multimodal information into specific words. Furthermore, due to the instantaneous nature of sarcasm, the connotative meanings of words post-multimodal integration generally deviate from their denotative meanings. Therefore, Word Weight Calculation (WWC) is presented to compute the weight of specific words based on CAAF's fusion nuances, illustrating the inconsistency between connotation and denotation. We evaluate our method on the MUStARD dataset, achieving an accuracy of 76.9 and an F1 score of 76.1, which surpasses the current state-of-the-art IWAN model by 1.7 and 1.6 respectively.

Keywords: Multimodal Sarcasm Detection, Context-Aware Self-Attention Fusion, Word Weight Calculation

1. Introduction

As a complex linguistic phenomenon, sarcasm often conceals hostility while enhancing the effects of sarcasm or humor (Tay et al., 2018), and typically conveys a negative meaning contrary to its literal interpretation in practice(Ding et al., 2022).

Emotion recognition tasks have been done quite well in recent years (Wen et al., 2023; Yu et al., 2024; Jiang et al., 2024; Tu et al., 2022; Jiang et al., 2022), but when sarcasm is featured inside a dialogue, the accuracy of the ERC model is greatly discounted. So sarcasm detection has received increasing attention. Early methods such as rule-based (Riloff et al., 2013; Van Hee, 2017) required predefined rules or manual feature extraction. More recent methods have focused on extracting multimodal features (Castro et al., 2019) and leveraging attention mechanisms for cross-modal relationships (Chauhan et al., 2020).

However, existing multimodal sarcasm detection methods often resort to simple concatenation of features or modelling of in-consistencies between modalities while overlooking critical nuances, such as textual word information, audio tone variations, facial expressions, and body postures in images, etc., which fall short of exploiting the full potential of multimodal information. In the meanwhile, the connotative meaning (the implied or suggested meaning) of these details often differs from the literal denotative meaning of text words, and this incongruity is crucial for multimodal sarcasm detection.

Our motivation is illustrated in Figure 1. For instance, when Sheldon utters "privilege" (text information) with a flat expression in the video frame (visual information), a small waveform and clinical tone in the audio frame (audio information), it becomes evident that the term "privilege" carries a negative connotation in Sheldon's discourse. This incongruity between connotation and denotation signifies sarcasm, serving as a vital cue for multimodal sarcasm detection.

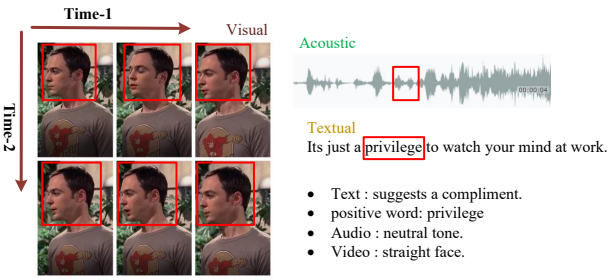

Figure 1: Sample from the MUStARD dataset. Sheldon's comment (Text) with a flat face (Visual) and a neutral tone (Acoustic) makes the sample sarcastic.

Based on the recent research (Yang et al., 2019; Kumar et al., 2022), a Context-Aware Self-Attention Fusion (CAAF) module is presented to intricately integrate multimodal information into the textual context to enhance its effectiveness. Additionally, a Word Weight Calculation (WWC) module is designed to assign weights to words based on the degree of multimodal incongruity, allowing our model to focus on words with incongruous information (im-

*Corresponding author

LREC-COLING 2024, pages 2493–2503
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

plying sarcasm). Experiments on the MUStARD multimodal sarcasm dataset demonstrate the efficacy of our methods, achieving an F1 score of 76.1, surpassing the state-of-the-art methods by 1.6 percentage points. Our primary contributions are as follows.

1. We present a pioneering context-aware self-attention mechanism for fine-grained word-level fusion of audio-visual cues.

2. By analyzing the connotative meaning of words, we effectively model inconsistencies between modalities, improving detection accuracy and interpretability of outcomes.

3. We conduct a comprehensive comparative analysis with state-of-the-art methods, exploring the impact of different modal inputs on the performance of our method.

2. Related Work

2.1. Sarcasm and Text

Sarcasm detection initially focused on textual content. There exist rule-based methods, such as those employed by (Riloff et al., 2013; Maynard and Greenwood, 2014). (Van Hee, 2017) utilized statistical machine learning techniques, necessitating the manual extraction of special symbols and syntax. In recent years, deep neural networks (DNNs) have gained prominence for text-based sarcasm detection. (Zhang et al., 2016; Poria et al., 2016) utilized various pretrained models to extract features, including emotions and personality traits. (Babanejad et al., 2020) was pretrained to extract emotionally enriched expressions to aid sarcasm detection. Notably, literature in psychology and linguistics highlights the significance of paralinguistic cues in understanding sarcasm and humor (Attardo et al., 2003; Tabacaru and Lemmens, 2014). This underscores the insufficiency of relying solely on textual data for sarcasm detection.

2.2. Sarcasm and Multimodality

Given the challenges of single-mode sarcasm detection, researchers have explored multimodal methods. Researchers have acknowledged that contextual sarcasm in text can manifest in other modalities, providing additional clues in common or comparative forms, and have ventured into multimodal sarcasm detection. In conversational settings, (Castro et al., 2019) provided the MUStARD dataset. (Chauhan et al., 2020) devised a multi-task and multimodal sarcasm detection framework, leveraging the intrinsic correlation between emotions and sarcasm. In bimodal scenarios, (Cai et al., 2019; Xu et al., 2020; Pan et al., 2020) utilized tweets containing images to identify sarcasm.

However, these multimodal sarcasm detection methods often underutilize the rich multimodal information, particularly local and momentary nuances. These multimodal nuances, including tone, voice modulation, facial expressions, and actions, often generate connotations that are inconsistent with the denotative meanings of text words. The incongruity between these connotations and the literal meanings of text words constitutes a pivotal factor in multimodal sarcasm detection. Hence, addressing this issue becomes paramount in our research.

In summary, recent advancements in sarcasm detection encompass both text-based and multimodal methods. However, the effective utilization of multimodal nuances for improved sarcasm detection remains an ongoing challenge and is the focal point of our proposed method.

3. Methodology

3.1. Overview

Our methodology begins with multimodal feature extraction, followed by the introduction of two critical components: Context-Aware Self-Attention Fusion (CAAF) and Word Weight Calculation (WWC). Figure 2 illustrates the overall structure of our model. This method optimizes the integration of multimodal nuances and textual context, addressing the challenge of inconsistent information across modalities. It combines advanced techniques in multimodal fusion, context-aware attention mechanisms, and word-level weighting to comprehensively address multimodal sarcasm detection.

3.2. Multimodal Feature Extraction

Visual and textual features are extracted according to (Castro et al., 2019). Utterance-level acoustic features can be obtained through OpenSmile (Eyben et al., 2010). Ultimately, the utterance-level features of vision, text, and acoustics are expressed as $m_u \in R^{d_m}$, $m \in \{a, v, t\}$. However, it should be noted that these features are incapable of offering nuanced multimodal details, such as variations in tone and subtle facial expressions. As a remedial measure, word-level multimodal features are employed to provide localized information.

In the given textual context, the initial step is to use the GENTLE [1] system to facilitate the alignment process, thereby enabling the audio segments to be aligned with the uttered words $\{w_1, w_2, \ldots, w_n\}$. Subsequently, we obtain video frames and audio clips that correspond to specific words. Furthermore, in consideration of the vital significance of

[1] https://github.com/lowerquality/gentle.git

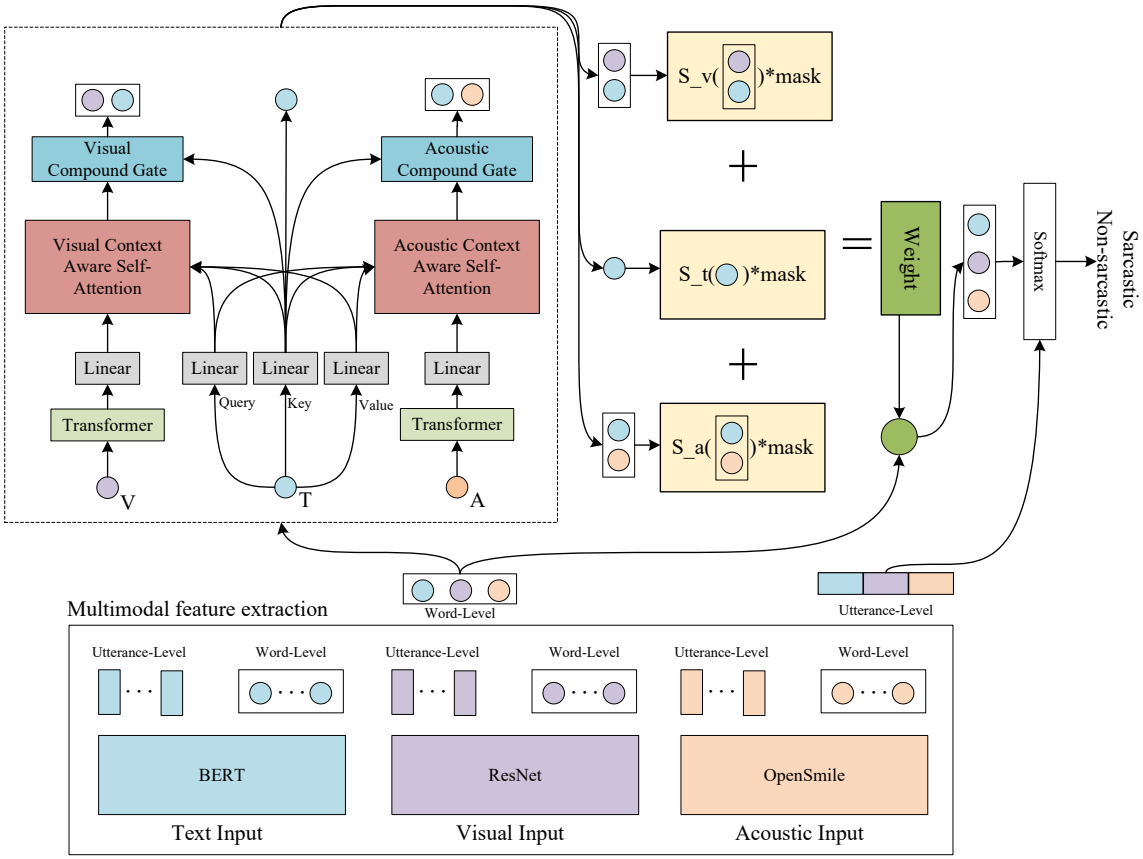

Figure 2: The overall structure of our model, in which the CAAF module is in the dashed box, the WWC module is on the right of the dashed box, and the feature extraction part is below the dashed box.

visual information pertaining to the speaker, the MTCNN (Zhang et al., 2019) (Multi-task Cascaded Convolutional Networks) is employed to extract the facial features. The results are subjected to clipping, followed by the utilization of the Perception ResnetV1 (kaiming he et al., 2016) model to designate the facial image of the speaker. Ultimately, the Resnet50 (kaiming he et al., 2016) model is employed to extract feature from the facial images. These features are then averaged, and the resultant values are employed to represent the visual features $\{v_{w_1}, v_{w_2}, \ldots, v_{w_n}\}$, where $v_{w_i} \in R^{d_v}$. For the text data, BERT (Devlin et al., 2018) is employed for processing. The results $\{t_{w_1}, t_{w_2}, \ldots, t_{w_n}\}$ are used as the representation of each word in the utterance, where $t_{w_i} \in R^{d_t}$. For the extraction of audio features, OpenSmile has been selected to derive Mel Frequency Cepstrum Coefficients (MFCCs), as well as other fundamental acoustic attributes, such as zero-crossing rate, from each audio clip. These are then averaged to produce $\{a_{w_1}, a_{w_2}, \ldots, a_{w_n}\}$. The averaged results are used as the representation of audio features, where $a_{w_i} \in R^{d_a}$.

3.3. Context-Aware Self-Attention Fusion

To achieve the fusion of multi-modal information, we propose a unique multimodal CAAF scheme. Using audio-visual information, the proposed attention method adjusts key vectors and value vectors, thus enhancing profound semantic interaction between multi-modal signals and text representations. These modified vectors are used to perform point product attention. The traditional cross channel attention scheme, using dot product, results in the direct interaction between text representation and other channels. Here, the text representation acts as a query to learn the multimodal representation, while the multimodal representation acts K and V. Because the embedded subspace corresponding to each mode is different, directly fusing multimodal information may lose some context information. Inspired by (Yang et al., 2019; Kumar et al., 2022), this study utilizes K and V vectors to generate multimodal information. Traditional scaling points are then applied to derive the attentions. The succeeding section will elaborate on this process in depth.

Information representation is acquired through the transformation of word-level features via a se-

ries of nonlinear layers. As shown in equation (1).

$$r_{w_i}^m = tanh(W_m m_{w_i} + b_m) \tag{1}$$

where d is the number of hidden units, $W_m \in R^{d \times d_m}, b_m \in R^d$ are trainable weights, $m \in \{a, v, t\}$.

The $Q \in R^{n \times d}, K \in R^{n \times d}$ and $V \in R^{n \times d}$ are calculated in accordance with the representation of textual feature information. As shown in equation (2), where $W_Q \in R^{d \times d}, W_K \in R^{d \times d}$ and $W_V \in R^{d \times d}$ are learnable parameters. $H = \{r_{w_1}^t, r_{w_2}^t, ..., r_{w_n}^t\}, n$ represents the maximum sequence length of text.

$$\begin{bmatrix} Q \\ K \\ V \end{bmatrix} = H \begin{bmatrix} W_Q \\ W_K \\ W_V \end{bmatrix} \tag{2}$$

Let C represent the vector obtained from the audio or visual representation, $C \in \{r^a, r^v\}, r^a = \{r_{w_1}^a, r_{w_2}^a, ..., r_{w_n}^a\}, r^v = \{r_{w_1}^v, r_{w_2}^v, ..., r_{w_n}^v\}, U_K$ and $U_V \in R^{n \times d}$ are learnable matrices. In order to effectively use this information and decide how much information to integrate from multimodal sources and how much information to retain from text modes, gated scalars are assigned to learn factors $\{\lambda_K, \lambda_V\} \in R^{n \times 1}$. Thus, the contribution of each representation and context vector to the prediction of attention weight are clearly quantified. On this basis, generate key and value vectors $\hat{K}$ and $\hat{V}$ for multimodal information representation according to (Yang et al., 2019). The related formulation is shown in equation (3).

$$\begin{bmatrix} \hat{K} \\ \hat{V} \end{bmatrix} = (1 - \begin{bmatrix} \lambda_K \\ \lambda_V \end{bmatrix}) \begin{bmatrix} K \\ V \end{bmatrix} + \begin{bmatrix} \lambda_K \\ \lambda_V \end{bmatrix} (C \begin{bmatrix} U_K \\ U_V \end{bmatrix}) \tag{3}$$

(Vaswani et al., 2017; Britz et al., 2017) pointed out that a large number of Q and K may push the SoftMax function to the region with minimal gradient. Similarly, inspired by previous research on multimodal networks (Xu et al., 2015; Calixto et al., 2017; Yang et al., 2017), a large number of our K and V may also cause the same problem. To address this issue, we regard $\{\lambda_K, \lambda_V\}$ as a factor regulating the size of $\hat{K}$ and $\hat{V}$. Thus, a gated scalar is assigned to learn the factor. The calculation process is shown in equation (4), where $W_{k_1}, W_{k_2}, W_{v_1}$ and $W_{v_2} \in R^{d \times 1}$ are the trainable parameter matrices, $\sigma(\cdot)$ represents the sigmoid function.

$$\begin{bmatrix} \lambda_K \\ \lambda_V \end{bmatrix} = \sigma(\begin{bmatrix} K \\ V \end{bmatrix} \begin{bmatrix} W_{k_1} \\ W_{v_1} \end{bmatrix} + C \begin{bmatrix} U_k \\ U_v \end{bmatrix} \begin{bmatrix} W_{k_2} \\ W_{v_2} \end{bmatrix}) \tag{4}$$

The fused multimodal information $\hat{K}$ and $\hat{V}$ are used to calculate the traditional scaling point product attention. Using the context-aware attention mechanism, the vector H_a and H_v are obtained by injecting audio information and visual information respectively.

$$H_a = Softmax(\frac{Q\hat{K}_a^T}{\sqrt{d_k}})\hat{V}_a \tag{5}$$

$$H_v = Softmax(\frac{Q\hat{K}_v^T}{\sqrt{d_k}})\hat{V}_v \tag{6}$$

where $\sqrt{d_k}$ is the scaling factor.

Because the information from audio and video modes needs to be combined, we control the amount of information transmitted by each channel through the auditory gate (g_a) and visual gate (g_v). As shown in equations (7) and (8).

$$g_a = [H \oplus H_a]W_m + b_m \tag{7}$$

$$g_v = [H \oplus H_v]W_n + b_n \tag{8}$$

where $\oplus$ indicates concatenation, $W_m, W_n \in R^{1 \times 2d}$; $b_m, b_n \in R^1$ are the trainable parameters.

3.4. Word Weight Calculation

The inconsistency between the denotative meaning and connotative meaning of text words is a strong sign of sarcasm. To identify these inconsistencies, our calculation weight module scrutinises the audio and video information to determine the weight of each positive word. It attempts to capture the irrelevant parts between two vectors. Specifically, inconsistencies arise when positive spoken language (text) is contrasted with negative facial expressions (vision) or negative tone (hearing). After these two cross modal comparison, any differences that arise during these cross modal comparisons may indicate an inconsistency between the denotative and connotative meanings of the current text words.

The following two equations are used to identify inconsistencies, with greater weight assigned to words that carry more inconsistent information. This means that the connotative meaning is generated and is opposite to the denotative meaning.

$$S_v = ReLu(W_m g_v + b_m) \tag{9}$$

$$S_a = ReLu(W_n g_a + b_n) \tag{10}$$

In the literature (Wu et al., 2021), sarcasm is defined as the conveying positive emotions only in negative situations, thus we only need to calculate the inconsistent weights of positive words, that is, the inconsistent weights between positive words (text) and negative facial expressions (visual) or negative tones (auditory). Therefore, it is necessary to acquire positive words, so SentiWords (Gatti et al., 2016) is used to identify positive words. The emotional weight of each word in the utterance is obtained from an emotional dictionary. Words are recognised as positive if they exceed a pre-set

threshold, ranging from 0 to 1. Equation (11) is used to calculate the importance of non-positive words.

$$S_w = WH + b \tag{11}$$

S_w is used to get the weight of each word and use S_v and S_a to calculate the weight of positive words. Equation (12) is used to get the final attention weight p_{w_i}. The value of $mask_{w_i}$ is 1 or 0, corresponding to the current word being a positive word or other word respectively.

$$p_{w_i} = Softmax(S_w + S_v \times mask_{w_i} + S_a \times mask_{w_i}) \tag{12}$$

3.5. Final Representation and Classification

As shown in equations (13) - (16), we obtain the overall representation of word level, utterance level and global information features through some weighted sum operations and cascade operations. Then the final representation r_c is obtained r_w, r_u, r_g through equation (17).

$$r_{w_i} = t_{w_i} \oplus v_{w_i} \oplus a_{w_i} \tag{13}$$

$$r_w = \sum_{i=1}^{n} p_{w_i} r_{w_i} \tag{14}$$

$$r_u = t_u \oplus v_u \oplus a_u \tag{15}$$

$$r_g = t_g \oplus v_g \oplus a_g \tag{16}$$

$$r_c = tanh(r_w \oplus r_u \oplus r_g) \tag{17}$$

r_c is used as the input of the classification function to predict the results, i.e. SoftMax. Cross entropy loss is used as loss function.

4. Experiments

4.1. Dataset

In experiments, MUStARD (Castro et al., 2019) is utilized as the primary dataset for sarcasm detection. The MUStARD is a diverse collection of dialogues sourced from popular TV shows, including Friends, The Big Bang Theory, The Golden Girls, and Ironic Humor Anonymous. This dataset comprises a total of 690 conversation samples, evenly split into 345 sarcasm-laden utterances and 345 non-sarcasm utterances. Each sample encompasses a tri-modal representation, encompassing visual, auditory, and textual elements.

4.2. Setup

We conduct sarcasm detection experiments in two distinct settings: speaker-dependent and speaker-independent.

Speaker-Independent Setting: In this setting, our model is trained on segments of the MUStARD dataset pertaining to The Big Bang Theory, The Golden Girls, and Ironic Humor Anonymous. We then evaluate the model's performance on the portion of the dataset corresponding to the Friends TV show. Speaker-independent settings present a particular challenge, as they require our model to generalize effectively without relying on specific speaker information.

Speaker-Dependent Setting: For the speaker-dependent setting, we adopt a five-fold cross-validation methods. The MUStARD dataset is partitioned into five folds. Each fold takes turns as the test set while the others are used for training. The final evaluation metric is the average performance across all five iterations. This setup allows to assess the model's performance while considering the influence of different speakers within the dataset.

To assess the performance of our sarcasm detection model, three standard evaluation metrics are obtained: precision (P), recall (R), and F1-score (F1). These metrics provide a comprehensive view of our model's effectiveness in correctly identifying sarcasm and non-sarcasm utterances.

4.3. Implementation Details

In both speaker dependent and speaker independent experimental settings, the same hyper-parameters are used. Adam is employed when choosing optimisers. For other parameters, the dimensions of d are set as as 200, d_t, d_v, d_a, d_{aw} are 768, 2048, 1000 and 65 respectively, dropout rate is 0.5, and learning rate is 1×10^{-4}. To identify positive words within our text data, we introduce a threshold value. Words with weights above this threshold are considered positive words. Empirically, a threshold of 0.5 may provide the best results for sarcastic language detection tasks. This threshold selection is a crucial aspect of our method, as it significantly impacts the model's ability to identify words associated with sarcasm. Due to the small size of the dataset, Support Vector Machine (SVM) classifier is used instead of SoftMax classifier during testing.

4.4. Selection of Threshold Value

The presented approach involves the use of an emotion dictionary to identify positive words by setting a threshold value on word weights. The choice of this threshold value significantly impacts

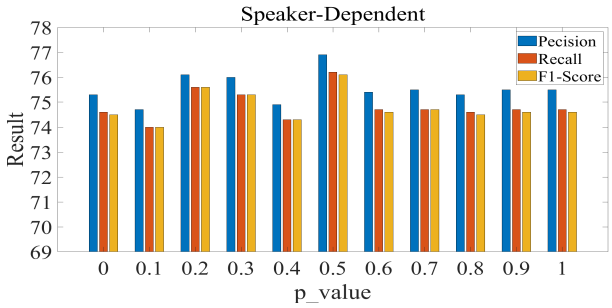

Figure 3: The experimental results of using different positive word thresholds in speaker dependent settings.

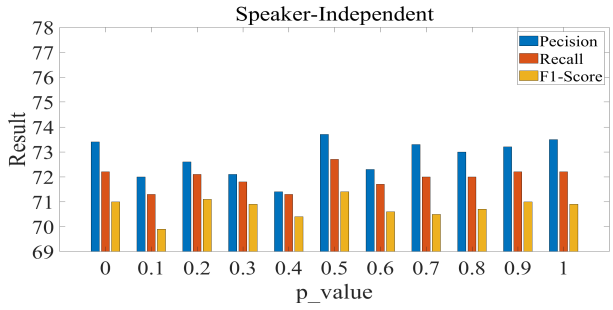

Figure 4: The experimental results of using different positive word thresholds in speaker independent settings.

our model's performance, as it dictates which words are classified as positive, while WWC module exclusively computes weights for these positive words. It is worth emphasizing that the weights generated by the WWC module are essential in the final sarcasm classification, underscoring the importance of threshold selection.

For threshold selection, the range from 0 to 1. Setting a threshold value too low may erroneously categorise words with low emotional intensity as positive, potentially introducing noise into experiments. Conversely, if the threshold is set too high, there may be an insufficient number of positive words per utterance, resulting in insufficient training data.

In this study, we systematically experiment with various threshold values to identify the most suitable one. The outcomes of this threshold exploration are visually depicted in Figure 3 and Figure 4. These visualizations offer a lucid comprehension of the influence varying threshold values have on our model's efficacy.

Notably, our findings consistently demonstrate that the model achieves optimal performance when assigning a threshold of 0.5 for identifying positive words, irrespective of the experimental settings. Hence, we adopt a threshold of 0.5 for the selection of positive words, as it strikes an ideal balance between capturing emotional cues and maintaining a sufficient quantity of training data for robust sarcasm detection.

4.5. Results and Analysis

4.5.1. Model Comparison

In terms of model comparison, our model is mainly compared with two types of comparison models: text models and multimodal models. Table 1 lists the performance of the comparison models and our model. These models are conducted on the MUStARD dataset, and are evaluated in both speaker dependent and speaker independent situations. The competing methods are listed as follows: **SMSD** (Xiong et al., 2019), **MIARN** (Tay et al., 2018), **BERT** (Devlin et al., 2018), **LSTM (A)** (Chen et al., 2017), **MFN** (Zadeh et al., 2018), **RAVEN** (Wang et al., 2019), **EF-Concat** (Castro et al., 2019), **IAIE** (Chauhan et al., 2020) and **IWAN** (Wu et al., 2021).

The comparison results are shown in Table 1. From the overall perspective, the performance of sarcasm detection based on text models is worse than that based on multi-modal models. The reason is that audio mode and video mode provide rich supplementary information for text mode, which is critical for sarcasm detection. As a result, multimodal sarcasm detection produces better results than single text mode.

This is due to the speakers being independent in both the training and test sets of the former, resulting in difficulties in detecting sarcasm and ultimately yielding poorer results than in the latter. However, it is worth noting from Table 1 that our model can still achieve the best results in the former experimental setup. The F1 score surpasses that of the current leading model (IWAN) by 1.4%, and that of the second-best model (IAIE) by 5.8%. The results show that our model can capture more critical sarcasm information during the process of multimodal sarcasm detection. Moreover, it is evident that there are marked differences between the text based SMSD model and the multimodal RAVEN model in two experimental settings. The potential explanation is that they pay too much attention to the specific scenario of the speaker in the process of sarcasm detection, which does not happen in our model.

Finally, the Table 1 clearly indicates that our model is much better than MFN, which captures the interactions between different modalities and feature segments. This is due to our model can provide more weight to words with modal conflict, enabling maximum utilization of multimodal information. In addition, our model demonstrated a 1.6% and 1.4% performance increase compared to IWAN in the two experimental settings, respectively, which emphasizes the effectiveness of our model.

Input	Method	S-D			S-I		
		P	R	F1	P	R	F1
T	SMSD	61.6	61.0	61.1	51.7	48.2	47.0
	MIARN	64.7	64.0	63.9	60.4	55.2	54.0
	BERT	67.5	66.9	66.8	58.2	56.7	57.0
T,A,V	RAVEN	69.1	67.5	67.1	53.8	50.4	49.7
	LSTM(A)	67.3	66.7	66.3	56.7	54.1	54.0
	MFN	70.1	69.6	69.7	66.0	62.5	62.4
	EF-Concat	71.2	70.8	70.8	64.3	63.1	63.3
	IAIE	72.1	71.6	72.0	66.0	65.5	65.6
	IWAN	75.2	74.6	74.5	71.9	71.3	70.0
T,A,V	**Ours(C&W)**	**76.9**	**76.2**	**76.1**	**73.7**	**72.7**	**71.4**

Table 1: The experimental results of nine comparison models and our proposed model. S-D means Speaker-Dependent; S-I means Speaker-Dependent.

4.5.2. Modality Comparison

In order to evaluate the effectiveness of each modality, we take different combinations of Text (T), Acoustic (A) and Visual (V) modes as the input for our model. However, all combinations include text modes, as our study centres on multimodal sarcasm detection, which necessitates at least two modal inputs. The experimental results of modal comparisons are shown in Table 2.

It is recorded in the literature of psychology and linguistics that there are obvious paralinguistic clues, which aid in the comprehension of sarcasm and humor (Attardo et al., 2003; Tabacaru and Lemmens, 2014). It is also evident from Table 2 that the performance of the model decreases in the absence of acoustic or visual modal. Removing a certain mode may result in the loss of some visual or auditory information, which directly affects the work of CAAF(*CAAF module can gradually integrates auditory and visual modes to help the model understand the connotative meaning of text words*). Consequently, our WWC module would not operate correctly, resulting in our model unable to capture the inconsistency between different modalities. This, in turn, hampers the ability of our model to accurately comprehend the connotations associated with text words, ultimately leading to a significant decline in its performance.

4.5.3. Ablation analysis

The ablation experiments on the MUStARD dataset were conducted to verify the validity of the CAAF and WWC modules. **C&W** is our complete model. **Ours w/o CAAF** is C&W remove C (i.e. CAAF) and retain W (i.e. WWC). **Ours w/o WWC** is C&W remove W (i.e. WWC) and retain C (i.e. CAAF). **Ours w/o C&W** is C&W remove C (i.e. CAAF) and W (i.e. WWC). Each ablation experiment was divided into speaker-dependent setting and speaker-independent setting.

Firstly, Table 3 indicates that our model's effectiveness is inferior to that of the complete model following the removal of CAAF or WWC. Secondly, after removing the CAAF module, the F1 score decreased by 0.5% and 2.5% respectively in the two experimental settings, which indicates that the latter experimental settings are more dependent on the CAAF module than the former. This phenomenon occurs because the former experimental arrangement requires less information integration compared to the latter. In addition, upon removal of the WWC module downstream of the task, the F1 scores in the two experimental settings decreased significantly by 3.3% and 3.8% respectively. This proves the necessity of the weight we calculated for positive words. Because this weight represents the degree of inconsistency between the denotative meaning and the connotative meaning of text words. It also conforms to our view above: the denotative meaning of text words, audio clips and image key frames, as well as the text connotative meaning after multimodal integration, are most likely to be the source of sarcasm. Lastly, our model's performance notably worsened upon removal of both the CAAF and WWC modules.

4.5.4. Case Study

To gain an intuitive understanding our proposed model, we visualized the attention weight of sarcasm utterance in Figure 6. The specific data can be found in Figure 5. Intuitively, the positive word "sweet" in the text is accompanied by a disappointed voice and a frown expression, resulting in a conflict between modes. It is reasonable that this time point is the best time to carry out the sarcasm detection task. It is evident from Figure 5 that our model also accurately assigns a weight of 0.39 to the word "sweet," which receives the high-

Method	Input	S-D			S-I		
		P	R	F1	P	R	F1
Ours	T,V	72.6	71.8	71.8	67.9	68.2	67.9
	T,A	72.2	71.8	71.8	66.7	67.0	66.3
	T,A,V	**76.9**	**76.2**	**76.1**	**73.7**	**72.7**	**71.4**

Table 2: Different modal inputs and their corresponding experimental results.

Method	S-D			S-I		
	P	R	F1	P	R	F1
C&W	**76.9**	**76.2**	**76.1**	**73.7**	**72.7**	**71.4**
Ours w/o CAAF	74.4	73.7	73.6	72.9	72.1	70.9
Ours w/o WWC	73.6	72.8	72.8	71.7	69.9	67.6
Ours w/o C&W	73.0	72.4	72.4	70.6	69.6	67.8

Table 3: According to the test data divided in MUStARD dataset, we performed ablation analysis on our model.

est weight among all the words in this sentence. This is consistent with the intuition, indicating that our model can correctly complete this multi-mode sarcasm detection task. In addition, four cases in Figure 5 illustrate that the simple text mode can not detect the word "sweet" smoothly. Our model can focus on the word "sweet" smoothly with the gradual inclusion of acoustic and visual information, while ignoring other words. After all acoustic and visual information are added, our model can accurately pay attention to the word "sweet" while simultaneously paying attention to other words. This indicates that integrating multimodal information into text information facilitates our model to focus on the inconsistent information in details between modes.

5. Conclusion

In this paper, we propose a model C&W (CAAF&WWC) to conduct multimodal sarcasm detection, consisting of two sub modules, CAAF and WWC. In detail, CAAF carefully integrates nuanced multi-modal information into the text mode to facilitate WWC to calculate the weight of positive words according to the degree of inconsistency between the denotative meaning and connotative meaning of positive words. The more modal inconsistencies present in positive language, the greater the inconsistency between its denotative and connotative meanings. Then the WWC assigns it a higher weight to accurately detect sarcasm. Regarding the choice of positive words, we quantitatively analyzed the impact of the threshold selection of positive words on the C&W. Our experiments indicate that C&W performs best at a threshold of 0.5. We compared C&W with the other nine baseline models on the MUStARD dataset, where the experimental results demonstrate the superiority of C&W. We believe that a future study could include the speaker's personality characteristics, so that the sarcasm detection task could be completed according to the contrast between the speaker's personality characteristics and the speaker's various performances in the current situation.

6. Acknowledgements

The authors would like to respect and thank all reviewers for their constructive and helpful review. This research is funded by the National Natural Science Foundation of China (62372283, 62206163), Science and Technology Major Project of Guangdong Province(STKJ2021005, STKJ202209002, STKJ2023076), Natural Science Foundation of Guangdong Province (2024A1515010239).

7. Bibliographical References

Salvatore Attardo, Jodi Eisterhold, Jennifer Hay, and Isabella Poggi. 2003. Multimodal markers of irony and sarcasm.

Nastaran Babanejad, Heidar Davoudi, Aijun An, and Manos Papagelis. 2020. Affective and contextual embedding for sarcasm detection. In *Proceedings of the 28th international conference on computational linguistics*, pages 225–243.

Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc Le. 2017. Massive exploration of neural machine translation architectures. *arXiv preprint arXiv:1703.03906*.

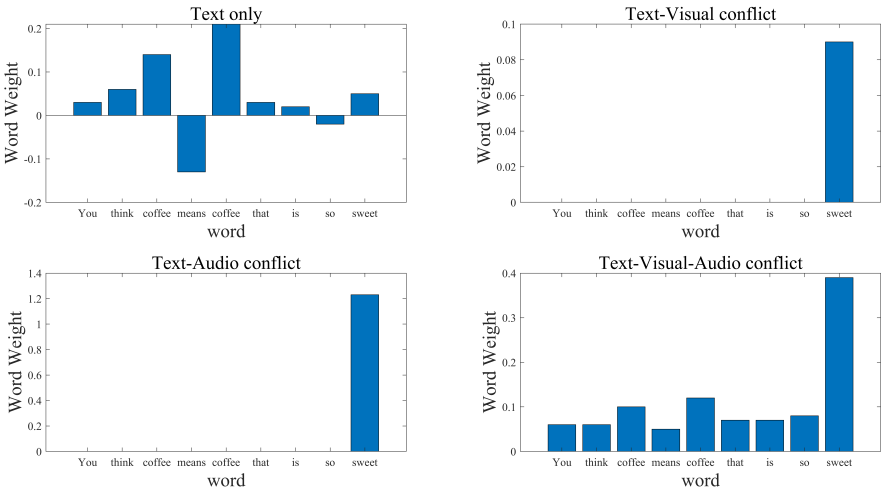

Figure 5: Visualization of weight calculated by WWC module of our model.

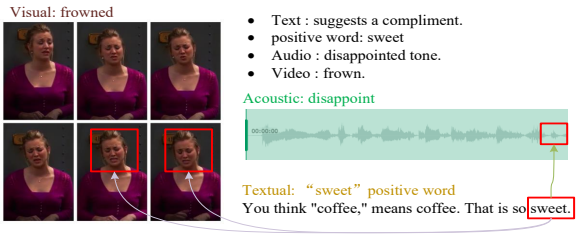

Figure 6: Sample from the MUStARD dataset.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. *arXiv preprint arXiv:1702.01287*.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv preprint arXiv:1906.01815*.

Dushyant Singh Chauhan, SR Dhanush, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4351–4360.

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM international conference on multimodal interaction*, pages 163–171.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ning Ding, Sheng-wei Tian, and Long Yu. 2022. A multimodal fusion method for sarcasm detection based on late fusion. *Multimedia Tools and Applications*, 81(6):8597–8616.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1459–1462, New York, NY, USA. Association for Computing Machinery.

Lorenzo Gatti, Marco Guerini, and Marco Turchi. 2016. Sentiwords: Deriving a high precision and high coverage lexicon for sentiment analysis. *IEEE Transactions on Affective Computing*, 7(4):409–421.

Dazhi Jiang, Hao Liu, Geng Tu, Runguo Wei, and Erik Cambria. 2024. Self-supervised utterance order prediction for emotion recognition in conversations. *Neurocomputing*, page 127370.

Dazhi Jiang, Runguo Wei, Jintao Wen, Geng Tu, and Erik Cambria. 2022. Automl-emo: Automatic knowledge selection using congruent effect for emotion identification in conversations. *IEEE Transactions on Affective Computing*.

kaiming he, xiangyu zhang, shaoqing ren, and jian sun. 2016. Deep residual learning for image recognition. abs/1512.03385:770–778.

Shivani Kumar, Atharva Kulkarni, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. When did you become so smart, oh wise one?! sarcasm explanation in multi-modal multi-party dialogues. *arXiv preprint arXiv:2203.06419*.

Diana G Maynard and Mark A Greenwood. 2014. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *Lrec 2014 proceedings*. ELRA.

Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. 2020. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1383–1392.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. 2016. A deeper look into sarcastic tweets using deep convolutional neural networks. *arXiv preprint arXiv:1610.08815*.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 704–714.

Sabina Tabacaru and Maarten Lemmens. 2014. Raised eyebrows as gestural triggers in humour: The case of sarcasm and hyper-understanding. *The European Journal of Humour Research*, 2(2):11–31.

Yi Tay, Luu Anh Tuan, Siu Cheung Hui, and Jian Su. 2018. Reasoning with sarcasm by reading in-between. *arXiv preprint arXiv:1805.02856*.

Geng Tu, Bin Liang, Dazhi Jiang, and Ruifeng Xu. 2022. Sentiment-emotion-and context-guided knowledge selection framework for emotion recognition in conversations. *IEEE Transactions on Affective Computing*.

Cynthia Van Hee. 2017. *Can machines sense irony?: exploring automatic irony detection on social media*. Ph.D. thesis, Ghent University.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7216–7223.

Jintao Wen, Geng Tu, Rui Li, Dazhi Jiang, and Wenhua Zhu. 2023. Learning more from mixed emotions: A label refinement method for emotion recognition in conversations. 11:1485–1499.

Yang Wu, Yanyan Zhao, Xin Lu, Bing Qin, Yin Wu, Jian Sheng, and Jinlong Li. 2021. Modeling incongruity between modalities for multimodal sarcasm detection. *IEEE MultiMedia*, 28(2):86–95.

Tao Xiong, Peiran Zhang, Hongbo Zhu, and Yihui Yang. 2019. Sarcasm detection with self-matching networks and low-rank bilinear pooling. In *The world wide web conference*, pages 2115–2124.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Nan Xu, Zhixiong Zeng, and Wenji Mao. 2020. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3777–3786.

Baosong Yang, Jian Li, Derek F Wong, Lidia S Chao, Xing Wang, and Zhaopeng Tu. 2019. Context-aware self-attention networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 387–394.

Baosong Yang, Derek F Wong, Tong Xiao, Lidia S Chao, and Jingbo Zhu. 2017. Towards bidirectional hierarchical representations for attention-based neural machine translation. *arXiv preprint arXiv:1707.05114*.

Weilun Yu, Chengming Li, Xiping Hu, Wenhua Zhu, Erik Cambria, and Dazhi Jiang. 2024. Dialogue emotion model based on local–global context encoder and commonsense knowledge fusion attention. pages 1–15.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Li Zhang, Guan Gui, Abdul Mateen Khattak, Minjuan Wang, Wanlin Gao, and Jingdun Jia. 2019. Multi-task cascaded convolutional networks based intelligent fruit detection for designing automated robot. 7:56028.0–56038.0.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2016. Tweet sarcasm detection using deep neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: technical papers*, pages 2449–2460.