*applied sciences*

MDPI

*Article*

# EFAFN: An Efficient Feature Adaptive Fusion Network with Facial Feature for Multimodal Sarcasm Detection

**Yukuan Sun [1,2,†], Hangming Zhang [3,†], Shengjiao Yang [4] and Jianming Wang [3,*]**

[1] Center for Engineering Intership and Training, Tiangong University, Tianjin 300384, China
[2] Department of Artificial Intelligence Convergence Network, Ajou University, Suwon 16499, Korea
[3] School of Computer Science and Technology, Tiangong University, Tianjin 300384, China
[4] School of Psychology, Guizhou Normal University, Guiyang 550025, China
* Correspondence: wangjianming@tiangong.edu.cn
† These authors contributed equally to this work.

**Abstract:** Sarcasm often manifests itself in some implicit language and exaggerated expressions. For instance, an elongated word, a sarcastic phrase, or a change of tone. Most research on sarcasm detection has recently been based on text and image information. In this paper, we argue that most image data input to the sarcasm detection model is redundant, for example, complex background information and foreground information irrelevant to sarcasm detection. Since facial details contain emotional changes and social characteristics, we should pay more attention to the image data of the face area. We, therefore, treat text, audio, and face images as three modalities and propose a multimodal deep-learning model to tackle this problem. Our model extracts the text, audio, and image features of face regions and then uses our proposed feature fusion strategy to fuse these three modal features into one feature vector for classification. To enhance the model's generalization ability, we use the IMGAUG image enhancement tool to augment the public sarcasm detection dataset MUStARD. Experiments show that although using a simple supervised method is effective, using a feature fusion strategy and image features from face regions can further improve the F1 score from 72.5% to 79.0%.

**Keywords:** sarcasm detection; multimodal analysis; deep learning

check for updates

**Citation:** Sun, Y.; Zhang, H.; Yang, S.; Wang, J. EFAFN: An Efficient Feature Adaptive Fusion Network with Facial Feature for Multimodal Sarcasm Detection. *Appl. Sci.* **2022**, *12*, 11235. https://doi.org/10.3390/app122111235

Academic Editors: Min Yang, Hao Liu, Shanxiong Chen and Yinong Chen

Received: 25 September 2022
Accepted: 31 October 2022
Published: 5 November 2022

## 1. Introduction

Sarcasm is widely used in life. Irish playwright Oscar Wilde defined sarcasm as a low-level joke with high intelligence. In many cases, the sarcastic person is expressing their anger and disgust. Still, they try to use humor to weaken the unpleasantness caused to the other party and make it easier for the other party to accept it. Sarcasm is very common on today's social platforms, and the progress of this research helps the review department to complete their tasks better. Moreover, the automatic detection of sarcasm has important practical significance in emotional analysis, harassment detection [1], and various tasks that require knowledge of people's honest thoughts.

The general problem of sarcasm detection is how to mine the inconsistent information of modal context and the conflicting information between different modalities. For example, there are two sarcasm cases in Figure 1. The sarcasm in Figure 1a is represented by inconsistencies between modalities, where both the text and tone indicate a downbeat, while the expression is surprised. Changes in human facial expressions manifest the sarcasm in Figure 1b. Therefore, sarcasm detection needs to discover these inconsistencies.

Sarcasm is an essential way to express emotion and it is an important research direction of artificial intelligence sentiment analysis. Before the U.S. election, Facebook's AI research team used language models to identify disinformation and hate speech on the web. In contrast, the subtlety of sarcasm makes it a challenging task. Identifying sarcasm is more complex than identifying false political discourse, so an efficient sarcasm detection model has significant practical significance.

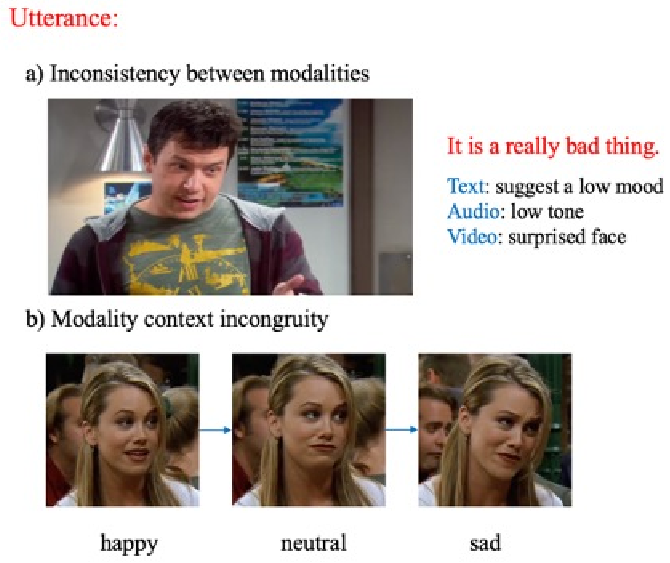
Utterance:
a) Inconsistency between modalities
It is a really bad thing.
Text: suggest a low mood
Audio: low tone
Video: surprised face
b) Modality context incongruity
happy
neutral
sad

**Figure 1.** Inconsistency in sarcasm. The text and tone in Figure (**a**) show the same feelings, while the expression shows the opposite. In daily life, a low style usually expresses boredom, while a surprised expression expresses exaggeration and wonder. The feelings expressed by the two are inconsistent, which is a kind of sarcasm. In Figure (**b**), the character's mood changes from happiness to sadness, indicating an inconsistency of feelings, so this is another expression of sarcasm.

Traditional sarcasm detection methods can be divided into three categories. The first category is based on text data. By capturing the inconsistency between texts, many sarcasm detection methods have been proposed [2–5]. However, these methods only detect sarcasm through textual data, ignoring the incongruity of information between different modalities. The second category includes methods based on images, text, and data seen from images (attributes, objects, text, descriptions, etc.) Cai et al. [6] constructed a multimodal dataset and proposed a new hierarchical fusion model to fully utilize the information of three modalities of image, text, and image attributes to solve the problem of sarcasm detection. Yao et al. [7] proposed a multi-modal, multi-interactive, multi-level neural network. This model takes images, texts in images, image descriptions, and Twitter text data as inputs and detects sarcasm by stimulating the brain's first-order and second-order understanding of sarcasm. Pan et al. [8] proposed an architecture based on the BERT model and found inconsistencies among multimodalities. However, such methods rely heavily on the clarity of the image and the data detected from the image. However, with the advent of the short video era, people are more willing to share short videos to record their lives on the social media platforms Twitter and TikTok. Therefore, only modeling text and image data are not enough to detect sarcasm. From this, the third category of sarcasm detection methods is derived—video, text, and speech-based methods. Castro et al. [9] proposed the first sarcasm detection dataset based on text, speech, and video data. They extract text, speech, and video features separately and concatenate the features of the three modalities as the input of the SVM classifier. However, the SVM algorithm is difficult to implement for large-scale training samples, and the algorithm is also sensitive to missing data and parameters, as well as the choice of kernel functions. In addition, taking vision as a modality and simply concatenating the features of each modality is unfavorable for sarcasm detection. Wu et al. [10] argue that previous work did not explicitly model the incongruity between modalities. Therefore, they propose an incongruity-aware attention network (IWAN) to detect sarcasm through a scoring mechanism for word-level incongruities between modalities. Chauhan et al. [11] believe that contextual and multimodal information is sometimes unhelpful for sarcasm detection, so they extended the dataset proposed by Castro et al. [9] using emojis and offered an emoji-aware multitask deep learning framework. However, such methods have a disadvantage in that the training speed is slow, and it ignores essential information, namely, that there is a large amount of data irrelevant to sarcasm detection in the video modality, such as the background information in the video.

In this paper, we propose an efficient feature adaptive fusion network with a facial feature for multimodal sarcasm detection (EFAFN). We utilize three types of components, namely, text, speech, and face image features, and use an adaptive feature fusion strategy to fuse the three types of features into a single vector for prediction. Our model is divided into three stages: multimodal feature extraction, adaptive feature fusion, and feature classification. First, use BERT [12] and Librosa [13] to extract the features of text and speech modalities, and then use the face detection tool provided by DLIB to cut out all the face images in each frame for horizontal stitching operation. The obtained stitched image is used as the input of ResNet-152 [14] to obtain the face image features of each video. Second, three types of parts are fused using an adaptive fusion strategy, which is different from other fusion strategies because it uses a fusion weight parameter to control the incongruity of information between different modalities. Furthermore, high performance is another manifestation of the fusion strategy. Finally, the fused vector is sent to the fully connected layer for prediction. Our results show that image features from face regions are more helpful for model performance. Furthermore, our fusion strategy is more effective than simply concatenating the three types of components.

To sum up, the adaptive fusion network with facial features is effective for sarcasm detection and it has the advantages of fast fusion speed and high performance. In addition, the model proposed in this paper solves the existence of information irrelevant to sarcasm detection in video modalities, improves the efficiency of multimodal fusion strategies, and promotes the practical implementation requirements of multimodal sarcasm detection. It has significant reference significance for other tasks in multimodal fields.

The rest of this paper is organized as follows: Section 2 mainly introduces the related research work. Section 3 presents the architecture of our proposed efficient adaptive fusion network with facial features and the feature fusion strategy. Section 4 offers our proposed data augmentation method, baseline model, experimental setup, and evaluation metrics. Section 5 presents extensive experimental results and demonstrates the effectiveness of our network. Section 6 presents the conclusions of this paper and ideas for future research work.

Our main contributions are summarized as follows:

- We propose a new multimodal deep learning sarcasm detection model, which aims to solve the problem that the existing multimodal sarcasm detection models based on three modalities of text, speech, and image only take the whole image as input, which will bring the model a sea of redundant image data, thus affecting the classification accuracy of the model. At the same time, because facial information contains emotional information related to sarcasm, we believe that we should pay attention to the image features of the facial region. Therefore, the face recognition operation is performed on the image first, and then the detected face regions are stitched horizontally to obtain the image data of the final input model;
- Traditional feature fusion methods connect each modality's features or add each modality's characteristics. Since the parts of the speech modality are numerically different from those of the other two modalities, they are easily filtered out as noise in the network, so we propose an adaptive feature fusion strategy. The characteristic feature of this fusion strategy is that the fusion weights between the three modalities can be adjusted adaptively to simulate the inconsistency between multiple modalities;
- We use a data augmentation method based on the MUStARD dataset to address the overfitting problem during deep learning training. Additionally, a series of experiments are conducted to demonstrate the effectiveness of our model; our model achieves a 6.5% improvement in F1-score over the method using a baseline for sarcasm detection.

## 2. Related Works

### 2.1. Sarcasm in Text

Sarcasm detection methods based on text data are mainly divided into three categories: rule-based methods, machine learning-based methods, and deep learning-based methods. The rule-based approach judges sarcasm by detecting some common phenomena of sarcasm. Riloff et al. [2] developed an algorithm that iteratively expanded positive and negative phrases and then used the learned words for sarcasm detection. However, the rule-based method is challenging to use to identify sarcasm due to undiscovered rules. Therefore, some scholars began to extract text features and use machine learning methods to detect sarcasm. Ghosh et al. [15] proposed to treat sarcasm detection as a word sense disambiguation problem and used SVM as their classifier. However, feature extraction is very time-consuming, and recent work has been mainly based on the deep learning method because deep learning can automatically extract features. Poria et al. [16] used a pre-trained CNN-SVM to extract sentiment, emotion, and personality features for sarcasm detection. Tay et al. [4] and Xiong et al. [5] use an attention mechanism to model incongruities between words to detect sarcasm. Felbo et al. [17] propose a distant supervision scheme and detect sarcasm through layerwise training. Hazarika et al. [18] detect sarcasm using a content- and context-driven approach. Ilic et al. [19] utilize character-level word representations in ELMo to indicate contextual sarcasm.

### 2.2. Sarcasm in Speech

The sarcasm detection method based on speech data mainly focuses on identifying the acoustic features associated with sarcasm. Features studied include speech rate, amplitude range, average amplitude, and the harmonic-to-noise ratio [20]. Bryant [21] and Woodland et al. [22] found that prosodic features such as intonation and stress are essential cues for sarcasm. Tepperman et al. [23] proposed using prosodic, spectral, and contextual sound cues to detect sarcasm. Rockwell [24] found that a slower speech rate and higher pitch may be a sign of sarcasm.

### 2.3. Multimodal Sarcasm

Single-modality-based sarcasm detection methods are no longer sufficient to explore the mystery of sarcasm. Most of the work on sarcasm detection in the last five years has been based on multimodality. Schifanella et al. [25] propose two multimodal frameworks to fuse visual and textual modalities, the first attempt to use multimodal information for sarcasm detection. Mishra et al. [26] proposed a cognitive NLP system for sarcasm detection. They introduced a model to extract eye movement features from eye movement data and then used a CNN to encode text features and eye movement features for classification. Chauhan et al. [27] extend the dataset proposed by Castro et al. [9] and present a multi-task framework to identify sarcasm, sentiment, and emotion. Firdaus et al. [28] assign sentiment labels to each instance in the MUStARD dataset and use two attention mechanisms to model dynamic information between modalities. Cai et al. [6] regard image, image attributes, and text as three modalities and propose a hierarchical fusion model to fuse the features of the three modalities for prediction. Liang et al. [29] identified the incongruity of specific intra-modal and cross-modal graphs by constructing heterogeneous intra-modal and cross-modal graphs for each multimodal instance, on which they explored an interactive graph convolutional network architecture to jointly and interactively learn the dissonance relations within and across modality graphs. Zhang et al. [30] model the context sequence of the dialogue and speaker, and mainly deal with conflicting information between modalities. Liang et al. [31] constructed a novel cross-modal graph convolutional neural network to understand the incongruity between multimodal sarcasm. The model achieved state-of-the-art results on the dataset proposed by Cai et al. [25]

In addition to using multiple modalities in this paper, we focus on extracting helpful information for sarcasm detection in video modality. For example, in video images, the background information of the picture is not intended for sarcasm detection. At the same time, facial expressions contain emotional information related to sarcasm, so we pay more attention to the feature extraction of facial information.

## 3. The Proposed EFAFN Model

In this section, we first formulate our problem, then introduce multimodal feature extraction, and finally detail our proposed model structure. As shown in Figure 2, our network architecture can be divided into three parts: (a) multimodal feature extraction, (b) adaptive feature fusion, and (c) feature classification.

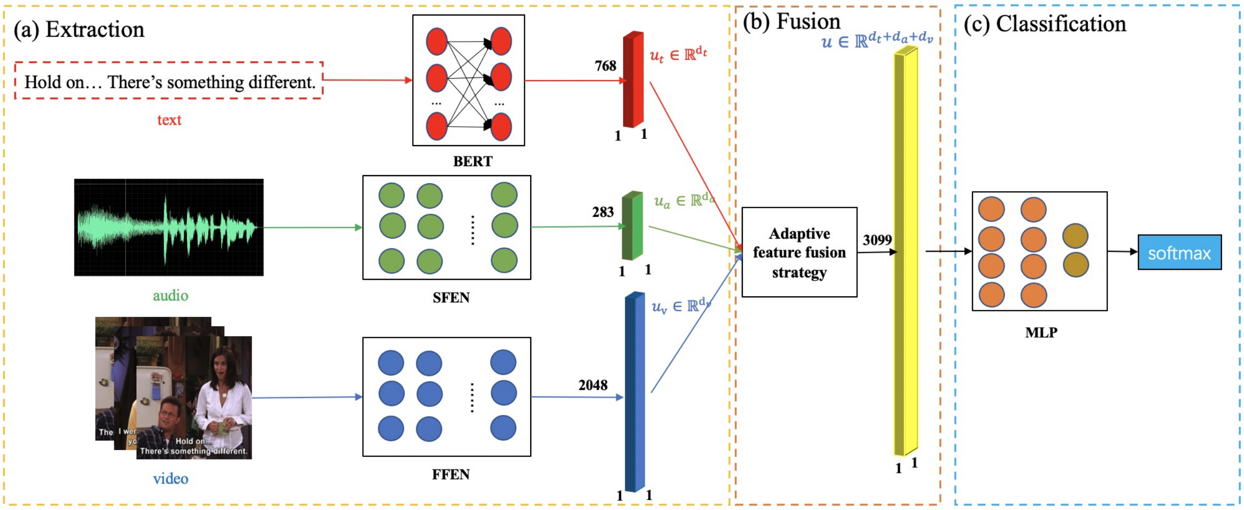
(a) Extraction
Hold on… There's something different.
text
BERT
768
$u_t \in \mathbb{R}^{d_t}$
1 1
audio
SFEN
283
$u_a \in \mathbb{R}^{d_a}$
1 1
video
FFEN
2048
$u_v \in \mathbb{R}^{d_v}$
1 1
(b) Fusion
$u \in \mathbb{R}^{d_t+d_a+d_v}$
Adaptive feature fusion strategy
3099
1 1
(c) Classification
MLP
softmax

**Figure 2.** The whole architecture of the network we proposed. It includes three key steps: (**a**) multimodal feature extraction, (**b**) adaptive feature fusion, and (**c**) feature classification. Among them, BERT represents a self-encoding language model presented by Devlin et al. [12], SFEN represents the speech feature extraction method presented by Castro et al. [9], and FFEN represents the facial feature extraction network we use, which we will describe in detail in Section 3.2. The adaptive feature fusion strategy is introduced, in fact, in Section 3.3. MLP stands for Multilayer Perceptron and is described in detail in Section 3.4.

### *3.1. Problem Formulation*

Multimodal sarcasm detection aims to identify whether a speech and text associated with a video are sarcastic. Formally, given a set of multimodal dataset $D$, for each sample $d \in D$, it contains a video $V$ with $n$ frames $\{v_1, v_2, v_3, \ldots, v_n\}$, a sentence $T$ with $m$ words $\{t_1, t_2, t_3, \ldots, t_m\}$, and an associated speech $A$. Our model's target is to classify unseen samples as sarcastic or non-sarcastic precisely.

### *3.2. Multi-Modal Feature Extraction*

Text Features: we use BERT [12] to process the text data in the dataset into a unique feature vector. First, we remove the data with empty text in the dataset, then input the text into the $BERT_{base}$ model and average the outputs of the last $L = 4$ transformer layers in the $BERT_{base}$ model as the feature vector of each text. Finally, each piece of text will be represented by a $d_t = 768$ dimensional feature vector $u_t$.

$$\{u_i^t\}_{i=1}^{L} = BERT_{base}(T) \tag{1}$$

$$u_t = \frac{1}{L}\left(\sum_{i=1}^{L} u_i^t\right) \in \mathbb{R}^{d_t} \tag{2}$$

where $u_i^t$ represents the output of the last i-th transformer layer in the BERT-base model, and $T$ represents a piece of text.

Speech Features: we adopt the method of extracting features from speech data proposed by Castro et al. [9], and use the speech processing library Librosa, which extracts basic features from audio data. First, we load audio samples into a time series signal with a sampling rate of 22,050 Hz; we use a heuristic sound extraction method to remove background noise. Finally, we divide the audio signal into $d_w$ non-overlapping windows to extract local features, including MFCC, Mel-spectogram, spectral centroid, and their associated temporary derivatives (delta). All the elements are connected to form a joint feature vector $\{u_i^a\}_{i=1}^{d_w}$with $d_a = 283$ dimensions for each window. Finally, the joint feature $u_a$ of each piece of audio is obtained by calculating the average value of all windows.

$$u_i^a = u_i^{MFCC} \oplus u_i^{MFCC\ delta} \oplus u_i^{Mel} \oplus u_i^{Mel\ delta} \oplus u_i^{spec} \tag{3}$$

$$u_a = \frac{1}{d_w}(\Sigma_i u_i^a) \in \mathbb{R}^{d_a} \tag{4}$$

where $\oplus$ is the concatenation operator, $u_i^{MFCC}$, $u_i^{MFCC\ delta}$, $u_i^{Mel}$, $u_i^{Mel\ delta}$, and $u_i^{spec}$ represent MFCC features, MFCC associated temporary derivatives, Mel-spectogram features, Mel-spectogram associated temporary derivatives, and spectral centroid of each window, respectively.

Face Image Features: as facial information contains emotional information related to sarcasm detection, we believe that the video modality should focus on the person's news rather than the image's background information. The network architecture of facial feature extraction is shown in Figure 3. First, we use the face detection model provided by DLIB (http://dlib.net/face_detection_ex.cpp.html, accessed on 13 August 2021.) to detect the face $f_i$ of each frame in the video conversation. Let $D_i$ denote the maximum height of the face image seen from each image frame. Since the size of each face detected in the face detection process is inconsistent, the black block $B_i$ is used to fill the vacant part in the splicing process. After the filling operation, the heights of all face images are uniform. Then, a horizontal stitching operation is performed on all faces to obtain the final image input to the network. The formulation is as follows.

$$d_{B(f_i)} = (3, Len(f_i), D_i - Hei(f_i)) \tag{5}$$

$$padding(f_i) = \begin{cases} f_i, Hei(f_i) = D_i \\ f_i \oplus B_i \in \mathbb{R}^{d_{B(f_i)}}, Hei(f_i) < D_i \end{cases} \tag{6}$$

$$f_i^{face} = stitching(padding(f_i)) \tag{7}$$

where $Len()$ represents the width of the image, $Hei()$ represents the height of the image, $d_{B(f_i)}$ represents the dimension of the black block that the face needs to be supplemented with, $\oplus$ represents the vertical stitching operation, and $padding()$ represents the padding operation, $stitching()$ represents a horizontal stitching operation.

We use $f_i^{face}$ to represent the stitched image. Then, we preprocess each frame of $f_i^{face}$ by normalization, and then redefine the last layer of the ResNet-152 [14] image classification model pre-trained by ImageNet to 2048 dimensions, using this model to extract features for $f_i^{face}$ representation. To obtain the visual representation of each sentence of text, we calculate the average value of the feature vector $u_i^{f_i^{face}}$ with $d_v = 2048$ dimensions obtained in each frame.

$$u_i^{f_i^{face}} = ResNet - 152(f_i^{face}) \tag{8}$$

$$u_v = \frac{1}{F}\left(\Sigma_i u_i^{f_i^{face}}\right) \in \mathbb{R}^{d_v} \tag{9}$$

where $F$ represents the number of video frames.

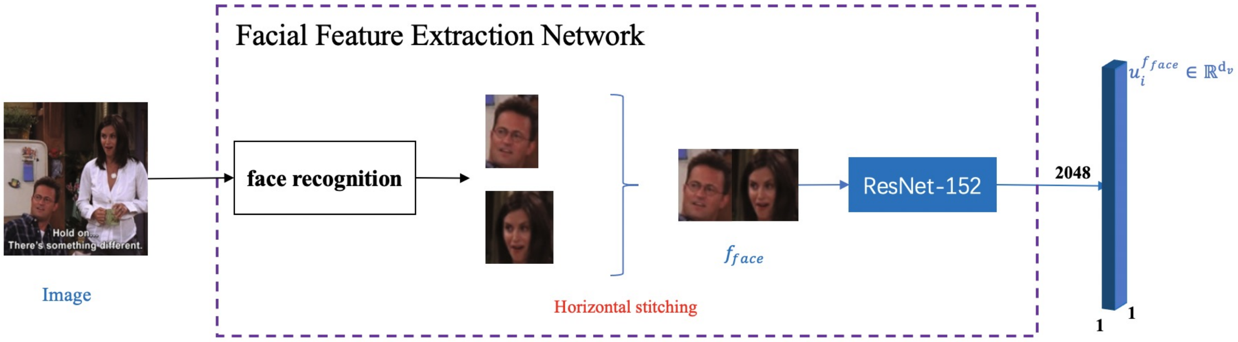
Facial Feature Extraction Network
face recognition
Hold on.
There's something different.
Image
$f_{face}$
Horizontal stitching
ResNet-152
2048
$u_i^{f_{face}} \in \mathbb{R}^{d_v}$
1
1

**Figure 3.** Facial feature extraction network architecture diagram. In the above, face recognition uses the face detection model provided by DLIB. ResNet-152 is a residual learning network proposed by He et al. [14]. This network is one of the most commonly used backbone networks in image classification tasks. The objects processed in our dataset are videos, so we use this architecture to average the features obtained from each frame in the video as the final facial feature vector.

### *3.3. Adaptive Feature Fusion*

Typical feature fusion methods are divided into early fusion and late fusion. Early fusion refers to the fusion of features and input into a model for training. For example, the most common method is to perform a simple concatenation operation on the parts. Early fusion methods learn to exploit the correlations and interactions between low-level features of each modality. The late fusion method trains a model for each modality separately and then uses a fusion mechanism to integrate the results of all single-modality models. Commonly used fusion mechanisms include averaging, voting, and training fusion models. Since late fusion methods train different models for different modalities, each modality can be better modeled, allowing for better flexibility. It is worth noting, however, that the late fusion approach ignores low-level interactions between modalities. In theory, early fusion should achieve better results for the sarcasm detection task than late fusion because the corresponding features have a specific index relationship and less feature abstraction in reality. Beyond that, sarcasm detection is all about finding interactions between modalities.

After the above analysis, to better fuse the features of the three modalities together, we propose an adaptive feature fusion strategy with an early fusion method, as shown in Figure 4. Given the initial text feature $u_t$, speech feature $u_a$, and face image feature $u_v$, the final prediction vector $u$ is obtained according to the strategy $\pi$. We approximate the policy by a deep policy network $\pi(\cdot;\theta)$:

$$u \sim \pi(u_t, \varphi(u_a;\alpha), \phi(u_v;\beta);\theta) \tag{10}$$

where $\alpha$, $\beta$, and $\theta$ are the parameters of the network, which can be updated according to gradients. Since text features, speech features, and face image features come from three different modalities, we design two sub-network branches $\varphi(\cdot;\alpha)$ and $\phi(\cdot;\beta)$ to combine speech features and face image features are mapped into a feature vector that combines text features.

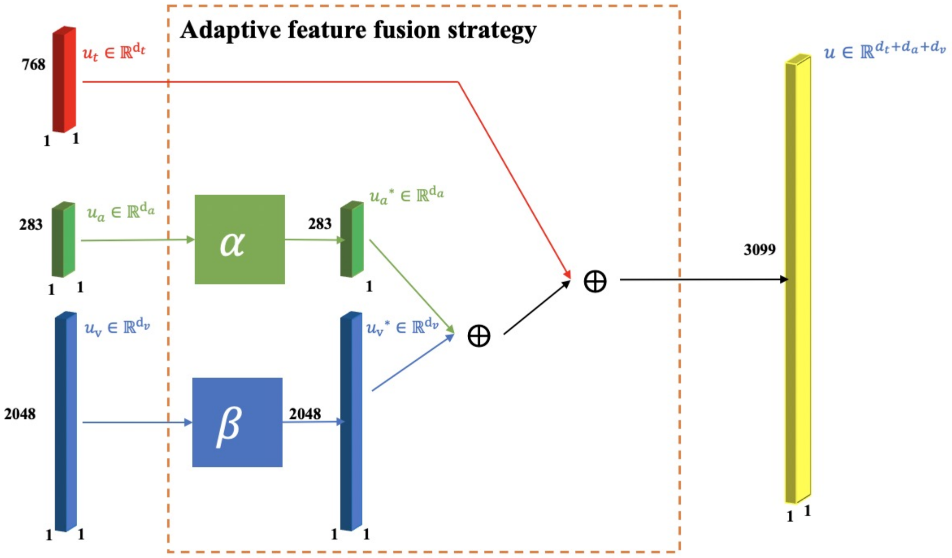
Adaptive feature fusion strategy
768
$u_t \in \mathbb{R}^{d_t}$
283
$u_a \in \mathbb{R}^{d_a}$
α
$u_a{}^* \in \mathbb{R}^{d_a}$
2048
$u_v \in \mathbb{R}^{d_v}$
β
$u_v{}^* \in \mathbb{R}^{d_v}$
⊕
⊕
3099
$u \in \mathbb{R}^{d_t+d_a+d_v}$
1 1

**Figure 4.** Adaptive feature fusion strategy diagram. Where $\alpha$ and $\beta$ are the parameters of the subnetworks $\varphi(\cdot;\alpha)$ and $\phi(\cdot;\beta)$, $\oplus$ represents the join operation, and $u_a^*$ and $u_v^*$ are the subnetworks $\varphi(\cdot;\alpha)$ and $\phi(\cdot;\beta)$ output.

### 3.4. Feature Classification

Inspired by the VGG16 [32] network structure, we designed a Multilayer Perceptron (MLP) and Softmax layers for feature classification, as shown in Figure 5. The activation function of the hidden layer and output layer is LeakyReLU, and the loss function uses cross entropy.

$$J = -\sum_{i=1}^{N}[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)] + \lambda R \tag{11}$$

where $J$ is the cost function. $\hat{y}_i$ is the prediction result of our model for sample $i$, and $y_i$ is the true label for sample $i$. $N$ is the size of the training data. $R$ is the standard L2 regularization, and $\lambda$ is the weight of $R$.

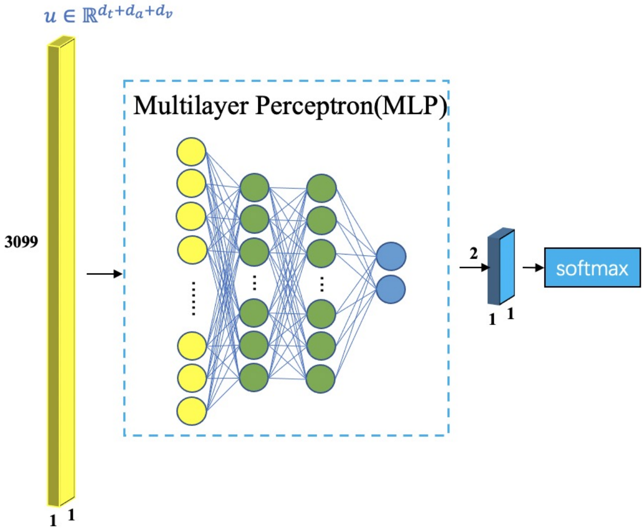

**Figure 5.** Multilayer perceptron block diagram. It consists of a three-layer fully connected neural network. The number of neurons in the input layer and output layer is 3099 and 2, respectively, and the number of neurons in the hidden layer is 328.

## 4. Experiment

To explore the role of each modality in sarcasm detection, we conducted multiple experiments to evaluate the performance of each modality and modal combinations in the model separately. In addition, to address the overfitting problem during training, we propose a data augmentation method. Finally, we performed a series of ablation experiments to verify the effectiveness of the components in our model. We performed 5-fold cross-validation experiments on the dataset and averaged the cross-validation results to evaluate the proposed classifier. In each fold, we ran the data augmentation method only on the data in the training set.

### *4.1. Dataset and Data Augmentation*

We evaluated our model on a publicly available multimodal sarcasm detection dataset, MUStARD (https://github.com/soujanyaporia/MUStARD, accessed on 26 July 2021), which was collected by Castro et al. [9]. The dataset contains 690 videos with a total duration of about 9626 s. The data come from funny American TV series on YouTube, such as *The Big Bang Theory*, *Friends*, etc. The data contain sarcastic expressions, speakers, video, audio, sarcastic tags, etc. They can be used in many aspects of research.

In addition, we used a data augmentation method to augment the dataset by a factor of 15. The purpose is to solve the problem of overfitting during training due to the small dataset. We randomly perform data augmentation on $f_{face}$ with 0∼5 out of 13 methods. All methods were implemented using the image enhancement tool IMGAUG ∣ (https://github.com/CrazyVertigo/awesome-data-augmentation, accessed on 3 November 2021). The random mode was used in the data augmentation method mainly because the data augmentation cannot be overwhelming in the deep learning model training. Just imagine that using Gaussian blur for each image will destroy the characteristics of the original data that have been changed, and the consequences are enormous. The augmented dataset details are listed in Table 1.

**Table 1.** Labels distribution in the MUStARD and augmented MUStARD.

| Dataset | Train | Test | Labels | |
|---|---|---|---|---|
| | | | Sarcastic | Non-Sarcastic |
| MUStARD | 552 | 138 | 345 | 345 |
| augmented MUStARD | 8280 | 138 | 4209 | 4209 |

The 13 methods of data augmentation are as follows:

- Use superpixel enhancement;
- Use one of the Gaussian, mean, and median to blur the picture;
- Sharpen the image;
- Add emboss effect to the image;
- Perform edge detection on the original image, assign the detected edge to 0 or 255 and then superimpose it on the original image;
- Add Gaussian noise to the image;
- Set 1%∼10% of the pixels to black or replace 3%∼15% of the pixels with black squares 2%∼5% of the original size;
- Set 5% probability inversion pixel intensity;
- Each pixel in the image randomly adds or subtracts a number between −10 and 10;
- Multiply each pixel in the image by a number between 0.5 and 1.5;
- Halve or double the contrast of the entire image;
- Distort the local area of the image;
- Move the pixels around.

### *4.2. Baselines*

The experiments were mainly conducted using five baseline methods:

**Random:** This baseline will randomly classify the test samples.

**Castro et al. [9]:** This baseline uses a Support Vector Machine (SVM) as the classifier of the model. SVM is a robust classifier for small data sets. Castro et al. [9] use an SVM with RBF kernel and a scalable gamma, and the penalty term C is adjusted to Castro et al. [9] according to the parameters set by each experiment (we choose between 1, 10, 30, 500, and 1000).

**Two-attention-based encoder [28]:** This model is used to simultaneously detect sarcasm, sentiments, and emotions and is a multi-task learning framework. The input to the model is the concatenation of each modality's focused utterance and its historical contextual utterances. It then uses inter-segment and intra-segment inter-modal attention to model inter-modal dynamics.

**Sequential Context Encoder [30]:** Zhang et al. [30] propose a sequential context encoder and a contrastive attention-based encoder. However, the input of the contrastive attention-based encoder is only bi-modal, different from the tri-modal input of our model, so we only choose the sequential context encoder as the baseline model. This baseline first extracts three features of text, speech, and image modalities, then encodes these three features using this baseline, and then uses a linear decoder for classification.

**ViViT-VAT:** The ViViT [33] model was proposed by Arnab et al. [33] to solve the video classification task and achieve state-of-the-art results on multiple datasets. This baseline uses the ViViT [33] model as the backbone network and then linearly transforms the text and speech modality into the space and temporal transformers, respectively.

### *4.3. Experimental Setup*

Our model was built on PyTorch [34] and ran on NVIDIA GeForce RTX 3090 GPU. We used SGD as our optimizer and set the initial learning rate to $1 \times 10^{-5}$. Due to the small amount of data, in the experiment, we put all the training data into the neural network for training at one time and only updated the parameters of the feature fusion layer and the classification layer during the training process. In addition, we set the initial values of the neural network parameters $\alpha$ and $\beta$ in the adaptive feature fusion strategy to 0.01 and 1. These parameters will be optimized according to the gradient descent algorithm. All the experimental hyper-parameters are listed in Table 2.

**Table 2.** Hyper-parameters table. The parameters marked with * indicate that they can be updated during the training process, and those listed in the table indicate their initial values at the beginning of training.

| Hyper-Parameters | Value |
|---|---|
| Learning rate | $1 \times 10^{-5}$ |
| Early stop patience | $5 \times 10^{5}$ |
| Dropout rate | 0.5 |
| $\alpha^*$ | 0.01 |
| $\beta^*$ | 1.0 |
| Number of iterations | $3 \times 10^{6}$ |
| Fully connected input size | 3099 |
| Fully connected hidden size | 328 |

### *4.4. Evaluation Metrics*

In this section, we evaluate the performance of sarcasm detection using three metrics: (1) precision; (2) recall; (3) balanced F-score (F1-score). Among these three metrics, F1-score is commonly used when dealing with binary classification problems [9,28,30]. Therefore, we chose F1-score as our primary evaluation metric and others as our secondary evaluation metrics. The three metrics are calculated as follows:

Let TP represent that a sample is a positive class and is also judged to be a positive class, FP represents a sample that is a false class but is judged to be a positive class, and FN means that a sample is positive but judged to be incorrect.

(1) Precision indicates the proportion of correct predictions that are positive to all positive predictions:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{12}$$

(2) Recall indicates the proportion of correct predictions that are positive to all actual positives:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{13}$$

(3) F1-score can be seen as a harmonic mean of model precision and recall:

$$\text{F1-score} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{14}$$

## 5. Results and Discussion

### *5.1. Model Comparison*

First, we show the difference between our method and other baseline models in Table 3, then we compare the performance of our model with five baseline models and list the experimental results in Table 4.

**Table 3.** The differences between the proposed model and the baseline model are compared. Two attention mechanisms refer to inter-segment inter-modal attention and intra-segment inter-modal attention, SVM refers to support vector machine, and MLP refers to multi-layer perceptron.

| Model | Fusion Method | Classification Method |
|---|---|---|
| Castro el al. [9] | Concat | SVM |
| Two-attention-based encoder [28] | Two attention mechanisms | Linear decoder |
| Sequential Context Encoder [30] | Sequential context encoder | Linear decoder |
| ViViT-VAT | Add | MLP |
| Ours | Adaptive feature fusion | MLP |

**Table 4.** Sarcasm detection results of different models.

| Model | Precision | Recall | F1-Score |
|---|---|---|---|
| Random | 49.5 | 49.5 | 49.5 |
| Castro et al. [9] | 72.5 | 72.5 | 72.5 |
| Two-attention-based encoder [28] | 71.5 | 71.4 | 70.5 |
| Sequential Context Encoder [30] | 72.3 | 72.3 | 72.3 |
| ViViT-VAT | 69.3 | **88.4** | 77.7 |
| Ours | **79.0** | 79.0 | **79.0** |

In Table 4, our model performs better than the five baseline models. Specifically, our model achieves a 6.5% (from 72.5% to 79.0%) improvement in F1-score compared to the SVM classification based method proposed by Castro et al. [9]. This shows that although SVM is a good classification model, since sarcasm detection needs to mine inconsistent information between multiple modalities, the SVM method cannot better solve the automatic classification of sarcasm. Compared with the Two-attention-based encoder method proposed by Firdaus et al. [28], our model offers an improvement of 8.5% (from 70.5% to 79.0%). This shows that fusing multimodal features for prediction is more effective than using an attention mechanism to simulate dynamic information between modalities. Compared with the Sequential Context Encoder proposed by Zhang et al. [30], our model achieves an improvement of 6.7% (from 72.3% to 79.0%). This shows that using the encoder–decoder architecture is insufficient to reflect the inconsistent information between modalities, and the deep learning network using the feature fusion strategy performs better. Compared with the ViViT-VAT model using ViViT [33] as the backbone network, our model improves by 1.3% (from 77.7% to 79.0%), which shows that although the ViViT [33] network achieves state-of-the-art results in single-modal video classification, for the multimodal video classification task, the ViViT network is still not able to fuse multimodal features well. However, the network achieves the best results in the baseline model. Furthermore, the best performance obtained by our model is 79.6%, with a standard deviation of 0.54.

### *5.2. Modality Comparison*

To further explore the influence of three modalities on our proposed model, we used all possible inputs to evaluate our model: uni-modal (T, A, V/$V^*$), bi-modal (T + A, T + V/$V^*$, A + V/$V^*$), and tri-modal (T + A + V/$V^*$), the performance is shown in Table 5.

**Table 5.** Results on the multimodal sarcasm detection augmented dataset. The best results are shown in bold. P means precision, R means recall, F1 means F1-score, SCE means sequential context encoder, T means text, A means voice, V means face image, and $V^*$ means the whole image.

| Model | Modality | T | A | V/$V^*$ | T + A | T + V/$V^*$ | A + V/$V^*$ | T + A + V/$V^*$ |
|---|---|---|---|---|---|---|---|---|
| P | Castro et al. [9] (V) | 65.1 | 65.9 | 65.2 | 66.6 | 69.6 | 68.1 | 72.5 |
| | SCE [30] ($V^*$) | 53.4 | 68.0 | 71.4 | 60.4 | 70.4 | 73.2 | 72.3 |
| | Ours (V) | 73.9 | 65.2 | 74.6 | 69.0 | 77.7 | 72.5 | **79.0** |
| R | Castro et al. [9] (V) | 64.6 | 64.6 | 65.2 | 66.2 | 69.6 | 68.1 | 72.5 |
| | SCE [30] ($V^*$) | 53.4 | 67.2 | 71.2 | 60.5 | 70.3 | 72.0 | 72.3 |
| | Ours (V) | 73.9 | 65.2 | 74.6 | 68.8 | 77.5 | 72.5 | **79.0** |
| F1 | Castro et al. [9] (V) | 64.6 | 64.6 | 65.2 | 66.2 | 69.6 | 68.1 | 72.5 |
| | SCE [30] ($V^*$) | 53.4 | 66.5 | 71.2 | 60.4 | 70.4 | 71.9 | 72.3 |
| | Ours (V) | 73.9 | 65.2 | 74.6 | 68.8 | 77.5 | 72.5 | **79.0** |

As can be seen from Table 5, whether it is the SVM classifier-based method proposed by Castro et al. [9] or our method, the model using only face image features (V) or text features (T) performs well, but the combination of them (T + V) improves the F1-score by 4.4% (from 65.2% to 69.6%) and 2.9% (from 74.6% to 77.5%), respectively. Therefore, for sarcasm detection, the blessing of the two is beneficial to improving the model's performance. It is worth noting that the blessing of speech modality (A) did not improve our model greatly. On the contrary, the blessing of speech modality improves the performance of the baseline method. Specifically on the method proposed by Castro et al. [9], the blessing of speech modality (T + A, A + V) improves the performance of the model by 1.6% (from 64.6% to 66.2%) and 2.9% (from 65.2% to 68.1%), respectively. On the sequential context encoder proposed by Zhang et al. [30], the blessing of speech modality (T + A, A + $V^*$) improves the performance of the model by 7.0% (from 53.4% to 60.4%) and 0.7% (from 71.2% to 71.9%), respectively. This shows that using the baseline method can better integrate the features of speech modalities. In addition, in the model with three modalities (T + A + V/$V^*$) as input, our method outperforms the baseline method by 6.5% (from 72.5% to 79.0%) and 6.7% (from 72.3% to 79.0%), respectively. This shows that our proposed model can better fuse the features of the three modalities.

### *5.3. Ablation Study*

To evaluate the effectiveness of the components in our model, we conducted a series of ablation experiments.

#### 5.3.1. Ablation 1

We removed the data augmentation method to get the model (w\o A), which is only trained using the original dataset. Then, we eliminated the feature fusion strategy to obtain the model (w\o F), which concatenates the features of the three modalities and performs classification. The experimental results are presented in Table 6. The lack of a data augmentation method leads to a drop in the F1-score (Dropped from 79.0% to 74.8%), proving that our data augmentation method is effective. No feature fusion strategy affects the performance of the model (a drop from 79.0% to 74.2%), proving that our feature fusion strategy is more necessary than simply concatenating the features of the three modalities.

**Table 6.** Ablation experiment results. w\o means removing this component. A represents the data enhancement method, and F represents the feature fusion strategy.

| Model | Precision | Recall | F1-Score | Standard Deviation(F1-Score) |
|---|---|---|---|---|
| Model (w\o A) | 74.8 | 74.8 | 74.8 | 0.57 |
| Model (w\o F) | 74.9 | 74.3 | 74.2 | 1.31 |
| Ours | **79.0** | **79.0** | **79.0** | **0.54** |

### 5.3.2. Ablation 2

To evaluate our adaptive feature fusion strategy, we fixed the values of $\alpha$ and $\beta$ to 0.01 and 1 without adaptive adjustment. We used this strategy to perform feature fusion and train model(X). After that, we set up three adaptive feature fusion strategies. The formulas are as follows:

$$u \sim \pi(u_t, \varphi(u_a; \alpha), u_v; \theta) \tag{15}$$

$$u \sim \pi(\psi(u_t; \gamma), \varphi(u_a; \alpha), u_v; \theta) \tag{16}$$

$$u \sim \pi(\psi(u_t; \gamma), \varphi(u_a; \alpha), \phi(u_v; \beta); \theta) \tag{17}$$

where $\psi(\cdot; \gamma)$ represents a sub-network branch, and $\gamma$ is the parameter of the network. In this ablation experiment, the initial value of $\gamma$ is set to 1.0.

The models trained according to the three feature fusion strategies are model($A_l$), model($TA_l$), and model($TAV_l$), and the model performance is shown in Table 7. In Table 7, fixing the values of $\alpha$ and $\beta$ will reduce the F1-score, but compared with the features of the three modalities directly concatenated, the F1-score is improved by 3.1% (from 74.2% to 77.3%). This shows that using our proposed feature fusion strategy leads to better model performance. Using the other three feature fusion strategies will cause different degrees of decline. Specifically, using the feature fusion strategy only on the speech modality will cause a 2.2% (from 79.0% to 76.8%) drop, and using the feature fusion strategy on the text and voice modalities will cause a 2.0% (from 79.0% to 77.0%) decrease. Using a feature fusion strategy on three modalities results in a drop of 1.4% (from 79.0% to 77.6%). These results show that our proposed feature fusion strategy can better fuse the features of the three modalities.

**Table 7.** Ablation experiment results. X represents the fixed $\alpha$ and $\beta$ values of 0.01 and 1.0.

| Model | Precision | Recall | F1-Score | Standard Deviation(F1-Score) |
|---|---|---|---|---|
| Model(w\o F) | 74.9 | 74.3 | 74.2 | 1.31 |
| Model(X) | 77.5 | 77.4 | 77.3 | 0.52 |
| Model($A_l$) | 76.9 | 76.9 | 76.8 | 1.12 |
| Model($TA_l$) | 77.1 | 77.0 | 77.0 | 0.82 |
| Model($TAV_l$) | 77.8 | 77.6 | 77.6 | 0.56 |
| Ours | **79.0** | **79.0** | **79.0** | **0.54** |

### 5.3.3. Ablation 3

In addition, we used the contextual video features collected by Castro et al. [9] as the feature of the video modality, which takes the whole image as the input of ResNet-152 [14]. Then, the text, speech, and video features were fused using the adaptive feature fusion strategy to train the model ($V^*$). To evaluate the speech modalities' features, we overlapped each window's speech features and used the unaveraged local features to represent the features of the speech modalities. Then, we used MLP to transform the speech elements to the exact dimensions as video and text, and used the unaveraged local features, text features, and face image features to train to obtain Model(A_N). In addition, we extracted global speech features using WavLM [35], a speech model pre-trained by Libri-Light, GigaSpeech, and VoxPopuli, which employs a gated relative positional bias for the Transformer to better capture sequences relation with more giant time steps in the time series. Then, we used this global speech feature and the features of the other two modalities to train the Model(A_O). The experimental results in Table 8 show that replacing face image features with whole image features will reduce the F1-score score (a drop from 79.0% to 77.3%), proving that we should pay more attention to people's facial features information. In addition, we find that the performance of the model($V^*$) is 5.0% (from 72.3% to 77.3%) higher than that of the Sequential Context Encoder proposed by Zhang et al. [30]. This shows that our adaptive feature fusion strategy is better at discovering inconsistent information between modalities than the encoder-decoder architecture. Using unaveraged speech local features reduces the model's performance by 1.1% (a drop from 79.0% to 77.9%) compared to the averaged features, proving that the averaging operation can highlight sarcasm better. Additionally, using global speech features causes severe performance degradation (Dropped from 79.0% to 73.7%). This shows that global speech features are unfavorable for sarcasm detection because sarcastic speech features have been submerged in global speech features.

**Table 8.** Ablation experiment results. $V^*$ represents the contextual video features collected by Castro et al. [9].

| Model | Precision | Recall | F1-Score | Standard Deviation (F1-Score) |
|---|---|---|---|---|
| SCE [30] | 72.3 | 72.3 | 72.3 | - |
| Model($V^*$) | 77.5 | 77.4 | 77.3 | 1.10 |
| Model(A_N) | 78.5 | 78.0 | 77.9 | 1.03 |
| Model(A_O) | 73.9 | 73.8 | 73.7 | 0.72 |
| Ours | **79.0** | **79.0** | **79.0** | **0.54** |

## 6. Conclusions

We studied the redundant information problem of video modalities in multimodal sarcasm detection. We constructed an EFAFN multimodal sarcasm detection model that used the open-source sarcasm detection dataset MUStARD for experiments and proposed a data augmentation method to solve the overfitting problem during model training. In addition, a ViViT-VAT model was also constructed for comparative experiments. The final experimental results show that the EFAFN multimodal sarcasm detection model outperforms the ViViT-VAT model, proving the effectiveness of our model. The model based on facial features outperforms the model based on whole image features, demonstrating the importance of facial information in sarcasm detection. Ablation experiment results show that not using data augmentation methods and adaptive fusion strategies will cause model performance degradation, illustrating the necessity of our proposed two components. The use of global speech features and unaveraged local speech features also degrades the model performance, indicating that the operation of averaging features will highlight the sarcasm features of speech modalities.

Although the model's effectiveness in this paper, the usefulness of the three modalities, and the necessity of the adaptive fusion strategy are proved by experiments, there is still room for further optimization. First, this paper only removes information irrelevant to sarcasm detection for video modalities, and then we can discuss redundant information removal algorithms for speech and text modalities. Second, some multimodal information may not be helpful for sarcasm detection. We will consider designing better processing methods for text and speech modalities in the future. For example, the speech data in the dataset are augmented with additional data, such as emojis [11], to add valuable information for sarcasm detection. Third, the dataset used in this paper only contains three modalities, and the influence of other modalities' data, such as text in images, on sarcasm detection is not considered. Other existing public datasets can be used in the future.

**Author Contributions:** Data curation, H.Z.; Formal analysis, S.Y.; Funding acquisition, Y.S. and J.W.; Methodology, Y.S. and H.Z.; Software, Y.S.; Supervision, Y.S. and J.W.; Writing—original draft, H.Z. and S.Y.; writing—review and editing, Y.S. and J.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by The National Natural Science Foundation of China (62072335) and The Tianjin Science and Technology Program (19PTZWHZ00020).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The dataset can be obtained from https://github.com/soujanyaporia/MUStARD, accessed on 26 July 2021.

**Acknowledgments:** We really appreciate the valuable comments from the anonymous reviewers, which have made the paper more comprehensive.

**Conflicts of Interest:** The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## References

1. Pang, B.; Lee, L. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.* **2008**, *2*, 1–135. [CrossRef]
2. Riloff, E.; Qadir, A.; Surve, P.; De Silva, L.; Gilbert, N.; Huang, R. Sarcasm as contrast between a positive sentiment and negative situation. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 704–714.
3. Joshi, A.; Sharma, V.; Bhattacharyya, P. Harnessing context incongruity for sarcasm detection. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Beijing, China, 26–31 July 2015; pp. 757–762.
4. Tay, Y.; Tuan, L.A.; Hui, S.C.; Su, J. Reasoning with sarcasm by reading in-between. *arXiv* **2018**, arXiv:1805.02856.
5. Xiong, T.; Zhang, P.; Zhu, H.; Yang, Y. Sarcasm detection with self-matching networks and low-rank bilinear pooling. In Proceedings of the World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 2115–2124.
6. Cai, Y.; Cai, H.; Wan, X. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 2506–2515.
7. Yao, F.; Sun, X.; Yu, H.; Zhang, W.; Liang, W.; Fu, K. Mimicking the Brain's Cognition of Sarcasm From Multidisciplines for Twitter Sarcasm Detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2021** . [CrossRef] [PubMed]
8. Pan, H.; Lin, Z.; Fu, P.; Qi, Y.; Wang, W. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 16–20 November 2020; pp. 1383–1392.
9. Castro, S.; Hazarika, D.; Pérez-Rosas, V.; Zimmermann, R.; Mihalcea, R.; Poria, S. Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv* **2019**, arXiv:1906.01815.
10. Wu, Y.; Zhao, Y.; Lu, X.; Qin, B.; Wu, Y.; Sheng, J.; Li, J. Modeling incongruity between modalities for multimodal sarcasm detection. *IEEE MultiMedia* **2021**, *28*, 86–95. [CrossRef]
11. Chauhan, D.S.; Singh, G.V.; Arora, A.; Ekbal, A.; Bhattacharyya, P. An emoji-aware multitask framework for multimodal sarcasm detection. *Knowl.-Based Syst.* **2022**, *257*, 109924. [CrossRef]
12. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
13. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.P.; McVicar, M.; Battenberg, E.; Nieto, O. Librosa: Audio and music signal analysis in python. In Proceedings of the 14th Python in Science Conference, Austin, TX, USA, 6–12 July 2015; Volume 8, pp. 18–25.
14. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
15. Ghosh, D.; Guo, W.; Muresan, S. Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1003–1012.
16. Poria, S.; Cambria, E.; Hazarika, D.; Vij, P. A deeper look into sarcastic tweets using deep convolutional neural networks. *arXiv* **2016**, arXiv:1610.08815.
17. Felbo, B.; Mislove, A.; Søgaard, A.; Rahwan, I.; Lehmann, S. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv* **2017**, arXiv:1708.00524.
18. Hazarika, D.; Poria, S.; Gorantla, S.; Cambria, E.; Zimmermann, R.; Mihalcea, R. Cascade: Contextual sarcasm detection in online discussion forums. *arXiv* **2018**, arXiv:1805.06413.
19. Ilić, S.; Marrese-Taylor, E.; Balazs, J.A.; Matsuo, Y. Deep contextualized word representations for detecting sarcasm and irony. *arXiv* **2018**, arXiv:1809.09795.

20. Cheang, H.S.; Pell, M.D. The sound of sarcasm. *Speech Commun.* **2008**, *50*, 366–381. [CrossRef]
21. Bryant, G.A. Prosodic contrasts in ironic speech. *Discourse Process.* **2010**, *47*, 545–566. [CrossRef]
22. Woodland, J.; Voyer, D. Context and intonation in the perception of sarcasm. *Metaphor. Symb.* **2011**, *26*, 227–239. [CrossRef]
23. Tepperman, J.; Traum, D.; Narayanan, S. "Yeah Right": Sarcasm Recognition for Spoken Dialogue Systems. In Proceedings of the Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, 17–21 September 2006.
24. Rockwell, P. Lower, slower, louder: Vocal cues of sarcasm. *J. Psycholinguist. Res.* **2000**, *29*, 483–495. [CrossRef]
25. Schifanella, R.; de Juan, P.; Tetreault, J.; Cao, L. Detecting sarcasm in multimodal social platforms. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 1136–1145.
26. Mishra, A.; Dey, K.; Bhattacharyya, P. Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 377–387.
27. Chauhan, D.S.; Dhanush, S.; Ekbal, A.; Bhattacharyya, P. 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, Online, 5–10 July 2020; pp. 4351–4360.
28. Firdaus, M.; Chauhan, H.; Ekbal, A.; Bhattacharyya, P. MEISD: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; pp. 4441–4453.
29. Liang, B.; Lou, C.; Li, X.; Gui, L.; Yang, M.; Xu, R. Multi-Modal Sarcasm Detection with Interactive In-Modal and Cross-Modal Graphs. In Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021; pp. 4707–4715.
30. Zhang, X.; Chen, Y.; Li, G. Multi-modal Sarcasm Detection Based on Contrastive Attention Mechanism. In Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing, Qingdao, China, 13–17 October 2021; pp. 822–833.
31. Liang, B.; Lou, C.; Li, X.; Yang, M.; Gui, L.; He, Y.; Pei, W.; Xu, R. Multi-Modal Sarcasm Detection via Cross-Modal Graph Convolutional Network. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 22–27 May 2022; pp. 1767–1777.
32. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
33. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; Schmid, C. Vivit: A video vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 6836–6846.
34. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32* , 8026–8037.
35. Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1505–1518. [CrossRef]