Multimedia Tools and Applications (2022) 81:8597–8616
https://doi.org/10.1007/s11042-022-12122-9

A multimodal fusion method for sarcasm detection based on late fusion

Ning Ding[1] · Sheng-wei Tian[2] · Long Yu[3]

Received: 27 April 2021 / Revised: 27 July 2021 / Accepted: 3 January 2022
Published online: 4 February 2022

Abstract
Information on social media is multi-modal, most of which contains the meaning of sarcasm. In recent years, many people have studied the problem of sarcasm detection. Many traditional methods have been proposed in this field, but the study of deep learning methods to detect sarcasm is still insufficient. It is necessary to comprehensively consider the information of the text,the changes of the tone of the audio signal,the facial expressions and the body posture in the image to detect sarcasm. This paper proposes a multi-level late-fusion learning framework with residual connections, a more reasonable experimental data-set split and two model variants based on different experimental settings. Extensive experiments on the MUStARD show that our methods are better than other fusion models. In our speaker-independent experimental split, the multi-modality has a 4.85% improvement over the single-modality, and the Error rate reduction has an improvement of 11.8%. The latest code will be updated to this URL later: https://github.com/DingNing123/m_fusion

Keywords Multi-modal fusion · Sarcasm detection · Dialogue emotion recognition

✉ Sheng-wei Tian
tianshengwei@163.com

Ning Ding
dn1992@stu.xju.edu.cn

Long Yu
yul@xju.edu.cn

[1] School of Information Science and Engineering, Xinjiang University, No.666, Shengli Road, Tianshan District, Urumqi, Xinjiang 830046, People's Republic of China

[2] School of Software, Xinjiang University, No.499, Xibei Road,Saybagh District, Urumqi, Xinjiang 830008, People's Republic of China

[3] Network Center, Xinjiang University, No.666, Shengli Road, Tianshan District, Urumqi, Xinjiang 830046, People's Republic of China

1 Introduction

In daily life, people often express negative meaning with sarcasm, but the literal meaning is affirmative. For example, "You are so right", if this sentence is said with smile and sincere tone, it is the meaning of affirmative praise. However, if it is accompanied by an exaggerated expression, contemptuous smile, slow speaking speed and abnormally high tone, it is likely to be negative and disagreeable. It is flooded with a large number of comments on products, policies and so on in social media. Calculating the true intentions of these data has far-reaching significance. If there is an error in the sarcasm true intention detection, it will have the opposite effect. For example, the original opinion on the product is negative, but the test result is a compliment to the product. In this case, it will only make the situation worse. Therefore, The use of advanced and efficient methods to study sarcasm detection is of great significance.

Sarcasm has metaphorical characteristics [14] and is often associated with wisdom [-34],which makes it difficult to detect [21]. A lot of research on single-modal irony detection has been carried out [19, 30–32], For example, detecting the sarcasm of the text modal by capturing the inconsistency of the text context [34, 39], and detecting whether the facial expression is a sincere smile or a contemptuous fake smile through the analysis of facial expression units. Intuitively, the detection effect of a single modal cannot be better than the effect of comprehensive consideration of multiple modalities [35, 42]. As long as the inconsistency of text, intonation and facial expressions can be detected, the accuracy of detection should be greatly improved [25]. However, the inconsistency between the various modes increases the difficulty of information fusion.

The previous sarcasm detection method was based on rules and statistical knowledge, by manually extracting special vocabulary and punctuation as features to detect sarcasm. However, this method requires professional domain knowledge and is not robust, which means this method needs to design different rules and feature extraction methods for different scenarios, so the cost is high. Deep learning has been successful in other fields. From an intuitive point of view, using deep learning methods can automatically learn useful features for sarcasm detection And when the training samples are comprehensive enough, the deep learning method should be better. Therefore, this paper proposes a multi-level multi-modal fusion network with residual connections on the later fusion method based on deep learning, which improves the accuracy of irony detection on some data sets. The contribution of our work are as follows:

(a) We Proposed a network fusion model with residual connections based on late fusion;
(b) The proposed model was first applied to the data set MUStARD and competitive performance was obtained;
(c) A more reasonable speaker-independent experimental setup is proposed on the data set MUStARD, which laid a foundation for exploring deep multi-modal fusion on a small data set.

The main work of this article is to transfer a variety of multimodal fusion methods to the detection of the MUStard data set, and do sufficient ablation experiments. The data set is small, and the irony detection is very difficult. "It is difficult for a clever woman to cook without rice." Especially in the speaker-independent experimental setting, the improved post-fusion network with residual connections has achieved good results. Next, The rest of this paper is organized into seven sections: The second part is a literature review of irony detection methods. Part 3 is a description of our proposed model. Part 4 introduces the data set and

data preprocessing methods we use, and the method of feature extraction. Part 5 introduces the details of the 3 groups of experimental settings and the baseline models we used. Part 6 shows the detection results of multiple models in 3 sets of experimental settings. Part 7 shows the detection results on multiple other data sets and the analysis of detection errors in the experiment. Part 8 is about future work prospects and improvement directions.

2 Related work

2.1 Multi-modal sarcasm data set

Available multi-modal irony data sets are very rare. Schifanella et al. [33] collected pictures and texts from three social platforms to create a data set, and used SVM to explore multi-modal irony detection for the first time. Castro et al. [5] produced a satirical TV series data set MUStARD, and used SVM for classification after stitching text, audio and image features. However, This method is only an early fusion, and it will inevitably lose the semantic information of different modalities. Moreover, the support vector machine has limited ability to model complex satirical semantics.

2.2 Single mode detection methods

Early sarcasm detection work was mainly focused on a single mode [11, 13], where detection methods were divided into traditional machine learning methods [18] and deep learning methods [48]. Many text modal detection methods are based on recurrent neural networks, such as LSTM [1] and GRU. Until the emergence of Bert [10] achieved the best performance on multiple natural language processing tasks. Through training on a large number of samples, Bert can better capture the deep semantic features of the text. Many image modal detection methods are based on CNN networks [16]. Wang K et al. [37] proposes a simple yet efficient Self-Cure Network to effectively improve the accuracy of facial expression detection. However, the detection of single modalities cannot better solve the irony detection in the actual scene for lacking of other modalities information, so the multi-modal detection method is the trend of future research.

2.3 Multimodal fusion method

The existing multi-modal fusion methods include early fusion [38] [26], which directly cascade the single modal features and send them to the classifier for classification [33]. This method ignores the inconsistencies between the modes. Direct cascading in different semantic spaces will cause the loss of information, and cannot effectively solve the problem of information redundancy and complementarity. Some researchers have proposed a strategy called late fusion [4]. That is, the features of each modal are selected and extracted separately, and then projected to a unified semantic space by a fusion net respectively. This step effectively resolves the inconsistency between the various modes. However, some useful information will inevitably be lost in the respective fusion process. Pereira et al. [28] proposed a new multi-modal method that integrates the facial expressions, audio features, and transcripts of the host and reporter to estimate the degree of tension in news videos, and achieved good results, but the model's ability to fuse semantic information in multiple modalities is limited.

Cai et al. [3] proposed a hierarchical fusion model that combines text features, image features and image attributes. Xu et al. [40] proposed a cross-modal attention semantic comparison and relational network fusion model, in which the decomposition of relational networks can provide relevant evidence for interpreting sarcasm detection. However, these two models do not comprehensively consider the collaborative fusion of the original modal semantic space and the unified semantic space after fusion. Therefore, we propose an improved residual connection fusion algorithm [15]. Our idea is that the internal representation of the subsequent fusion continues to be spliced with the original single-modal feature, and then fused to obtain the best detection performance. Intuitively, the results of multi-modal comprehensive analysis need to be compared with the feature of a single modal, so that collaborative analysis can draw better conclusions. A large number of experiments on the MUStARD data set show that our idea is feasible.

2.4 Multi-task and multi-modal fusion method

Chauhan et al. [7] proposed a multi-modal multi-task learning framework, which uses the attention between tasks and the attention between classes to simulate the relationship between different types of tasks. Similarly, Chauhan et al. [8] used the data set MUStARD and proposed a multi-task learning framework and two attention mechanisms under the condition of marking additional explicit and implicit emotions, proving that emotions and sentiment help to improve the effect of satire detection. Yu et al. [41] used film and television works as a multi-modal data set and proposed a multi-modal multi-task learning framework. By individually labeling each mode, the differences and complementary features between the learning modes are learned. However, the multi-task fusion framework requires additional annotations, which greatly increases the cost of data processing, and these models are not suitable for the detection of a small number of irony data set. Therefore, this paper proposes an improved multi-modal multi-level fusion network with residual connections to improve the accuracy of sarcasm detection.

3 Multi-modal learning framework

As shown in Fig. 1, there are 3 modalities as the input of the model. We take text input as an example. The text is divided into two parts, composed of final sentences and context. We use the pre-trained model Bert [10] to get the word embedding feature of each word and the word embedding of the [CLS] token that representing the semantics of the entire sentence. After that we get the feature vector of (batch_size, seq_length_text, 768). This feature is transformed into a feature vector of (batch_size, 128) by Text Sub-Net. This process means a nonlinear mapping from the text semantic space to the unified first-level fusion space. As shown in Fig. 1, there are a total of 3 levels fusion network. The feature dimension after fusion should be less than that before, so as to achieve the purpose of feature selection and learning deep semantics. In the first-level fusion network, we use a design idea similar to residuals. The 768-dimensional vector that Bert extracted is spliced with the fused 128-dimensional vector. On the one hand, it can avoid over-fitting due to the deepening of the network. On the other hand, this method is intuitively explained by considering the original semantic space of the text modal and the semantic space after fusion will better capture the inconsistency between other two modalities, and improve the accuracy of sarcasm detection. The method in the second-level

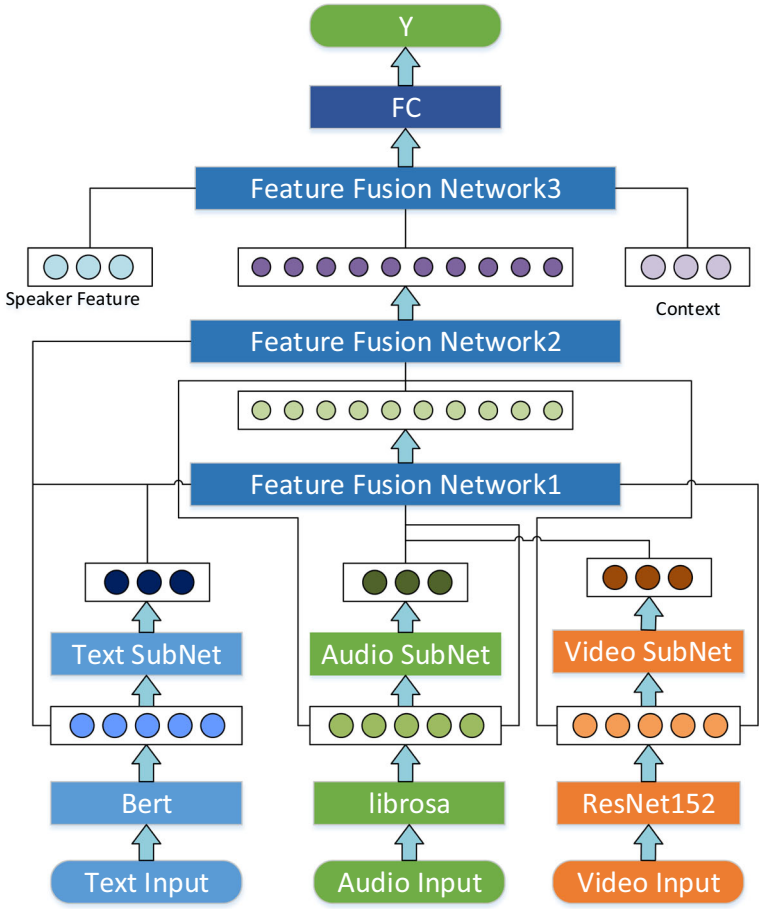

Fig. 1 multimodal fusion framework

fusion is similar to the first-level fusion. In the third level of fusion, we combine contextual features [17] and speaker identity features to further improve the accuracy of sarcasm detection. Obviously, the semantics of sarcasm will be easier to detect with the context of final utterance.

The method of audio and image processing is similar to that of text. The difference is that the feature extraction of audio uses the librosa library [24] to extract (zero_crossing_rate, mfcc, chroma_cqt) which is a total of 33-dimensional features. The video feature extraction step is to first segment the video to multi-frame pictures which are then used to extract image features using the pre-trained model ResNet-152 [15] on ImageNet [9] to obtain a feature vector of (batch_size, seq_len_video, 2048). In order to speed up the training during the experiment, the average of 10 picture frames was used as a window to reduce the sequence length of the feature vector. The experiment was carried out on a rtx3080 GPU. Each model trains taking about a few minutes.

The sub-network of each mode is to get the internal representation of the mode, which can be formalized as:

$$O_u = F_u(I_u)$$

where $I_u \in R^{B\times D_i^u}$, $O_u = R^{B\times D_o^u}$. $F(\bullet)$ is the feature extractor network for modal u.I_u is the input of mode u,where $u \in \{t,\ v,\ a\}$.It is worth noting that the video modal and audio modal are averaged over time steps before being input to the feature extraction sub-net.

As shown in Fig. 2, we show the model diagram of the text sub-network and the first-level fusion network. The fusion network of audio and picture is similar to the text sub-network, and the other two-level fusion network is similar to the first-level fusion network. In the text sub-network, We can obtain the feature vector of (768,) by obtaining the mean value of the word embedding of the sentence. We use BatchNorm to accelerate the convergence of training, use Dropout to avoid overfitting, and use 3-layer DNN and ReLU activation function in the hidden layer of the model for nonlinear features mapping. After that the final feature dimension is reduced to 128 dimensions, and then we feed the 768-dimensional features and 128-dimensional features into the first-level fusion network. It can be seen from the figure that the structure of the first-level fusion network is similar to the text subnet, but the difference is that the post fusion feature is 256 dimensions which is bigger than 128.

The feature fusion network is to obtain the representation after the fusion of the three modes. We use the idea of residual network, which can be formalized as:

$$O_{m1} = G_1(O_t, O_v, O_a, I_t, I_v, I_a)$$

where $O_t, O_v, O_a \in R^{B\times D_o^u}$ are the unimodal representations. $G_1(\bullet)$ is the feature fusion network and O_{m1} is the first-level fusion representation.

$$O_{m2} = G_2(O_{m1}, I_t, I_v, I_a)$$

where O_{m2} is the second-level fusion representation..$G_2(\bullet)$ is the feature fusion network.

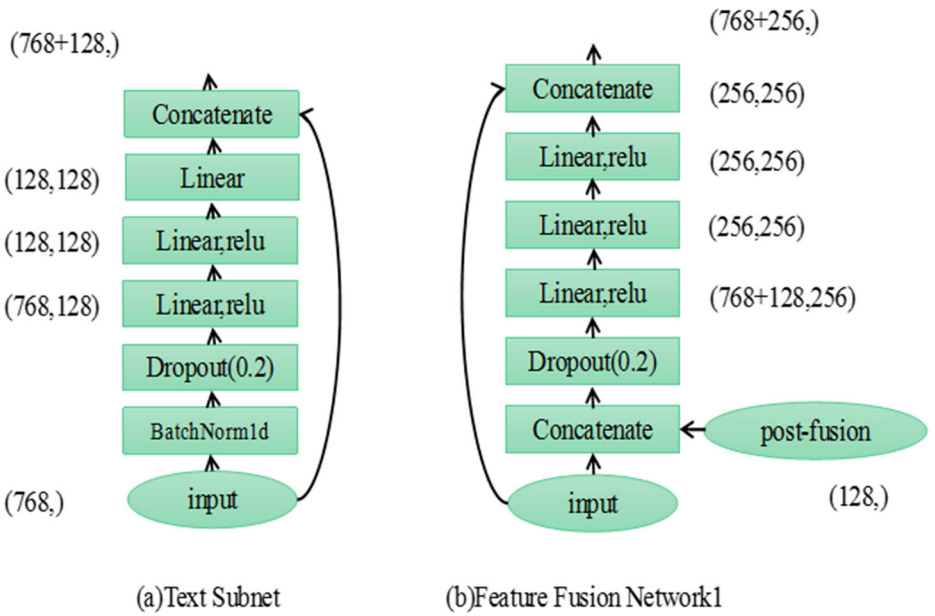

Fig. 2 Sub-Net and Feature Fusion Network Model diagram

$$O_{m3} = G_3(O_{m2}, I_s, O_{tc})$$

where O_{m3} is the last-level fusion representation.I_s is one-hot representation of speaker identifiers,O_{tc} is the context text representation.

4 Feature extraction and dataset

Following the work of [5], the features of the text are extracted using the Bert model, and finally a 768-dimensional vector representation is extracted for each target sentence. The method is that we use the average of the internal representations of the last four transformer [12, 20] layers of the first token[CLS] as the final representation of the sentence. In the later stage of the experiment, we use the last [CLS] token to get the representation of the whole sentence, which is better.

In order to obtain the audio representation, we use the popular librosa library to obtain a fixed-length representation by dividing d_w non-overlapping windows, otherwise the length of the speech will be variable in the data set. In the early stage of the experiment, each window gets a 283 dimensional vector representation, including MFCC, MFCC Delta, Mel spectrogram, Mel spectrogram Delta and spectral central. The final representation of the whole discourse is obtained by the mean of all window vectors. Later, we adopted the second processing method, the effect of detection is better, and set hop_ Length = 5120, select zero_ crossing_ Rate, MFCC, and chroma_ CQT constitutes a 33 dimensional eigenvector.

In order to extract the features of the visual modal, we treat the f frame as a window, and use the representation of the pool5 layer u_i^v of the ResNet-152 model pre-trained on ImageNet to take the average value to obtain a 2048-dimensional vector,where $u_v = \frac{1}{f}\left(\sum_i u_i^v\right) \in R^{d_v}$.

After the feature extraction of text, audio and video modes, in order to batch training, we truncated and padding the training set of 690 data. The final length is composed of the sum of the mean and standard deviation of all data lengths,which is final_len = mean(seq_len) + std.(seq_len)。 In order to visualize the visual model and fine tune the model, we try to design an end-to-end training scheme, that is, extract features from the original video and send them directly to the model training. We found that if the features are not saved to the hard disk in the form of files, the speed of training will be greatly reduced. Because the performance of our training equipment is limited, we choose to extract the features and save them as files. However, the end-to-end scheme can be used to verify the model, that is, inputting a video can quickly judge whether the video contains ironic meaning.

We use the data set MUStARD, which is composed of video clips of 4 TV shows, Friends, The Golden Girls, Sarcasmaholics Anonymous and The Big Bang Theory. There are a total of 690 videos in the data set, of which about half are sarcastic, and about half are not sarcastic. As described in previous work, the detection of sarcasm often requires comprehensive consideration of complementary information from multiple modalities. At the same time, because some characters themselves are set as satirical characters, the speaker's identity characteristics are also conducive to sarcasm detection. In addition, the context of the target sentence will also help the detection of sarcasm [5]. The data set is divided into two parts, final utterance and context. The statistical information is shown in Table 1 [5]. There are 18 speakers in the whole data set.

Table 1 Data set statistics by final utterance and context

statistics	Final utterance	context
Unique words	1991	3205
Avg tokens	14	10
Max tokens	73	71
Avg duration(seconds)	5.22	13.95

5 Experiments

5.1 Experimental setup

Following the work of [5], we set up 4 sets of experiments, the first is speaker-dependent set, where we used a random five-fold crossover experiment, and the sample allocation is random. In other words, the speakers in the training set may also appear in the test set.

The second experimental setting is consistent with the setting in the previous work [5]. All samples belonging to the TV series Friends are used as the test set, and the samples of the other three TV shows are used as the training set, which guarantees the speakers of the training set and the test set will not overlap. However, this division has caused imbalance in training, that is, the ratio of the ironic and non-satiric categories in the training set and the test set is quite different, and the precious samples are not fully utilized.

In the third set of experiments, we also set up a speaker-independent data set partition, and obtained a more reasonable setting to verify the generalization ability of the model.

In the fourth group of experiments, we added the information fusion of the features of contextual text and speaker identity information to test the fusion ability of the model.

For all experimental settings we use weighted Precision, Recall and F-Score values as measurement indicators,where the weights are obtained in proportion to the number of sarcasm and non-sarcasm categories. In the second group of experiments, it is easy to over fit the model because of the imbalance of data set partition and the small proportion of training set and validation set. In order to avoid over fitting, we have reduced the number of layers of the model, removed the components of nonlinear activation function, and used the weighted F-Score as a measure indicator of early stop. In other cases, Because the setting of five fold cross validation experiment means that the division of data set is balanced, there is little difference in whether the measure indicators is weighted or not.

The early stop epoch is set to 20, that is, if the highest weighted F-Score value is still not updated after 20 epochs, the training procedure is stopped. The optimizer, the weight decay value and the learning rate value refer to the work of [41], and have been adjusted and tested during the experiment process. In the training process, because the number of samples in the data set is relatively small, the training has certain fluctuations, but the average training effect is better than the result of the early fusion of SVM.

5.2 Baselines

Majority Following the work of [5],This baseline assigns all the instances to the majority class,i.e.,non-sarcastic.

Random Following the work of [5],This baseline make random predictions across test set.

SVM Following the work of [5],We use early fusion to splice three modal features and feed them into (Support Vector Machines) SVM classifier [27].

EF-LSTM The early fusion lstm model stitches the original input features of the three modalities, and feeds them into LSTM to capture deep semantic features. This method requires the modalities to be aligned in time step.

MFN The memory fusion network [45] first feeds the three modalities into the LSTM unit separately, and obtains the cross-attention internal state of the three modalities with a window of two time steps, and uses a special gating mechanism to obtain the entire semantic features of the sentence.

LMF The Low-rank Multimodal Fusion model adds one-dimensional features to each mode, and uses low rank factor to perform nonlinear transformation of features.

MULT The Multimodal Transformer model uses the other two modalities as a reference which is mapped into the third modality and finally the information interaction of the three modes is obtained by Transformer Encoder.

LF_DNN The model used in the speaker-dependent experimental setting only contains the first-level fusion model and does not use the residual network.

LF_DNNv1 Following the work of [5], because the data set split is imbalanced, in order to avoid over-fitting, only the first-level fusion model is retained, which reduces the number of layers of the fusion network and cancels the nonlinear activation function. The model degenerates to a fully connected neural network without nonlinear activation.

LF_DNNv2 In the speaker-independent experimental setting that we designed, the model needs stronger generalization ability. We propose a three-level fusion network with residual connections. A competitive performance has been achieved, The structure of the model is shown in Fig. 1.

6 Ablation study

As shown in Table 1, among the speaker-dependent settings, Majority's baseline is the worst because of its arbitrary decision. The Random baseline is consistent with our intuition. The prediction accuracy is about 0.5. Generally speaking, the late fusion LF_DNN model is superior to the early fusion model of SVM either in terms of single mode or multi-modal. Overall,the combination of visual,audio and textual signals significantly improves over the uni-modal variants,with a relative error rate reduction up to 17.8%.

For single-mode signals, the amount of information of visual signals is the largest. Because the training set and the test set are randomly disrupted, and the visual features include the scene layout in the TV program, the characters' faces and postures [49], and these features may be well memorized by the neural network in the training set. Therefore,it has better performance than other modalities in the test set. The 2048-dimensional features contain a wealth of

information. Therefore, the learning effect of video modal is fully demonstrated in the later fusion of neural network.

The multi-modal fusion effect of the model is generally higher than the single-modal detection effect, which indicates that the single-modal learning is lacking in the information source. Only by comprehensively considering the three modalities can the detection F-Score value be improved. Let's look at the fusion of T + V mode and the fusion effect of T + V + A mode. This may be because the sound mode has very little supplementary information for the other two modes.what it provides is redundant information. It can also be said that even if it is a silent movie with subtitles, the audience can clearly feel whether the dialogue is ironic or not, which is intuitive.

We use t-SNE [23] to visualize the internal representation of single-mode and multi-mode in the later fusion model. As can be seen from Fig. 3, in the single-mode internal representation, The internal representation of the text modal is not conducive to the detection of sarcasm, while the internal representation of the visual modal is more conducive to the detection of sarcasm. After the fusion of the three modal features, the model eliminates redundant features, merges complementary features, and learns more useful rules to detect sarcasm.

As shown in Table 3, in the speaker-independent experimental settings, the effects of SVM and LF_DNNv1 are generally lower than the performance in Table 2. This is also intuitive, because the division of the data set reduces the number of training sets to 334, while the test set is 356, which is different from the number of divisions of one-fifth of the data in the five-fold cross as the test set. So this is more challenging for the model. And unbalanced data may have little effect on SVM, but it will not give full play to the advantages of neural networks and

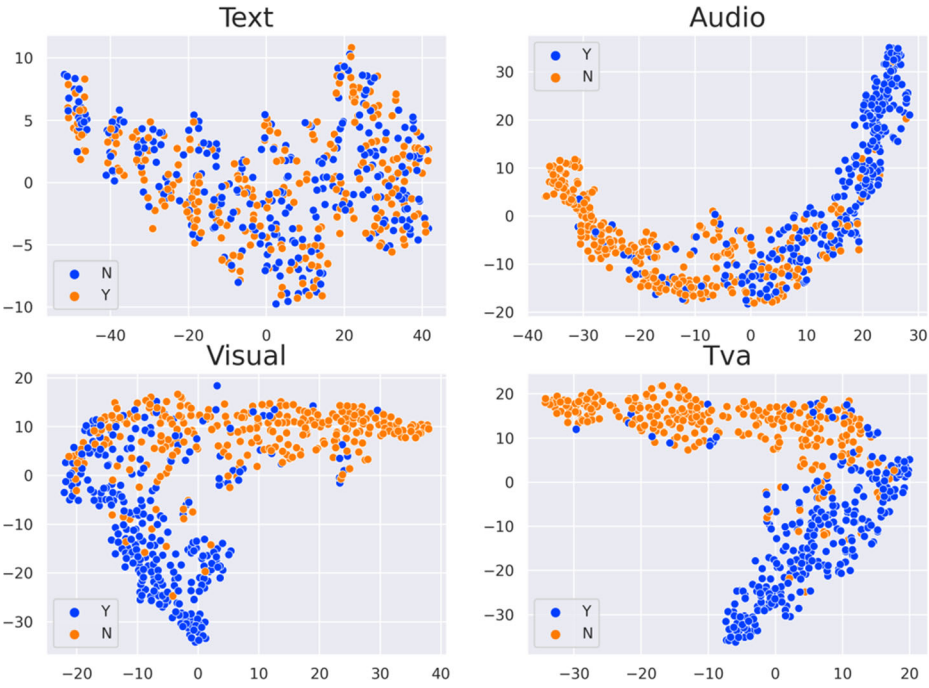

Fig. 3 Visualization of multi-modal fusion, Each sub-picture shows the internal representation of the text, audio, visual and the fusion of the three modes. The Speaker-dependent setup shown in Table 2 is used

Table 2 Speaker-dependent setup. We use five-fold cross-validation, run 5 times and take the average value. Precision, Recall and F-Score are all weighted values. Two values are separated by commas in each cell, the first is the average and the second is the standard deviation. This is the case for the values in the following tables. Δmulti−unimodal is the difference between the highest value of multi-mode and the highest value of single-mode. The Error rate reduction is the reduction of the relative Error rate. Whose value is the result of the difference between the two divided by the absolute error of the single mode, such as 5.7/(100–68.1)

Modality	Algorithm	Precision	Recall	F-Score
–	Majority	25.0	50.0	33.3
–	Random	49.5	49.5	49.8
T	SVM	65.1, 0.0	64.6,0.0	64.6,0.0
	EF-LSTM	69.53,3.85	68.99,3.19	68.83,3.04
	LF_DNN	67.17, 1.69	67.1, 1.75	67.06, 1.78
	LF_DNNV2	67.73,1.76	67.68,1.75	67.66,1.75
A	SVM	65.9,0.0	64.6,0.0	64.6,0.0
	EF-LSTM	74.15,0.91	74.06,0.85	74.03,0.83
	LF_DNN	**73.22, 2.71**	**73.04, 2.57**	**73.0, 2.54**
	LF_DNNV2	**73.22,1.81**	**72.61,1.48**	**72.44,1.44**
V	SVM	**68.1,0.0**	**67.4,0.0**	**67.4,0.0**
	EF-LSTM	71.38,2.46	70.58,2.36	70.3,2.4
	LF_DNN	72.06, 1.9	71.88, 1.86	71.83, 1.85
	LF_DNNV2	71.6,1.67	71.01,1.02	70.84,0.88
T+A	SVM	66.6,0.0	66.2,0.0	66.2,0.0
	EF-LSTM	75.23,1.5	74.06,1.41	73.75,1.53
	MFN	71.38,3.14	70.43,2.35	70.14,2.33
	LMF	70.73,1.7	70.14,1.61	69.94,1.63
	MULT	72.28,2.85	71.59,1.74	71.42,1.5
	LF_DNN	68.65, 1.63	68.55, 1.56	68.51, 1.54
	LF_DNNV2	71.19,2.2	70.87,1.97	70.77,1.92
T+V	SVM	72.0,0.0	71.6,0.0	71.6,0.0
	EF-LSTM	72.6,1.81	72.32,1.91	72.23,1.96
	MFN	70.55,2.51	70.14,2.22	70.01,2.15
	LMF	70.09,3.81	69.71,3.65	69.58,3.62
	MULT	72.19,2.32	71.59,2.35	71.39,2.46
	LF_DNN	72.67, 2.49	72.61, 2.57	72.58, 2.6
	LF_DNNV2	73.44,2.84	73.33,2.91	73.3,2.94
A+V	SVM	66.2,0.0	65.7,0.0	65.7,0.0
	EF-LSTM	75.31,1.48	75.07,1.63	75.01,1.67
	MFN	72.89,1.97	72.61,1.97	72.52,1.99
	LMF	71.55,3.18	71.16,3.02	71.04,3.0
	MULT	72.33,2.44	71.88,2.26	71.75,2.25
	LF_DNN	72.1, 1.5	72.03, 1.42	72.01, 1.4
	LF_DNNV2	73.81,0.88	73.33,1.06	73.19,1.17
T+A+V	SVM	71.9,0.0	71.4,0.0	71.5,0.0
	EF-LSTM	74.34,1.05	74.2,0.98	74.17,0.97
	MFN	72.65,1.02	72.46,0.92	72.41,0.9
	LMF	71.81,2.25	71.3,1.81	71.16,1.7
	MULT	73.07,1.61	72.32,0.96	72.11,0.9
	LF_DNN	**73.82, 3.07**	**73.77, 3.09**	**73.75, 3.1**
	LF_DNNV2	**73.8,1.85**	**73.62,1.75**	**73.58,1.73**
Δmulti−unimodal	–	↑5.7%	↑6.4%	↑6.3%
Error rate reduction	–	↑17.8%	↑19.6%	↑19.3%

deep learning. And because none of the speakers in the test set appeared in the training set, the model is more prone to overfitting, so we used a variant of the model LF_DNNv1 to reduce the number of layers of the fusion network and cancel the Relu activation function. The final experimental effect is better than SVM.

In single modality, the effects of visual modalities are comparable to those of audio modalities, and they are both superior to text modalities, which shows that sound and animation are more expressive than text. The textual modality alone cannot predict sarcasm accurately. The effect of multi-modal fusion is similar to the speaker-dependent experimental settings, and the contribution of audio modalities in the fusion process is relatively low. This may be because the complementary information of the text modality and the picture modality has been better fused by the model, while the information of the audio modality has less complementary information and much redundant information to the other two modalities.

In general, the model variant LF_DNNv1 improves the performance compared with the single-mode best result of SVM, but it still does not reach the speaker-dependent experimental setting. In order to better verify the generalization ability of the model, we analyzed the data set, as shown in Table 4 and Table 5.In the distribution of the data set, the total number of samples belonging to HOWARD and SHELDON in The Big Bang Theory TV show accounted for exactly one-fifth of the total number of data sets.

If the samples belonging to two persons are used as the test set, the ratio of the number of sarcastic samples to the number of non-satiric samples in the test set is approximately 1:1, which happens to be a balanced division. And the ratio of training set to test set is about 4:1, which is similar to the speaker-dependent 5-fold crossover experiment setting. At the same time, HOWARD and SHELDON in the test set will not appear in the training set. This also satisfies the original intention of the speaker-independent experimental setup.

As shown in Table 6, after a more reasonable division of the data set, with sufficient training data, the 3-level fusion model variant LF_DNNv2 with residuals obtained better performance, with a relative error reduction of up to 11.8%. However, there is still a certain gap in comparison with the speaker-dependent experimental settings in Table 2. This is because predicting speakers who have never appeared in the training set are inherently challenging, and the characteristics of the visual modality are the target features for the entire picture and scene, not the facial expressions and postures for the sarcasm.

Compared with the results in Table 3, it can be seen that even for the more complex and fusion-capable LF_DNNv2 model variants, the more reasonable data set division failed to significantly improve the performance of speaker-independent sarcasm detection. This may be because HOWARD and SHELDON have distinctive character settings, The irony habits and methods of scientists and ordinary people are quite different. This also shows that the speaker-independent experiments are challenging, but our proposed model is still better than the early fusion model of SVM.

We report the effects of contextual features and speaker identity features on the best single-modal and multi-modal models, as shown in Table 7. In the Speaker Dependent setting, the visual mode with the best performance is selected, and the multi-modal variant is the combination of textual, audio and visual, which is determined by the results shown in Table 2. In the Speaker Independent setting, we use the experimental results in Table 6, select textual for single mode, and the same setting for multimodal as Speaker-dependent setup.

For the addition of context feature, the performance improvement of the model is not obvious except for the experiment of multi-modal variants in the Speaker Dependent setting. This may be because the features are input into the model after averaging. This process loses the timing information of the context and the target sentence. Because the length of the context and the final sentence is quite different, and we truncate and padding the features respectively, we lose some semantic information. If we change the way to splice the context and the final

Table 3 Speaker-independent setup. Run 5 times and average the results. Following the work of [5] for comparison. But this data set split is very imbalanced

Modality	Algorithm	Precision	Recall	F-Score
–	Majority	32.8	57.3	41.7
–	Random	51.1	50.2	50.4
T	SVM	60.9,0.0	59.6,0.0	59.8,0.0
	EF-LSTM	57.07,1.98	56.85,2.47	56.67,2.22
	LF_DNNv2	58.18, 1.17	57.75, 0.92	57.84, 0.98
	LF_DNNv1	59.44, 1.32	57.25, 1.32	57.31, 1.34
A	SVM	**65.1,0.0**	**62.6,0.0**	**62.7,0.0**
	EF-LSTM	56.33,3.12	54.55,3.82	54.23,4.09
	LF_DNNv2	72.93, 6.17	71.4, 7.34	71.21, 7.27
	LF_DNNv1	44.41, 21.94	54.16, 9.89	43.52, 15.85
V	SVM	54.9,0.0	53.4,0.0	53.6,0.0
	EF-LSTM	64.99,3.17	64.78,3.1	64.59,3.14
	LF_DNNv2	70.37, 0.96	70.56, 0.91	70.13, 1.0
	LF_DNNv1	71.67, 0.67	71.69, 0.76	71.3, 0.81
T+A	SVM	64.7,0.0	62.9,0.0	63.1,0.0
	EF-LSTM	58.17,2.8	56.01,3.64	55.93,3.93
	MFN	56.47,1.98	55.9,1.52	56.02,1.67
	LMF	58.85,3.93	57.81,4.44	57.68,4.34
	MULT	69.77,4.83	69.04,4.4	68.55,4.87
	LF_DNNv2	62.75, 6.66	61.24, 7.47	61.33, 7.39
	LF_DNNv1	58.69, 0.49	56.91, 0.42	57.05, 0.46
T+V	SVM	62.2,0.0	61.5,0.0	61.7,0.0
	EF-LSTM	70.06, 1.71	69.55, 2.03	69.62, 2.01
	MFN	56.65, 1.54	55.9, 1.4	56.06, 1.4
	LMF	70.73, 0.68	70.9, 0.63	70.68, 0.68
	MULT	68.5,4.82	67.75,5.57	67.58,5.32
	LF_DNNv2	64.7, 1.19	64.72, 0.88	63.98, 1.24
	LF_DNNv1	69.83, 2.11	69.78, 2.12	69.01, 2.44
A+V	SVM	64.1,0.0	61.8,0.0	61.9,0.0
	EF-LSTM	64.13, 2.37	63.82, 2.35	63.4, 2.13
	MFN	59.92, 4.28	59.89, 4.62	59.72, 4.25
	LMF	71.35, 0.26	71.4, 0.33	71.04, 0.47
	MULT	65.73, 4.17	65.45, 4.23	64.93, 3.92
	LF_DNNv2	72.44, 1.31	71.57, 1.87	70.27, 2.76
	LF_DNNv1	70.03, 0.75	70.22, 0.71	69.79, 0.83
T+A+V	SVM	64.3,0.0	62.6,0.0	62.8,0.0
	EF-LSTM	67.47, 4.83	66.74, 5.3	66.62, 5.05
	MFN	57.32, 1.34	55.96, 1.66	56.04, 1.7
	LMF	70.46, 2.14	70.34, 2.33	69.92, 2.85
	MULT	65.51,3.3	64.78,4.1	64.49,3.93
	LF_DNNv2	65.95, 4.03	63.88, 1.51	62.3, 1.87
	LF_DNNv1	**71.55, 2.49**	**71.52, 2.19**	**70.99, 2.14**
Δmulti−unimodal	–	↑6.5%	↑8.9%	↑8.3%
Error rate reduction	–	↑18.6%	↑23.8%	↑22.2%

Table 4 The data set division in [5] is unbalanced, so we set our own division

splits	train	test
Non-Sarcastic	141	204
Sarcastic	193	152
total	334	356

Table 5 We choose the samples with two speakers of "HOWARD" and "SHELDON" as the test set to obtain a more reasonable speaker-independent data set division. This division is reasonable and balanced

splits	train	test
Non-Sarcastic	277	68
Sarcastic	277	68
total	554	136

sentence first, and then truncate and fill the length, the effect may be improved. We can do this work in the future.

Table 6 Speaker-independent setup. We ran the experiment 5 times and reported the average of the results. Use our own set of data set division, that is, "HOWARD" and "SHELDON" are used as the test set, and the others are used as the training set for prediction. We extended the work of [5] to prove the generalization ability of our proposed model. The LF_DNNv2 model variant here uses a three-level late fusion structure with residual connections

Modality	Algorithm	Precision	Recall	F-Score
–	Majority	32.8	57.3	41.7
–	Random	51.1	50.2	50.4
T	SVM	58.8,0.0	58.1,0.0	57.3,0.0
	EF-LSTM	60.08, 3.18	59.41, 2.34	58.93, 1.95
	LF_DNNv2	60.03, 2.8	59.85, 2.74	59.67, 2.75
A	SVM	**59.0,0.0**	**56.6,0.0**	**53.5,0.0**
	EF-LSTM	64.79, 2.44	62.5, 2.33	60.98, 2.77
	LF_DNNv2	67.17, 2.78	59.71, 4.06	54.51, 7.29
V	SVM	50.7,0.0	50.7,0.0	50.7,0.0
	EF-LSTM	62.28, 2.96	61.91, 2.65	61.68, 2.5
	LF_DNNv2	**61.83, 0.83**	**61.47, 0.59**	**61.19, 0.68**
T+A	SVM	57.6,0.0	54.4,0.0	49.1,0.0
	EF-LSTM	62.21, 2.2	59.26, 1.28	56.75, 1.77
	MFN	62.1, 2.09	61.18, 1.43	60.51, 1.08
	LMF	57.11, 2.37	56.18, 2.01	54.77, 2.19
	MULT	61.29, 3.94	60.59, 3.85	59.95, 4.13
	LF_DNNv2	62.04, 2.69	61.62, 2.3	61.33, 2.16
T+V	SVM	60.3,0.0	58.8,0.0	57.3,0.0
	EF-LSTM	62.23, 1.44	62.06, 1.58	61.91, 1.73
	MFN	62.73, 2.27	62.35, 2.34	62.04, 2.53
	LMF	59.66, 3.09	58.82, 2.5	58.06, 2.23
	MULT	64.3, 4.04	62.21, 1.84	61.07, 1.15
	LF_DNNv2	63.32, 1.9	62.21, 1.0	61.49, 0.7
A+V	SVM	61.0,0.0	56.6,0.0	51.8,0.0
	EF-LSTM	**65.61, 2.49**	**64.26, 1.65**	**63.57, 1.42**
	MFN	**65.86, 3.45**	**63.53, 2.35**	**62.28, 2.98**
	LMF	61.91, 2.14	60.74, 1.44	59.84, 1.24
	MULT	69.72, 3.6	65.0, 2.26	62.9, 4.17
	LF_DNNv2	63.51, 1.05	63.24, 1.04	63.05, 1.07
T+A+V	SVM	59.9,0.0	56.6,0.0	52.7,0.0
	EF-LSTM	64.38, 1.54	63.53, 1.19	63.01, 1.14
	MFN	59.99, 1.02	59.85, 1.0	59.72, 1.0
	LMF	57.46, 2.26	56.91, 2.16	56.06, 2.41
	MULT	63.59, 2.96	62.79, 2.49	62.32, 2.28
	LF_DNNv2	**63.85, 2.64**	**63.09, 3.1**	**62.45, 3.58**
Δmulti−unimodal	–	↑4.85%	↑6.49%	↑8.95%
Error rate reduction	–	↑11.8%	↑14.9%	↑19.2%

Table 7 Role of text context and Speaker identifier features, Speaker-independent setup that we use our own data set split

Setup	Features	Precision	Recall	F-Score
Speaker Dependent	A	73.22,1.81	72.61,1.48	72.44,1.44
	+ context	68.94, 2.36	68.55, 2.36	68.39, 2.42
	+ speaker	69.9, 3.34	69.71, 3.29	69.64, 3.29
	Best(T+A+V)	**73.8,1.85**	**73.62,1.75**	**73.58,1.73**
	+ context	73.0, 2.4	72.9, 2.49	72.86, 2.53
	+ speaker	72.49, 1.56	72.17, 1.26	72.09, 1.2
Speaker Independent	V	61.83, 0.83	61.47, 0.59	61.19, 0.68
	+ context	59.51, 2.68	59.41, 2.61	59.32, 2.57
	+ speaker	59.16, 2.24	58.97, 2.15	58.79, 2.12
	Best(T+A+V)	**63.85, 2.64**	**63.09, 3.1**	**62.45, 3.58**
	+ context	61.68, 0.96	61.18, 1.08	60.74, 1.27
	+ speaker	61.32,3.89	60.74,3.5	60.31,3.34

For the addition of speaker identification features, the performance of the model in the Speaker Dependent setting has been improved. This is because the identities of the speakers in the training set and the test set overlap, and the model has learned the law of the speaker's sarcasm tendency. In the Speaker Independent setting, because the speaker identification information in the test set has not appeared in the training set, the addition of the speaker identification feature cannot effectively improve the detection effect.

In the speaker-independent setup, we reported some fluctuations in the results of the experiment. This may be because the number of data sets is limited, and samples with scientists' satirical features like 'HOWARD' and 'SHELDON' are less in the training set.

7 Supplementary experiment

In this section, in order to verify the generalization ability of the model, we supplemented the experiments on several data sets. They are:the CMU-MOSI [43, 46], MELD [29], CH-SIMS [41]. Because multimodal data sets are very rare, we use the closest sentiment classification

Table 8 Test results of 4 data sets in 7 models. All results are the average of 5 runs. The best results are shown in bold. There are two values in each cell, separated by commas. The first is the mean value of the binary classification accuracy and the second is the standard deviation. In the speaker-dependent experimental setting, we use five-fold cross-validation, and in the speaker-independent experimental setting, we use our own data set split

dataset	MUStARD		MOSI	MELD	CH-SIMS
model	dependent	independent			
EF-LSTM	**75.07, 2.03**	65.0, 1.36	**76.47, 0.96**	81.77, 2.13	69.37, 0.0
MFN	70.72, 2.54	63.09, 1.27	74.39, 1.08	**83.98, 0.84**	77.86, 0.4
LMF	70.58, 2.62	59.85, 2.31	72.91, 2.91	79.86, 1.89	79.34, 0.4
TFN	72.46, 2.29	60.0, 1.36	77.58, 1.04	83.62, 0.84	**80.66, 1.4**
MULT	69.28, 2.96	59.71, 1.5	75.62, 2.27	83.83, 0.92	77.94, 0.9
LF_DNN	73.0, 2.4	59.41, 2.15	75.8, 0.92	82.29, 2.14	79.87, 0.6
LF_DNNv2	74.45, 2.07	**66.1, 2.15**	71.77, 1.77	83.95, 0.36	77.24, 1.04

Table 9 Example of misclassification corresponding to fig. 5.TRUE means sarcasm, and FALSE means non-sarcasm

segments	right_label	text
1_1180	TRUE	Oh, please. The only way she could make a contribution to science would be if they resumed sending chimps into space.
1_2354	TRUE	So? Do cocaine smugglers write “cocaine” on the box?
1_1189	FALSE	I’m sorry, I am not going back to the Renaissance fair.
1_1484	FALSE	The big deal is that nobody touches food on my plate.

data set for verification. In order to verify the comparison results between the deep learning models, we have added the experimental results of several models, they are: EF-LSTM [38], MFN [47], LMF [22], TFN [44] and MULT [36]. The results are shown in Table 8.

As can be seen from Table 8, no model can perform best on all data sets. Although our model is simple, it has good performance. It is worth noting that MOSI’s data set is time-aligned, so EF-LSTM has played its due advantage. The MULT model is the most time-consuming, but the effect is not good.

As shown in Fig. 4, on the MUSTARD data set, the performance of LF_DNNv2 is better than other models except EF_LSTM. The large fluctuation of the curve is because the data set is too small. We adopt an early stopping strategy, that is, if the test accuracy does not improve after 20 iterations, the training is stopped.

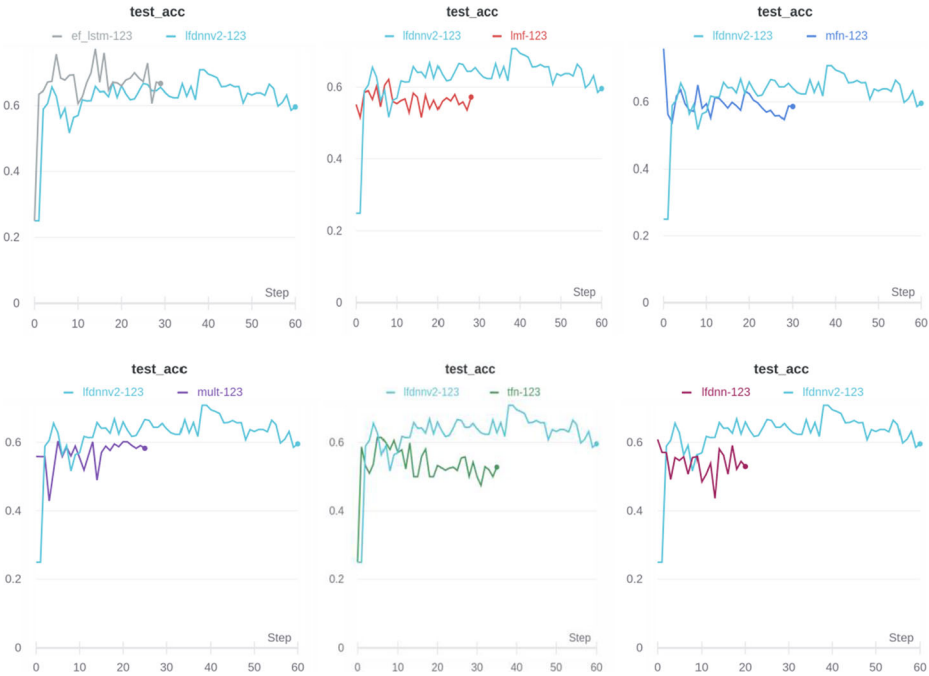

Fig. 4 Comparison of training curves between LF_DNNv2 and the remaining six models in the speaker-independent experimental setting of our split,. The y-axis represents the accuracy of the test set, and the x-axis represents the epoch of the iteration.’123′ represents the random number seed of one of the 5 experiments. We used Weights & Biases [2] for experiment tracking and visualizations to develop insights for this paper

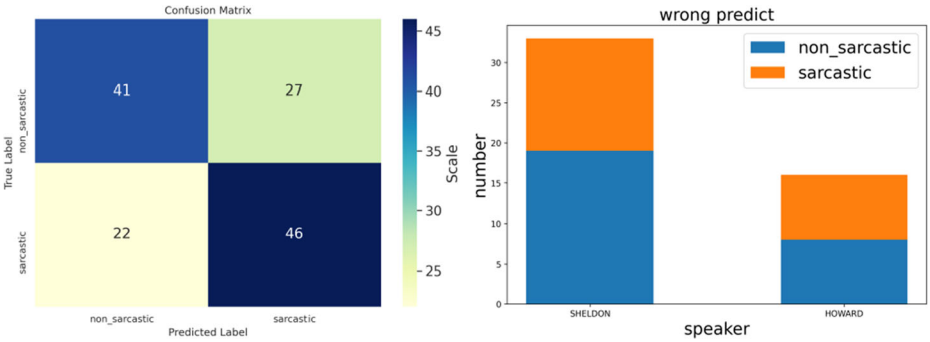

Fig. 5 The left sub graph shows the classification results of the test set of LF_DNNv2 in the speaker-independent experimental setting of our split. The subgraph on the right shows the distribution of classification errors among the two speakers

As shown in Fig. 5, the best LF_DNNv2 still has a lot of room for improvement in the speaker-independent experimental setting of our split. 22 sarcasm samples were misclassified as non-sarcasm, and 27 non-sarcasm samples were misclassified as sarcasm. SHELDON's sample had more classification errors than HOWARD, reaching 33. The number of misclassifications in HOWARD's sample is 16. For SHELDON, more non-satiric samples were misclassified as satire, reaching 19, which may be due to SHELDON's distinctive satire style.

As shown in Fig. 6 and Table 9, in the 4 samples that the model classified incorrectly, SHELDON was always expressionless, speaking faster, and had no change in tone. The meaning of satire in the text is also more abstract and requires additional common sense [6]. Therefore, it is so difficult to classify satire on this small data set.

8 Conclusion and future work

This paper proposes a post-fusion model with a three-level fusion structure and residual network. Experimental results on the MUSTARD data set show that this post-fusion model can better integrate three modalities into a unified semantic space to improve the detection effect of sarcasm. The application of neural network gives full play to the advantages of late fusion, which is better than the SVM model based on early fusion. According to different experimental environments, we propose a later fusion model variant and a more reasonable Speaker-independent data set split, which will help make full use of this data set for research in the future. We analyzed methods to avoid over-fitting and ideas for adjusting model variants according to data set split.

Fig. 6 Examples of misclassification

In the future, we can consider trying more advanced fusion methods, such as adding timing features, and performing feature alignment and fusion on timing. For the model, we can try to change it to end-to-end, based on the pretrained fine-tuned model to further improve the effect of the model. For image features, we can extract the features of face and pose to further improve the effect of the model, instead of using a more general features of the whole screen. The expression of sarcasm may involve the intention and relationship of multiple representatives of the speaker, and the modeling of this intention and relationship can also be added to the multi-modal fusion model. We can also use knowledge graphs to model sarcasm external knowledge bases, and introduce external knowledge into the model fusion process to further improve the performance of multi-modal sarcasm detection. In order to give full play to the advantages of deep learning, we can collect and expand the sarcasm data set for further research. Sarcasm detection is closely related to emotion. We can use deep learning to learn deep semantic knowledge in the data set of emotion detection, and then transfer this knowledge to the current field of sarcasm detection.

Acknowledgements This work was supported by the National Natural Science Foundation of China (61962057), Key Program of National Natural Science Foundation of China (U2003208), the major scientific and technological project of the Autonomous Region, "Research and Development of Key Technologies for Public Health and Epidemics Prevention System in Xinjiang", project number: 2020A03004-4, and Key research and development project of Xinjiang Uygur Autonomous Region (2021B01002). We thank the creators of MUSTARD for their efforts and contributions.

References

1. Baziotis C, Athanasiou N, Papalampidi P, et al. (2018) Ntua-slp at semeval-2018 task 3: Tracking ironic tweets using ensembles of word and character level attentive rnns [J]. arXiv preprint arXiv:1804.06659
2. Biewald L "Experiment Tracking with Weights and Biases," Weights & Biases. [Online]. Available: http://wandb.com/
3. Cai Y, Cai H, Wan X (2019) Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics : 2506–2515
4. Cambria E, Hazarika D, Poria S, et al. (2017) Benchmarking multimodal sentiment analysis [C]//international conference on computational linguistics and intelligent text processing. Springer, Cham, 166–179
5. Castro S, Hazarika D, Pérez-Rosas V, et al. (2019) Towards Multimodal Sarcasm Detection (An _Obviously_ Perfect Paper)[J]. arXiv preprint arXiv:1906.01815
6. Chakrabarty T, Ghosh D, Muresan S, et al. (2020) R3: Reverse, Retrieve, and Rank for Sarcasm Generation with Commonsense Knowledge [J]. arXiv preprint arXiv:2004.13248
7. Chauhan DS, Dhanush SR, Ekbal A, et al. (2020) All-in-One: A Deep Attentive Multi-task Learning Framework for Humour, Sarcasm, Offensive, Motivation, and Sentiment on Memes [C]//Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing : 281–290.
8. Chauhan D S, Dhanush S R, Ekbal A, et al. (2020) Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : 4351–4360.
9. Deng J, Dong W, Socher R et al (2009) Imagenet: a large-scale hierarchical image database [C]//2009 IEEE conference on computer vision and pattern recognition. Ieee:248–255
10. Devlin J, Chang M W, Lee K, et al. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv:1810.04805
11. Diao Y, Lin H, Yang L, Fan X, Chu Y, Xu K, Wu D (2020) A multi-dimension question answering network for sarcasm detection [J]. IEEE Access 8:135152–135161
12. Farha I A, Magdy W (2021) Benchmarking Transformer-based Language Models for Arabic Sentiment and Sarcasm Detection [C]//Proceedings of the Sixth Arabic Natural Language Processing Workshop : 21–31.

13. Felbo B, Mislove A, Søgaard A, et al. (2017) Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm [J]. arXiv preprint arXiv:1708.00524
14. Grice HP (1975) Logic and conversation [M]//speech acts. Brill:41–58
15. He K, Zhang X, Ren S, et al. (2016) Deep residual learning for image recognition [C]//Proceedings of the IEEE conference on computer vision and pattern recognition : 770–778.
16. Jain D, Kumar A, Garg G (2020) Sarcasm detection in mash-up language using soft-attention based bi-directional LSTM and feature-rich CNN [J]. Appl Soft Comput 91:106198
17. Jena A K, Sinha A, Agarwal R (2020) C-net: Contextual network for sarcasm detection [C]//Proceedings of the Second Workshop on Figurative Language Processing : 61–66.
18. Joshi A, Tripathi V, Patel K, et al. (2016) Are word embedding-based features useful for sarcasm detection?[J]. arXiv preprint arXiv:1610.00883
19. Joshi A, Bhattacharyya P, Carman MJ (2017) Automatic sarcasm detection: a survey [J]. ACM Computing Surveys (CSUR) 50(5):1–22
20. Kumar A, Narapareddy VT, Srikanth VA et al (2020) Sarcasm detection using multi-head attention based bidirectional LSTM [J]. Ieee Access 8:6388–6397
21. Liu B (2010) Sentiment analysis and subjectivity [J]. Handbook of natural language processing 2(2010): 627–666
22. Liu Z, Shen Y, Lakshminarasimhan V B, et al. (2018) Efficient low-rank multimodal fusion with modality-specific factors [J]. arXiv preprint arXiv:1806.00064
23. Maaten L, Hinton G (2008) Visualizing data using t-SNE [J]. J Mach Learn Res 9(Nov):2579–2605
24. Brian McFee, Matt McVicar, Stefan Balke, Carl Thomé, Vincent Lostanlen, Colin Raffel, Dana Lee, Oriol Nieto, Eric Battenberg, Dan Ellis, Ryuichi Yamamoto, Josh Moore, WZY , Rachel Bittner, Keunwoo Choi, Pius Friesch, Fabian-Robert Stöter, Matt V ollrath, Siddhartha Kumar, nehz, Simon Waloschek, Seth, Rimvydas Naktinis, Douglas Repetto, Curtis "Fjord" Hawthorne, CJ Carr,João Felipe Santos, JackieWu, Erik, and Adrian Holovaty (2018) librosa/librosa: 0.6.2.
25. Mishra A, Kanojia D, Bhattacharyya P (2016) Predicting readers' sarcasm understandability by modeling gaze behavior [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 30(1).
26. Paul S, Saha S, Hasanuzzaman M (2020) Identification of cyberbullying: a deep learning based multimodal approach. Multimed Tools Appl. https://doi.org/10.1007/s11042-020-09631-w
27. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in Python [J]. The. J Mach Learn Res 12:2825–2830
28. Pereira MHR, Pádua FLC, Dalip DH, Benevenuto F, Pereira ACM, Lacerda AM (2019) Multimodal approach for tension levels estimation in news videos. Multimed Tools Appl 78:23783–23808. https://doi.org/10.1007/s11042-019-7691-4
29. Poria S, Hazarika D, Majumder N, et al. (2018) Meld: A multimodal multi-party dataset for emotion recognition in conversations [J]. arXiv preprint arXiv:1810.02508
30. Poria S, Majumder N, Mihalcea R, Hovy E (2019) Emotion recognition in conversation: research challenges, datasets, and recent advances [J]. IEEE Access 7:100943–100953
31. Ren L, Xu B, Lin H, Liu X, Yang L (2020) Sarcasm detection with sentiment semantics enhanced multi-level memory network [J]. Neurocomputing 401:320–326
32. Sarsam SM, Al-Samarraie H, Alzahrani AI et al (2020) Sarcasm detection using machine learning algorithms in twitter: a systematic review [J]. Int J Mark Res 62(5):578–598
33. Schifanella R, de Juan P, Tetreault J, et al. (2016) Detecting sarcasm in multimodal social platforms [C]// Proceedings of the 24th ACM international conference on Multimedia : 1136–1145.
34. Tay Y, Tuan L A, Hui S C, et al. (2018) Reasoning with sarcasm by reading in-between [J]. arXiv preprint arXiv:1805.02856
35. Tembhurne JV, Diwan T (2020) Sentiment analysis in textual, visual and multimodal inputs using recurrent neural networks. Multimed Tools Appl 80:6871–6910. https://doi.org/10.1007/s11042-020-10037-x
36. Tsai YHH, Bai S, Liang PP et al (2019) Multimodal transformer for unaligned multimodal language sequences [C]//proceedings of the conference. Association for Computational Linguistics Meeting NIH Public Access 2019:6558
37. Wang K, Peng X, Yang J, et al. (2020) Suppressing uncertainties for large-scale facial expression recognition [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. : 6897–6906.
38. Williams J, Kleinegesse S, Comanescu R, et al. (2018) Recognizing emotions in video using multimodal DNN feature fusion [C]//Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML) : 11–19.
39. Xiong T, Zhang P, Zhu H, et al. (2019) Sarcasm Detection with Self-matching Networks and Low-rank Bilinear Pooling [C]//The World Wide Web Conference : 2115–2124

40. Xu N, Zeng Z, Mao W (2020) Reasoning with Multimodal Sarcastic Tweets via Modeling Cross-Modality Contrast and Semantic Association [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : 3777–3786
41. Yu W, Xu H, Meng F, et al. (2020) Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : 3718–3727
42. Yu W, Xu H, Yuan Z, et al. (2021) Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis [J]. arXiv preprint arXiv:2102.04830
43. Zadeh A, Rowan Zellers, Eli Pincus, and LouisPhilippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos.arXiv preprint arXiv:1606.06259.
44. Zadeh A, Chen M, Poria S, et al. (2017) Tensor fusion network for multimodal sentiment analysis [J]. arXiv preprint arXiv:1707.07250
45. Zadeh A, Liang P P, Mazumder N, et al. (2018) Memory fusion network for multi-view sequential learning [C]//Proceedings of the AAAI Conference on Artificial Intelligence. 32(1)
46. Zadeh A A B, Liang P P, Poria S, et al. (2018) Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). : 2236–2246.
47. Zadeh A, Liang P P, Mazumder N, et al. (2018) Memory fusion network for multi-view sequential learning [C]//Proceedings of the AAAI Conference on Artificial Intelligence. 32(1).
48. Zhang M, Zhang Y, Fu G (2016) Tweet sarcasm detection using deep neural network [C]//Proceedings of COLING 2016, The 26th International Conference on Computational Linguistics: Technical Papers : 2449–2460.
49. Zhang K, Zhang Z, Li Z, Qiao Y (2016) Joint face detection and alignment using multitask cascaded convolutional networks [J]. IEEE Signal Processing Letters 23(10):1499–1503