Advances in Artificial Intelligence, Big Data and Algorithms
G. Grigoras and P. Lorenz (Eds.)

doi:10.3233/FAIA230853

Multi-Modal Sarcasm Detection via Cross-Modal Attention Mechanism

Yueying LI, Hui CAO[1], Xiaotian XIA and Quan SONG
Northwest University for Nationalities, China

Abstract. With the popularity of social media, the sarcasm figure of speech has become a common phenomenon on social media platforms, and many studies have utilized text and visual information for multimodal sarcasm detection. This paper use a method based on cross modal attention mechanism. Specifically, the paper extract multimodal features firstly, use attention mechanism to focus on the inconsistent information between modalities, and then use the inconsistent information for prediction. The experimental results show that the performance of this paper is improved on the Twitters sarcasm dataset.

Keywords. Sarcasm detection, Multi-modal analysis, Attention mechanism

1. Introduction

Sarcasm is a kind of ironic tone or writing technique in speech or writing, which cannot be understood from the words alone, but in fact its original meaning is exactly the opposite of the meaning that can be understood literally, and usually needs to be understood from the context and background. sarcasm recognition is important in public opinion mining, social opinion analysis and other sentiment analysis tasks. With the development of social science and technology, people can easily express their opinions on social media by choosing to use text, audio and video. People post many multimodal messages combining text and images on social media. The study of irony recognition using multimodal information has also become more important. Therefore, many methods have been proposed to detect irony using multimodal information.

The use of multimodal information for sarcasm detection has been the focus of research in recent years. At this stage, the multimodal sarcasm task mainly faces two problems, one is the integrity and authority of the dataset, and the other is how to better fuse the information features from different modalities. In 2019, Cai et al.[1] proposed a new multimodal sarcasm dataset containing both image and text modal information. Also in the paper, experiments on multimodal sarcasm are constructed, and a multimodal hierarchical fusion model is proposed experimentally. The model first extracts the features of three modalities: text, image features, and image attributes, then reconstructs the features of these three modalities, and finally fuses them into a feature vector for prediction. Xu et al.[2] proposed a method to model the comparison between different modalities in relevant contexts. The method models the contrast between two modalities and the association between semantics by constructing a decomposition relation network,

[1] Corresponding author: Hui CAO, Northwest University for Nationalities;
Telephone number: 13893694060, e-mail: caohui@xbmu.edu.cn

respectively. The decomposition network is able to obtain commonalities as well as differences between images and texts, while the relational network is able to obtain semantic associations in contextual association contexts. Sangwan et al.[3] proposed a recursive neural network-based approach, which mainly uses textual and visual information for multimodal sarcasm detection. Pan et al.[4] proposed a model based on the BERT architecture, which first extracts image features and text features, and then input both features into the improved BERT model as a way to extract the inconsistent information between different modalities. A common attention mechanism is also used to extract the inconsistent information within the text, and finally these inconsistent feature information is combined for prediction. Liu et al.[5] proposed a new hierarchical framework for sarcasm detection. The framework uses a multi-headed cross-attention mechanism and graph neural network for extracting atomic-level consistent information and combinatorial-level consistent information, respectively.

2. Method

This paper comprises four main sections: feature representation of images and text, text-image attention network layer, image-text attention network layer, and feature fusion and classification output part. The specific model framework is depicted in Figure 1.

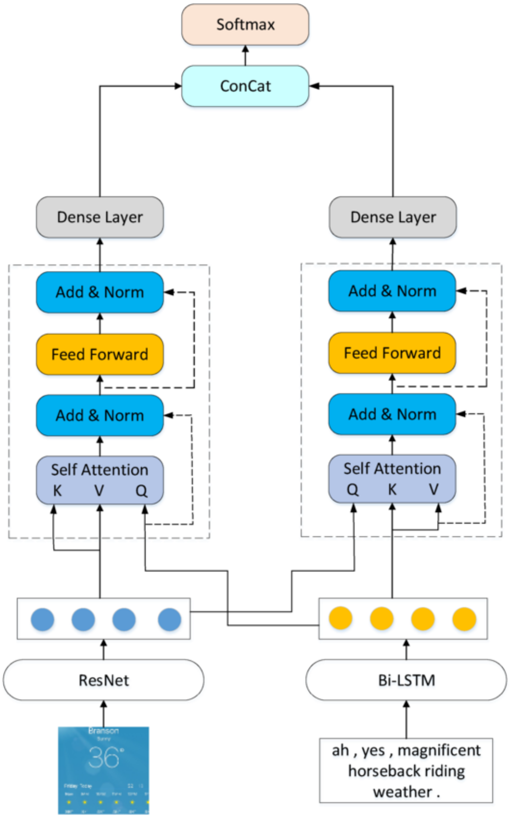

Figure 1. Model diagram

2.1. Feature Representation of Images and Text

To process text, considering a sequence of words $X = \{x_1, x_2, \ldots, \ x_n\}$, where $x_i \in \mathbb{R}^d$ represents the combined embedding of words, segments, and positions, n represents the maximum length of the sequence, and d denotes the embedding size. Initially, the pre-trained BERT[6] model is utilized to acquire the textual representation, followed by employing Bi-LSTM[7] to capture contextual semantic information. The resulting encoded text can be represented as $H^T \in \mathbb{R}^{2dn}$.

Regarding image processing, for a given image I, the first step is to resize it to 224 * 224 pixels. Subsequently, the ResNet-152[8] model is employed to extract the image representation.

$$ResNet(I) = \{r_i | r_i \in R^{2048}\} \tag{1}$$

each $r_i$corresponds to a 2048-dimensional vector that represents a specific region within the image. Consequently, an image I can be represented as $ResNet(I) \in \mathbb{R}^{2048*49}$. Eventually, the encoded image representation $ResNet(I)$ undergoes a linear transformation to project the visual features onto the same dimensions as the text features:

$$H^I = W_v ResNet(I) \tag{2}$$

$W_v \in \mathbb{R}^{d*2048}$ represents a trainable parameter, where d denotes the dimensionality of the text features. $H^I \in \mathbb{R}^{d*49}$is an encoded representation of a visual feature.

2.2. Text-image attention network layer

Inconsistency between modalities plays a vital role in sarcasm detection. Hence, inspired by the concept of self-attention mechanism, we devised a text-image attention network layer to capture such inconsistency between text and images. This network layer employs text features H^T as the query (Q) and image features as the key (K) and value (V). Consequently, the text features guide the model's attention towards the inconsistent regions of the image. For the ith head of the text-image attention network layer, this can be mathematically represented as:

$$head_i = Att_i(H^T, H^I) \tag{3}$$

$$Att_i(H^T : H^I) = softmax\left(\frac{\left[W_i^Q H^T\right]^T \left[W_i^K H^I\right]}{\sqrt{d_k}}\right)[W_i^V H^I]^T \tag{4}$$

where $d_k \in \mathbb{R}^{d/h}$, $head_i \in \mathbb{R}^{N*d_k}$, $\{W_i^Q, W_i^K, W_i^V\} \in \mathbb{R}^{d_k*d}$, is the matrix of parameters to be learned under the attention vector $head_i$.The outputs from the h heads are subsequently concatenated and subject to a linear transformation, resulting in the following expression:

$$Z = [head_1, head_2, \ldots, head_h]W^o \tag{5}$$

where $W^o \in \mathbb{R}^{d*d}$, is the matrix of parameters to be learned. After the vector Z has been calculated by the self-attentive mechanism network, it still needs to go through a series of calculations such as the normalisation layer within the Transformer encoder and the feed-forward layer network layer.The final expression can be described as:$H_{TI} \in \mathbb{R}^d$。

2.3. *Image-text Attention Network Layer*

The image-text attention network layer also borrows ideas from the self-attention mechanism, with the text-image attention network layer accepting image features as Q and text features as K and V. This allows the image features to guide the model to focus more on regions of text that do not match it. The specific computation process is similar to that of the text-image attention network layer. The resulting representation can be described as $H_{IT} \in \mathbb{R}^d$.

2.4. *Feature Fusion and Classification Output*

After the processing of text-image matching layer and image-text matching layer, two inconsistency vectors H_{TI} and H_{IT} are obtained,They are connected for the prediction of the output. The prediction part mainly consists of a linear layer for dimensionality reduction and a $Softmax$ classifier, and the probability distribution $\hat{p}$ is obtained after the operation, which is calculated as follows:

$$\hat{p} = Softmax(W[H_{TI}:H_{IT}] + b) \tag{6}$$

where $W \in \mathbb{R}^{2d}$, is the matrix of parameters to be learned in the model and b is the bias term to be learned.

In this paper, the cross-entropy loss function is used to optimize the model, and the calculation process is as follows:

$$J = -\sum_{i=1}^{N}[p_i \log \widehat{p_i} + (1 - p_i) \log(1 - \widehat{p_i})] + \lambda R \tag{7}$$

where $\widehat{p_i}$ represents the predicted result, p_i denotes the true result.

3. Experimental Setup

3.1. *Dataset*

The sarcasm dataset utilized for the experiments is derived from the Twitters platform. This dataset comprises both ironic and unironic categories, with each sample containing text and associated images. To facilitate experimentation, the dataset is split into training, test, and validation sets in an 8:1:1 ratio. Table 1 illustrates the distribution of specific categories in terms of the number of samples:

Table 1. Distribution of data sets

	Training	Testing	Development
Positive	8642	959	959
Negative	11174	1451	1450
All	19816	2410	2409

3.2. Baseline Models

The model proposed in this paper will be compared with the following model, which is detailed as follows:

(1) Text modality models:

BERT: BERT vectors are used as pretraining and hyperparameters are set to represent the text.

Bi-LSTM: Learning features of text with bidirectional LSTM networks and then performing classification of sarcasm.

(2) Image modality models:

ResNet: Feature extraction using ResNet for sarcasm detection.

(3) Multimodal:

HFM: A hierarchical multimodal sarcasm detection model based on multimodal feature fusion.

D&R Net: Decomposition and relational network modeling of cross-modal comparisons and semantic associations.

4. Experimental Results

4.1. Analysis of Results

Table 2. Algorithm performance comparison results

Modality	Method	Accuracy(%)	Precision(%)	Recall(%)	F1-score(%)
Text	BERT	83.85	78.72	82.27	80.22
	Bi-LSTM	81.90	76.66	78.42	77.53
Image	ResNet	64.76	54.41	70.80	61.53
Text+Image	HFM	83.44	76.57	84.15	80.18
	D&R Net	84.02	77.97	83.42	80.60
	This paper	**86.40**	**81.22**	**85.43**	**83.27**

Table 2 shows this paper's experiments compared with other models proposed with different models containing text, images and multimodality. The experimental results prove that the results of this study reach the optimum and outperform other models. Specifically, the F1 value of the model proposed in this paper is improved by 2.67% compared to the D&R Net model. These results validate the superiority of the model proposed in this paper in capturing multimodal features, and it makes full use of the cross-modal intermodal features of the model to improve the recognition of sarcasm.

4.2. Ablation Experiments

Table 3. Results of ablation experiments

Model	Accuracy(%)	Precision(%)	Recall(%)	F1-score(%)
Model(w\o T-I)	85.15	79.77	83.93	81.80
Model(w\o I-T)	86.15	80.97	85.18	83.02
This paper	**86.40**	**81.22**	**85.43**	**83.27**

Based on the cross-modal deep learning model for sarcasm recognition proposed in this paper, ablation experiments are conducted in this paper for the different effects of the image-text attention network layer and the text-image attention network layer on the model, respectively. Table 3 shows the results of the ablation experiments. It can be observed that the result of subtracting the text-image-attention layer in cross-modality is 81.80%, indicating that the text-image-attention layer plays an important role in the extraction of important features in multimodality, focusing on capturing the interaction information between text and images. Removing the image-text attention network layer also has a bad effect on the model results, which indicates that the image-text attention network layer also contributes to the acquisition of inter-modal inconsistencies.

5. Conclusion

In this paper, a multimodal sarcasm detection model based on cross-modal attention is proposed to extract inconsistent information features between text and images using the cross-modal attention mechanism. The experimental results show that the structure proposed in this study is superior compared to the baseline approach and shows excellent performance in the irony detection task. In addition, the model is also applicable to other multimodal sentiment analysis tasks, and can provide a reliable theoretical basis for the government in public opinion research and judgment of hot events, and valuable reference opinions for enterprises to improve their services.

Acknowledgement

The paper is supported by the project “Graduate Research and Innovation Fund of China National Information Technology Research Institute (31920230047). “ of Fundamental Research Funds for the Central Universities.

References

[1] Cai Y, Cai H, Wan X. Multi-modal sarcasm detection in twitter with hierarchical fusion model[C]//Proceedings of the 57th annual meeting of the association for computational linguistics. 2019: 2506-2515.

[2] Xu N, Zeng Z, Mao W. 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. 2020: 3777-3786.

[3] Sangwan S, Akhtar M S, Behera P, et al. I didn’t mean what I wrote! Exploring Multimodality for Sarcasm Detection[C]//2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020: 1-8.

[4] Pan H, Lin Z, Fu P, et al. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. 2020: 1383-1392.

[5] Liu H, Wang W, Li H. Towards Multi-Modal Sarcasm Detection via Hierarchical Congruity Modeling with Knowledge Enhancement[J]. arXiv preprint arXiv:2210.03501, 2022.

[6] Kenton J D M W C, Toutanova L K. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of naacL-HLT. 2019, 1: 2.

[7] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.

[8] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.