Fusion and Discrimination: A Multimodal Graph Contrastive Learning Framework for Multimodal Sarcasm Detection

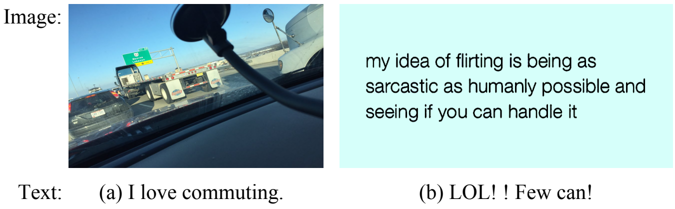
Bin Liang, Lin Gui, Yulan He, Erik Cambria, *Fellow, IEEE*, and Ruifeng Xu*, *Member, IEEE*

Abstract—Identifying the sarcastic cues from both textual and visual information has become an important research issue, called Multimodal Sarcasm Detection. In this paper, we investigate multimodal sarcasm detection from a novel perspective, where a multimodal graph contrastive learning strategy is proposed to fuse and distinguish the sarcastic cues from both text- and image-modality. Specifically, we first utilize object detection to derive the crucial visual regions accompanied by their captions of the images, which allows better learning of the image-modality and better modeling of the relations between visual regions. In addition, to make full use of the semantic information of the image-modality, we employ optical character recognition to extract the textual descriptions in the images. Then, based on image regions and the textual descriptions of the image-modality and the context of the text-modality, we build a multimodal graph for each instance to model the intricate relations and incongruous sentiment expressions between modalities. Furthermore, we devise a graph-oriented contrastive learning strategy to leverage the correlations in the same label and differences between different labels, so as to capture better multimodal representations for multimodal sarcasm detection. Extensive experiments and in-depth analysis show that our method outperforms the state-of-the-art models in multimodal sarcasm detection.

Index Terms—Multimodal sarcasm detection, sarcasm detection, graph model, contrastive learning.

1 Introduction

Sarcasm is a peculiar form and sophisticated linguistic phenomenon of language behavior, where people express ironic sentiment or intention that is opposite to the authentic/apparent sentiment [1], [2], [3], [4]. Merriam-Webster[1] defines sarcasm as *"the use of words that mean the opposite of what you really want to say especially in order to insult someone, to show irritation, or to be funny."*. Sarcasm is popular on social media platforms, which may mislead the prediction of sentiment analysis methods. Therefore, detecting the sarcastic expression of the social media data is a recommendable strategy to improve the performance of sentiment analysis and opinion mining. Early research work usually focused on the problem of text-only stance detection [5], [6], [7]. In recent years, with the development of social media, more and more people tend to post multimodal messages on social platforms.



my idea of flirting is being as sarcastic as humanly possible and seeing if you can handle it

(a) I love commuting. (b) LOL! ! Few can!

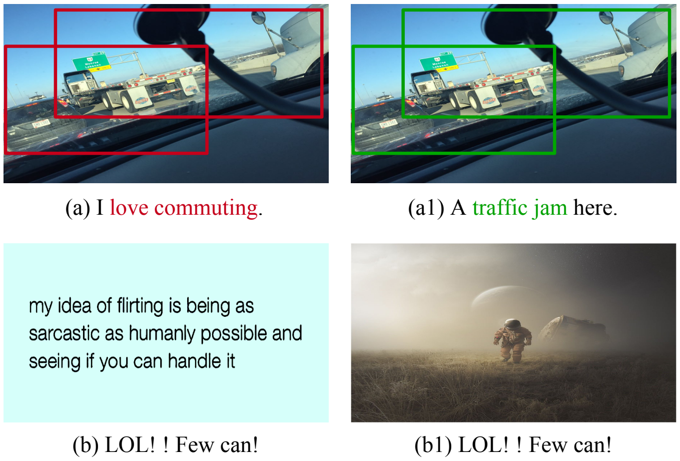
Fig. 1. Examples of multimodal sarcastic expressions.

Therefore, many recent studies have been carried out learning multimodal features for the social media data. Corresponding to the sarcasm detection, study on multimodal sarcasm detection for text and image has received much research attention [8], [9]. Unlike sarcasm detection purely based on text, multimodal sarcasm detection towards text- and image-modality requires a reasoning of the sarcastic cues from both the two modalities [8]. As two sarcastic examples shown in Figure 1, owing to the the phrase *"love commuting"* in the context of Figure 1 (a) and the word *"LOL"* in the context of Figure 1 (b), text-only sarcasm detection methods may erroneously identify the two examples as a positive sentiment polarity. However, combining the information in the image-modality, both these two cases contain a sarcastic expression with negative sentiment, because they are accompanied by an image with a visual information of *"traffic jam"* or an image with textual descriptions of *"flirting is being as sarcastic as humanly possible"*. Therefore, understanding the relations and information from both the text and image modalities is a key issue for multimodal sarcasm detection.

- *B. Liang is with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China, Joint Lab of HITSZ and China Merchants Securities, Shenzhen 518055, China. E-mail: bin.liang@stu.hit.edu.cn;*
- *G. Lin, Y. He are with the Department of Computer Science, University of Warwick, Coventry CV47AL, U.K. E-mail: {lin.gui, Yulan.He}@warwick.ac.uk;*
- *E. Cambria is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. E-mail: cambria@ntu.edu.sg;*
- *R. Xu (*corresponding author) is with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China, Peng Cheng Laboratory, Shenzhen 518055, China. E-mail: xuruifeng@hit.edu.cn.*

Manuscript received April 19, 2005; revised August 26, 2015.

[1] http://www.merriam-webster.com/dictionary/sarcasm



(a) I love commuting.

(a1) A traffic jam here.

my idea of flirting is being as sarcastic as humanly possible and seeing if you can handle it

(b) LOL! ! Few can!

(b1) LOL! ! Few can!

Fig. 2. Examples of two multimodal sarcastic instances accompanied with the crucial visual regions ((a) and (b)). Examples of (a1) and (b1) are the variants of non-sarcastic instances with respect to (a) and (b), respectively. Boxes and words in the same color are highly correlated sarcastic/non-sarcastic expressions.

To deal with multimodal sarcasm detection on multimodal instances that are composed of text and image modalities, existing research studies generally attempt to combine the features from text and image modalities to learn the multimodal sarcastic clues based on features concatenation [8], incorporate external knowledge, such as image attributes [9], adjective-noun pairs (ANPs) [10] or hash tags [11] to implicitly fuse the features of different modalities with attention mechanism, or explore graph neural model to learn the relationships between different modalities [12]. Despite promising progress made by the above research studies, they still suffer from the following limitations:

- For image-modality, modeling the image purely from the visual information may be inadequate to reason the semantic information of satirical expression. As the sarcastic example shown in Figure 2 (a), accurately extracting the crucial information in the red boxes and identifying it as "*a traffic jam*" is the key to correctly detecting sarcasm of this example.
- Simply understanding the visual information of the images may lead to a poor performance, since as the example shown in Figure 1 (b), crucial satirical clues in some images may be presented in the form of text description. Therefore, leveraging the textual descriptions of the image-modality is significant to improve the learning of visual information.
- As the example shown in Figure 2 (a), the key visual information, which is associated with different words in the text-modality, may be overlapping or scattered in an image. Therefore, it is imperative to model the intricate relationships between modalities and learn the incongruent sentiment expressions between the key visual information "*a traffic jam*" and the crucial textual words "*love commuting*".
- The combination of the same/similar image with different texts (Figure 2 (a) → Figure 2 (a1)), or the same/similar text with different images (Figure 2 (b) → Figure 2 (b1)), may lead to different sarcasm detection results. Therefore, discriminating the representations of different multimodal instances in the latent space may lead to improved learning of sarcastic clues in multimodal sarcasm detection.

To address the above limitations, we propose a novel Multimodal Graph Contrastive Learning (`MMGCL`) framework to deal with the multimodal sarcasm detection task, in which the crucial information of image-modality can be explicitly linked with the associated words of text-modality to model the multimodal relations with a multimodal graph structure and distinguish the multimodal representations with a graph-oriented contrastive learning strategy To be specific, **1) for the limitation#1**, we employ the object detection model proposed by [13] to capture the crucial visual regions and the corresponding *attribute-object* pairs for each image. **2) for the limitation#2**, we utilize an awesome open-source optical character recognition (OCR) toolkit PaddleOCR[2] to extract the textual descriptions presented in the images, allowing more effective learning of images' semantic information for sarcastic clues extraction. **3) for the limitation#3**, we build a multimodal graph for each instance based on the image regions derived by object detection (called visual mode), the textual descriptions derived by OCR (called OCR mode), and the context from the text-modality. More concretely, based on the results of Part-of-Speech Tagging, we explore a novel strategy to assign weights of the edges in the multimodal graph. On the one hand, we compute the word similarity across the contextual nouns of text-modality, the nouns of OCR mode, and the *object* caption of visual mode to effectively model the intricate relations between modalities. On the other hand, to leverage the incongruous sentiment expressions of sarcastic clues, we compute the sentiment inconsistency scores across the contextual adjectives of text-modality, the adjectives of OCR mode, and the *attribute* caption of visual mode based on the external affective knowledge (SenticNet [14]). **4) for the limitation#4**, we devise a contrastive loss towards the graph-oriented multimodal representations of the multimodal instances to pull together the clusters of representations belonging to the same label in embedding space, and simultaneously push apart representations of samples from different labels. This essentially allows the learning of the correlation of features in the same sarcasm label and the difference of features between sarcasm and non-sarcasm expressions, enabling model to derive better multimodal representations for multimodal sarcasm detection.

The main contributions of our work can be summarized as follows:

- To the best of our knowledge, we are the first to explore the use of multimodal graph contrastive learning for modeling the intricate relations between modalities and distinguishing the multimodal representations between classes in the multimodal sarcasm detection task.
- We exploit the visual regions' captions of object detection and the textual descriptions of OCR results to improve the learning of image-modality, and then model the multimodal relations and contradictory

[2] https://github.com/PaddlePaddle/PaddleOCR

sentiments between textual and visual information based on the external knowledge sources for learning sarcastic clues.

- We explore a multimodal graph-oriented contrastive learning strategy to distinguish the similarity between examples in one class and difference between examples in other classes, so as to improve the multimodal representations for extracting sarcastic features in multimodal sarcasm detection.
- Extensive experiments and in-deep analysis on a public multimodal sarcasm detection benchmark dataset show that our proposed method outperforms the state-of-the-art baselines in the multimodal sarcasm detection task.

2 Related Work

2.1 Textual Sarcasm Detection

Textual sarcasm detection aims to detect the sarcastic expression of a context purely from the text-modality [6]. Some early studies extracted the contextual incongruity expression with feature engineering approaches, such as searching a set of positive verbs and negative situations [5], [15] or building lexical indicators [16], [17], [18] to produce the sarcastic patterns. Further, neural network-based methods are widely used in textual sarcasm detection [7], [19] due to less feature engineering. Such as [20], [21] adopted a pre-trained convolutional neural networks (CNNs) architecture to extract sentiment, emotion and personality features for sarcasm detection. [19] utilize a bi-directional gated recurrent neural network and a pooling neural network to respectively capture tweet content and contextual information for tweet sarcasm detection. [22] modeled the contrast and incongruity sentiment of the context with an attention-based neural model, [7] learned the incongruity expressions between the word pairs with a deep neural architecture fused by self-matching network, bi-directional long short-term memory (Bi-LSTM), and Low-rank Bilinear Pooling method. In addition, [23] designed an affective dependency graph network based on dependency tree and affective knowledge to leverage the long range inconsistent sentiment relations in detecting sarcastic expression. For more powerful BERT-based [24] method, [25] produced affective and contextual feature embeddings to extend the architecture of BERT for textual sarcasm detection.

2.2 Multimodal Sarcasm Detection

Different from sarcasm detection purely based on text, multimodal sarcasm detection aims to detect the sarcastic expression of the instances composed of different modalities [8], [26]. Detecting the sarcastic expression for data composed of text- and image-modality has achieved much research attention in recent years. Among them, [8] firstly presented the research of multimodal sarcasm detection towards text and image modalities, and proposed a model based on manually designed features to deal with multimodal sarcasm detection. Further, [9] created a new dataset composed by text- and image-modality for multimodal sarcasm detection, and proposed a hierarchical fusion model with multimodal features to improve the learning of multimodal sarcasm detection. [10] explored decomposition and relation network to devise a fused model for learning both cross-modality contrast and semantic association. For BERT-based methods, [11] proposed to model the intra-modality and inter-modality incongruous sentiment expressions with the inter-modality attention and co-attention mechanism based on BERT [24]. Further, [12] devised an interactive in-modal and cross-modal graph architecture to model the relationships within and across modalities in learning the sarcastic clues for multimodal sarcasm detection.

2.3 Graph Neural Networks

Recently, study methods based on graph neural networks, including graph convolutional network (GCN) [27] and graph attention network (GAT) [28], have achieved promising progress in many research domain. Such as computer vision [29], [30], [31], natural language processing [23], [32], [33], [34], [35], [36], and recommendation systems [37], [38], [39], etc. Correspondingly, there are also a series graph network-based research studies in the multimodal learning domain, such as multimodal sentiment analysis [40], multimodal neural machine translation [41], multimodal video moment retrieval [42], and multimodal named entity recognition [43], etc. Furthermore, in the task of multimodal sarcasm detection, there is also a graph neural network-based model proposed by [12], in which an interactive in-modal and cross-modal graph architecture is devised to learn the relationships within and across modalities for learning the incongruous sentiment expressions of sarcasm. Inspired by the remarkable performance achieved by [12] in multimodal sarcasm detection and by [23] in the textual sarcasm detection, in our work, we explore a novel multimodal GCN architecture to leverage the inconsistent sentiment implications among the text-modality and visual and OCR modes, allowing the understanding of intricate multimodal relationships to reason the sarcasm expression.

2.4 Contrastive Learning

Contrastive learning in the latent space, which aims to automatically make the representation of a given anchor example to be similar to its positive pairs and dissimilar to its negative pairs, has recently shown promising progress in many research fields [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57]. Further, a series of contrastive learning methods have been explored to deal with multimodal learning tasks, such as multimodal representation learning [58], visual representation learning [59], and image registration [60], etc. In addition, there are also some research studies attempted to explore contrastive learning based on graph models, so as to improve the learning of graph representations. [61] proposed a graph contrastive learning framework based on graph data augmentation, which improves the graph representations for better generalizability and robustness. [62] devised a graph contrastive representation learning method with adaptive augmentation that incorporates various priors for topological and semantic aspects of the graph, so as to highlight important connective structures and enforce the model to recognize underlying semantic information. Furthermore, supervised contrastive learning is proposed by [48], which is built on the contrastive self-supervised literature [45], [46], [63], [64]

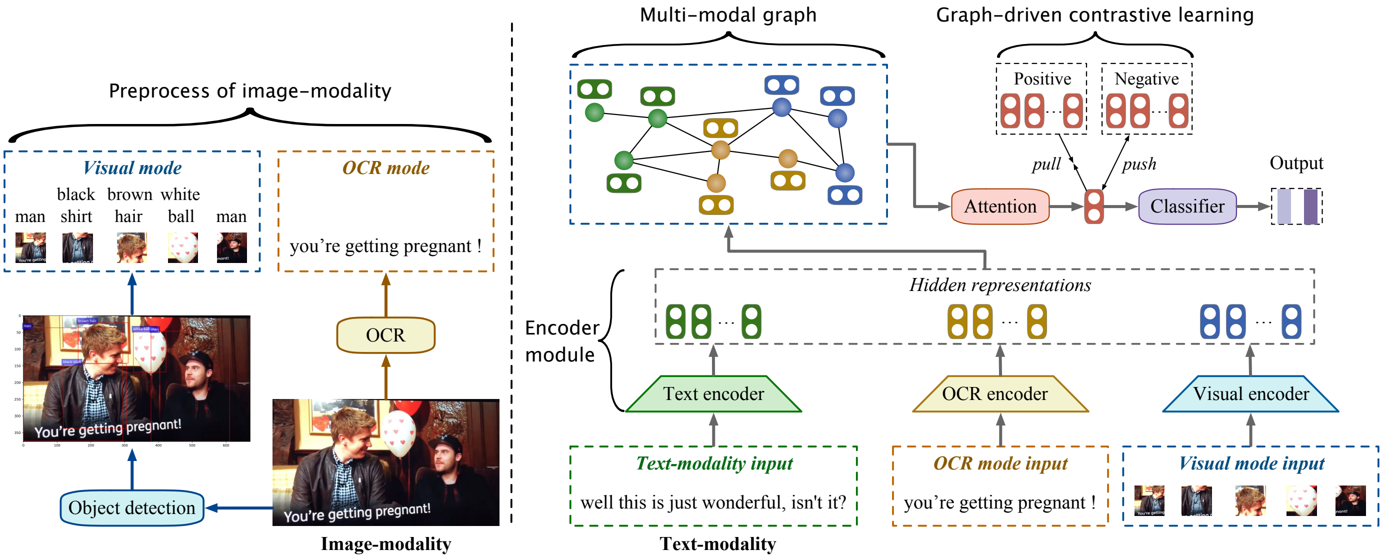


Fig. 3. The architecture of our Multimodal Graph Contrastive Learning (MMGCL) framework.

and allows us to effectively leverage the label information of the data. In which, clusters of points belonging to the same class are pulled together in embedding space, while simultaneously pushing apart clusters of samples from different classes. Inspired by [48], [62], in this work, we devise a graph contrastive loss based on the supervised signal from sarcasm labels, aiming to capture the similarity between examples in one class and contrast them with examples in other classes, so as to derive better multimodal representation for multimodal sarcasm detection.

3 Methodology

In this section, we describe our proposed Multimodal Graph Contrastive Learning (MMGCL) framework in details. The architecture of our (MMGCL) framework is shown in Figure 3. There are four main components in the (MMGCL) framework: 1) ***Preprocess of image-modality***, in which we use object detection to capture the crucial visual regions accompanied by their *attribute-object* captions of the image and exploit optical character recognition (OCR) to derive the textual description of the image; 2) ***Encoder module***, in which we utilize the pre-trained uncased BERT-base model [24] as the text encoder and OCR encoder to capture the hidden representations of the inputs of text-modality and OCR mode, and employ the pre-trained Vision Transformer (ViT) [65] as the visual encoder to capture the hidden representation of the input of visual mode; 3) ***Multimodal graph***, in which we build a multimodal graph for each input example based on the hidden representations, the Part-of-Speech Tagging results of text-modality and OCR mode, the *attribute-object* captions of visual mode, and the affective information from an external knowledge source; 4) ***Graph-oriented contrastive learning***, in which we devise a contrastive loss for the graph-oriented multimodal representations based on the supervised signal from sarcasm labels to leverage the correlation of features in the same sarcasm label and the difference of features between sarcasm and non-sarcasm expressions.

3.1 Task Description

Given an example composed by a sentence consists of N_t words $S^T = \{w_i\}_{i=1}^{N_t}$, where w_i represents the i-th contextual word, and an image I. The aim of multimodal sarcasm detection is to predict the sarcasm label (*Sarcasm* or *Non-sarcasm*) of the given example.

3.2 Preprocess of Image-Modality

In this section, we describe how to capture the crucial visual information of the image-modality by exploring the merit of object detection and optical character recognition (OCR), so as to make full use of the sarcastic cues from the image-modality in multimodal sarcasm detection.

3.2.1 Object Detection

Previous studies generally attempt to understand the semantic information of the image simply by modeling the visual features of the image-modality, which may lead to a poor understanding of the image-modality in learning sarcastic information. Therefore, we utilize a trained object detection model proposed by [13] to derive a series of visual regions of objects $\{v_i\}_{i=1}^{N_v}$ paired with their *attribute-object* pairs, as the bounding boxes in the image shown in Figure 3. N_v is the number of visual regions of the image. Then the visual regions can be regarded as the visual mode input to represent the visual information of the image. Further, the *attribute-object* pairs (such as "*man*", "*black shirt*", etc.) are served as a bridge to link the representations of visual mode with the text-modality and the OCR mode for building the multimodal graph.

3.2.2 Optical Character Recognition

The intention of using optical character recognition (OCR) to extract the textual descriptions from images is to better learn the textual semantic information presented in the images for reasoning sarcastic cues. As the example shown in Figure 1 (b), extracting the textual description in the image is imperative to understand the ironic expression of the image-modality, allowing a better learning of semantic information for reasoning sarcastic clues than simply modeling the visual features. Therefore, we exploit a public awesome open-source optical character recognition toolkit PaddleOCR[3] to extract the textual description of each image:

$$S^O = \{o_i\}_{i=1}^{N_o} = \mathrm{OCR}(I) \tag{1}$$

where o_i represents the i-th word in the textual description of the OCR result, N_o is the length of the textual description. Note that if the images contain no text, the length of OCR results is 0: i.e. $N_o = 0$.

3.3 Encoder Module

In this section, we describe the three encoder modules used in our MMGCL framework in details.

3.3.1 Text Encoder

For the processing of text-modality, given a sequence of words $S^T = \{w_i\}_{i=1}^{N_t}$, N_t is the length of S^T. We employ the pre-trained uncased BERT-base model [24] to encode each word w_i into a d^T-dimensional embedding:

$$\boldsymbol{X}^T = [\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, \cdots, \boldsymbol{x}_{N_t}^T] = \mathrm{BERT}([\mathrm{CLS}]S^T[\mathrm{SEP}]) \tag{2}$$

where $\boldsymbol{X}^T$ is the embedding representations of the text-modality. Note that the representations of tokens [CLS] and [SEP] are not adopted in building the multimodal graph, since the graph structure aims to model the multimodal relationships for the context words of the text-modality. Then, to unify the dimensions of representations between different modalities and capture the sequential relations of the context, we utilize a bidirectional LSTM [66] (Bi-LSTM) to learn the hidden representation of the text-modality:

$$\boldsymbol{T} = \{\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_{N_t}\} = \mathrm{Bi\text{–}LSTM}(\boldsymbol{X}^T) \tag{3}$$

where $\mathrm{Bi\text{–}LSTM}(\cdot)$ is the computation of the Bi-LSTM layers, $\boldsymbol{t}_j \in \mathbb{R}^{2d_h}$ denotes the hidden state vector at time step j from the bidirectional LSTM, d_h denotes the dimensionality of the text-modality hidden state representation.

3.3.2 OCR Encoder

For the processing of OCR mode, for each textual description of the OCR mode $S^O = \{o_i\}_{i=1}^{N_o}$. Corresponding to the Text Encoder, we utilize the pre-trained uncased BERT-base model [24] to encode each word o_i into a d^O-dimensional embedding:

$$\boldsymbol{X}^O = [\boldsymbol{x}_1^O, \boldsymbol{x}_2^O, \cdots, \boldsymbol{x}_{N_o}^O] = \mathrm{BERT}([\mathrm{CLS}]S^O[\mathrm{SEP}]) \tag{4}$$

where $\boldsymbol{X}^O$ is the embedding matrix of the OCR mode input. As processed in Section 3.3.1, the representations of tokens [CLS] and [SEP] are not utilized for constructing the multimodal graph. Then, we utilize a bidirectional LSTM (Bi-LSTM) to learn the hidden representation of the input of OCR mode:

$$\boldsymbol{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_{N_o}\} = \mathrm{Bi\text{–}LSTM}(\boldsymbol{X}^O) \tag{5}$$

where $\boldsymbol{c}_j \in \mathbb{R}^{2d_h}$ denotes the hidden state vector at time step j from the bidirectional LSTM, and d_h denotes the dimensionality of the hidden state representation of each word in OCR mode.

3.3.3 Visual Encoder

For the processing of visual mode, for each visual region $v_i \in \mathbb{R}^{L_h \times L_w}$, following [10], we first resize it to 224×224, i.e. $L = L_h = L_w = 224$. Then, following [65], we reshape the region $v_i \in \mathbb{R}^{L \times L}$ into a sequence $v_i = \{\boldsymbol{p}_j \in \mathbb{R}^{L/p \times L/p}\}_{j=1}^{r}$, where $r = p \times p$ is the number of patches. Further, we flatten and embed each patch to a d^V-dimensional vector with a trainable linear projection: $\boldsymbol{z}_j = \boldsymbol{p}_j \boldsymbol{E}$.

Based on the sequence of the image patches, a [class] token embedding $\boldsymbol{z}_{[class]} \in \mathbb{R}^{d^V}$ is prepended for the sequence of embedded patches, and simultaneously position embeddings are added to retain positional information. Therefore, the input of each visual region v_i is represented as:

$$\boldsymbol{Z}_i = [\boldsymbol{z}_{[class]}; \boldsymbol{z}_1; \boldsymbol{z}_2; \cdots; \boldsymbol{z}_r] + \boldsymbol{E}_{pos} \tag{6}$$

where $\boldsymbol{Z}_i \in \mathbb{R}^{(r+1) \times d^V}$ represents the embedding matrix of the image patches, and $\boldsymbol{E}_{pos} \in \mathbb{R}^{(r+1) \times d^V}$ represents the position embedding matrix. Then, we feed the matrix $\boldsymbol{Z}_i$ into the pretrained ViT to derive the representation $\boldsymbol{x}_i^V$ of the visual region v_i:

$$\boldsymbol{X}_i^V = \mathrm{ViT}(\boldsymbol{Z}_i),\ \boldsymbol{x}_i^V = \boldsymbol{X}_{i,[class]}^V \tag{7}$$

Note that, to capture a vector representation for each visual region, the representation of the [class] token embedding is used to represent the visual region. As such, the representation of the image I is defined as:

$$\boldsymbol{X}^V = \{\boldsymbol{x}_1^V, \boldsymbol{x}_2^V, \cdots, \boldsymbol{x}_{N_v}^V\} \tag{8}$$

Subsequently, to unify the dimensions of representations between different modalities, we employ a trainable Linear Projection to map each $\boldsymbol{v}_i$ to a $2d_h$-dimensional vector:

$$\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_{N_v}\} = \boldsymbol{X}^V \boldsymbol{W}^V \tag{9}$$

where $\boldsymbol{W}^V \in \mathbb{R}^{d^V \times 2d_h}$ is a trainable weight matrix.

3.4 Multimodal Graph

In this section, we describe how to construct a multimodal graph for each instance and how to fuse the multimodal features by modeling the multimodal graphs.

3.4.1 Multimodal Graph Construction

In the light of the merit of graph model, we devise a multimodal graph for each instance to explicitly link the textual words with the associated visual regions and OCR tokens, so as to leverage the intricate sarcastic relations between multimodal features. To be specific, the nodes $\boldsymbol{H}$ of the multimodal graph are the hidden representations from text-modality, OCR mode and visual mode: $\boldsymbol{H} = \{\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_N\} = \{\boldsymbol{t}_1, \cdots, \boldsymbol{t}_{N_t}, \boldsymbol{c}_1, \cdots, \boldsymbol{c}_{N_o}, \boldsymbol{v}_1, \cdots, \boldsymbol{v}_{N_v}\}$. Here, $N = N_t + N_o + N_v$ is the length of the hidden representations. The edges of the graph reflect the association of nodes.

As previously discussed by [23], [33], [40], the weights of the edges are crucial in graph information aggregation. Therefore, the key issue of constructing a multimodal graph boils down to the computation of the edge weights in the graph. Here, corresponding to the hidden representations

[3] https://github.com/PaddlePaddle/PaddleOCR

of the nodes $\boldsymbol{H}$, we define $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ as the adjacency matrix of the multimodal graph. Note that the elements in $\boldsymbol{A}$ are initialized by 0. To describe the construction of multimodal graph more clearly, we split the procedure of constructing multimodal graph as 1) setting edges for text-modality; 2) setting edges for OCR mode; 3) setting edges for visual mode; 4) setting edges between modalities.

3.4.1.1 **Setting edges for text-modality:** As in the examples shown in Figure 2, the sarcastic information of text-modality may be expressed by multiple words, such as "*love commuting*". Therefore, we incorporate the syntax-aware relations over the dependency tree of the sentence into the multimodal graph to advance the learning of the contextual dependencies[4], so as to better learn the relations between words. Here, we employ the dependency tree of the text-modality to set the weights of edges for the sub-graph of text-modality $\boldsymbol{A}^T \in \mathbb{R}^{N_t \times N_t}$ as:

$$A^T_{i,j} = \begin{cases} 1 & \text{if } \mathcal{D}_{i,j} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where $\mathcal{D}_{i,j}$ represents there is a relation between words w_i and w_j in the dependency tree of the sentence.

3.4.1.2 **Setting edges for OCR mode:** Corresponding to the processing of text-modality, we also compute the weights of edges for OCR mode based on the dependency tree of the textual description of the image extracted by the OCR, aiming to better learn the syntactical relations between words in the light of the graph structure. The sub-graph $\boldsymbol{A}^O \in \mathbb{R}^{N_o \times N_o}$ of OCR mode is defined as:

$$A^O_{i,j} = \begin{cases} 1 & \text{if } \mathcal{D}_{i,j} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

3.4.1.3 **Setting edges for visual mode:** For the visual mode, since there is no given information to present the relations between the representations of visual regions, we explore a novel solution to explicitly link the representations in the graph by exploiting the *attribute-object* pairs of the visual regions to compute the weights of edges. Specifically, we set the weights of edges for the sub-graph of visual mode by computing the word similarities between the *object* words. Since intuitively, the greater the similarity of words, the more likely the two visual regions are to present similar information. The adjacency matrix $\boldsymbol{A}^V \in \mathbb{R}^{N_v \times N_v}$ of the visual mode graph is defined as:

$$A^V_{i,j} = Sim(obj_i, obj_j) \tag{12}$$

where $Sim(\cdot)$ denotes the computation of word similarity[5]. obj_i and obj_j represent the two *object* words of visual region v_i and v_j. We set $Sim(\cdot) = 0$ if the return value is $None$.

3.4.1.4 **Setting edges between modalities:** Based on the adjacency matrices of the sub-graphs of text-modality, OCR mode and visual mode, we then set the weights of edges between modalities to build the complete adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ of the multimodal graph regarding the hidden representations of the nodes $\boldsymbol{H}$. As described in the examples of Figure 1, the key of understanding the sarcastic clues of multimodal sarcasm detection is to capture the relations and understand the incongruous sentiment expression between modalities. As such, the main issue of constructing multimodal graph evolves into how to link the related nodes and express the incongruous sentiment relationships between different modalities by the graph structure. That is, to model the related nodes and capture the incongruous sentiment relationships between different modalities, we need to 1) track the related description between modalities; 2) reason the sentiment expression between the related description.

Intuitively, it is easy to understand the similarity of words and the sentiment expression of words for text-modality and OCR mode. Therefore, for the visual mode, we regard the *attribute-object* pairs of the image regions as a bridge to track the relations and sentiment expressions. Based on it, we explore a novel solution of setting the weights of edges between modalities based on the word similarities and the affective clues from external knowledge. In the real world scenario, nouns usually express objects or entities, while adjectives are usually rich in sentiment expressions[6]. Therefore, based on the results of Part-of-Speech Tagging, 1) similar to the strategy described in Section 3.4.1.3, we compute the word similarities between each *noun pair* of text-modality, OCR mode, and visual mode. Here, the word in each word pair could be a noun in text-modality or OCR mode, or a *object* caption in visual mode; 2) with the help of the affective clues extracted from the SenticNet [14], we then capture the sentiment incongruity between *adjectives* of text-modality or OCR mode and *attribute* captions of visual mode by computing the difference of sentiment scores. The complete adjacency matrix $\boldsymbol{A}$ is defined as:

$$A_{i,j} = \begin{cases} Sim(w_i, obj_j) & \text{if } w^n_i,\, i < N_{to},\, j \geq N_{to} \\ Sim(w_i, w_j) & \text{if } w^n_i,\, w^n_j,\, i < N_t,\, N_{to} > j \geq N_t \\ |\boldsymbol{\omega}(w_i) - \boldsymbol{\omega}(att_j)| & \text{if } w^a_i,\, i < N_{to},\, j \geq N_{to} \\ |\boldsymbol{\omega}(w_i) - \boldsymbol{\omega}(w_j)| & \text{if } w^a_i,\, w^a_j, i < N_t,\, N_{to} > j \geq N_t \\ A^T_{i,j} & \text{if } i < N_t,\, j < N_t \\ A^O_{i-N_t, j-N_t} & \text{if } N_{to} > i \geq N_t,\, N_{to} > j \geq N_t \\ A^V_{i-N_{to}, j-N_{to}} & \text{if } N > i \geq N_{to},\, N > j \geq N_{to} \end{cases} \tag{13}$$

$$N_{to} = N_t + N_o \tag{14}$$

where w^n_i represents w_i is a *noun*, w^a_i represents w_i is an *adjective*, att_j represents the *attribute* of visual region j.

$\boldsymbol{\omega}(w_i) \in [-1, 1]$ represents the affective weight of word w_i retrieved from SenticNet [**?**]. We set $\boldsymbol{\omega}(w_i) = 0$ if w_i cannot be found in SenticNet. $|\cdot|$ represents absolute value calculation. The intention of computing the absolute value of the sentiment score between a word pair is that the greater the sentiment difference, the more likely it is to illustrate the sentiment inconsistency of sarcastic expression. We hence need to set greater weights to these edges in the graph, so as to focus on the relationships of them in the graph learning. Further, inspired by [27], we construct the multimodal graph as an undirected graph, $A_{i,j} = A_{j,i}$, and set a self-loop for each node, $A_{i,i} = 1$.

[4] We utilize the spaCy toolkit (https://spacy.io/) to derive the dependency tree of a sentence.

[5] We use the NLTK toolkit (http://www.nltk.org/) to compute the similarity of a word pair based on the WordNet.

[6] For text-modality and OCR mode, we utilize NLTK toolkit for Part-of-Speech Tagging to find nouns and adjectives in sentences. For visual mode, we use the *attribute* word as the adjective of a visual region.

3.4.2 Multimodal Features Fusion

Based on the adjacency matrix $\boldsymbol{A}$ derived in Section 3.4.1, for each multimodal example, we explore a graph architecture to update the multimodal representations and extract the crucial sarcastic clues by aggregating the correlation of nodes in the multimodal graph. To be specific, we feed the adjacency matrix of the multimodal graph $\boldsymbol{A}$ and the corresponding nodes' representations $\boldsymbol{H}$ of each multimodal instance into a multi-layers GCNs architecture to capture the graph representation of the input instance. For the graph convolutional operation, each node in the l-th GCN layer is updated according to the hidden representations of its neighborhoods in the adjacency matrices of the multimodal graph, which is defined as:

$$\boldsymbol{G}^l = \mathrm{ReLU}(\tilde{\boldsymbol{A}}\boldsymbol{G}^{l-1}\boldsymbol{W}^l + \boldsymbol{b}^l) \tag{15}$$

where $\tilde{\boldsymbol{A}} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}$ is the normalized symmetric adjacency matrix. $\boldsymbol{D}$ is the degree matrix of $\boldsymbol{A}$, where $D_{ii} = \sum_j A_{i,j}$. $\boldsymbol{G}^{l-1}$ is the hidden graph representation evolved from the preceding GCN layer. $\boldsymbol{W}^l \in \mathbb{R}^{2d_h \times 2d_h}$, $\boldsymbol{b}^l \in \mathbb{R}^{2d_h}$ are the trainable parameters of the l-th GCN layer. The nodes input of the first GCN layer are the concatenation of text-modality and image-modality representations, i.e. $\boldsymbol{G}^0 = \boldsymbol{H}$.

Then, as previously discussed by [67], we employ a retrieval-based attention mechanism to capture the graph-oriented multimodal representation based on the concatenation of the representations of text-modality, OCR mode, and visual mode: $\boldsymbol{H} = \{\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_N\}$ according to the graph representation $\boldsymbol{G}^L = \{\boldsymbol{g}_i\}_{i=1}^N$ derived from the final GCN layer. The aim is to retrieve significantly correlated multimodal features where nodes are associated with other nodes in the multimodal graph. The attention weights are defined as:

$$\alpha_t = \frac{\exp(\beta_{\mathrm{t}})}{\sum_{i=1}^{N}\exp(\beta_i)} \tag{16}$$

$$\beta_t = \sum_{i\in\mathcal{I}} \boldsymbol{h}_t^\top \boldsymbol{g}_i \tag{17}$$

where $\mathcal{I}$ denotes a set of indices in which nodes contain an edge with a weight other than 0 in the graph. $\top$ represents the matrix transposition. The final representation of an input instance i is defined as:

$$\boldsymbol{z}_i = \sum_{t=1}^{N} \alpha_t \boldsymbol{h}_t \tag{18}$$

3.5 Graph-Oriented Contrastive Learning

Supervised contrastive learning is proposed by [48], which is built on the contrastive self-supervised literature [45], [46], [63], [64] and allows us to effectively leverage the label information of the data. In which, clusters of points belonging to the same class are pulled together in embedding space, while simultaneously pushing apart clusters of samples from different classes. More concretely, the supervised contrastive loss is an extension of the self-supervised contrastive loss that adapting contrastive learning to the fully supervised setting. The supervised contrastive loss is defined as follow:

$$\mathcal{L}^{sup} = \sum_{i\in I} \frac{-1}{|P(i)|} \sum_{p\in P(i)} \log \frac{\exp(z_i, z_p/\tau)}{\sum_{a\in A(i)}\exp(z_i, z_a/\tau)} \tag{19}$$

Here, i is the anchor, $P(i) \equiv \{p \in A(i) : \hat{y}_p = \hat{y}_i\}$ is the set of indices of all positives in the batch distinct from i, and $|P(i)|$ is its cardinality. $A(i) \equiv I \setminus \{i\}$. I is the set of batch indices. τ denotes the temperature parameter.

In our work, inspired by the supervised contrastive learning proposed by [48], we devise a graph-oriented supervised contrastive learning strategy to pull together the clusters of graph-oriented multimodal representations belonging to the same label in embedding space, and simultaneously push apart representations of samples from different labels, allowing the learning of the correlation of features in the same sarcasm label and the difference of features between sarcasm and non-sarcasm classes. Given the final multimodal representations derived by graph-oriented attention mechanism $\{\boldsymbol{z}_i\}_{i=1}^{N_b}$ in a mini-batch $\mathcal{B}$ (here, N_b is the size of mini-batch), and an *anchor* $\boldsymbol{z}_i$. Here, $\boldsymbol{z}_i, \boldsymbol{z}_j \in \mathcal{B}$ with the same stance label is considered as a *positive* pair, i.e. $y^i = y^j$, where y^i and y^j are the labels of $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$, respectively, while the samples $\{\boldsymbol{z}_k \in \mathcal{B}, k \neq i\}$ are treated as *negative* representations with respect to the *anchor*. Then the contrastive loss is computed across all *positive* pairs, in a mini-batch:

$$\mathcal{L}_{con} = \frac{-1}{N_b} \sum_{\boldsymbol{z}_i\in\mathcal{B}} \ell(\boldsymbol{z}_i) \tag{20}$$

$$\ell(\boldsymbol{z}_i) = \log \frac{\sum_{j\in\mathcal{B}\setminus i} \mathbb{1}_{[y^i=y^j]}\exp(f(\boldsymbol{z}_i,\boldsymbol{z}_j)/\tau)}{\sum_{j\in\mathcal{B}\setminus i}\exp(f(\boldsymbol{z}_i,\boldsymbol{z}_j)/\tau)} \tag{21}$$

where $\mathbb{1}_{[i=j]} \in \{0, 1\}$ is an indicator function evaluating to 1 iff $i = j$. $f(\boldsymbol{u}, \boldsymbol{v}) = sim(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top\boldsymbol{v}/\|\boldsymbol{u}\|\|\boldsymbol{v}\|$ denotes the cosine similarity between vectors $\boldsymbol{u}$ and $\boldsymbol{v}$.

3.6 Sarcasm Detection

For each instance i, we feed the final representation into a fully-connected layer with a softmax function to capture a probability distribution $\hat{\boldsymbol{y}} \in \mathbb{R}^{d_y}$ in the sarcasm decision space:

$$\hat{\boldsymbol{y}} = \mathrm{softmax}(\boldsymbol{W}_o\boldsymbol{z}_i + \boldsymbol{b}_o) \tag{22}$$

where d_y is the dimensionality of sarcasm labels. $\boldsymbol{W}_o \in \mathbb{R}^{d_y\times 2d_h}$ and $\boldsymbol{b}_o \in \mathbb{R}^{d_y}$ are trainable parameters. Then, we employ a cross-entropy loss between predicted distribution $\hat{\boldsymbol{y}}_i$ and ground-truth distribution $\boldsymbol{y}_i$ of instance i to train the classifier:

$$\mathcal{L}_{sar} = -\sum_{i=1}^{N_b}\sum_{j=1}^{d_y} y_i^j \log \hat{y}_i^j \tag{23}$$

3.7 Learning Objective

The learning objective of the proposed `MMGCL` is to train the framework by jointly minimizing the two losses derived by sarcasm detection and graph-oriented contrastive learning. The overall loss $\mathcal{L}$ is formulated as:

$$\mathcal{L} = \mathcal{L}_{sar} + \mathcal{L}_{con} + \lambda||\Theta||^2 \tag{24}$$

where Θ denotes all trainable parameters of the model, λ represents the coefficient of L_2-regularization.

TABLE 1
Statistics of the experimental data.

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

4 Experimental Setup

4.1 Dataset

To evaluate the effectiveness of our MMGCL framework, we conduct experiments on a publicly available multimodal sarcasm detection benchmark dataset collected by [9]. This dataset contains English tweets, where examples express *sarcasm* information are *Positive* examples and express *non-sarcasm* information are *Negative* examples. Each instance consists of a text and an associated image. The dataset is divided into a training set, a development set, and a testing set with a ratio of 80%:10%:10%. The statistics of the dataset are shown in Table 1.

4.2 Experimental Settings

Following [9], the tweets containing *sarcasm*, *sarcastic*, *irony*, *ironic*, *jokes*, *humor*, *exgag* as regular words and URLs are discarded during data pre-processing, and the NLTK toolkit is adopted to separate words, emoticons and hashtags. We set the maximum number of visual regions as 10 for object detection results. Note that we select the top 10 bounding boxes with highest scores if the number of output bounding boxes is greater than 10. We utilize the pre-trained uncased BERT-base [24] mode as the Text Encoder module and OCR Encoder module to embed each word from text-modality and OCR mode as a 768-dimensional embedding[7]. For visual mode, we employ the pre-trained ViT[8] proposed by [65] to embed each visual region patch into a 768-dimensional embedding. That is, $d^T = d^O = d^V = 768$. The resolution of visual region patch is set to $L_p = 32$, correspondingly, $p = 7, r = 49$. In the preliminary experiments, we also tried other division resolutions, and found that the fluctuation of performance is negligible over different resolutions of image patches. The number of GCN layers is set to 2, which is the optimal depth in the pilot experiments. The dimensionality of hidden representations is set to $d_h = 512$. For the graph-oriented contrastive loss, we set the temperature parameter τ to 0.07. For the complete loss function, we set $\gamma_s = 1$ and $\gamma_c = 1$, which are the optimal coefficients in the pilot experiments. The coefficient λ is set to 0.00001. Adam is utilized as the optimizer with a learning rate of 0.00002, and the mini-batch size is 32. The dropout rate with 0.1 is utilized to avoid overfitting. We use early-stopping with patience of 5.

Following [9], we use *Accuracy*, *Precision*, *Recall*, and *F1-score* to measure the performance of models. In addition, since the label distribution of the dataset is imbalanced, following [11], we also report the Macro-average results to evaluate the performance of models. The experimental results of our models are averaged over 10 runs with different random seeds to ensure the final reported results are statistically stable.

4.3 Comparison Models

We compare our proposed MMGCL framework with a series of related baseline models, which are summarized as follow:

- **Image-modality methods**: These models only use the visual information of the image-modality for sarcasm detection, including **Image** [9], which utilizes the image vector after the pooling layer of ResNet [70] to train a classifier for sarcasm detection; and **ViT** [65], which employs the '[class]' token representation of the pre-trained ViT model to detect the sarcasm of the input instance.
- **Text-modality methods**: These models only utilize the textual information for sarcasm detection, including **TextCNN** [69], a convolutional neural network-based text classification for sarcasm detection; **Bi-LSTM** [66], a sarcasm detector based on the bidirectional LSTM network; **SIARN** [22], which uses an inner-attention for textual sarcasm detection; **SMSD** [7], which explores a self-matching network to learn the textual incongruity information for sarcasm detection; and **BERT** [24], the vanilla pre-trained uncased BERT-base model that takes '[CLS] text [SEP]' as input for detecting sarcasm.
- **Multimodal methods**: These models consider the information from both text and image modalities. Models in this category includes **HFM** [9], which adopts a hierarchical multimodal features fusion strategy for multimodal sarcasm detection; **D&R Net** [10], which explores Decomposition and Relation Network to model both cross-modality contrast and semantic association; **Res-BERT** [11], which combines image features and BERT-based text features for sarcasm detection; **Att-BERT** [11], which is a BERT-based model based on an inter-modality attention and a co-attention for modeling the incongruity of multimodal sarcasm detection; and **InCrossMGs** [12], a graph neural network-based model that leverages the sarcastic relations from both intra- and inter-modal perspectives for multimodal sarcasm detection.

Furthermore, we devise several variants of our MMGCL to analyze the impact of different components of the framework in the ablation study:

- **w/o** $\mathcal{G} + \mathcal{C}$. This variant represents without multimodal graph contrastive learning. It only concatenates the representations of '[CLS]' tokens from the BERT encoder modules of text-modality and OCR mode and '[class]' token from the ViT encoder module of the visual mode to derive the final representation.
- **w/o** $\mathcal{G}$. This variant represents without multimodal graph. It concatenates the representations of '[CLS]' tokens from the BERT encoder modules of text-modality and OCR mode and '[class]' token from

[7] We also tried BERT-Large and RoBERTa [68] in the experiments of Section 5.3 and found that they achieved slightly better performance than BERT-base. Since the baselines are BERT-base based, we construct our model based on BERT-base for a fair comparison.

[8] https://github.com/lukemelas/PyTorch-Pretrained-ViT

TABLE 2
Main comparison results on unimodal (only image-modality or text-modality) and multimodal scenarios. The results of the baselines with $\natural$ are retrieved from [12], others are run by the open source codes. Best scores of each group are in bold. Results with $\star$ denote the significance tests of the proposed MMGCL over the baseline models at $p-\text{value} < 0.05$.

MODALITY	METHOD	*Accuracy (%)*	*F1-score*			*Macro-average*		
			Precision (%)	*Recall (%)*	*F1-score (%)*	*Precision (%)*	*Recall (%)*	*F1-score (%)*
image	Image [9]	64.76$^\natural$	54.41$^\natural$	70.80$^\natural$	61.53$^\natural$	60.12	73.08	65.97
	ViT [65]	67.83	57.93	70.07	63.43	65.68	71.35	68.40
text	TextCNN [69]	80.03$^\natural$	74.29$^\natural$	76.39$^\natural$	75.32$^\natural$	78.03	78.28	78.15
	Bi-LSTM [66]	81.90$^\natural$	76.66$^\natural$	78.42$^\natural$	77.53$^\natural$	80.97	80.13	80.55
	SIARN [22]	80.57$^\natural$	75.55$^\natural$	75.70$^\natural$	75.63$^\natural$	80.34	78.81	79.57
	SMSD [7]	80.90$^\natural$	76.46$^\natural$	75.18$^\natural$	75.82$^\natural$	80.87	78.20	79.51
	BERT [24]	83.85$^\natural$	78.72$^\natural$	82.27$^\natural$	80.22$^\natural$	81.31	80.87	81.09
image+text	HFM [9]	83.44$^\natural$	76.57$^\natural$	84.15$^\natural$	80.18$^\natural$	79.40	82.45	80.90
	D&R Net [10]	84.02$^\natural$	77.97$^\natural$	83.42$^\natural$	80.60$^\natural$	-	-	-
	Res-BERT [11]	84.80$^\natural$	77.80	84.15	80.85	78.87$^\natural$	84.46$^\natural$	81.57$^\natural$
	Att-BERT [11]	86.05$^\natural$	78.63	83.31	80.90	80.87$^\natural$	85.08$^\natural$	82.92$^\natural$
	InCrossMGs [12]	86.10$^\natural$	81.38$^\natural$	84.36$^\natural$	82.84$^\natural$	85.39$^\natural$	85.80$^\natural$	85.60$^\natural$
	MMGCL (ours)	**88.57***	**84.36***	**85.31***	**84.83***	**87.89***	**87.72***	**87.80***

the ViT encoder module of the visual mode to derive the representation of multimodal example, and then combines the cross-entropy loss and contrastive learning loss to train the model.

- **w/o** $\mathcal{C}$. This variant represents without graph-oriented contrastive learning. That is, the loss function (Eq. 24) is replaced with $\mathcal{L} = \mathcal{L}_{sar} + \lambda||\Theta||^2$.
- **w/o** $\mathcal{V}$. This variant represents without deriving visual regions by object detection. That is, the whole image is input into the visual encoder, and the edge weights related to the modes of visual mode are set to 1 in the multimodal graphs.
- **w/o** $\mathcal{O}$. This variant represents without deriving the textual description form the images by OCR. That is, we build the multimodal graph only based on the features from text-modality and visual mode.
- **w/o** $\mathcal{S}$. This variant represents without using word similarity and external affective knowledge to compute the relationships and incongruous sentiment between tokens. That is, all the weights of edges are set to 1 in the multimodal graph. Further, **w/o** $\mathcal{S}^a$ represents without using affective knowledge, and **w/o** $\mathcal{S}^w$ represents without using word similarity.
- **w/o** $\mathcal{D}$. This variant represents without using syntax-aware information to learn the phrasal relations of text-modality and OCR mode in the multimodal graph construction.

In addition, to investigate the effectiveness and generalization of our proposed multimodal graph contrastive learning framework regarding different pre-trained methods, we also experiment with the following variations:

- **-GloVe+ResNet**: We replace the BERT-base encoders in our MMGCL framework with GloVe [71] to initialize each word into a 300-dimensional embedding for the text-modality and the OCR mode and replace ViT with ResNet-152 [70] to embed each image patch as a 2048-dimensional vector for the visual mode.
- **-GloVe+ViT**: We employ GloVe for deriving hidden representations for text-modality and OCR mode and use ViT as the encoder of the visual mode.
- **-BERT+ResNet**: We use BERT-base as the encoders for text-modality and OCR mode and use ResNet-152 as the encoder of the visual mode.
- **-BERT-Large+ViT**: We replace the BERT-base encoders in our MMGCL framework with BERT-Large.
- **-RoBERTa+ViT**: We replace the BERT-base encoders in our MMGCL framework with RoBERTa [68].

5 EXPERIMENTAL RESULTS

5.1 Main Results

Table 2 shows the results of our proposed MMGCL and the comparison baselines regarding *Text-modality*, *Image-modality*, and *Text+Image modalities* (multimodal) on the evaluation metrics of *Accuracy*, *Precision*, *Recall*, *F1-score*, and *macro metrics* From the experimental results, we can draw the following conclusions. **1)** Our MMGCL framework performs consistently better than all the baseline models across all evaluation metrics. This verifies the effectiveness of our proposed MMGCL in multimodal sarcasm detection. **2)** We conduct significance tests of our MMGCL over the baseline models, the results of significance tests demonstrate that our MMGCL significantly outperforms the baseline models on all the evaluation metrics (with $p-\text{value} < 0.05$). **3)** From the unimodal results, we can see that the methods based on text-modality perform consistently better than the methods based on image-modality, which indicates that the expression of sarcastic/non-sarcastic information primarily resides in the text modality. This also verifies the feasibility and significance of exploring OCR to derive the textual descriptions in the images from another angle. **4)** Furthermore, BERT and ViT perform overall better than other baseline models in the unimodal scenario. This indicates that the learning of sarcastic information could be improved in the light of using a more powerful pre-trained model. **5)** Methods based on both image and text modalities (*image+text*) perform better than the unimodal baselines overall. This implies that leveraging the sarcastic information from both image and text modalities is more effective for multimodal

TABLE 3
Experimental results of ablation study.

Model	*Accuracy (%)*	*F1 (%)*	*Macro-F1 (%)*
MMGCL	**88.57**	**84.83**	**87.80**
w/o $\mathcal{G}+\mathcal{C}$	85.17	81.06	82.55
w/o $\mathcal{G}$	85.98	81.84	84.68
w/o $\mathcal{C}$	86.73	83.05	85.26
w/o $\mathcal{V}$	86.38	82.54	84.87
w/o $\mathcal{O}$	86.23	82.35	84.72
w/o $\mathcal{S}$	86.57	83.12	85.20
w/o $\mathcal{S}^a$	86.83	83.34	85.53
w/o $\mathcal{S}^w$	87.15	83.69	85.84
w/o $\mathcal{D}$	87.41	84.03	86.62

sarcasm detection. **6)** The results of macro metrics are better than the common F1-score, which implies that models may perform better in the "*negative*" (*non-sarcasm*) class due to the imbalanced distribution of classes. **7)** Compared with the previous state-of-the-art model (InCrossMG), our MMGCL achieves consistently outstanding performance on all the evaluation metrics. This verifies that modeling intricate relations and incongruous sentiment expressions between modalities in the graph based on the external knowledge sources can improve the understanding and fusion of multimodal information, and thus lead to improved performance of multimodal sarcasm detection.

5.2 Ablation Study

In this section, we analyze the impact of different components of our MMGCL on the performance. We conduct experiments of ablation study and report the results in Table 3. We can see that the removal of multimodal graph contrastive learning (w/o $\mathcal{G}+\mathcal{C}$) sharply degrades the performance on all evaluation metrics. This verifies the significance of the proposed multimodal graph contrastive learning strategy in understanding multimodal sarcastic clues for multimodal sarcasm detection. Further, the removal of multimodal graph (w/o $\mathcal{G}$) performs significantly poorer than the complete MMGCL. This verifies the importance and effectiveness of exploring graph structure to fuse multimodal features in learning sarcastic clues, which can lead to better multimodal sarcasm detection performance. The experimental results of w/o $\mathcal{C}$ show that the removal of graph-oriented contrastive learning leads to considerable performance degradation. This implies that the graph-oriented contrastive learning can derive better representation of multimodal features by capturing the similarity between examples in one class and contrasting them with examples in other classes, and thus leads to better performance of multimodal sarcasm detection.

In addition, the removal of using object detection to derive visual regions of the image-modality (w/o $\mathcal{V}$) leads to a serious performance degradation. This demonstrates that utilizing object detection to extract crucial visual regions and their *attribute-object* captions of the image-modality is important for better understanding the visual features of multimodal sarcastic cues. Further, the removal of OCR (w/o $\mathcal{O}$) also leads to very serious performance degradation. This verifies that employing OCR to extract the textual description of the image-modality can lead to a better understanding of the visual information, so as to achieve a better performance of multimodal sarcasm detection.

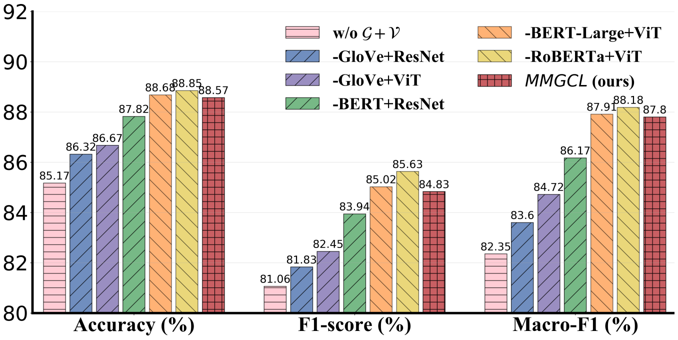


Fig. 4. Comparison results of using different pre-trained methods.

For the multimodal graph construction, from the experimental results of w/o $\mathcal{S}$, w/o $\mathcal{S}^a$ and w/o $\mathcal{S}^w$, we can conclude that both affective knowledge and word similarity are significant for linking the nodes in building multimodal graph. This also verifies that exploiting the *attribute-object* pairs of visual regions as a bridge to set the weights of edge based on word similarity and affective scores is effective in building multimodal graphs. Further, compared the experimental results of w/o $\mathcal{S}^a$ and w/o $\mathcal{S}^w$, we can see that the removal of affective knowledge leads to poorer performance. This indicates that exploiting external affective knowledge to compute the sentiment inconsistency is significant to reason the sarcastic expression in multimodal sarcasm detection. In addition, the removal of syntax-aware information (w/o $\mathcal{D}$) of text-modality and OCR mode leads to considerable performance degradation, which indicates that incorporating syntactic information in the graph makes better learning of dependency relations of textual words and thus improves the performance of sarcasm detection.

5.3 Generalizability of Multimodal Graph Contrastive Learning

To investigate the generalizability and effectiveness of our multimodal graph contrastive learning MMGCL framework with regard to different pre-trained methods, we conduct comparison experiments with five variants of our MMGCL by using different encoder modules and the model of without multimodal graph contrastive learning (w/o $\mathcal{G}+\mathcal{C}$). The comparison results are shown in Figure 4. Compared with w/o $\mathcal{G}+\mathcal{C}$, all the five variants of our MMGCL consistently achieve better performance on all evaluation metrics. This verifies that the proposed multimodal graph contrastive learning framework can directly work with various pre-trained models and achieve outstanding performance. In addition, according to the performance of the variants of "-GloVe+ResNet", "-GloVe+ViT", "-BERT+ResNet" and ours ("-BERT+ViT"), we can conclude that superior performance is obtained when using more powerful pre-trained methods, such as ViT and BERT. Furthermore, both "-BERT-Large+ViT" and "-RoBERTa+ViT" perform better than the proposed MMGCL ("-BERT+ViT"). This further proves that our proposed multimodal graph contrastive learning strat-

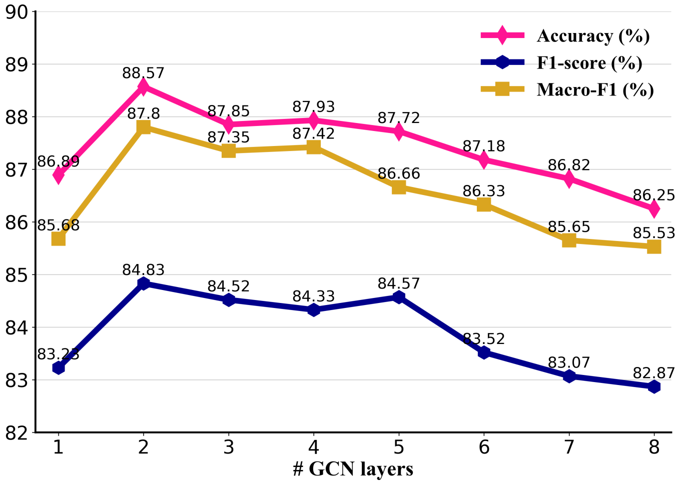


Fig. 5. Experimental results of using different GCN layers.

egy can achieve better performance with the enhancement of the pre-trained methods.

5.4 Impact of GCN Layers

To analyze the impact of the number of GCN layers on the performance of our MMGCL framework, we vary the number of layers from 1 to 8 and show the experimental results in Figure 5. We can see that the 2-layer GCN architecture consistently performs better than other numbers on all evaluation metrics, and thus we set the number of GCN layers to 2 in our MMGCL. The GCN architecture with one layer performs worse on all evaluation metrics, which indicates that a shallow graph network structure is inadequate to make full use of the sarcastic features in multimodal sarcasm detection. When the number of layers is between 2 and 5, the performance of the MMGCL fluctuates in all evaluation metrics. This implies that the performance of the model is influenced by the deeper layers of the GCN architecture. Furthermore, when the number of layers is greater than 5, the performance considerably tends to decline. This potentially indicates that further increasing the number of layers beyond 5 degrades the model performance possibly due to the inappropriate increase of parameters in the graph architecture.

5.5 Analysis of Graph-Oriented Contrastive Learning

In this section, we analyze how the graph-oriented contrastive learning in our MMGCL helps the model improve the quality of representation in the training process, and thus leads to improved performance of multimodal sarcasm detection. We track the checkpoints from the two variants of "w/o $\mathcal{G}+\mathcal{C}$" and "w/o $\mathcal{C}$", and our complete MMGCL during the training process. Then, following [72], we visualize the alignment and uniformity metrics of the checkpoints in Figure 6. As previously discussed by [72], models attain both lower *alignment* and *uniformity* will achieve better performance. From the results, we can see that our MMGCL shows lower $\mathcal{L}_{align}$ and $\mathcal{L}_{uniform}$ during the training, which verifies that our MMGCL attain strong ability in contrastive learning. It can be seen that both "w/o $\mathcal{G}+\mathcal{C}$" and "w/o $\mathcal{C}$" present the worst alignment and uniformity, which indicates that contrastive learning can advance a better latent space for the learned representations. That is, the graph-oriented contrastive learning strategy explored in our MMGCL is effective and significant in learning sarcastic clues for multimodal sarcasm detection.

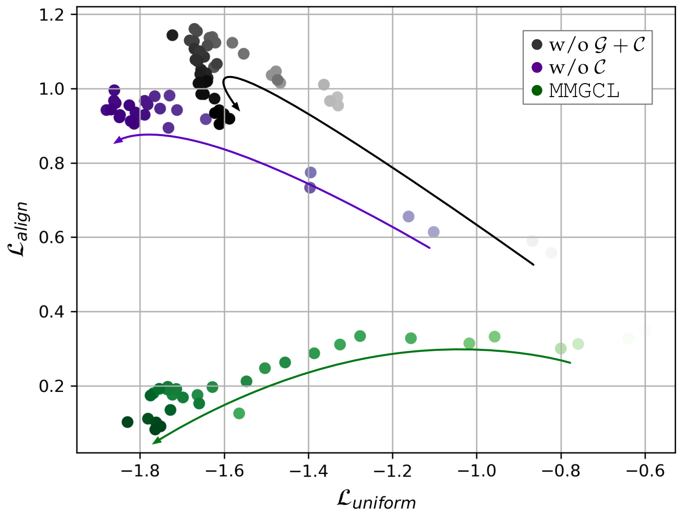
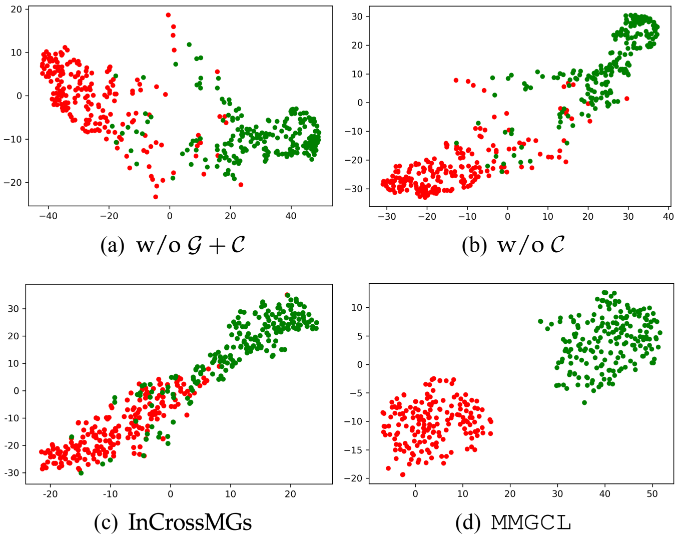


Fig. 6. Visualization of the contrastive representation of checkpoints in every 50 training steps. The darker the color of the point, the greater the accuracy. The arrows present the training direction. As previously discussed by [72], models with low $\mathcal{L}_{align}$ and $\mathcal{L}_{uniform}$ consistently perform well (lower left corner).

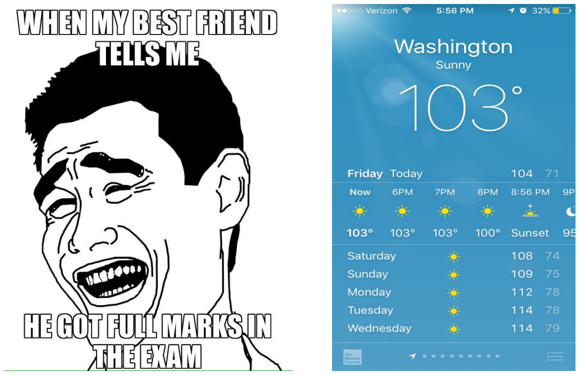

(a) w/o $\mathcal{G}+\mathcal{C}$ (b) w/o $\mathcal{C}$

(c) InCrossMGs (d) MMGCL

Fig. 7. Visualization of the intermediate vectors learned by the variants of "w/o $\mathcal{G}+\mathcal{C}$" and "w/o $\mathcal{C}$", the previous state-of-the-art model "InCrossMGs" [12], and our MMGCL. Red dots denote *sarcasm* examples, green dots denote *non-sarcasm* examples.

5.6 Visualization of Intermediate Vectors

In this section, we qualitatively demonstrate how the proposed MMGCL achieves better sarcastic feature representations in multimodal sarcasm detection. We randomly select 200 test instances for each label from the dataset, and then show the t-SNE [73] visualization of intermediate embeddings learned by the variants of "w/o $\mathcal{G}+\mathcal{C}$" and "w/o $\mathcal{C}$",


WHEN MY BEST FRIEND TELLS ME
HE GOT FULL MARKS IN THE EXAM
●●○○○ Verizon 5:56 PM 32%
Washington
Sunny
103°
Friday Today 104 71
Now 6PM 7PM 8PM 8:56 PM 9P
103° 103° 103° 100° Sunset 95
Saturday 108 74
Sunday 109 75
Monday 112 78
Tuesday 114 78
Wednesday 114 79

(a) When my best friend tells me he got full marks in the exam!

(b) It was nice enjoying this cooler weather for a change today.

Fig. 8. Examples of the misclassified instances.

the previous state-of-the-art model "InCrossMGs" [12] and our MMGCL framework. The results are shown in Figure 7. We can observe that the distributions of the intermediate representations derived from "w/o $\mathcal{G}+\mathcal{C}$" or "w/o $\mathcal{C}$" largely diffuse and overlap between different labels. But there are clear separations between different labels produced by our proposed MMGCL. This further verifies the effectiveness of our proposed multimodal graph contrastive learning strategy, and qualitatively demonstrates that our MMGCL can learn better feature representations with the help of multimodal graph contrastive learning. Furthermore, the distributions of the intermediate representations derived from the previous graph-based model ("InCrossMGs") are more separated between different labels compared with both "w/o $\mathcal{G}+\mathcal{C}$" and "w/o $\mathcal{C}$" but still partially overlap. This indicates that the novel multimodal graph contrastive learning strategy in our MMGCL can better separate representations from different labels than the previous "InCrossMGs" with the help of the solution of multimodal construction and the merit of multimodal graph-oriented contrastive learning strategy, and thus improves the performance of multimodal sarcasm detection.

5.7 Error Analysis

In this section, we conduct an error analysis about the erroneously predicted examples. We find that most of the misclassified examples can be broadly categorized as: 1) the image-modality contains a facial expression or emoji. We need to understand the meaning of the expression to judge whether the instance is sarcastic or not (as the example shown in Figure 8 (a)); 2) the image-modality contains some data metrics. We need to understand the data metrics to get the answer of the multimodal sarcasm detection (as the example shown in Figure 8 (b)). This may be because both object detection and OCR are inadequate to make sense of the realistic visual information from the images. Therefore, our future research can consider exploring background knowledge of the image-modality, such as facial expression source, emoji dictionary, visual commonsense knowledge, etc., to further improve the performance of the multimodal sarcasm detection task.

6 Conclusion and Future Work

In this paper, we propose a novel multimodal graph contrastive learning MMGCL framework for multimodal sarcasm detection, in which a novel multimodal graph structure is devised to leverage the multimodal sarcastic cues between modalities and a graph-oriented contrastive learning strategy is explored to improve the multimodal representations for sarcasm information extraction. To be specific, we utilize object detection to derive the crucial visual regions accompanied by their *attribute-object* captions for the image-modality and employ OCR to make sense of the textual descriptions in the images. This essentially allows the model more easily to understand the significant visual information and semantic expression of the image-modality, so as to better reason the sarcastic expressions of the images. Then, to leverage the intricate multimodal sarcastic relations, we build a multimodal graph for each instance based on the context of text-modality, the textual descriptions extracted by OCR, and the crucial visual regions derived by object detection. In the multimodal graph, we explore word similarity to explicitly link the associated tokens for learning the correlative semantic relations between modalities and exploit an external affective knowledge source to present the sentiment inconsistency for leveraging the incongruous sentiment expression of sarcasm. Afterwards, a graph-oriented contrastive learning is devised to capture the similarity between examples in the same sarcasm class and contrast them with examples in other classes, allowing better multimodal representations for sarcasm detection. Extensive experimental results on a public benchmark dataset show that the proposed MMGCL framework significantly outperforms state-of-the-art baseline methods in multimodal sarcasm detection.

As described in Section 5.7, there are still two typical misclassified examples regarding the proposed MMGCL. Therefore, future research can consider exploiting practicable approaches to improve the learning of facial expression/gesture and statistics/data metrics for the image-modality. Such as incorporating facial expression source, emoji dictionary, or visual commonsense knowledge into the graph structure to fuse the multimodal features for multimodal sarcasm detection.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (61876053, 62006062, 62176076), UK Engineering and Physical Sciences Research Council (grant no. EP/V048597/1, EP/T017112/1), Shenzhen Foundational Research Funding (JCYJ20180507183527919 and JCYJ20200109113441941), Joint Lab of Lab of HITSZ and China Merchants Securities. Yulan He is supported by a Turing AI Fellowship funded by the UK Research and Innovation (UKRI) (grant no. EP/V020579/1).

References

[1] R. W. Gibbs, "On the psycholinguistics of sarcasm." *Journal of experimental psychology: general*, vol. 115, no. 1, p. 3, 1986.

[2] S. Dews and E. Winner, "Muting the meaning a social function of irony," *Metaphor and Symbol*, vol. 10, no. 1, pp. 3–19, 1995.

[3] R. W. Gibbs, "On the psycholinguistics of sarcasm," *Irony in language and thougt: A cognitive science reader*, pp. 173–200, 2007.

[4] E. Cambria, R. Mao, M. Chen, Z. Wang, and S.-B. Ho, "Seven pillars for the future of artificial intelligence," *IEEE Intelligent Systems*, vol. 38, no. 6, pp. 62–69, 2023.

[5] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang, "Sarcasm as contrast between a positive sentiment and negative situation," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, 2013, pp. 704–714. [Online]. Available: https://aclanthology.org/D13-1066

[6] A. Joshi, V. Sharma, and P. Bhattacharyya, "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: Association for Computational Linguistics, 2015, pp. 757–762. [Online]. Available: https://aclanthology.org/P15-2124

[7] T. Xiong, P. Zhang, H. Zhu, and Y. Yang, "Sarcasm detection with self-matching networks and low-rank bilinear pooling," in *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, L. Liu, R. W. White, A. Mantrach, F. Silvestri, J. J. McAuley, R. Baeza-Yates, and L. Zia, Eds. ACM, 2019, pp. 2115–2124. [Online]. Available: https://doi.org/10.1145/3308558.3313735

[8] R. Schifanella, P. de Juan, J. R. Tetreault, and L. Cao, "Detecting sarcasm in multimodal social platforms," in *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*, 2016, pp. 1136–1145. [Online]. Available: https://doi.org/10.1145/2964284.2964321

[9] Y. Cai, H. Cai, and X. Wan, "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: Association for Computational Linguistics, 2019, pp. 2506–2515. [Online]. Available: https://aclanthology.org/P19-1239

[10] N. Xu, Z. Zeng, and W. Mao, "Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020, pp. 3777–3786. [Online]. Available: https://aclanthology.org/2020.acl-main.349

[11] H. Pan, Z. Lin, P. Fu, Y. Qi, and W. Wang, "Modeling intra and inter-modality incongruity for multi-modal sarcasm detection," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, 2020, pp. 1383–1392. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.124

[12] B. Liang, C. Lou, X. Li, L. Gui, M. Yang, and R. Xu, "Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs," in *Proceedings of the 29th ACM International Conference on Multimedia*. Association for Computing Machinery, 2021, pp. 4707–4715. [Online]. Available: https://doi.org/10.1145/3474085.3475190

[13] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 6077–6086. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Anderson_Bottom-Up_and_Top-Down_CVPR_2018_paper.html

[14] E. Cambria, Q. Liu, S. Decherchi, F. Xing, and K. Kwok, "SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis," in *LREC*, 2022, pp. 3829–3839.

[15] D. Bamman and N. A. Smith, "Contextualized sarcasm detection on twitter," in *Ninth international AAAI conference on web and social media*, 2015, pp. 574–577.

[16] D. Davidov, O. Tsur, and A. Rappoport, "Semi-supervised recognition of sarcasm in Twitter and Amazon," in *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*. Uppsala, Sweden: Association for Computational Linguistics, 2010, pp. 107–116. [Online]. Available: https://aclanthology.org/W10-2914

[17] R. González-Ibáñez, S. Muresan, and N. Wacholder, "Identifying sarcasm in Twitter: A closer look," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, 2011, pp. 581–586. [Online]. Available: https://aclanthology.org/P11-2102

[18] E. Lunando and A. Purwarianti, "Indonesian social media sentiment analysis with sarcasm detection," in *2013 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. IEEE, 2013, pp. 195–198.

[19] M. Zhang, Y. Zhang, and G. Fu, "Tweet sarcasm detection using deep neural network," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, 2016, pp. 2449–2460. [Online]. Available: https://aclanthology.org/C16-1231

[20] S. Poria, E. Cambria, D. Hazarika, and P. Vij, "A deeper look into sarcastic tweets using deep convolutional neural networks," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, 2016, pp. 1601–1612. [Online]. Available: https://aclanthology.org/C16-1151

[21] N. Majumder, S. Poria, H. Peng, N. Chhaya, E. Cambria, and A. Gelbukh, "Sentiment and sarcasm classification with multitask learning," *IEEE Intelligent Systems*, vol. 34, no. 3, pp. 38–43, 2019.

[22] Y. Tay, A. T. Luu, S. C. Hui, and J. Su, "Reasoning with sarcasm by reading in-between," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 1010–1020. [Online]. Available: https://aclanthology.org/P18-1093

[23] C. Lou, B. Liang, L. Gui, Y. He, Y. Dang, and R. Xu, "Affective dependency graph for sarcasm detection," in *the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*, 2021, pp. 1844–1849.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[25] N. Babanejad, H. Davoudi, A. An, and M. Papagelis, "Affective and contextual embedding for sarcasm detection," in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020, pp. 225–243. [Online]. Available: https://aclanthology.org/2020.coling-main.20

[26] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, "Towards multimodal sarcasm detection (an _Obviously_ perfect paper)," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 4619–4629. [Online]. Available: https://aclanthology.org/P19-1455

[27] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=SJU4ayYgl

[28] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. [Online]. Available: https://openreview.net/forum?id=rJXMpikCZ

[29] J. Wu, L. Wang, L. Wang, J. Guo, and G. Wu, "Learning actor relation graphs for group activity recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 9964–9974. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Wu_Learning_Actor_Relation_Graphs_for_Group_Activity_Recognition_CVPR_2019_paper.html

[30] K. Zhang, Y. Li, J. Wang, E. Cambria, and X. Li, "Real-time video emotion recognition based on reinforcement learning and domain knowledge," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[31] G.-S. Xie, J. Liu, H. Xiong, and L. Shao, "Scale-aware graph neural

network for few-shot semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 5475–5484.

[32] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 7370–7377.

[33] B. Liang, R. Yin, L. Gui, J. Du, and R. Xu, "Jointly learning aspect-focused and inter-aspect relations with graph convolutional networks for aspect sentiment analysis," in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020, pp. 150–161. [Online]. Available: https://aclanthology.org/2020.coling-main.13

[34] B. Liang, Y. Fu, L. Gui, M. Yang, J. Du, Y. He, and R. Xu, "Target-adaptive graph for cross-target stance detection," in *the Web Conference 2021 (WWW '21)*, 2021.

[35] B. Liang, H. Su, R. Yin, L. Gui, M. Yang, Q. Zhao, X. Yu, and R. Xu, "Beta distribution guided aspect-aware graph for aspect category sentiment analysis with affective knowledge," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 208–218. [Online]. Available: https://aclanthology.org/2021.emnlp-main.19

[36] B. Liang, H. Su, L. Gui, E. Cambria, and R. Xu, "Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks," *Knowledge-Based Systems*, vol. 235, p. 107643, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705121009059

[37] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec, "Graph convolutional neural networks for web-scale recommender systems," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, Y. Guo and F. Farooq, Eds. ACM, 2018, pp. 974–983. [Online]. Available: https://doi.org/10.1145/3219819.3219890

[38] H. Wang, F. Zhang, M. Zhang, J. Leskovec, M. Zhao, W. Li, and Z. Wang, "Knowledge-aware graph neural networks with label smoothness regularization for recommender systems," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, A. Teredesai, V. Kumar, Y. Li, R. Rosales, E. Terzi, and G. Karypis, Eds. ACM, 2019, pp. 968–977. [Online]. Available: https://doi.org/10.1145/3292500.3330836

[39] Q. Tan, N. Liu, X. Zhao, H. Yang, J. Zhou, and X. Hu, "Learning to hash with graph neural networks for recommender systems," in *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, Y. Huang, I. King, T. Liu, and M. van Steen, Eds. ACM / IW3C2, 2020, pp. 1988–1998. [Online]. Available: https://doi.org/10.1145/3366423.3380266

[40] X. Yang, S. Feng, Y. Zhang, and D. Wang, "Multimodal sentiment detection based on multi-channel graph neural networks," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, 2021, pp. 328–339. [Online]. Available: https://aclanthology.org/2021.acl-long.28

[41] Y. Yin, F. Meng, J. Su, C. Zhou, Z. Yang, J. Zhou, and J. Luo, "A novel graph-based multi-modal fusion encoder for neural machine translation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020, pp. 3025–3035. [Online]. Available: https://aclanthology.org/2020.acl-main.273

[42] Y. Zeng, D. Cao, X. Wei, M. Liu, Z. Zhao, and Z. Qin, "Multi-modal relational graph for cross-modal video moment retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2215–2224.

[43] D. Zhang, S. Wei, S. Li, H. Wu, Q. Zhu, and G. Zhou, "Multi-modal graph fusion for named entity recognition with targeted visual guidance," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 16, 2021, pp. 14 347–14 355.

[44] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[45] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 3733–3742. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Wu_Unsupervised_Feature_Learning_CVPR_2018_paper.html

[46] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 2020, pp. 776–794.

[47] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 1597–1607. [Online]. Available: http://proceedings.mlr.press/v119/chen20j.html

[48] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html

[49] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/fcbc95ccdd551da181207c0c1400c655-Abstract.html

[50] D. Zhang, F. Nan, X. Wei, S.-W. Li, H. Zhu, K. McKeown, R. Nallapati, A. O. Arnold, and B. Xiang, "Supporting clustering with contrastive learning," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 2021, pp. 5419–5430. [Online]. Available: https://aclanthology.org/2021.naacl-main.427

[51] Z. Wang, X. Wang, X. Han, Y. Lin, L. Hou, Z. Liu, P. Li, J. Li, and J. Zhou, "CLEVE: Contrastive Pre-training for Event Extraction," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, 2021, pp. 6283–6297. [Online]. Available: https://aclanthology.org/2021.acl-long.491

[52] B. Gunel, J. Du, A. Conneau, and V. Stoyanov, "Supervised contrastive learning for pre-trained language model fine-tuning," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=cu7IUiOhujH

[53] M. Kachuee, H. Yuan, Y.-B. Kim, and S. Lee, "Self-supervised contrastive learning for efficient user satisfaction prediction in conversational agents," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 2021, pp. 4053–4064. [Online]. Available: https://aclanthology.org/2021.naacl-main.319

[54] Y. Qin, Y. Lin, R. Takanobu, Z. Liu, P. Li, H. Ji, M. Huang, M. Sun, and J. Zhou, "ERICA: Improving entity and relation understanding for pre-trained language models via contrastive learning," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, 2021, pp. 3350–3363. [Online]. Available: https://aclanthology.org/2021.acl-long.260

[55] N. Yang, F. Wei, B. Jiao, D. Jiang, and L. Yang, "xMoCo: Cross momentum contrastive learning for open-domain question answering," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, 2021, pp. 6120–6129. [Online]. Available: https://aclanthology.org/2021.acl-long.477

[56] Y. Liu and P. Liu, "SimCLS: A simple framework for contrastive learning of abstractive summarization," in *Proceedings of the 59th Annual Meeting of the Association for Computational*

Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Online: Association for Computational Linguistics, 2021, pp. 1065–1072. [Online]. Available: https://aclanthology.org/2021.acl-short.135

[57] B. Liang, W. Luo, X. Li, L. Gui, M. Yang, X. Yu, and R. Xu, *Enhancing Aspect-Based Sentiment Analysis with Supervised Contrastive Learning*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 3242–3247. [Online]. Available: https://doi.org/10.1145/3459637.3482096

[58] Y. Liu, Q. Fan, S. Zhang, H. Dong, T. Funkhouser, and L. Yi, "Contrastive multimodal fusion with tupleinfonce," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 754–763.

[59] X. Yuan, Z. Lin, J. Kuen, J. Zhang, Y. Wang, M. Maire, A. Kale, and B. Faieta, "Multimodal contrastive training for visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6995–7004.

[60] N. Pielawski, E. Wetzer, J. Öfverstedt, J. Lu, C. Wählby, J. Lindblad, and N. Sladoje, "Comir: Contrastive multimodal image representation for registration," *Advances in neural information processing systems*, vol. 33, pp. 18 433–18 444, 2020.

[61] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen, "Graph contrastive learning with augmentations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5812–5823, 2020.

[62] Y. Zhu, Y. Xu, F. Yu, Q. Liu, S. Wu, and L. Wang, "Graph contrastive learning with adaptive augmentation," in *Proceedings of the Web Conference 2021*, 2021, pp. 2069–2080.

[63] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018. [Online]. Available: https://arxiv.org/abs/1807.03748

[64] O. J. Hénaff, "Data-efficient image recognition with contrastive predictive coding," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 4182–4192. [Online]. Available: http://proceedings.mlr.press/v119/henaff20a.html

[65] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[66] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[67] C. Zhang, Q. Li, and D. Song, "Aspect-based sentiment classification with aspect-specific graph convolutional networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 4568–4578. [Online]. Available: https://aclanthology.org/D19-1464

[68] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[69] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1746–1751. [Online]. Available: https://aclanthology.org/D14-1181

[70] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

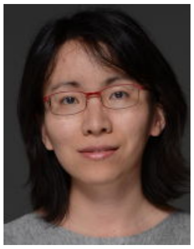
[71] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1532–1543. [Online]. Available: https://aclanthology.org/D14-1162

[72] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9929–9939.

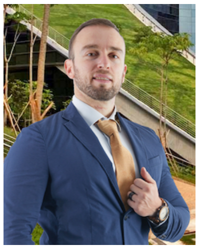
[73] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

Bin Liang is a Ph.D. Student at the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China. His current research interests include natural language processing, sentiment analysis, deep learning and machine learning.

Lin Gui is a Research Fellow at the Department of Computer Science in the University of Warwick. He received Ph.D. degree in Computer Science from Harbin Institute of Technology, China, specializing in Machine Learning, text mining, natural language processing. His research interests include text classification algorithms, natural language understanding, sentiment analysis, stance detection, emotion cause detection, and topic modeling.



Yulan He is a Professor of Computer Science in the University of Warwick, U.K. She is a Turing AI Fellow. She has published over 170 papers in the areas of natural language understanding, sentiment analysis and opinion mining, question-answering, topic/event extraction from text, biomedical text mining, and social media analytics. Yulan received her Ph.D. degree in spoken language understanding from the University of Cambridge, U.K.



Erik Cambria is the Founder of SenticNet, a Professor at NTU, where he also holds the appointment of Provost Chair in Computer Science and Engineering. Prior to joining NTU, he worked at Microsoft Research Asia (Beijing) and HP Labs India (Bangalore) and earned his PhD through a joint programme between the University of Stirling and MIT Media Lab. His research focuses on neurosymbolic AI for explainable natural language processing in domains like sentiment analysis, dialogue systems, and financial forecasting.

Ruifeng Xu received the Ph.D. degree in computer science from The Hong Kong Polytechnic University. He is currently a Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China. He has published more than 150 papers in natural language processing, sentiment analysis, and social media analysis.