F-042

Multimodal Sarcasm Analysis: Sarcastic Cues Capture from Both Emotional Polarity and Semantic Information

Ziwei Yang†,‡ Jiajia Tang† Feiwei Zhou† Teruki Toya‡ Kenji Ozawa‡ Wanzeng Kong†

Abstract

Multimodal sarcasm analysis exhibits distinct characteristics compared to other emotions. The intended meaning of a speaker with sarcasm is typically opposite to the literal meaning. Most of the existing studies try to capture the semantic satirical cues from modalities to improve the ability of sarcasm analysis. However, sarcasm is a complex emotion including different categories and expression forms, which makes the satirical features of semantic content diverse. Consequently, it is difficult to accurately extract the semantic satirical cues of different sarcasm. In this paper, we capture satirical cues from both commonality (emotional polarity) and individuality (semantic information) aspects to cope with this problem.

1. INTRODUCTION

The development of sentiment analysis has greatly facilitated many fields. For instance, enterprises can use it to analyze customers' online reviews to obtain their preferences and improve their products [1]. As a kind of human emotion, sarcasm exhibits unique characteristics. When an individual conveys sarcasm, the literal meaning of their expression is contrary to their actual intentions [2]. Sarcasm analysis is more complex and challenging than regular sentiment analysis. Sarcasm has diverse means of expression and different types, moreover, ordinary emotions typically contain only one emotional polarity. For example, happiness contains positive factors, while anger contains negative factors. In contrast, sarcasm contains both positive and negative emotional factors [3], because sarcasm is incongruous across modalities [4]. However, it is not only the distinguishing attribute of sarcasm from other ordinary emotions but also the commonness of different types and expressions of sarcasm.

Most of the early research on sarcasm analysis focuses on unimodality, especially textual modality whose data are usually derived from social platforms such as Twitter. For example, Carvalho et al. [5] have proposed a method of sarcasm analysis using features such as emoticons and special punctuation, mainly to do feature analysis on the utterance-level. However, for sarcasm, due to the incongruity between modalities, different modalities carry crucial unique information, which makes multimodal methods more advantageous compared to unimodal methods. Since Schifanella et al. [6] first proposed the multimodal sarcasm analysis task, many researchers have focused on this aspect. For example, Liang et al. [7] utilize a cross-modal graph convolutional network model for the fusion of textual modality and visual modality, and they are the first one to explore the use of the graph model based on auxiliary object detection for modeling the contradictory sentiments between key textual and visual information in multimodal sarcasm analysis. Most of the previous methods are devoted to extracting semantic sarcastic cues from different modalities for sarcasm analysis, but for different types and expressions of sarcasm, the cues are very different and sometimes even fuzzy, so it is very difficult to accurately extract semantic sarcastic cues from different sarcasm.

In order to overcome these problems, we proposed a new method to extract sarcasm cues in both commonality and individuality aspects. Specifically, at the commonality level, i.e., the emotional polarity level, we use the emotional polarity factor capture components to capture the positive and negative factors of the fused representations based on the commonality that all kinds of sarcasm contain both positive and negative emotional factors; at the individuality level, i.e., the semantic information level, each utterance has its own semantic information which contains specific sarcastic cues, so we use the specific information capture component to obtain the specific semantic information of each fused representations as an information supplement to make up for the omission of the content and details. Owing to the combination of commonality and individuality, we not only use the semantic sarcastic cues of each utterance to prevent the omission of information but also control the whole from a broader and more general level, which reduces the difficulty of detecting different types and expressions of sarcasm to a certain extent.

2. RELATED WORK

2.1 Multimodal sentiment analysis

As the amount of multimodal data in social media continues to increase, sentiment analysis using multimodal data has gradually attracted people's attention. Ha et al. [8] used both feature and decision-level fusion methods to fuse the information extracted from the different modalities. Chaturvedi et al. [9] introduced a joint model that combined CNN and fuzzy logic, which predicts the degree of a particular emotion. Xu et al. [10] proposed a progressive dual attention module to capture the correlations between image and text, and then learn a joint image-text representation from the perspective of content information. Hazarika et al. [11] presented a new framework named MISA, leveraging two subspaces to represent each modality. One subspace is modality-invariant, intended for learning cross-modal common features, while the other subspace is modality-specific, capturing individual characteristics. Yu et al. [12] designed a label generation module based on the self-supervised learning strategy to acquire independent unimodal supervision. Then, joint training of the multimodal and uni-modal tasks to learn the consistency and difference, respectively. Recently, Mai et al. [13] proposed a new framework HyCon for intra-modal and inter-modal contrastive learning and semi-contrastive learning of trimodal representation to fully explore cross-modal interactions, learn inter-sample and inter-class relationships, and reduced

† Hangzhou Dianzi University
‡ University of Yamanashi

Copyright © 2024 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

modal gaps.

2.2 Multimodal sarcasm analysis

With the rapid development of social media platforms, multimodal sarcasm analysis has received increasing attention. Schifanella et al. [6] first tackled the task of multimodal sarcasm analysis by simply concatenating features from textual and visual modalities. Cai et al. [14] created a multimodal sarcasm dataset and proposed a hierarchical fusion model to integrate features from both textual and visual modalities. Xu et al. [15] introduced a Decomposition and Relation Network (D&R Net) to model the cross-modal contrast and semantic associations between multimodal information. Pan et al. [16] proposed a model based on the BERT architecture to capture the intra-modality and inter-modality incongruity for multi-modal sarcasm detection. Liang et al. [17] explored the incongruity within and between modalities by constructing heterogeneous intra-modal and cross-modal graphs for each multimodal instance. Most recently, Tian et al. [18] employed a dynamic network design for the first time to model and capture cross-modal incongruity dynamically.

However, their methods didn't explore sarcastic cues in terms of commonalities across different types and expressions of sarcasm. It made their methods not comprehensive enough to detect sarcasm, which led to low accuracy.

3. METHODOLOGY

3.1 Preliminaries

The public multimodal sarcasm analysis benchmark MUStARD++ [19] consists of three modalities vision, audio and text, represented as $X_v \in R^{T_v \times d_v}$, $X_a \in R^{T_a \times d_a}$, $X_t \in R^{T_t \times d_t}$. T_i ($i \in$ {a, v, t}) refers to the number of utterances, and the feature dimension is denoted as $d_i (i \in$ {a, v, t}). Note that, each word utterance stands for each time stamp. We also set emotional polarity label for the samples in the benchmarks according to their valence label.

3.2 Modality Encoding and Fusion

Due to the high multimodal dependence and the incongruity between modalities of sarcasm, their emotional polarity and semantic sarcastic cues come not only from the unique information of each modality itself, but also from the new information obtained from intermodal information interactions. Therefore, we perform modal fusion before emotional polarity factor and semantic information capture.

Before fusion, we perform initial feature extraction for each modality. For text modality, we use pre-trained BERT [20] for feature extraction. And for video and audio modalities, we use single-directional Long-Short-Term Memory (sLSTM) networks [21] to extract temporal features. Due to the different sampling rates for video and audio modalities, we set up a convolutional fusion block for 2D temporal convolutional operations, which maps them to the same dimension by controlling the size of the corresponding convolutional kernel. The above procedures are formulated as follows:

$$F_v = sLSTM(I_v; \theta_v^{lstm}) \in R^{d_a} \quad (2)$$

$$F_a = sLSTM(I_a; \theta_a^{lstm}) \in R^{d_v} \quad (3)$$

$$X_{\{v,a\}} = Conv2D(\{v, a\}, k_{\{v,a\}}) \in \mathbb{R}^{T_{\{v,a\}} \times d} \quad (4)$$

where $k_{\{v,a\}}$ is the size of the convolutional kernels for the modalities $\{v, a\}$, θ_v^{lstm} and θ_a^{lstm} are learnable parameters and d is a common dimension.

Then we use multimodal fusion method based on Transformer [22] to fuse text, audio, video modalities and obtain fused utterance vector U $\in R^{T_u \times d_u}$ as follows:

$$U = Fusion(X_v, X_a, X_t; \theta^{pri}). \quad (5)$$

3.3 Emotional polarity factor capture

This process captures positive and negative emotional factor in each utterance representation. Capturing positive emotional factor and negative emotional factor enables the extraction of sarcastic cues at the emotional polarity level. These cues are generalizable to analyzing different types and expressions of sarcasm.

We input each of the fused utterance vectors U into the positive emotional factor capture component E_{pos} and the negative emotional factor capture component E_{neg}. Both of these two components are encoders project the utterance vector U to two different representations respectively to obtain the positive emotional factor vector and the negative emotional factor vector：

$$h_{pos} = E_{pos}(U; \theta^{pos}) \quad (1)$$

$$h_{neg} = E_{neg}(U; \theta^{neg}). \quad (2)$$

3.4 Semantic Information capture

Semantic information capture component extracts semantic information of the fused utterance vectors other than emotional polarity.

Semantic information capture component is also an encoder represented as:

$$h_{pri} = E_{pri}(U; \theta^{pri}). \quad (4)$$

We use loss function to ensure that this component capture different aspects of fused utterance vector U with emotional polarity factor capture component.

Finally, in order to combine the common sarcastic cues and specific sarcastic cues to work together in sarcasm analysis, we federate the obtained emotional polarity factor representations and semantic information representations by multi-head self-attention based on Transformer and feed it into a classifier for prediction.

4. EXPERIMENTS

4.1 Dataset

The MUStARD++ was organized based on the MUStARD dataset by Castro et al. [4] There are 1202 samples in the dataset. And it consists of three modalities: text, video and audio. The videos are clips from the TV series Friends, The Golden Girls, Sarcasmaholics Anonymous and The Big Bang Theory and so on. The dataset doubled the sample size while keeping the ratio of sarcastic to non-sarcastic samples at one to one. It labelled the samples in the entire dataset on the affective dimensions of valence and arousal. And it added explicit emotion labels and implicit emotion labels (the explicit emotion of sarcastic samples is different from the implicit emotion).

Copyright © 2024 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

4.2 Multimodal Feature Extraction

Before the experiments, we perform feature extraction on the data of all three modalities in MUStARD++.

First, we use the same method as Castro et al. [4] to extract the features of video and audio modalities. Specifically, for the video modality, after resizing, centre-cropping and normalizing each frame in the video, visual features are extracted using a pool5 layer of an ImageNet [23] pretrained ResNet-152 [24] image classification model; for audio modality, we use the speech-processing library Librosa [25] to extract the features in the audio data stream, including MFCC, melspectogram, and spectral centre and so on. Finally, for textual modality, we use the the BERT-base-uncased pre-trained model for feature extraction of textual utterance.

4.3 Evaluation Metrics

In this paper, we use the following evaluation metrics to analyze the task performance: binary accuracy (Acc), F-Score (F1), precision and recall. Higher values of these metrics indicate better task performance.

4.4 Baseline

In order to demonstrate the superiority of our model for sarcasm analysis, we introduce state-of-the-art models as baselines: CGMF (the collaborative gating-based multimodal fusion) [19], BERTEmo [26], RoBERTaEmo [26], VADProjWBal [27], PredGaze [28] and SSS [29].

4.5 Experiments Results

We compared the performance of the state-of-the-art models on the benchmark MUStARD++ with our proposed model, and it exceeds the previous best model on metric F1. This result shows that our proposed model can indeed help us improve the ability to analyze complex sarcastic sentiments. By extracting sarcastic cues from both the commonality (emotional polarity) and individuality (semantic information) levels of sarcasm and jointly analyzing them can indeed reduce the problem of analytical difficulties caused by the diversity of sarcastic expressions.

5. CONCLUSION

We proposed an emotional polarity factor capture component and a semantic information capture component to capture emotional sarcastic cues and semantic sarcastic cues in the fused utterance representation respectively. This allows us to simultaneously analyze sarcasm in terms of both commonality and individuality of different types and expressions of sarcasm, which reduces the difficulty of sarcasm analysis brought by content complexity and expression diversity of sarcasm.

Acknowledgement

We are very grateful to some students including Jun Cao, Tianyi Zhao, Haitao Long of Hangzhou Dianzi University for their help during the experiment.

Reference

[1] Z. Wang, S. J. Hu, and W. D. Liu, “Product feature sentiment analysis based on gru-cap considering chinese sarcasm recognition,” Expert Systems with Applications, vol. 241, p. 122512, 2024.

[2] A. T. Zuhri, R. W. Sagala et al., “Irony and sarcasm detection on public figure speech,” Journal of Elementary School Education, pp. 41–45, 2022.

[3] D. S. Chauhan, G. V. Singh, A. Arora, A. Ekbal, and P. Bhattacharyya, “An emoji-aware multitask framework for multimodal sarcasm detection,” Knowledge-Based Systems, vol. 257, p. 109924, 2022

[4] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, “Towards multimodal sarcasm detection (an obviously perfect paper),” arXiv preprint arXiv:1906.01815, 2019.

[5] P. Carvalho, L. Sarmento, M. J. Silva, and E. De Oliveira, “Clues for detecting irony in user-generated contents: oh...!! it’s” so easy”;-,” in Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion, 2009, pp. 53–56.

[6] R. Schifanella, P. De Juan, J. Tetreault, and L. Cao, “Detecting sarcasm in multimodal social platforms,” in Proceedings of the 24th ACM international conference on Multimedia, 2016, pp. 1136–1145.

[7] B. Liang, C. Lou, X. Li, M. Yang, L. Gui, Y. He, W. Pei, and R. Xu, “Multi-modal sarcasm detection via cross-modal graph convolutional network,” in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1. Association for Computational Linguistics, 2022, pp. 1767–1777.

[8] H.-N. Tran and E. Cambria, “Ensemble application of elm and gpu for real-time multimodal sentiment analysis,” Memetic Computing, vol. 10, pp. 3–13, 2018.

[9] I. Chaturvedi, R. Satapathy, S. Cavallari, and E. Cambria, “Fuzzy commonsense reasoning for multimodal sentiment analysis,” Pattern Recognition Letters, vol. 125, pp. 264–270, 2019.

[10] J. Xu, Z. Li, F. Huang, C. Li, and S. Y. Philip, “Social image sentiment analysis by exploiting multimodal content and heterogeneous relations,” IEEE Transactions on Industrial Informatics, vol. 17, no. 4, pp. 2974–2982, 2020.

[11] D. Hazarika, R. Zimmermann, and S. Poria, “Misa: Modality-invariant and-specific representations for multimodal sentiment analysis,” in Proceedings of the 28th ACM international conference on multimedia, 2020, pp. 1122–1131.

[12] W. Yu, H. Xu, Z. Yuan, and J. Wu, “Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis,” in Proceedings of the AAAI conference on artificial intelligence, vol. 35, no. 12, pp. 10 790–10 797, 2021.

[13] S. Mai, Y. Zeng, S. Zheng, and H. Hu, “Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis,” IEEE Transactions on Affective Computing, 2022.

[14] 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, 2019, pp. 2506–2515.

[15] 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, 2020, pp. 3777–3786.

[16] H. Pan, Z. Lin, P. Fu, Y. Qi, and W. Wang, “Modeling intra and intermodality incongruity for multi-modal sarcasm detection,” in Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 1383–1392.

[17] 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, 2021, pp. 4707–4715.

[18] Y. Tian, N. Xu, R. Zhang, and W. Mao, “Dynamic routing transformer network for multimodal sarcasm detection,” in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 2468–2480.

[19] A. Ray, S. Mishra, A. Nunna, and P. Bhattacharyya, “A multimodal corpus for emotion recognition in sarcasm,” arXiv preprint arXiv:2206.02119, 2022.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

Copyright © 2024 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

[21] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems, vol. 30, 2017.
[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in 2009 IEEE conference on computer vision and pattern recognition. Ieee, pp. 248–255, 2009.
[24] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.
[25] B. McFee, M. McVicar, S. Balke, C. Thomé, V. Lostanlen, C. Raffel, D. Lee, O. Nieto, E. Battenberg, D. Ellis et al., “Wzy,” Rachel Bittner, Keunwoo Choi, Pius Friesch, Fabian-Robert Stter, Matt Vollrath, Siddhartha Kumar, nehz, Simon Waloschek, Seth, Rimvydas Naktinis, Douglas Repetto, Curtis Fjord Hawthorne, CJ Carr, Joo Felipe Santos, JackieWu, Erik, and Adrian Holovaty, librosa/librosa: 0.6, vol. 2, 2018.
[26] S. Shah, S. Reddy, and P. Bhattacharyya, “Retrofitting light-weight language models for emotions using supervised contrastive learning,” arXiv preprint arXiv:2310.18930, 2023.
[27] S. Shah, S. Reddy, and P. Bhattacharyya, “Affective retrofitted word embeddings,” in Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2022, pp. 550–561.
[28] D. Tiwari, D. Kanojia, A. Ray, A. Nunna, and P. Bhattacharyya, “Predict and use: Harnessing predicted gaze to improve multimodal sarcasm detection,” in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 15 933–15 948.
[29] S. Bhosale, A. Chaudhuri, A. L. R. Williams, D. Tiwari, A. Dutta, X. Zhu, P. Bhattacharyya, and D. Kanojia, “Sarcasm in sight and sound: Benchmarking and expansion to improve multimodal sarcasm detection,” arXiv preprint arXiv:2310.01430, 2023.

Copyright © 2024 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.