Abstract

Over the past few years, most of the research in sarcasm detection has been based on text and image information. However, sarcasm often exhibits itself through some implicit language and overemphasized expressions such as a sarcastic phrase, an elongated word, or a change in tone. Many traditional methods have been proposed in this field, but the study of deep learning methods to detect sarcasm is still insufficient. Our work argue that most of the image data entering the sarcasm detection model is redundant, such as complex background information and foreground information that is irrelevant to sarcasm detection. Since the details of the face contain emotional changes and social characteristics, we should pay more attention to the image data of the face region. Therefore, considering text, audio, and facial frames as three modalities we propose a multimodal deep learning model to solve this problem. Our model extracts text, audio, and image features of facial regions and then uses our proposed feature fusion strategy to combine these three modal features into a single feature vector for classification. In order to enhance the generalization ability of the model, we use the IMGAUG image enhancement tool to augment the MUStARD, an open source sarcasm detection dataset. Experiments unveils that although a simple supervised method is effective, but using feature fusion strategy and image features from facial regions can further improve the F1 score.

Keywords: Machine Learning, Sarcasm Detection, Multimodal Framework, Irony, Deep Learning