1213Li at SemEval-2021 Task 6: Detection of Propaganda with Multi-modal Attention and Pre-trained Models

Peiguang Li [1,2,3,4], **Xuan Li** [1,2,3,4,*], **Xian Sun** [1,2,3,4,†]
[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China
[2] Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China
[3] University of Chinese Academy of Sciences, Beijing 100190, China
[4] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China
{lipeiguang17, lixuan173}@mails.ucas.ac.cn, sunxian@aircas.ac.cn

Abstract

This paper presents the solution proposed by the *1213Li* team for subtask 3 in SemEval-2021 Task 6: identifying the multiple persuasion techniques used in the multi-modal content of the meme. We explored various approaches in feature extraction and the detection of persuasion labels. Our final model employs pre-trained models including RoBERTa and ResNet-50 as a feature extractor for texts and images, respectively, and adopts a label embedding layer with multi-modal attention mechanism to measure the similarity of labels with the multi-modal information and fuse features for label prediction. Our proposed method outperforms the provided baseline method and achieves 3rd out of 16 participants with 0.54860/0.22830 for Micro/Macro F1 scores.

GET US VACCINATED \n\n NOBODY CARES WHATS IN IT	PICK YOUR ROLE\ n\n OR TAKE THE CHOISE MADE BY THEM
Name calling/Labeling Slogans Smears	Appeal to fear/prejudice Black-and-white Fallacy/Dictatorship

Figure 1: Examples of multi-modal samples, we rewrite the sentences on our own and collect the images from‡, §, ¶, respectively. The first two rows illustrate the visual and the textual content, and in the last row, each line reveals the label (techniques) of the sample.

1 Introduction

The development of the Internet and Information Technology promotes the generation and dissemination of information, but also fuels the proliferation of disinformation. As one of the most popular types of content in disinformation, memes attract readers easily and brought further challenges to the detection of disinformation (Martino et al., 2020; Dimitrov et al., 2021).

Specifically, memes employ a number of techniques to influence users, which can be divided into the use of logical fallacies and appealing to the emotions of the audience (Dimitrov et al., 2021). In practice, the former misuses logical rules to disguise wrong conclusions as correct and objective, while the latter utilizes emotional language to induce the audience to agree with the speaker emotionally and prevent their rational analysis of the argumentation.

Identifying the techniques used in memes contributes to the understanding of user-generated content and further helps to the detection of disinformation. The subtask 3 of SemEval-2021 Task 6 (Dimitrov et al., 2021) is organized to stimulate the study of computational methods to detect persuasion techniques in memes that inhere in texts and images.

As shown in Figure 1, each sample consists of a set of textual sentences and an attached image. According to the task description (Dimitrov et al., 2021), the image and the sentence could convey the modality-specific persuasion techniques, respectively, and at the same time, images can be combined with sentence to express some techniques, which we named global techniques. Based on the understanding of the task, we attribute the main challenges of subtask 3 to the following three as-

*Co-author.
†Corresponding author.
‡https://www.163.com/dy/article/F0HKK63D0511EPAO.html
§http://www.zhujia120.com/fenxi/202103/318716.html
¶https://www.163.com/dy/article/G56M8U7R05521HYB.html

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 1032–1036
Bangkok, Thailand (online), August 5–6, 2021.

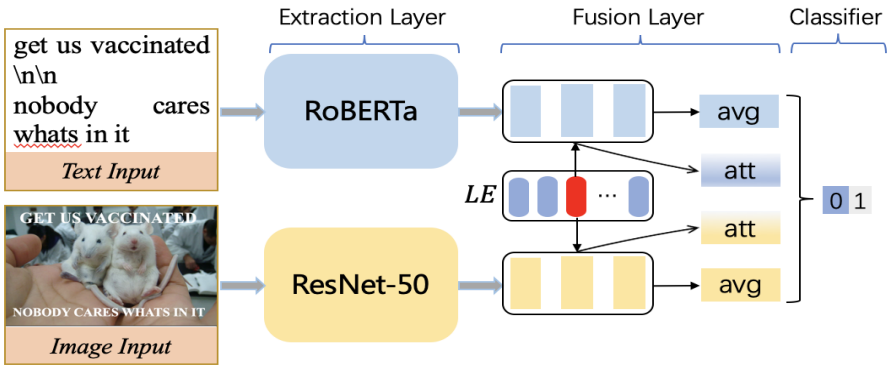

Figure 2: Overview of our proposed method. Our method takes the textual sentences and image as inputs and predicts a binary result for each label. Notably, "LE" in the figure denotes the label embedding module, and this figure illustrates the prediction for the label highlighted in red.

pects: 1) extracting essential features from each modality to predict modality-specific labels, 2) fusing multi-modal features to understand the content fully for predicting global labels, and 3) capturing the connections among multi labels.

Correspondingly, we present the methods to handle these challenges. Specifically, our method employs the powerful feature extractor including the pre-trained RoBERTa (Liu et al., 2019) and ResNet-50 (He et al., 2016) to extract textual and visual features, respectively. Besides, inspired by the work Augenstein et al. (2019), our method adopts a label embedding layer to learn how semantically close the labels are to one another implicitly, and the embedding layer maps each label to a learnable fixed-size vector. Before the label prediction, multi-modal features are fused according to their relevance with each label useing attention mechanism (Bahdanau et al., 2015), and the final prediction is based on the fused features.

2 Related Work

Subtask 3 is a multi-label classification task based on multi-modal data. As for the multi-label classification tasks, it was earlier handled by many machine learning methods. Zhang et al. (Zhang and Zhou, 2005) used a k-nearest neighbor-based algorithm to conduct experiments on real-world multi-label bioinformatic data. Vens et al. (Vens et al., 2008) proposed a hierarchical multi-label classification method based on Decision trees. With the rapid development of deep learning, multi-label classification methods based on deep neural networks have become mainstream. Wang et al. (Wang et al., 2016) introduced and multi-label image classification network with the fusion of CNN and RNN. Chernyavskiy et al. (Chernyavskiy et al., 2020) used a RoBERTa-based network combined with additional CRF (Lafferty et al., 2001) layers and transfer learning mechanism (Pan and Yang, 2010) to address a multi-label classification task in SemEval-2020. However, these previous multi-label classification tasks were often based on single modality data. These approaches fall short when the task requires the use of multiple modal data.

Moreover, in the field of multi-modal tasks, we focus on task of multi-modal fake news detection. Recent work (Jin et al., 2017; Wang et al., 2018; Khattar et al., 2019) mainly concern the fusion of multi-modal features and adopt a binary classifier, which is not applicable to current multi-label classification scenarios.

3 Methodology

3.1 Task Formulation

The task of identifying the techniques used in memes is defined as a multi-label classification problem of given multi-modal sample. We refer the textual sentences as $\mathbf{S}$ and the attached image as $\mathbf{I}$, and use $\mathcal{M}$ to denote the multi-modal model which map inputs $\mathbf{S}$ and $\mathbf{I}$ into a set of N binary values that represent the corresponding label. The task is formulated as follows:

$$\mathcal{M}(\mathcal{F}(\phi(\mathbf{S}), \phi(\mathbf{I}))) \longrightarrow \{0, 1, \ldots, 1\} \quad (1)$$

In the Equation 1, ϕ denotes the multi-modal feature extractor for textual and visual content, respectively, and $\mathcal{F}$ denotes the fusion of the multi-modal features. The length of predicted results is the same as the number of labels and 1 indicates the corresponding label is predicted.

3.2 Method

In this section, we demonstrate the method used by our team for subtask 3. As shown in Figure 2, our method consists of three main layers: Extraction Layer, Fusion Layer, and the final Classifier. In the rest of this subsection, we describe each layer in detail.

3.2.1 Extraction Layer

In the Extraction Layer, the pre-trained RoBERTa is used to extract textual features. Specifically, given that RoBERTa receives at most two sentences as input while some samples may contain multiple pieces of sentences, we splice all sentences into a single sentence and retain the character "\n\n" as the separator. As for the outputs of RoBERTa, we merely reserve the representation of each token as sequential features $\mathbf{T}$ for the post-processing.

For the image input, we use the ResNet-50 pre-trained on ImageNet to extract visual features. Before the image is input to the ResNet-50 network, it needs to be normalized and cropped into 3*224*224. Afterward, we select the last convolution layer's feature maps with size 2048*7*7 as visual features and transform it into a sequential features $\mathbf{V}$ with size of 49 * 2048.

3.2.2 Fusion Layer

The Fusion Layer aims to select the features for the label prediction. As mentioned earlier, the labels implied in the memes include both modality-specific labels and global labels. To promote the prediction of modality-specific labels, we perform the average-pooling on both textual and visual features to extract the modality-specific features $\mathbf{T}_{avg}$ and $\mathbf{V}_{avg}$ ("avg" in Figure 2).

$$\mathbf{T}_{avg} = \text{AvgPooling}(\mathbf{T}) \quad (2)$$

$$\mathbf{V}_{avg} = \text{AvgPooling}(\mathbf{V}) \quad (3)$$

Meanwhile, to promote global labels' prediction, we adopt the attention mechanism to fuse multi-modal features. Particularly, As depicted in Equation 4-6, we first calculate the similarity between ith label embeddings and textual features. We then weighted-sum the textual features according to the similarity scores and obtain label-related representation $\mathbf{T}_{i,att}$ ("att" in Figure 2). The similar operation is applied to the Visual and produce $\mathbf{V}_{i,att}$.

$$S_{i,j} = \mathbf{L}_i \cdot \mathbf{T}_j^T, \forall j \in [1, \ldots, \ell_T] \quad (4)$$

$$\alpha_i = \text{Softmax}\left[S_{i,1}, \ldots, S_{i,n}\right] \quad (5)$$

$$\mathbf{T}_{i,att} = \sum_{j=1}^{\ell_T} \alpha_{i,j}\, \mathbf{T}_j \quad (6)$$

Finally, we concatenate the features obtained above as the final representation of the input and pass it into the Classifier.

$$\mathbf{R}_i = [\mathbf{T}_{avg}; \mathbf{V}_{avg}; \mathbf{T}_{i,att}; \mathbf{V}_{i,att}] \quad (7)$$

3.2.3 Classifier

We adopt a three-layers fully connected network as the classifier, which maps the final representation $\mathbf{R}_i$ obtained ahead into a scalar. Then we employ a sigmoid function to squeeze the scalar to the interval of 0-1. Notably, for each label, the process mentioned above is required and performed synchronously. Hence our model finally outputs a vector whose length is consistent with the number of labels.

4 Experimental Setup

4.1 Dataset

The dataset was provided by SemEval2021 Task6 subtask3, and the training set, development set, and test set contain 687, 63, and 200 samples, respectively. Each sample is combined with an image-text pair, id, and labels.

4.2 Evaluation Measures

The official evaluation measure for this technique classification is Micro-F1. The Macro-F1 is also reported, and we will consider both the performance of Micro-F1 and Macro-F1 during the experiment.

4.3 Parameter Settings

To train the model, we adopt the binary cross-entropy loss as the objective function and employ the Adam method(Kingma and Ba, 2015) with a learning rate of 0.0001 to optimize it. We set the minibatch size at 64 and the dimensions of label embeddings at 256. Based on experimental verification, we fixed the parameters of ResNet-50 while fine-tuning the parameters of RoBERTa during the training. Our methods are implemented with PyTorch and run on a single Nvidia 1080ti graphic card.

Model	Macro F1	Micro F1
RoBERTa+ ResNet-50	0.0812	0.1333
+ visual_att	0.1155	0.4140
+ textual_att	0.0814	0.5256
+ full_att	0.2040	0.5680

Table 1: Ablation results on validation set.

Rank	Team	Macro F1	Micro F1
1	Alpha	0.27315	0.58109
2	MinD	0.24389	0.56623
3	**1213Li**	**0.22830**	**0.54860**
4	AIMH	0.20729	0.53994
5	Volta	0.18877	0.52070
...	...	...	...
16	Baseline	0.05152	0.07062

Table 2: Evaluation results of top 5 teams on blind test set that reported on the official website.

4.4 Ensemble

We use an ensemble of 5 models with different development set to predict the training set. Among the five ensembled models, one model uses the original training set and development set, and the remaining four models use the 64 samples randomly divided from the combined data of the training set and development set as the new development set, and use the rest as the training set.

5 Results and Discussion

The result of the ablation study is shown in Table 1. As we can see, the baseline method is very ineffective since it utilizes only the average-pooling features of visual and textual information, indicating that the lack of the interaction between modality-specific features and label information hinder the model to select vital features for prediction and leads to poor performance.

So we introduce the attention mechanism to selectively choose valid information from visual features and textual features, respectively. As shown in the second group of Table 1, the use of the attention mechanism significantly improves the model's performance, especially the Micro F1 score.

Finally, the model that uses both visual features and textual features in combination with the attention mechanism has the optimal performance. During the test stage, we chose the model that performed best on the development set and got the final result through the ensemble. The final evaluation results are reported in Table 2.

6 Conclusion

This paper demonstrates the method that we proposed for subtask 3 in SemEval-2021 Task 6, which aims to identify which of 22 persuasion techniques are used in the textual and visual content of the specific meme. Our method uses RoBERTa and ResNet-50 to extract multi-modal features, introduces the attention mechanism to fuse multi-modal features, and adopts the label embeddings to learn the representation of labels. Our proposed model achieves noticeable improvements over the baseline method, and the official evaluation ranked our submission 3rd out of 16 teams.

References

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 4684–4696. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. 2020. Aschern at semeval-2020 task 11: It takes three to tango: Roberta, crf, and transfer learning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING 2020, Barcelona (online), December 12-13, 2020*, pages 1462–1468. International Committee for Computational Linguistics.

Dimiter Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Task 6 at semeval-2021: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation*, SemEval '21, Bangkok, Thailand.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Ve-*

gas, NV, USA, June 27-30, 2016, pages 770–778. IEEE Computer Society.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 795–816. ACM.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2915–2921. ACM.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289. Morgan Kaufmann.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. Semeval-2020 task 11: Detection of propaganda techniques in news articles. *CoRR*, abs/2009.02696.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359.

Celine Vens, Jan Struyf, Leander Schietgat, Saso Dzeroski, and Hendrik Blockeel. 2008. Decision trees for hierarchical multi-label classification. *Mach. Learn.*, 73(2):185–214.

Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. 2016. CNN-RNN: A unified framework for multi-label image classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2285–2294. IEEE Computer Society.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 849–857. ACM.

Min-Ling Zhang and Zhi-Hua Zhou. 2005. A k-nearest neighbor based algorithm for multi-label classification. In *2005 IEEE International Conference on Granular Computing, Beijing, China, July 25-27, 2005*, pages 718–721. IEEE.