MinD at SemEval-2021 Task 6: Propaganda Detection using Transfer Learning and Multimodal Fusion

Junfeng Tian, Min Gui, Chenliang Li, Ming Yan, Wenming Xiao
Alibaba Group, China
{tjf141457, guimin.gm, lcl193798, ym119608, wenming.xiaowm}@alibaba-inc.com

Abstract

We describe our systems of subtask1 and subtask3 for SemEval-2021 Task 6 on *Detection of Persuasion Techniques in Texts and Images*. The purpose of subtask1 is to identify propaganda techniques given textual content, and the goal of subtask3 is to detect them given both textual and visual content. For subtask1, we investigate transfer learning based on pre-trained language models (PLMs) such as BERT, RoBERTa to solve data sparsity problems. For subtask3, we extract heterogeneous visual representations (i.e., face features, OCR features, and multimodal representations) and explore various multimodal fusion strategies to combine the textual and visual representations. The official evaluation shows our ensemble model ranks 1st for subtask1 and 2nd for subtask3.

1 Introduction

With the recent interest in "fake news", the detection of propaganda or highly biased texts has emerged as an active research area (Da San Martino et al., 2020, 2019; Chernyavskiy et al., 2020).

SemEval-2021 Task 6 (Dimitrov et al., 2021) provides three subtasks aiming to detect persuasion techniques in texts and images. We participate in subtask1 and subtask3, which are defined as follows:

- **subtask1:** Given only the "textual content" of a meme, identify which of the 20 techniques are used in it. This is a multilabel classification problem.
- **subtask3:** Given a meme, identify which of the 22 techniques are used both in the textual and visual content of the meme (multimodal task).

For subtask1, we focus on using transfer learning to tackle problems related to the scarcity of data since deep learning models require a whole lot of data while it is difficult to obtain vast amount of the labeled data. Especially, we first fine-tune the pre-trained language models on an external dataset from SemEval-2020 Task 11 (Da San Martino et al., 2020) and then continue to fine-tune them on the training dataset of SemEval-2021 Task 6. The probabilities of these tuned models are averaged to make the final prediction.

For subtask3, we concentrate on multimodal fusion to combine textual and visual representation. Heterogeneous visual representations are extracted, including face, OCR and multimodal representations. Face representation consists of recognized human faces and facial expressions. OCR representation can capture the relations among snippets in an image. Multimodal pre-trained model is capable of simultaneously processing multimodality inputs for joint visual and textual understanding. After that, we explore three multimodal fusion strategies (i.e., Average, Concat and MLP) to combine the textual and visual representations.

The experimental results show that transfer learning can leverage knowledge from source data to tackle problems related to the scarcity of data, and heterogeneous visual representation (i.e., face, OCR, and multimodal representation) can be used as complementary features to better detect persuasion techniques. Our ensemble model ranks 1st for subtask1 and 2nd for subtask3.

2 System Overview

In this section, we provide a general overview of our systems for the two subtasks. We consider the propaganda detection task as multimodal multi-class multi-label classification task, predicting one or more labels given an input text and an input image.

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 1082–1087
Bangkok, Thailand (online), August 5–6, 2021. ©2021 Association for Computational Linguistics

2.1 Model

Various pre-trained models are explored to extract textual and visual features, and these textual and visual features are fused to predict labels.

Textual Representation In this paper, five pre-trained language models (PLMs) are used. Representations of the special token `[CLS]` are passed to the classification layer. We briefly describe each PLM:

- **BERT** (Devlin et al., 2019) is a powerful transformer-based PLM and enables bidirectional training using a "masked language model" (MLM) pre-training objective. The masked language model randomly masks some input tokens and aims to predict the masked tokens. BERT also use next sentence prediction (NSP) objective during pretraining, which is a binary classification loss for predicting whether two segments follow each other in the original text. With tailored finetune objectives, BERT can improve performance on downstream tasks such as classification tasks.

- **RoBERTa** (Liu et al., 2019) proposes an improved recipe for training BERT models and boosts the performance on GLUE(Wang et al., 2019), RACE(Lai et al., 2017) and SQuAD(Rajpurkar et al., 2016). It shares the same model architecture with BERT, and mainly improves BERT by dynamic masking and a larger byte-level Byte-Pair Encoding (BPE)(Sennrich et al., 2016).

- **XLNet** (Yang et al., 2019) integrates the segment recurrence mechanism and relative encoding scheme of Transformer-XL(Dai et al., 2019) into pretraining with reparameterizing. It can capture the dependency between the masked positions and alleviate a pretrain-finetune discrepancy.

- **DeBERTa** (He et al., 2020) disentangles attention mechanism and encodes each word with two vectors representing content and position, respectively. An enhanced mask decoder is also used to incorporate absolute positions in the decoding layer to predict the masked tokens in model pre-training. These methods enables DeBERTa to obtain competitive performance of both natural language understand (NLU) and natural language generation (NLG) downstream tasks.

- **ALBERT** (Lan et al., 2019) replaces the next sentence prediction (NSP) loss with a sentence order prediction (SOP) loss to better model inter-sentence coherence. Besides, it equips two parameter reduction techniques to lower memory consumption and increase the training speed of BERT. With fewer parameters compared to BERT-large, ALBERT establishes new state-of-the-art results on the GLUE, RACE, and SQuAD benchmarks.

Visual Representation Three visual representations are adopted, including face representation, OCR representation and single-stream multimodal representation.

- **Face Representation** It is important to recognize faces and facial expressions for propaganda detection. We use a state-of-the-art *face recognition* model, which is a ResNet-34 network (He et al., 2016) with 29 conv layers. In an image, each face is encoded as a 128 dimensional vector using the published toolkit[1] and adopt mean pooling for the final face representation.

- **OCR Representation** For text in an image, the 2-D position of the text can capture the font size and the relationship among tokens within the image. Therefore, we use a 2-D position embedding to jointly model interactions between text and layout information across the image. We extract the bounding box 2-D position using the Microsoft OCR[2].

- **Multimodal Representation** Recent studies on vision-language pre-training have pushed the limits of a variety of Vision-and-Language (V+L) tasks, and both the image and text content can help understand the semantics of the meme for propaganda detection. Therefore, we also extract a region-based image features with Faster R-CNN (Ren et al., 2015) to represent the image. Then, we follow (Li et al., 2021) and use a pre-trained multi-modality model SemVLP to better learn the multi-modal fusion between the image and text.

Multimodal Fusion For multimodal propaganda detection, we employ 3 fusing methods to combine the textual and visual features.

[1] https://github.com/ageitgey/face_recognition

[2] https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/concept-recognizing-text

- **Average** The predicted probabilities of text and image features are averaged for prediction:

$$\hat{y_c} = (\text{sigmoid}(W_1(\tanh(W^t h^t) + b_1) + \text{sigmoid}(W_2(\tanh(W^v h^v) + b_2))/2$$

  where h_t and h_v stand for textual and visual representations, respectively.

- **Concat** The text and image features is concatenated to predict probabilities:

$$\hat{y_c} = \text{sigmoid}(W[h^t, h^v] + b)$$

- **MLP** Before making prediction, we map text and image features to the same semantic space:

$$\hat{y_c} = \text{sigmoid}(W(\tanh(W^t h^t + W^v h^v) + b))$$

2.2 Training

Multilabel Classifier We provide an additional label-wise feed-forward network(FFN) and a linear layer to extract label. At training time, we propose to minimize the binary cross-entropy (BCE) objective $\mathcal{L}$ as follows: $\mathcal{L}_{\text{BCE}}(\hat{y_c}, y_c) = -y_c \log \hat{y_c} - (1 - y_c) \log (1 - \hat{y_c})$ where y_c is the ground truth of class c and $\hat{y_c}$ is the predicted value. At test time, we predict the label as $\tilde{y_c} = \mathbb{I}(\hat{y_c} > T)$ where T is a probability threshold and $\mathbb{I}$ is the indicator function.

As for label imbalance problem, focal loss (FL) (Lin et al., 2017), which down-weights easy examples and focus training on hard negatives, is adopted during training.

Transfer Learning It is difficult to get vast amounts of labeled data for supervised models. Transfer Learning enables us to utilize knowledge from previously learned tasks and apply them to newer, related ones. We use transfer learning from the news articles domain: we first train the model using the news data, and then we continue training for this task. In preliminary experiments, we find that fine-tuning layers in the process is better than freezing them as feature extractors.

3 Experimental Setup

3.1 Dataset

We conduct experiments with the train, the dev and the test datasets provided by SemEval-2021 Task 6 (Dimitrov et al., 2021), which contains 687, 63 and 200 memes for subtask1 and subtask3, respectively.

External Resources We use the annotations of the PTC corpus (more than 20,000 sentences) from SemEval-2020 task 11 (Da San Martino et al., 2020) as external resource. Although its domain is news articles and fewer techniques are considered, the annotations are made using the same guidelines as SemEval-2021 task 6.

3.2 Evaluation Measures

Subtask1 and subtask3 are multi-label classification tasks. The official evaluation measure for both tasks is micro-F1. We also report macro-F1.

3.3 Parameter Settings

We adopt the large models and select hyper-parameters using validation on a subsample of the training data. The cased models are used because that upper cases contain strong emotion signals in this task. We use adamW optimizer(Loshchilov and Hutter, 2019) with 500 warm-up steps and train for 10 epochs with a 2e-5 learning rate and a 8 batch size. The last checkpoint is used for evaluation.

3.4 Submitted Systems

Post-processing *Repetition* means repeating the same message over and over again so that the audience will eventually accept it. Therefore, we assign a *Repetition* label in case if there exists a bigram appears more than 3 times.

Ensemble We use model ensemble for final submission. In particular, for subtask1 we explore 5 pre-trained models (using BCE Loss, Focal Loss and Transfer Learning, respectively), and for subtask3 we additionally explore face, OCR, multimodal representations and the fusion strategies. We take the probabilities of these settings and average them to make the final prediction.

3.5 Test Results

Table 1 and Table 2 list the results of the top-performing teams for subtask1 and subtask3. We can see that our proposed model is ranked 1st for subtask1 and 2nd for subtask3 among all teams.

4 Discussion

More thorough studies and analyses are conducted in this section, trying to answer two questions: (1) How is the performance of transfer learning on less data? (2) How is the performance of multimodal fusion on multimodal data? Moreover, we give

Rank	**Team**	F1-Macro	**F1-Micro**
1	**MinD**	**0.28993**	**0.59331**
2	Alpha	0.26218	0.57187
3	Volta	0.26621	0.56958
	Baseline	0.04427	0.06439

Table 1: Results of top 3 teams for **subtask1 (test)**.

Rank	**Team**	F1-Macro	**F1-Micro**
1	Alpha	**0.27315**	**0.58109**
2	**MinD**	0.24389	0.56623
3	1213Li	0.22830	0.54860
	Baseline	0.05152	0.07062

Table 2: Results of top 3 teams for **subtask3 (test)**.

error analyses on the test dataset to provide an overview of problematic labels.

4.1 Transfer Learning

We perform ablation study for each PLM (row) and each learning method (column) in Table 3 for subtask1. It shows that:

First, RoBERTa and DeBERTa were generally the best performing models. Given that RoBERTa and DeBERTa are carefully tuned models base on BERT, this result is reasonable.

Second, both Focal Loss and Transfer Learning help to alleviate data sparsity problems. Focal Loss help DeBERTa and ALBERT improve 0.7 and 0.6 points. Because Focal Loss assigns higher weights to sparse samples and reduces the weights to frequent samples. Transfer Learning helps BERT, RoBERTa, XLNet improve 1.0, 4.0, 5.7 points, respectively. RoBERTa with Transfer Learning achieves the best single model score. Transfer Learning help transfer the parameters trained on related data or task to the newer model. Instead of learning from scratch, the newer model can leverage knowledge to tackle problems related to the scarcity of data.

4.2 Multimodal Fusion

For subtask3, we compare different multimodal representations in Table 4 and fusion strategies in Table 5. We find that:

(1) both OCR Representation and Multimodal Representation models outperform the Text Representation model. OCR Representation can additionally capture the relative space relationship instead of sequential information among texts in an image.

Training PLM	BCE	FL	Transfer
BERT	0.5833	0.5552	**0.5941**
RoBERTa	0.6070	0.5950	**0.6478**
XLNet	0.5573	0.5418	**0.6148**
DeBERTa	0.6307	**0.6378**	0.6230
ALBERT	0.5251	0.5319	**0.5081**

Table 3: Results **(F1-Micro)** for **subtask1 (dev)**. **BCE**, **FL**, **Transfer** stand for models training using BCE Loss, Focal Loss and Transfer Learning, respectively.

Model	F1-Macro	**F1-Micro**
Text Representation	0.2481	0.5012
Face Representation	0.1956	0.2332
OCR Representation	**0.2722**	0.5208
Multimodal Representation	0.2355	**0.5876**

Table 4: Results for **subtask3 (dev)**. We explore various multimodal representations.

(2) Multimodal Representation model achieves the best single model performance since it jointly aligns the semantics between image and text and thus is effective for the vision-language understanding task.

(3) Table 5 lists the results of different fusion strategy. We combine the text and face representations since they are the minimal semantic elements in the image. *Concat* obtains the best result on both Macro-F1 and Micro-F1 metrics, though it is the simplest strategy for fusion.

4.3 Error Analysis

To provide an overview of problematic labels, We give error analysis in Table 6 and Table 7 . We find that: (1) *Loaded Language and Name Calling*, which are the most frequent labels, show reasonably good performance (0.8190 and 0.6667 F1 score).

(2) On the other hand, as to labels with fewer training samples (less than 20), the system tends not to predict. Additionally, we find rules for *Repetition* do not work and all the predicted label are wrongly classified.

(3) *Slogans, Glittering generalities and Smears* are relative hard to identify. Meanwhile, Recall values of *Transfer and Strong Emotions* for subtask3 are less than 0.1. It lacks enough training samples to well fit the network parameters.

Fusion	F1-Macro	F1-Micro
Average	0.3673	**0.6114**
MLP	0.3947	0.6094
Concat	**0.4218**	**0.6114**

Table 5: Results for **subtask3 (dev)**. We explore various multimodal representations.

Label	Precision	Recall	F1	#
Appeal to authority	-	-	-	13
Appeal to fear	0.4615	0.6000	0.5217	43
B&W	0.6667	0.2857	0.4000	18
Oversimplification	0.4000	0.6667	0.5000	27
Doubt	0.5294	0.3214	0.4000	48
Exaggeration	0.5238	0.5789	0.5500	52
Flag-waving	0.5714	0.6667	0.6154	27
Glittering generalities	0.6667	0.1818	0.2857	32
Loaded Language	0.7197	0.9500	0.8190	358
Straw Man	-	-	-	20
Name calling	0.5658	0.8113	0.6667	218
Obfuscation	-	-	-	4
Red Herring	-	-	-	1
Reductio ad hitlerum	-	-	-	9
Repetition	-	-	-	8
Slogans	0.2857	0.1053	0.1538	44
Smears	0.3864	0.7556	0.5113	200
Cliché	-	-	-	20
Whataboutism	0.5000	0.3000	0.3750	40
Bandwagon	-	-	-	2

Table 6: Precision, Recall and F1 of each label for **subtask1 (test)**. The last column (#) stands for the number of training samples.

5 Conclusion

In this paper, we adopt transfer learning to handle data sparsity problems for subtask1, and fuse heterogeneous multimodal representation for subtask3. The experimental results show that transfer learning can leverage knowledge from source data to tackle problems related to the scarcity of data, and heterogeneous visual representation (i.e., face, OCR, and multimodal representation) can extract complementary features.

In future work, we plan to explore fine-grained multimodal fusion with token representations in text and object features in images.

Label	Precision	Recall	F1	#
Appeal to authority	-	-	-	19
Appeal to fear	0.6667	0.2222	0.3333	66
B&W	1.0000	0.2857	0.4444	19
Oversimplification	0.3333	0.7500	0.4615	31
Doubt	0.6364	0.1707	0.2692	61
Exaggeration	0.6667	0.3871	0.4898	60
Flag-waving	0.5556	0.4167	0.4762	36
Glittering generalities	0.2857	0.0833	0.1290	84
Loaded Language	0.7333	0.8800	0.8000	360
Straw Man	-	-	-	32
Name calling	0.6329	0.7692	0.6944	252
Obfuscation	-	-	-	5
Red Herring	-	-	-	2
Reductio ad hitlerum	-	-	-	15
Repetition	-	-	-	10
Slogans	0.2857	0.0909	0.1379	45
Smears	0.5348	0.9524	0.6849	450
Cliché	-	-	-	20
Whataboutism	1.0000	0.1429	0.2500	47
Bandwagon	-	-	-	2
Transfer	0.6667	0.0870	0.1538	61
Strong Emotions	1.0000	0.0526	0.1000	68

Table 7: Precision, Recall and F1 of each label for **subtask3 (test)**. The last column (#) stands for the number of training samples.

References

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*, pages 1462–1468.

Giovanni Da San Martino, Alberto Barrón-Cedeno, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. Semeval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Viet Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2978–2988. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. RACE: large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 785–794. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Chenliang Li, Ming Yan, Haiyang Xu, Fuli Luo, Wei Wang, Bin Bi, and Songfang Huang. 2021. Sem{vlp}: Vision-language pre-training by aligning semantics at multiple levels.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

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. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.