BERT 4EVER@LT-EDI-ACL2022-Detecting signs of Depression from Social Media: Detecting Depression in Social Media using Prompt-Learning and Word-Emotion Cluster

Xiaotian Lin[1†], **Yingwen Fu**[1†*], **Ziyu Yang**[1†], **Nankai Lin**[1†], **and Shengyi Jiang**[2]
Guangdong University of Foreign Studies
[1]{20191002971,20201010002,20201002958,20191010004}@gdufs.edu.cn,
[2]shengyijiang@163.com,

Abstract

In this paper, we report the solution of the team BERT 4EVER for the LT-EDI-2022 shared task2: Homophobia/Transphobia Detection in social media comments in ACL 2022, which aims to classify Youtube comments into one of the following categories: no, moderate, or severe depression. We model the problem as a text classification task and a text generation task and respectively propose two different models for the tasks. To combine the knowledge learned from these two different models, we softly fuse the predicted probabilities of the models above and then select the label with the highest probability as the final output. In addition, multiple augmentation strategies are leveraged to improve the model generalization capability, such as back translation and adversarial training. Experimental results demonstrate the effectiveness of the proposed models and two augmented strategies.

1 Introduction

This paper includes a review and explanation of BERT4EVER's ideas and experiments for the LT-EDI-2022 shared task2 (Sampath et al., 2022): Homophobia/Transphobia Detection in social media comments in ACL 2022. In this task, participants would be given sentences from social media comments and then predict whether they contain any form of homophobia/transphobia. The Homophobia/Transphobia detection dataset (Chakravarthi et al., 2021), a collection of comments from Youtube, serves as the task's seed data. The comments were manually annotated to show whether the text contained homophobia/transphobia. The label annotation consists of three categories: no depression, moderate depression and severe depression. This is a comment/post level classification task. For this task, our solution consists of two main blocks.

- We model this task as a representative text classification task (abbreviated as "classification model"). Several works in literature have explored how to use linguistic and sentiment analysis to detect depression (Xue et al., 2014). For example, (Huang et al., 2014) proposed to explore linguistic features of these known cases using a psychological lexicon dictionary, and train an effective suicidal Weibo post detection model. Furthermore, in order to further investigate the latent information towards the social media, (Aragón et al., 2019) leveraged the model emotions in a fine-grained way to build a new representation for detecting the depression. Inspired by them, we utilize the clustering algorithm to obtain fine-grained emotion embedding of text to better detect depression based on the emotion dictionary.
- Inspired by Prompt-learning (Ding et al., 2021b), we model this task as a text generation task (abbreviated as "generation model"). Prompt-learning is a new paradigm in modern natural language processing to adapt pre-trained language models (PLMs) to downstream NLP tasks, which modifies the input text with a textual template and directly uses PLMs to conduct pre-trained tasks. It directly adapts PLMs to cloze-style prediction, autoregressive modeling, or sequence to sequence generation, resulting in promising performances on various tasks such as text classification (Liu et al., 2021; Gao et al., 2021), named entity typing (Ding et al., 2021a) and relation extraction (Han et al., 2021).

In addition, given the limited amount and category imbalance of training data available, we experiment with multiple augmentation techniques, including continual pre-training (Gururangan et al., 2020), back translation, adversarial training (Miyato et al., 2017), easy ensemble(Liu et al., 2009). Considering the combination of the knowledge

* Corresponding Author.
† Equal contribution.

Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 200 - 205

learned from the different models, we softly fuse the predicted probabilities of the two models above and then select the label with the highest probability as the final output.

We conduct extensive experiments on the given dataset and achieves comparable results which reaches 0.5818 on the test set and 0.5426 in the online evaluation. In summary, our contributions are as follow:

- We model the Detecting signs of Depression task in two dimensions, namely the classification model and the generation model, and then softly ensemble them.
- We adapt several augmentation techniques to alleviate the limited amount and category imbalance problems of training data available in this task.
- Experimental results indicate the effectiveness of the proposed method in this paper.

The rest of this paper is organized as follows: Section 2 describes the details of the proposed method in this paper. Section 3 elaborates the experimental setup and analyzes the experimental results. Finally, conclusions are drawn in Section 4.

2 Methods

2.1 Generation Model

As shown in Figure 1, the prompt-learning pipeline in our system consists of three main cores: a PLM, a template, and a verbalizer. Take a simple sentence in the dataset for example, the template is used to process the original text with some extra tokens, the PLM encodes the text to semantic vectors and the verbalizer projects original labels to words in the vocabulary for final prediction. Given a template "<text> The person has <mask> depression.", where the token <text> indicates the original text, and the verbalizer is "moderate": "moderate", "not depression": "no", "severe": "severe". The sentence "I don't want to live in this fucking world . "would be firstly wrapped by the template as "I don't want to live in this fucking world. The person has <mask> depression.". The wrapped sentence is then tokenized and fed into a PLM to predict the distribution over vocabulary on the <mask> token position. Since the label of this comment is "moderate", it is expected that the word "moderate" should have a larger probability than "severe"and "no".

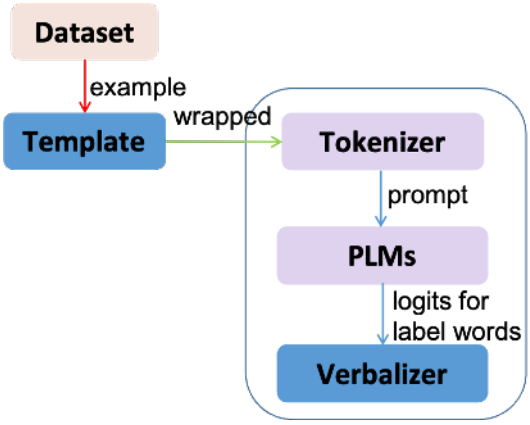

Figure 1: Prompt-learning-based Model

Class	Label Words
moderate	moderate, limited
not depression	no, little, paltry, inappreciable, insignificant, negligible
severe	severe, serious, critical, terrible, hard, high, heavy

Table 1: Label Words.

PLM. We use T5 as our base encoder which explores the landscape of transfer learning techniques for NLP by introducing a unified framework that converts every language problem into a text-to-text format.

Template. In this paper, we explore the effectiveness of two templates for the task. They are "*<text> The person has <mask> depression.*" and "*A comment with <mask> depression: <text>*".

Verbalizer. As for the label words for different class, taking the label imbalance problem into account, we introduce more label words for "not depression"and "severe" classes. The labels words are shown in Table 1.

2.2 Classification Model

Our classification model is composed of three following steps: (1) Generating word-emotion cluster representation; (2) Converting Text according to the word-emotion cluster representation.(3) combining multi-dimension information.

Generating Word-emotion Cluster Representation. Following (Aragón et al., 2019), to generate the fine-grained emotions, we use a lexical with eight recognized emotions, e.g., Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, Trust, Positive and Negative. Specially, provided that the emotions in the dictionary are represented as $E =$

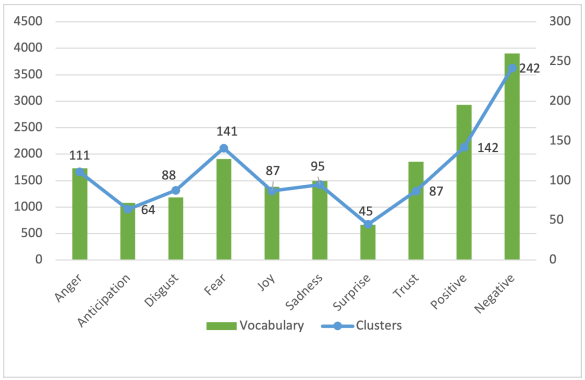

Figure 2: Generated Cluster Distribution

$E_1, E_2, E_3, \ldots, E_{10}$ where $E_i = w_1, w_2, \ldots w_n$ refers to the words set contained in the i-th emotion. For each word of the emotion, we use glove of size 300 to compute its word embedding. Then, we utilize Affinity Propagation (AP) clustering algorithm to cluster each emotion cluster to form multiple sub-emotion clusters. To better obtain the cluster representation for each sub-emotion, we further average the word embedding in each sub-emotion cluster, regarding it as the cluster representation of the sub-emotion cluster. To have an idea of how the vocabulary was distributed among emotions and the number of generated clusters after applying AP to the dictionary we present Figure 2.

Converting Text according to the Word-emotion Cluster Representation. After obtaining all the sub-emotion cluster representation, we can mask each word with the label of its closest sub-emotion cluster for the input sentence. Specially, given an input sentence $S_i = s_1^i, s_2^i, s_3^i, \ldots, s_n^i$, we first compute the vector representation of each word using word embedding from glove, then we measure the distance for each sentence S_i and all of the sub-emotion cluster representations by using Cosine similarity.

Eventually, we substitute each word by the label of its closest fine-grained emotion. For example, given an input sentence "Live in this fucking world", we can mask it as "joy49 positive28 trust0 negative97 anticipation54".

Combining Multi-dimension Information. Through the above steps, we obtain a converted sentence represented by word-emotion cluster representation for each input sentence. Provided a converted sentence is represented as $C_i = \{c_i^1, c_i^2, c_i^3, \ldots, c_i^n\}$ where $c_i^j R^m$ refers to a word-emotion cluster representation of the j-th word in the converted sentence C_i. We further leverage a CNN model with a Maxpooling layer to capture the emotion representation h_i^e for the converted sentence.

$$h_i^e = Maxpooling(Convid(C_i)) \quad (1)$$

where $Convid(C_i)$ represents the convolution operation of the CNN model.

After that, we utilize a PLM to obtain the general semantic information in terms of the input text and add the adversarial perturbations for the input words embedding (Miyato et al., 2017). Eventually, in order to fully learn the general semantic information and emotion information, we further spliced emotion representation output by the PLM to fuse multi-dimensional information and map it to the labels dimension by a fully connected layer.

$$h_i^g = PLM(S_i) \quad (2)$$

$$p = Softmax(W([h_i^g; h_i^e]) + b) \quad (3)$$

where W and b are parameters of the fully connected layer.

Class	Train	Dev	Test
Not depressed	1971	1830	-
Moderate	6019	2306	-
Severe	901	360	-
Total instances	8891	4496	3245

Table 2: Dataset Statistics.

3 Experiment

3.1 Dataset

In this paper, we conduct experiments on the dataset (Kayalvizhi and Thenmozhi, 2022) provided by the competition DepSign-LT-EDI@ACL-2022 which aims to detect the signs of depression of a person from their social media postings wherein people share their feelings and emotions. Across this dataset we have three different classes including "not depressed", "moderately depressed" and "severely depressed". The dataset statistics are shown in the Table 2.

3.2 Experimental Settings

Our models are all implemented with PyTorch[1]. We compare the performance of different pre-trained models for classification models such as

[1]https://pytorch.org/

Num.	Model Type	Model	Test	Online evaluation
0	Classification	Baseline	0.5123	-
1	Classification	Baseline-CP	0.5396	-
2	Classification	AT	0.5504	-
3	Classification	CM + AT	0.5491	-
4	Classification	CM + AT + BackT	0.5598	-
5	Classification	CM + AT + EE	0.5618	-
6	Generation	Template1	0.5409	-
7	Generation	Template2	0.5500	-
8	Generation	Template1 + BackT	0.5569	-
9	Generation	Template2 + BackT	0.5613	-
10	Generation	Template1 + EE	0.5450	-
11	Ensemble	4 +5 + 8 + 9	0.5818	0.5426

Table 3: Main Results. In this table, CP indicates continual pre-training, CM indicates classification model described in section 2.2, AT indicates adversarial training, BackT indicates back translation and EE indicates easy ensemble.

XLM-Base[2] (Conneau et al., 2020), RoBERTa-Base[3] (Conneau et al., 2020), and BERT-Base[4] (Devlin et al., 2019) and for generation models such as T5-base[5](Raffel et al., 2020), BART[6](Lewis et al., 2020). Finally, we respectively choose RoBERTa-Base and T5-Base as the base models for them. Following (Gururangan et al., 2020), in order to adapt the language models to the specific domain, we randomly sample 5 million sentences to continual pre-train the two PLMs. In addition, to fairly evaluate the effectiveness of different models, we leverage 5-fold cross-validation and soft ensemble the 5 models to present the their generalization performances.

Besides, given the limited amount and imbalance distribution of training data available, we introduce two strategies for training models, namely easy ensemble (Liu et al., 2009) and back translation. Specifically, we only conduct back translation on "not depressed" and "severe" samples using Google Translate.

As for evaluation metrics, we use Macro-F1 implemented by scikit-learn[7] for offline evaluation. In addition, we use the available dev data as the offline test set to represent the generation performance of different models.

[2]https://huggingface.co/xlm-roberta-base
[3]https://huggingface.co/roberta-base
[4]https://huggingface.co/bert-base-uncased
[5]https://huggingface.co/t5-base
[6]https://huggingface.co/facebook/bart-base
[7]https://scikit-learn.org/stable/

3.3 Results and Analysis

The main results are shown in the Table 3. Through these results, we can see that for classification models, the PLMs pre-trained for domain adaption outperforms that one without pre-training. This indicates that pre-train enables models to better learn domain-related language knowledge representations. In addition, the AT, AP, back translation and easy ensemble strategies all help to improve the model performance. Since different depression levels may match different sentiments, AP strategy can integrate sentiment knowledge to the model which can effectively enhance the model with the correlation of sentiment and depression level. AT strategy can improve the generalization performance of the model by adding perturbations to embeddings for model augmentation. Back translation and easy ensemble strategies can effectively address the problem of category imbalance and thus can enhance the performance of categories with small-size samples ("not depressed" and "severe"). And it seems that easy ensemble works better than back translation.

Besides, for generation models, similar to classification models, back translation and easy ensemble strategies can also improve the model performance. Interestingly, back translation is much more superior to easy ensemble. We hypothesize the reason is that since generation model require large-size data for training to guarantee effective performance and easy ensemble reduces training samples per fold while back translation increases training samples per fold to a certain extent, so back translation is more advantageous in genera-

tion models.

We ensemble and submit the four models with best offline results (with a macro-F1 score of 0.5818) and achieve an online macro-F1 score of 0.5426.

4 Conclusion

We present our submission to the Shared Task on Homophobia/Transphobia Detection in social media comments at LT-EDI 2022- ACL 2022. We model the problem as two different tasks and propose two novel methods respectively for the tasks. In addition, focusing on two problems of small data size and category imbalance in the original training set, we leverage multiple augmentation strategies to enhance the model performance. We softly fuse the predicted probabilities of different models and output the label with highest probability. We evaluate each single model on the validation set drawn from the training set and provide some explanations. To justify our design choices, we conduct an ablation study. Overall, we achieve the 4th macro averaged F-Score of the dataset on the online evaluation. In the further work, on one hand, we would explore how sentiment knowledge and prompting technique can further improve the model performance. On the other hand, we would expand our emotion lexicon and create more fine-grained representations of word-emotion clusters, allowing the classification model to learn more about emotions.

5 Acknowledgement

This work was supported by the Soft Science Research Project of Guangdong Province (No.2019A101002108), Science and Technology Program of Guangzhou (No.202002030227), and the Key Field Project for Universities of Guangdong Province (No. 2019KZDZX1016).

References

Mario Ezra Aragón, Adrián Pastor López-Monroy, Luis Carlos González-Gurrola, and Manuel Montes-y-Gómez. 2019. Detecting depression in social media using fine-grained emotions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1481–1486. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kayalvizhi Sampath, Durairaj Thenmozhi, Sathiyaraj Thangasamy, Rajendran Nallathambi, and John Philip McCrae. 2021. Dataset for identification of homophobia and transophobia in multilingual youtube comments. *CoRR*, abs/2109.00227.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Ning Ding, Yulin Chen, Xu Han, Guangwei Xu, Pengjun Xie, Hai-Tao Zheng, Zhiyuan Liu, Juanzi Li, and Hong-Gee Kim. 2021a. Prompt-learning for fine-grained entity typing. *CoRR*, abs/2108.10604.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. 2021b. Openprompt: An open-source framework for prompt-learning. *CoRR*, abs/2111.01998.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3816–3830. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8342–8360. Association for Computational Linguistics.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2021. PTR: prompt tuning with rules for text classification. *CoRR*, abs/2105.11259.

Xiaolei Huang, Lei Zhang, David Chiu, Tianli Liu, Xin Li, and Tingshao Zhu. 2014. Detecting suicidal ideation in chinese microblogs with psychological lexicons. In *2014 IEEE 11th Intl Conf on Ubiquitous Intelligence and Computing and 2014 IEEE 11th*

Intl Conf on Autonomic and Trusted Computing and 2014 IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops, Bali, Indonesia, December 9-12, 2014, pages 844–849. IEEE Computer Society.

S Kayalvizhi and D Thenmozhi. 2022. Data set creation and empirical analysis for detecting signs of depression from social media postings. *arXiv preprint arXiv:2202.03047*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. GPT understands, too. *CoRR*, abs/2103.10385.

Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. 2009. Exploratory undersampling for class-imbalance learning. *IEEE Trans. Syst. Man Cybern. Part B*, 39(2):539–550.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Kayalvizhi Sampath, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, and Jerin Mahibha C. 2022. Findings of the shared task on Detecting Signs of Depression from Social Media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Yuanyuan Xue, Qi Li, Li Jin, Ling Feng, David A. Clifton, and Gari D. Clifford. 2014. Detecting adolescent psychological pressures from micro-blog. In *Health Information Science - Third International Conference, HIS 2014, Shenzhen, China, April 22-23, 2014. Proceedings*, volume 8423 of *Lecture Notes in Computer Science*, pages 83–94. Springer.