# GUTS at SemEval-2022 Task 4: Adversarial Training and Balancing Methods for Patronizing and Condescending Language Detection

**Junyu Lu, Hao Zhang, Tongyue Zhang, Hongbo Wang, Haohao Zhu, Bo Xu, Hongfei Lin***
School of Computer Science and Technology, Dalian University of Technology, China
(dutljy,zh373911345,zty9818,hldwhb2017,zhuhh)@mail.dlut.edu.cn
(xubo,hflin)@dlut.edu.cn

## Abstract

Patronizing and Condescending Language (PCL) towards vulnerable communities in general media has been shown to have potentially harmful effects. Due to its subtlety and the good intentions behind its use, the audience is not aware of the language's toxicity. In this paper, we present our method for the SemEval-2022 Task4 titled "Patronizing and Condescending Language Detection". In Subtask A, a binary classification task, we introduce adversarial training based on Fast Gradient Method (FGM) and employ pre-trained model in a unified architecture. For Subtask B, framed as a multi-label classification problem, we utilize various improved multi-label cross-entropy loss functions and analyze the performance of our method. In the final evaluation, our system achieved official rankings of 17/79 and 16/49 on Subtask A and Subtask B, respectively. In addition, we explore the relationship between PCL and emotional polarity and intensity it contains. Our code is available on Github [1].

## 1 Introduction

Patronizing and Condescending Language (PCL) expresses a superior attitude towards vulnerable communities (e.g. women, refugees, poor families), and describes them or their situation in a charitable way that evokes feelings of compassion (Pérez-Almendros et al., 2022). Although it is generally used involuntarily and with good intentions, the use of PCL can potentially be very harmful, as it feeds stereotypes, routinizes discrimination and drives to greater exclusion. Due to the subtlety of PCL, PCL detection is difficult for both humans and NLP systems and has aroused broad attention.

To address the challenge of patronizing and condescending language detection in general media, Pérez-Almendros et al. (2022) introduce the Task 4 at SemEval-2022, and build a dataset with annotated paragraphs extracted from news articles in English. Given a paragraph, systems must predict whether it contains condescending language or not (Subtask A), and whether it contains any of the 7 subtypes identified in the PCL taxonomy (Subtask B).

For Subtask A, a binary classification task, we introduce adversarial training based on Fast Gradient Method (FGM) (Miyato et al., 2016), enhancing the robustness of the model. And in Subtask B, a multi-label classification problem, there is a long-tailed distribution of each label. To address the class imbalance problem, we utilize various improved multi-label cross-entropy loss functions: Focal loss (Lin et al., 2017), Class-balanced focal loss (Cui et al., 2019) and Distribution-balanced loss (Wu et al., 2020). We analyze the performance of our methods and demonstrate the contribution of each component of the architecture.

In addition to completing basic evaluation tasks, we also explore the relationship between PCL and emotional polarity and intensity it contains in official dataset. The experimental results demonstrate that the above two have relevance.

The structure of the paper is as follows: We first provide a brief overview of related research, and then introduce our proposed framework. Besides, experiments and evaluations as well as the analysis of results are given. Finally, we discuss the future directions of our work.

## 2 Related Work

Patronizing and condescending language has been studied extensively in sociolinguistics and the traits of PCL have been suggested by related research. PCL builds stereotypes (Fiske, 1993), which strengthen exclusion, discrimination, rumour spreading (Nolan and Mikami, 2013) and unbalanced power relations (Sap et al., 2019), relying on subtle language (Mendelsohn et al., 2020). It tends

[1] https://github.com/Nutpok/GUTS-at-SemEval-2022-Task-4.git

*Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 432 - 437
July 14-15, 2022 ©2022 Association for Computational Linguistics

to avoid stating the reasons for deep-rooted societal problems by concealing those responsible and proposes temporary solutions (Chouliaraki, 2010), which oversimplify the core problems (Head, 2008). The abuse of PCL exacerbates the difficulty of improving the lives of disadvantaged groups (Nolan and Mikami, 2013) and dehumanizes minorities in news media (Mendelsohn et al., 2020). Due to its hazard, PCL is classified as a milder form of toxic speech (Dale et al., 2021).

The increasingly social issue caused by PCL has attracted considerable attention of researchers in the natural language processing (NLP) field. Wang and Potts (2019) introduced the task of condescension detection in direct communication and built a dataset with annotated social media messages. Pérez-Almendros et al. (2020) proposed Don't Patronize Me!, an annotated dataset with PCL, and demonstrated the effectiveness of the model for PCL detection (Kenton and Toutanova, 2019).

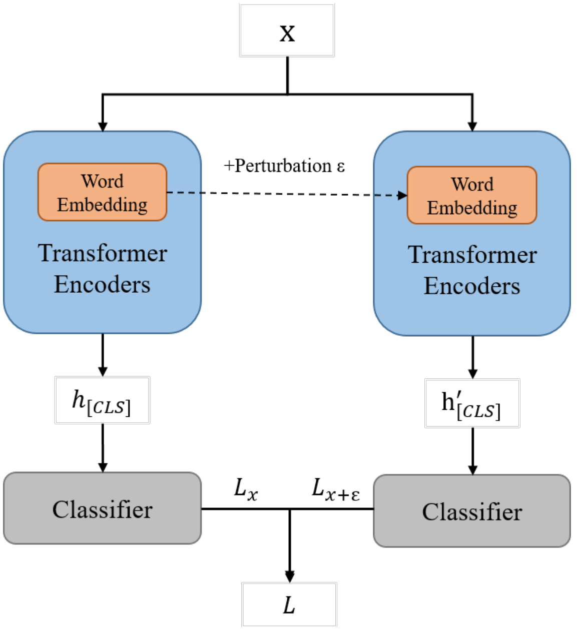

Figure 1: Model architecture for our proposed method in Subtask A.

## 3 Methodology

### 3.1 Preliminaries

We utilize a transformer-based pre-trained language model (PLM), such as BERT and RoBERTa, to represent the input sentences. Each sentence $x = [CLS, t_1, t_2, ..., t_T, SEP]$ is embedded as $s \in R^{n \times d_{emb}}$, where $n$ is the sequence length and $d_{emb}$ is the dimension of the embedding. We add a softmax classifier on the sentence-level embedding, such as the final hidden state $h_{CLS}$ of the $[CLS]$ in BERT:

$$p_i = softmax(Wh_{[CLS]}) \tag{1}$$

where $W \in \mathbb{R}^{C \times d_{emb}}$, and C denotes the number of classes.

### 3.2 Adversarial Training

Adversarial training (Goodfellow et al., 2015) is a effective regularization method for classifiers to improve robustness to small, approximately worst case perturbations. In SubtaskA, we introduce Fast Gradient Method (FGM) (Miyato et al., 2016), a novel approach in adversarial training, to improve the generalization ability of the model in PCL detection. Figure 1 shows the overall framework of our model.

According to FGM, we apply tiny perturbations to sentence embeddings rather than original input itself. The adversarial perturbation $r_{adv}$ on $s$ is defined as:

$$r_{adv} = \epsilon \cdot g/\|g\|_2 \; where \; g = \nabla_s L(s, y) \tag{2}$$

where $\epsilon$ is a hyperparameter limiting the size of the adversarial perturbations.

To integrate the information trained from original and adversarial samples, we use an overall loss function as follows:

$$L = L(s, y) + L_{adv}(s + r_{adv}, y) \tag{3}$$

### 3.3 Balancing Methods

Subtask B becomes a challenging multi-label text classification task because of its long-tailed distribution of labels, each training sample $\{(x^1, y^1), \ldots, (x^N, y^N)\}$ has a multi-label group $y^k = [y_1^k, \ldots, y_C^k] \in \{0, 1\}^C$, and a classification result $z^k = [z_1^k, \ldots, y_C^k]$. In this work, we use different balancing methods (Huang et al., 2021) re-weighting the binary cross entropy to address the class imbalance problem. And the sigmoid function is used for computing $p_i^k = \sigma(z_i^k)$. The codes of these several balanced loss functions are open source[2].

**Focal Loss (FL)** proposed by Lin et al. (2017) places a higher weight of loss on "hard-to-classify" instances, which are predicted with low probability. The FL can be formulated as follows:

$$L_{FL} = \begin{cases} -\alpha(1 - p_i^k)^\gamma \log(p_i^k) & if \; y_i^k = 1 \\ -\alpha(p_i^k)^\gamma \log(1 - p_i^k) & \text{otherwise} \end{cases} \tag{4}$$

[2] https://github.com/Roche/BalancedLossNLP

where $\gamma \geq 0$ is a non-negative tunable focusing parameter to differentiate between easy and difficult samples and $\alpha \in [0, 1]$ is a weighting factor to balance the training weights of positive and negative samples, $p_i^k$ is the $k$th choice of $p_i$.

**Class-balanced Focal Loss (CB)** (Cui et al., 2019) re-balances the loss according to the effective number of samples for each class. Data sampling can be viewed as a random coverage problem, therefore we assign weights to the different classes based on the number of effective samples. The class-balanced term is defined as:

$$r_{CB} = \frac{1-\beta}{1-\beta^{n_i}} \tag{5}$$

where $\beta \in (0, 1)$ controls the effect of effective number of samples on marginal benefit. And we can use this term to re-weight focal loss:

$$L_{CB} = \begin{cases} -r_{CB}(1-p_i^k)^\gamma \log(p_i^k) & if\ y_i^k = 1 \\ -r_{CB}(p_i^k)^\gamma \log(1-p_i^k) & \text{otherwise} \end{cases} \tag{6}$$

**Distribution-balanced loss (DB)** Wu et al. (2020) present DBloss to overcome the additional imbalance caused by label co-occurrence upon resampling. In the case of single label, the resampling probability of each instance can be defined as: $P_i^C = \frac{1}{C}\frac{1}{n_i}$; while under multi-label conditions, the instance is repeatedly sampled by each positive class it contains, thus the resampled probability can be defined as $P^I = \frac{1}{C}\sum_{y_i^k=1}\frac{1}{n_i}$. And we can obtain a balancing term: $r_{DB} = P_i^C/P^I$. With a smooth function $\hat{r}_{DB} = \alpha + \sigma(\beta \times (r_{DB} - \mu))$, mapping the weight $r_{DB}$ to a reasonable range, the re-balanced loss function is defined as:

$$L_{R-FL} = \begin{cases} -\hat{r}_{DB}(1-p_i^k)^\gamma \log(p_i^k) & if\ y_i^k = 1 \\ -\hat{r}_{DB}(p_i^k)^\gamma \log(1-p_i^k) & \text{otherwise} \end{cases} \tag{7}$$

To mitigate the over-suppression of negative labels, Wu et al. (2020) introduce a Negative Tolerant Regularization (**NTR**) in the loss function. NTR initializes a non-zero bias $v_i$ as a threshold, and linearly scales the negative logits before the original loss is computed negative, together with a regularization parameter $\lambda$ to constrain the gradient between 0 and 1. The distribution-balanced loss with NTR can be defined as:

$$L_{DB} = \begin{cases} -\hat{r}_{DB}(1-q_i^k)^\gamma \log(q_i^k) & if\ y_i^k = 1 \\ -\hat{r}_{DB}\frac{1}{\lambda}(q_i^k)^\gamma \log(1-q_i^k) & \text{otherwise} \end{cases} \tag{8}$$

where $q_i^k = \sigma(z_i^k - v_i)$ for positive instances and $q_i^k = \sigma(\lambda(z_i^k - v_i))$ for negative ones. Due to its strong applicability, NTR can also be utilized in Focal loss and DBloss to avoid over-suppression (Huang et al., 2021).

## 4 Experiments

### 4.1 Dataset and Evaluation

The dataset from the Task4 of SemEval2022 contains paragraphs about potentially vulnerable social groups[3]. The paragraphs have been extracted from the News on Web (NoW)[4] corpus (Davies, 2013). The total number of training set is 10469 and the final test set contains 2971 samples. The statistics of datasets are shown in Table 1 and the distribution of PCL categories is reported in Table 2.

| Label | Samples | Proportion |
|---|---|---|
| PCL | 993 | 9.49% |
| no PCL | 9476 | 90.51% |

Table 1: The distribution of labels in SubTaskA.

| PCL Categories | Samples | Proportion |
|---|---|---|
| Unb. power rel. | 716 | 6.84% |
| Shallow solu. | 196 | 1.87% |
| Presupposition | 224 | 2.14% |
| Authority voice. | 230 | 2.20% |
| Metaphor | 197 | 1.88% |
| Compassion | 469 | 4.48% |
| The p., the mer. | 40 | 0.04% |

Table 2: The distribution of labels in SubTaskB.

To estimate the performance of the system, the organizers used different metrics for subtask A and B. In Subtask A, a binary classification task, F1 over the positive class is applied as evaluation measure, while for Subtask B, framed as a multilabel classification problem, results are evaluated based on the macro-average F1 of seven PCL categories.

### 4.2 Experimental Settings

We utilize Roberta-base (Liu et al., 2019) as the pretrained language model for representing the input paragraphs. The AdamW optimizer is used for model training. In evaluation period, we perform five-folds cross-validation on training set and evaluate the performance of our model using average metrics over five-folds. We keep the model parameters for optimal performance. In test phase, we

[3] https://github.com/Perez-AlmendrosC/dontpatronizeme

[4] https://www.english-corpora.org/now/

utilize each fold of the optimal model to predict on the offical test set and vote on the results to obtain the final predictions.

Specially, we implement our model with `transformers`[5] package. During the training phase, we evaluate the performance of the model every 200 steps and retain the parameters of the model that performed best on the validation set. The hyperparameters settings adopted are shown in Table 3. All models are trained on NVIDIA Geforce GTX 3090 GPU.

| Hyperparameters | SubtaskA | SubtaskB |
|---|---|---|
| seed | 1234 | 1234 |
| epochs | 5 | 15 |
| batch size | 32 | 8 |
| learning rate | 2e-4 | 2e-4 |
| alpha | 0.6 | 0.95 |
| gamma | 2 | 4 |
| dropout | 0.25 | - |

Table 3: The hyperparameters of the experiment.

### 4.3 Results and Discussions

**The influence of adversarial training.** Table 4 shows the influence of adversarial training in Subtask A. Based on the experimental results, we observe that the introduction of FGM can improve the detection capability of the model in both evaluation phase and test phase. It shows that adversarial training can improve the robustness of the model.

| **Evaluation phase** | |
|---|---|
| **Model** | **F1(postive)** |
| RoBERTa | 0.5699 |
| RoBERTa+FGM | 0.5785 |
| **Test phase** | |
| **Model** | **F1(postive)** |
| RoBERTa | 0.5545 |
| RoBERTa+FGM | 0.5790 |

Table 4: The performance of our model in Subtask A.

**The influence of balancing methods.** Table 5 shows the results of our framework trained with various loss functions in Subtask B. It is observed that the performance after introducing the balancing methods is significantly more superior than BCE, while the effect is further improved after employing NTR.

In the period of test, we choose two models with the best performance during the evaluation phase to predict the samples, which are trained with FL+NTR and CB+NTR, respectively. Moreover, the model trained with FL+NTR performs better in the final test set. It is because CB is more sensitive to the assumed sample space size $\beta$. If there is a significant difference between the training set and the label distribution of the test set, the ability of the model to address label imbalance will be reduced. In the follow-up work, we will conduct more experiments to observe the impact of parameters on hyperparameters.

| **Evaluation Phase** | |
|---|---|
| **Loss Function** | **F1(macro)** |
| BCE | 0.2923 |
| FL | 0.3662 |
| DB | 0.3767 |
| CB | 0.3776 |
| FL+NTR | 0.3917 |
| CB+NTR | 0.3922 |
| **Test Phase** | |
| **Loss Function** | **F1(macro)** |
| FL+NTR | 0.3700 |
| CB+NTR | 0.3537 |

Table 5: The performance in Subtask B.

## 5 Emotional Polarity and Intensity of PCL

In this section, we conduct a further analysis to explore the relevance between PCL and emotional polarity and intensity it contains.

We employ NLTK[6], a natural language processing toolkit, to determine the emotional features of a paragraph. For a given text, parser of NLTK returns a sentiment score in a interval of [-1,1], which determines if sample is positive or negative and shows emotional intensity. We divide the sentiment score into 5 levels, and the mapping relationships reflecting motional polarity and intensity are shown in Table 6 and Table 7.

| Sentiment Score | Emotional Level |
|---|---|
| $[-1, -0.6]$ | -2 |
| $[-0.6, -0.2]$ | -1 |
| $[-0.2, 0.2]$ | 0 |
| $[0.2, 0.6]$ | 1 |
| $[0.6, 1]$ | 2 |

Table 6: Mapping between sentiment scores and emotional level of the polarity.

[5] https://huggingface.co/

[6] https://github.com/nltk/nltk

| Sentiment Score | Emotional Level |
|---|---|
| $[-0.2, 0.2]$ | 0 |
| $[0.2, 0.4] \cup [-0.4, -0.2]$ | 1 |
| $[0.4, 0.6] \cup [-0.6, -0.4]$ | 2 |
| $[0.6, 0.8] \cup [-0.8, -0.6]$ | 3 |
| $[0.8, 1] \cup [-1, -0.8]$ | 4 |

Table 7: Mapping between sentiment scores and emotional level of the intensity.

We divide the training set into 5 subsets based on the sentiment level and calculate the number of samples. Then we count the proportion of paragraph containing PCL in each subset. The experimental result is reported in Figure 2 and 3.

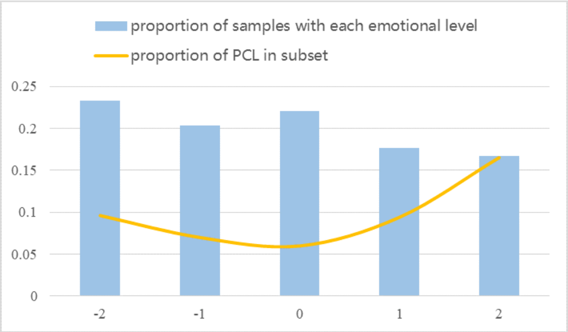

Figure 2: The emotional polarity level of PCL. Blue strip: proportion of samples with each level reflecting emotional polarity in the entire dataset, yellow line: proportion of PCL in subset with each emotional level.

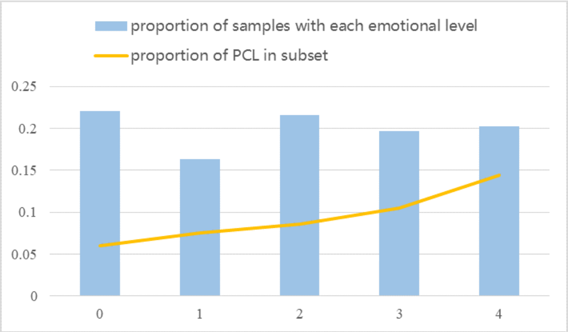

Figure 3: The emotional intensity level of PCL. Blue strip: proportion of samples with each level reflecting emotional intensity in the entire dataset, yellow line: proportion of PCL in subset with each emotional level.

From the results, we can observe that: a) The paragraph containing PCL is more likely to express positive emotions since the use of PCL is often with good intentions. b) Paragraphs with higher emotional intensity are more likely to contain PCL. This is because there are numerous excerpts of live speeches, speakers tend to express their opinions in a stronger tone, which is often condescending.

## 6 Conclusion and Future Work

In this work, we present our approach to the SemEval-2022 Task 4 to tackle the problem of patronizing and condescending language detection. We employ adversarial training and balancing methods for PCL classification with long-tailed class distribution and demonstrate the effectiveness of our methods.

Besides basic deep learning techniques, introducing multi-task learning in PCL detection, such as predicting the sentiment polarity of a paragraph, is also a problem worth discussing. We have found that PCL is associated with the emotional polarity and intensity of paragraphs. In the future, we will further explore the relationship between sentiment analysis and PCL detection and propose corresponding multitasking frameworks.

## Acknowledgement

We thank our anonymous reviewers for their helpful comments. This work was supported by grant from the Natural Science Foundation of China (No.62076046, 62006034), Natural Science Foundation of Liaoning Province (No.2021-BS-067) and the Fundamental Research Funds for the Central Universities (No.DUT21RC(3)015).

## References

Lilie Chouliaraki. 2010. Post-humanitarianism: Humanitarian communication beyond a politics of pity. *International journal of cultural studies*, 13(2):107–126.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277.

David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Text detoxification using large pre-trained neural models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7979–7996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mark Davies. 2013. Corpus of news on the web (now): 3+ billion words from 20 countries, updated every day. *Retrieved January*, 25:2019.

Susan T Fiske. 1993. Controlling other people: The impact of power on stereotyping. *American psychologist*, 48(6):621.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples.

Brian W Head. 2008. Wicked problems in public policy. *Public policy*, 3(2):101–118.

Yi Huang, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. 2021. Balancing methods for multi-label text classification with long-tailed class distribution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8153–8161.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

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.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. *Frontiers in artificial intelligence*, 3:55.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

David Nolan and Akina Mikami. 2013. 'the things that we have to do': Ethics and instrumentality in humanitarian communication. *Global Media and Communication*, 9(1):53–70.

Carla Pérez-Almendros, Luis Espinosa-Anke, and Steven Schockaert. 2020. Don't Patronize Me! An Annotated Dataset with Patronizing and Condescending Language towards Vulnerable Communities. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5891–5902.

Carla Pérez-Almendros, Luis Espinosa-Anke, and Steven Schockaert. 2022. SemEval-2022 Task 4: Patronizing and Condescending Language Detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2019. Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891*.

Zijian Wang and Christopher Potts. 2019. Talkdown: A corpus for condescension detection in context. *arXiv preprint arXiv:1909.11272*.

Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. 2020. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *European Conference on Computer Vision*, pages 162–178. Springer.