# PALI-NLP at SemEval-2022 Task 4: Discriminative Fine-tuning of Transformers for Patronizing and Condescending Language Detection

**Dou Hu** and **Mengyuan Zhou** and **Xiyang Du** and **Mengfei Yuan**
and **Meizhi Jin** and **Lianxin Jiang** and **Yang Mo** and **Xiaofeng Shi**
Ping An Life Insurance Company of China, Ltd.
{HUDOU470, ZHOUMENGYUAN425, DUXIYANG037, YUANMENGFEI854,
JINMEIZHI005, JIANGLIANXIN769, MOYANG853, SHIXIAOFENG309}
@pingan.com.cn

## Abstract

Patronizing and condescending language (PCL) has a large harmful impact and is difficult to detect, both for human judges and existing NLP systems. At SemEval-2022 Task 4, we propose a novel Transformer-based model and its ensembles to accurately understand such language context for PCL detection. To facilitate comprehension of the subtle and subjective nature of PCL, two fine-tuning strategies are applied to capture discriminative features from diverse linguistic behaviour and categorical distribution. The system achieves remarkable results on the official ranking, including 1st in Subtask 1 and 5th in Subtask 2. Extensive experiments on the task demonstrate the effectiveness of our system and its strategies.

## 1 Introduction

"*Don't worry, I know this is a mistake you usually make, we all make it sometimes, but I am bringing you a solution.*", which is a typical example of Patronizing and Condescending Language (PCL) (Giles et al., 1993; Huckin, 2002), shows a superior attitude and apparent kindness towards others, while is generally expressed unconsciously. The impact of PCL can potentially be very harmful, as it feeds inequalities and routinizes discrimination (Ng, 2007), especially if it is geared towards vulnerable communities in the media. If we are able to detect and identify when we are condescending or patronizing towards others, a corrective action (e.g., a more inclusive message) could be taken for a more responsible communication.

Recently, some works (Wang and Potts, 2019; Sap et al., 2020) on PCL are gradually emerging in NLP community. Remarkably, Pérez-Almendros et al. (2020) have shown that general pre-trained language models (Devlin et al., 2019; Liu et al., 2019) can achieve nontrivial performance. However, the behaviour of PCL is usually more unconscious, subtler, and subjective than other harmful types of discourse that are widely studied, i.e., hate speech (Basile et al., 2019), offensive language (Zampieri et al., 2019), intended sarcasm (Du et al., 2022), fake news (Zhang et al., 2021b) and rumor (Wei et al., 2021). These characteristics make PCL detection a difficult challenge, both for human judges and existing NLP systems.

To address this, we propose a novel Transformer-based model BERT-PCL (and its ensembles) with two **discriminative fine-tuning** strategies, to accurately understand such language context for PCL detection. The two strategies are grouped layer-wise learning rate decay (Grouped LLRD) and weighted random sampler (WRS), and both are beneficial for task-adaptive fine-tuning based on language models.

A brief description of these two strategies is as follows: **1)** As different layers capture different types of information (Yosinski et al., 2014), Grouped LLRD, a variant of LLRD (Howard and Ruder, 2018; Zhang et al., 2021a), is applied to group hidden layers of the pre-trained Transformer (Vaswani et al., 2017) into different sets and apply different learning rates to each in a certain extent. And then, we can make full use of different layers to capture more diverse and fine-grained linguistic features, which can boost understanding of the subtle and subjective nature of PCL. **2)** There is a quite common phenomenon that positive samples (patronizing or condescending) have a smaller number than the negative, which reflects usage rates of PCL in public forums (Wang and Potts, 2019; Pérez-Almendros et al., 2020). But the positive samples are more important when detecting PCL, due to the harmful impacts. To deal with imbalanced classes scenarios, we introduce WRS to place more emphasis on the minority classes. Under this strategy, our classifier can capture discriminative features from the categorical distribution and detect whether it contains PCL in an unbiased manner.

At SemEval-2022 Task 4 (Pérez-Almendros

*Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 335 - 343
July 14-15, 2022 ©2022 Association for Computational Linguistics

et al., 2022), our proposed system achieves **1st in Subtask 1** and **5th in Subtask 2** on the evaluation leaderboard[1]. Meanwhile, in the post-evaluation phase, we further verified the results of the system on the test set of both subtasks. For Subtask 1, the single model BERT-PCL and its ensembles obtained **63.69%** and **65.41%** performance in terms of F1 of positive class, respectively. For Subtask 2, the single model BERT-PCL and its ensembles obtained **43.28%** and **45.66%** performance in terms of macro-average F1, respectively. Moreover, a series of experiments are conducted on the two subtasks of PCL detection. Results consistently demonstrate that our model and its ensembles significantly outperform comparison methods and the effectiveness of two strategies used in our system.

## 2 Background

### 2.1 Task and Data Description

The aim of SemEval-2022 Task 4 (Pérez-Almendros et al., 2022) is to identify PCL, and to categorize the linguistic techniques (categories) used to express it, specifically when referring to vulnerable communities in the media.

This challenge is divided into two subtasks, each corresponding to a subset of the *Don't Patronize Me!* (DPM) dataset (Pérez-Almendros et al., 2020). The 10,469 annotated paragraphs (i.e., sentences in context) from the DPM corpus are used as training data, where each paragraph mentions one or several predefined vulnerable communities. These paragraphs are collected using a keyword-based strategy and cover English language news sources from 20 different countries. A short description of the two subtasks and training data is as follows:

- **Subtask 1: Binary classification**. Given a paragraph, a system must predict whether or not it contains any form of PCL. The training set consists of 10,469 paragraphs annotated with a label ranging from 0 to 4. Label 2, 3, and 4 means positive examples (condescending or patronizing) of PCL and the remaining labels means negatives.

- **Subtask 2: Multi-label classification**. Given a paragraph, a system must identify the categories of PCL that are present. The 993 unique paragraphs (positive examples) in the training set, totaling 2,760 instances of PCL, are labeled with one or more PCL categories: *Unbalanced power relations*, *Shallow solution*, *Presupposition*, *Authority voice*, *Metaphor*, *Compassion*, *The poorer, the merrier*.

In addition, the test set for the evaluation phase contains around 4,000 manually annotated paragraphs with the PCL annotation scheme. More details about the task can be found on the competition page[2].

### 2.2 Related Work

Harmful language detection/recognition has been widely studied in various forms of discourse, such as hate speech (Basile et al., 2019), offensive language (Zampieri et al., 2019), intended sarcasm (Du et al., 2022), fake news (Zhang et al., 2021b) and rumors (Wei et al., 2021; Hu et al., 2021). Unlike these works generally focused on explicit, aggressive and flagrant phenomena, the study of patronizing and condescending Language (PCL) (Giles et al., 1993; Huckin, 2002; Chouliaraki, 2006; Margić, 2017) has been almost ignored in NLP community until recently.

To encourage more research on PCL language, Wang and Potts (2019) present a condescension detection task and provides a *TALKDOWN* dataset in comment-reply pairs from Reddit. Besides, Pérez-Almendros et al. (2020) introduce a *Don't Patronize Me!* dataset and the challenge of PCL detection towards vulnerable communities (e.g. refugees, homeless people, poor families). These works establish several advanced baselines using pre-trained language models (Devlin et al., 2019; Liu et al., 2019), and suggest that detecting such language is a challenging task both for humans and NLP systems due to its subtle and subjective nature.

## 3 System Overview

In this section, we review our system adopted in SemEval-2022 Task 4, where we design a novel Transformer-based model BERT-PCL (and its ensembles) with two discriminative fine-tuning strategies for both subtasks of PCL detection.

### 3.1 Model Architecture

BERT (Devlin et al., 2019) uses masked language models to enable pretrained deep bidirectional representations, and can be fine-tuned to create task-

[1] https://sites.google.com/view/pcl-detection-semeval2022/ranking

[2] https://competitions.codalab.org/competitions/34344

specific models with powerful performance (Wei et al., 2020; Hu and Wei, 2020). Inspired by this, our system utilizes Transformers (Vaswani et al., 2017) to learn contextual representations of the input sentence under the BERT-like architecture.

Formally, given an input token sequence $x_{i1}, ..., x_{iN}$ where $x_{ij}$ refers to $j$-th token in the $i$-th input sample, and $N$ is the maximum sequence length, the model learns to generate the context representation of the input token sequences:

$$\mathbf{h}_i = \text{BERT}([\text{CLS}], x_{i1}, ..., x_{iN}, [\text{SEP}]), \quad (1)$$

where [CLS] and [SEP] are special tokens, usually at the beginning and end of each sequence, respectively. $\mathbf{h}_i$ indicates the hidden representation of the $i$-th input sample, computed by the representation of [CLS] token in the last layer of the encoder.

## 3.2 PCL Detection

### 3.2.1 Subtask 1: Binary Classification

Subtask 1, a binary classification task, aims to predict whether or not a paragraph contains any form of PCL. After encoding, we apply a fully connected layer with the Softmax function to predict whether or not the input contains any form of PCL:

$$\hat{\mathbf{y}}_i = Softmax(\mathbf{W}\mathbf{h}_i + \mathbf{b}), \quad (2)$$

where $\mathbf{W}$ and $\mathbf{b}$ are trainable parameters. We leverage Cross-entropy loss to optimize the system. The objective function of Subtask 1 is defined as:

$$L = -\frac{1}{N}\sum_i (\mathbf{y}_i \log(\hat{\mathbf{y}}_i) + (1 - \mathbf{y}_i)\log(1 - \hat{\mathbf{y}}_i)) \quad (3)$$

where $\mathbf{y}_i$ is the ground-truth label of PCL.

### 3.2.2 Subtask 2: Multi-Label Classification

Subtask 2 is a multi-label classification task. Its goal is to determine which PCL categories a paragraph expresses. After encoding, we also apply a fully connected layer with the sigmoid function to predict the probability of each PCL class:

$$\hat{\mathbf{y}}_i^c = \sigma(\mathbf{W}^c \mathbf{h}_i + \mathbf{b}^c), \quad (4)$$

where $\sigma$ is the sigmoid function. $\mathbf{W}^c$ and $\mathbf{b}^c$ are trainable parameters. We use Binary Cross Entropy (BCE) loss (Bengio et al., 2013) for the multi-label classification task, denoted as:

$$L = -\frac{1}{N}\sum_i \sum_{c=1}^{M} [\mathbf{y}_i^c \log(\hat{\mathbf{y}}_i^c) + (1 - \mathbf{y}_i^c)\log(1 - \hat{\mathbf{y}}_i^c)], \quad (5)$$

where $M$ is the number of classes, $\hat{\mathbf{y}}_i^c$ indicates the predicted probability that the $i$-th sample belongs to the $c$-th class.

## 3.3 Fine-tuning Strategies

For discriminative fine-tuning of the model, we introduce two strategies to boost the accurate understanding of PCL context, namely grouped layer-wise learning rate decay (Grouped LLRD) and weighted random sampler (WRS).

### 3.3.1 Grouped LLRD

As different layers capture different types of information (Yosinski et al., 2014), they should be fine-tuned to different extents. Therefore, instead of using the same learning rate for all hidden layers of the Transformer, we tune each layer with different learning rates. Layer-wise learning rate decay (LLRD) (Howard and Ruder, 2018; Zhang et al., 2021a) is a popular fine-tuning strategy that applies higher learning rates for top layers and lower learning rates for bottom layers. Inspired by this, we group layers into different sets and apply different learning rates to each, denoted as Grouped LLRD.

Formally, we split all hidden layers of the Transformer into $G$ sets with embeddings attached to the first set. The parameters of layers are denoted as $\{\theta^1, ..., \theta^G\}$, where $\theta^g$ refers to the $g$-th group. The corresponding learning rate values are denoted as $\{\eta^1, ..., \eta^G\}$, where $\eta^g$ indicates the learning rate of the $g$-th group. To capture discriminative features, a multiplicative decay rate $\lambda$ is used to change relative value of initial learning rates from adjacent groups in a controlled fashion. At time step $t$, the update of parameters $\theta$ is computed by:

$$\theta_t^g = \theta_{t-1}^g - \eta^g \cdot \nabla_{\theta^g} J(\theta), \quad (6)$$

where $\nabla_{\theta^g} J(\theta)$ is the gradient with regard to the model's objective function. The learning rate of the lower layer is applied as $\eta^{g-1} = \eta^g/\lambda$ during fine-tuning to decrease the learning rate group-by-group. In addition, same as LLRD, we use a learning rate that is slightly higher than the top hidden layer for the pooler head and classifier.

Under the above setting, we can capture more diverse and fine-grained linguistic features by flexibly optimizing different hidden layers of the Transformer. It can boost understanding of PCL's subtle and subjective nature.

#### 3.3.2 Weighted Random Sampler

The PCL dataset is highly imbalanced, which causes problems for training the above models. To alleviate this imbalanced classes problem, we use a Weighted Random Sampler (WRS) to place more emphasis on the minority classes. The samples are weighted and the probability of each sample being selected is determined by its relative weight.

For both subtasks, the sampling weight of the $i$-th sample is computed by:

$$s_i = \begin{cases} 1/\sqrt{\kappa_p}, & \text{if it contains PCL,} \\ 1/\sqrt{\kappa_n}, & \text{otherwise,} \end{cases} \tag{7}$$

where $\kappa_p$ and $\kappa_n$ refer to the ratio of positive and negative examples of PCL in training data, respectively. Then, the elements are sampled based on the passed weights. It is worth noting that the number of samples is equal to the length of the training set. During training, the sampler tends to select samples from positive examples with small data volume. In this way, we can have positive and negative classes with equal probability. And the classifier can capture discriminative features from categorical distribution in an unbiased manner.

### 3.4 Ensemble

For the final submissions, we apply a voter-based fusion technique (Morvant et al., 2014) to ensemble several BERT-PCL models. Concretely, we train the proposed BERT-PCL with five different random seeds. Then, we select Top-3 models according to average result of k-fold cross-validation on the training data. Finally, results of the test set predicted by the three optimal models are voted to get the final submission.

## 4 Experimental Setup

### 4.1 Comparison Methods

We compare BERT-PCL and its ensembles with the following several methods:

- **Random** is based on random guessing, choosing each class/label with an equal probability.

- **BERT** (Devlin et al., 2019) is a language model pre-trained in a self-supervised fashion based on deep bidirectional transformers. We use *bert-base-uncased*[3] to initialize BERT.

- **ALBERT** (Lan et al., 2020) presents two parameter-reduction techniques to lower memory consumption and increase the training speed of BERT. We use *albert-large-v2*[3] to initialize ALBERT.

- **ERNIE 2.0** (Sun et al., 2020) is a continual pre-training framework, which builds and learns incrementally pre-training tasks through constant multi-task learning. We use *nghuyong/ernie-2.0-large-en*[3] to initialize ERNIE 2.0.

- **RoBERTa** (Liu et al., 2019) optimizes the training procedure of BERT and removes the next sentence predict objective when pre-training. We use *roberta-large*[3] to initialize RoBERTa.

### 4.2 Implementation Details

For both subtasks, stratified k-fold cross validation (Kohavi, 1995; Sechidis et al., 2011) is performed to split limited training data into 5 folds. We choose the optimal hyperparameter values based on the the average result of validation sets for all folds, and evaluate the performance of systems on the test data. BERT-PCL is initialized with the *roberta-large*[3] parameters, due to the nontrivial and consistent performance in both subtasks. Following Pérez-Almendros et al. (2020, 2022), the evaluation metrics are F1 over the positive class for Subtask 1 and macro-average F1 for Subtask 2.

We group layers into 3 groups, i.e., $G = 3$. The learning rate for layers in the lower, median, and higher groups as $\eta/\lambda$, $\eta$, and $\eta * \lambda$, respectively, where $\eta$ is set to 1e-5. $\lambda$ is a hyperparameter searched from $\{0.6, 1.6, 2.6, 3.6, 4.6, 5.6, 6.6\}$. For the training of BERT-PCL, the optimal value of $lambda$ is 1.6 for Subtask 1 and 3.6 for Subtask 2. The experiments are conducted with batch size of 4, maximum length of 250, and dropout rate of 0.4. The number of epochs is set to 10 and the maximum patience number of early stopping is set to 50 batches. AdamW optimizer (Kingma and Ba, 2015) is used with a weight decay of 0.01. A cosine annealing schedule (Loshchilov and Hutter, 2017) is applied to decay the learning rate, with a linear warmup for the first 10% steps.

To effectively utilize the country term and search keyword term corresponding to each paragraph in the corpus, we concatenate these terms with the original paragraph as the input sequence. In the

[3] https://huggingface.co/

|  | P | R | F1 |
|---|---|---|---|
| Random | 8.98 | 55.21 | 15.45 |
| BERT | 56.20 | 48.58 | 52.12 |
| ALBERT | 59.43 | 32.81 | 42.28 |
| ERNIE 2.0 | 59.24 | 58.68 | 58.95 |
| RoBERTa | 60.65 | 64.67 | 62.60 |
| **BERT-PCL** | 64.31 | 63.09 | **63.69** |
| Ensemble 1.0 † | 64.60 | 65.62 | 65.10 |
| **Ensemble 2.0** | 65.20 | 65.62 | **65.41** |

Table 1: Results for the problem of detecting PCL, viewed as a binary classification problem (Subtask 1). The results are reported in terms of the precision (P), recall (R) and F1 score of the positive class. All compared pre-trained models are fine-tuned on the task dataset. † indicates the results on the official ranking.

implementation, two special token pairs (i.e., <e> and </e>) are introduced as the term boundary.

## 5 Result and Discussion

### 5.1 Overall Result

The overall results in both subtasks are summarized in Table 1 and 2. Unsurprisingly, all pre-trained models clearly outperform the random baseline. The proposed BERT-PCL and its ensembles (i.e., Ensemble 1.0 and Ensemble 2.0) consistently obtain the best performance than the comparison methods on both subtasks. Specifically, BERT-PCL gains 1.09% and 5.39% absolute improvements for Subtask 1 and 2, respectively. These results show the superiority of our models.

In Table 1, both Ensemble 1.0 and Ensemble 2.0 are fused by three optimal full BERT-PCL models with different seeds and obtain a better performance than BERT-PCL in Subtask 1. In Table 2, Ensemble 1.0 is fused by the three optimal BERT-PCL that removes WRS, since we found that it is worse when performing WRS according to weights of category label in Subtask 2. Different from it, in the post-evaluation phase, we perform WRS according to weights of positive samples (patronizing or condescending) and fuse three optimal full BERT-PCL as Ensemble 2.0. As shown in Table 2, Ensemble 2.0 obtains a better performance than BERT-PCL and Ensemble 1.0 in Subtask 2.

Then, we qualitatively analyze the performance of BERT-PCL and typical baselines on the validation set for both subtasks. The results are illustrated in Figure 1. From the figure, BERT-PCL consistently obtains the best performance on the validation set for both subtasks, which confirms again the superiority of the proposed method.

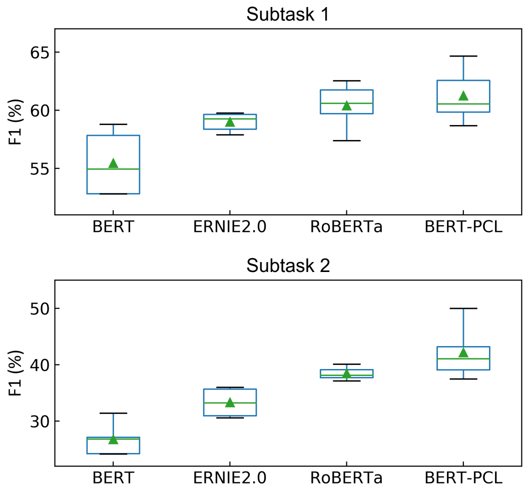

Figure 1: Results on the validation set for both subtasks. The box displays the distribution of results where the green triangle indicates the mean of results, the green line and two blue lines represent the 25%, 50%, and 75% quartiles, respectively, and black lines are the maximum and minimum values. For Subtask 1, we report F1 score of the positive class; and for Subtask 2, we list macro-average F1 score.

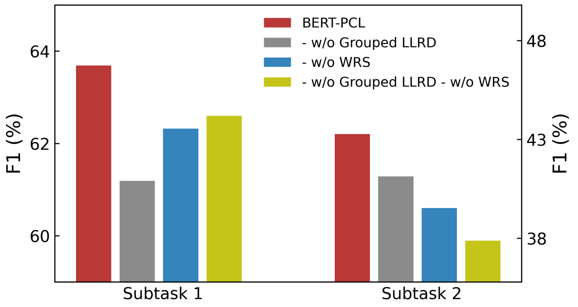

Figure 2: Results of ablation study for PCL detection. For Subtask 1, we report F1 score of the positive class; and for Subtask 2, we list macro-average F1 score.

### 5.2 Ablation Study

We conduct ablation studies by removing key components of BERT-PCL: 1) **- w/o Grouped LLRD** refers to removing the Grouped LLRD. 2) **- w/o WRS** refers to removing the WRS. 3) **- w/o Grouped LLRD - w/o WRS** refers to removing both Grouped LLRD and WRS, degenerated to RoBERTa.

Figure 2 shows results of ablation studies on two subtasks of PCL detection. The full model yields the best performance on both subtasks. When removing either Grouped LLRD or WRS, the results

|  | Random | BERT | ALBERT | ERNIE 2.0 | RoBERTa | **BERT-PCL** | Ensemble 1.0 † | **Ensemble 2.0** |
|---|---|---|---|---|---|---|---|---|
| *unb.* | 10.82 | 51.52 | 51.31 | 56.50 | 57.41 | 58.90 | 62.78 | 61.40 |
| *shal.* | 2.23 | 40.62 | 41.79 | 55.88 | 50.60 | 50.55 | 54.76 | 55.81 |
| *pres.* | 3.03 | 20.29 | 16.39 | 25.97 | 26.83 | 42.86 | 34.15 | 39.13 |
| *auth.* | 4.21 | 6.98 | 9.64 | 16.49 | 26.79 | 28.07 | 34.19 | 34.15 |
| *met.* | 1.91 | 8.70 | 9.52 | 28.07 | 40.51 | 40.00 | 33.33 | 43.84 |
| *comp.* | 5.98 | 42.48 | 40.00 | 45.90 | 48.28 | 49.24 | 50.82 | 49.61 |
| *merr.* | 1.22 | 0.00 | 0.00 | 0.00 | 14.81 | 33.33 | 8.70 | 35.71 |
| Average | 4.20 | 24.37 | 24.09 | 32.69 | 37.89 | **43.28** | 39.82 | **45.66** |

Table 2: Results for the problem of categorizing PCL, viewed as a paragraph-level multi-label classification problem (Subtask 2). We report the per-class F1 and macro-average F1. All compared pre-trained model are fine-tuned on the task dataset. † indicates the results on the official ranking. The considered seven categories are *Unbalanced power relations* (unb.), *Shallow solution* (shal.), *Presupposition* (pres.), *Authority voice* (auth.), *Metaphor* (met.), *Compassion* (comp.) and *The poorer, the merrier* (merr.).

of variants decline significantly on both subtasks. Specifically, when only removing Grouped LLRD, the model achieves 2.50% and 2.14% degradation of performance in Subtask 1 and 2, respectively. When only removing WRS, the results decline by 1.37% and 3.74% in terms of F1 scores in Subtask 1 and 2, respectively. The above results consistently indicate the effectiveness of the two components.

When removing both components, the performance also decreases on both subtasks. Note that **- w/o Grouped LLRD - w/o WRS** achieves a better F1 score of the positive class than **- w/o WRS** or **- w/o Grouped LLRD** in Subtask 1. This can be explained that ignoring Grouped LLRD limits to explore diverse features of positive samples, and further removing WRS may magnify this limitation due to the imbalanced class problem. Therefore, the model with two modules removed yields slightly better results than ablation models with only one module removed. Different from Subtask 1, Subtask 2 is a multi-label classification problem and we report the macro-average F1 score. Using Grouped LLRD can capture diverse features of each category label in positive samples, and WRS according to weights of positive and negative samples further promotes the model's attention to positive samples. Hence, removing both modules obtains the worst performance in Subtask 2.

### 5.3 Parameter Analysis

In this part, we explore the performance of BERT-PCL against different $\lambda$ in Grouped LLRD. Bottom groups often encode more general and broad-based information, while top groups closer to the output encode information more localized and specific to the task on hand. In our model, a suitable value $\lambda$ can control and balance these different layers of the Transformer to capture different kinds of features from diverse linguistic behaviour.

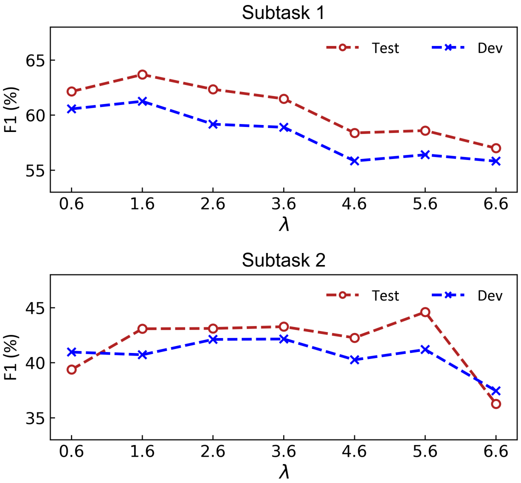

Figure 3: Results against different values of hyperparameter $\lambda$ in Grouped LLRD on both test and validation sets. We report F1 score of the positive class for Subtask 1, and list macro-average F1 score for Subtask 2.

The results are illustrated in Figure 3. Based on k-fold cross validation on training data, we find the local optimum of $\lambda$, 1.6 for Subtask 1 and 3.6 for Subtask 2, and the resulting model consistently performed excellently on the test set. Under the optimal setting, different layers in the model can capture more diverse and fine-grained linguistic features, enhancing the understanding of the subtle and subjective nature of PCL. However, larger $\lambda$ would make the model overfit to small datasets and suffer catastrophic forgetting during fine-tuning. Hence, the performance degrades as $\lambda$ increases.

It is worth noting that in when $\lambda$ is up to 5.6 for Subtask 2, the model achieves suboptimal results on the validation set but performs exceptionally well on the test set. This may be because the model

| No. | Para. | Gold | Pred. (BERT-PCL) | Pred. (RoBERTa) | Pred. (ERNIE 2.0) |
|---|---|---|---|---|---|
| 1 | *“Jesus is the Master Feminist because he championed the cause of women,” she said.* | pos. | pos. | neg. | neg. |
| 2 | *There is a saying that goes "A friend in need is a friend indeed. " This means, a good friend is the one who rescues a friend trapped in unsolved problems.* | neg. | neg. | neg. | pos. |
| 3 | *"The government is implementing several schemes that would change the economic position of poor families," she added.* | pos. | neg. | neg. | neg. |
| 4 | *Alexis and her family decided to donate more than 400 of those presents to children in need.* | neg. | pos. | pos. | pos. |

Table 3: Case studies in Subtask 1: Binary Classification. The table shows four examples of paragraphs, their gold labels and predictions by three methods (BERT-PCL, RoBERTa and ERNIE 2.0). The pos. means the positive class of PCL, i.e. as instances containing PCL. Likewise, the neg. means the negatives.

| No. | Para. | Gold | Pred. (BERT-PCL) | Pred. (RoBERTa) | Pred. (ERNIE 2.0) |
|---|---|---|---|---|---|
| 1 | *Through Gawad Kalinga, Meloto has proven to be a key player in the housing industry, helping provide decent homes and sustainable livelihood to the marginalized and homeless Filipinos.* | unb., comp. | unb., comp. | - | unb., comp. |
| 2 | *Pope Francis will visit a tiny Italian island to greet refugees and immigrants, pray for those who have lost their lives at sea and call for greater solidarity.* | unb., shal. | unb., shal. | - | - |
| 3 | *In South Africa, education is a right and not a privilege, but an unfavourable background can unconsciously infringe on this right.* | unb., pres., met. | unb., met. | unb., auth. | unb. |
| 4 | *Thankfully, while Krishna Tulasi can’t entirely escape from the trope of disabled persons with hearts of gold, it manages to do better than many previous films with disabled protagonists.* | merr. | - | - | - |

Table 4: Case studies in Subtask 2: Multi-Label Classification. The table shows four examples of paragraphs, their gold labels and predictions by three methods (BERT-PCL, RoBERTa and ERNIE 2.0). The categories stand for: *Unbalanced power relations* (unb.), *Shallow solution* (shal.), *Presupposition* (pres.), *Authority voice* (auth.), *Metaphor* (met.), *Compassion* (comp.) and *The poorer, the merrier* (merr.).

overfits some redundant features of the corpus.

### 5.4 Case Study

Table 3 and Table 4 show several typical examples in the training set from Subtask 1 and 2, respectively. Their gold labels and predictions by BERT-PCL, RoBERTa and ERNIE 2.0 are presented in the corresponding columns.

In Table 3, the first case is correctly classified by BERT-PCL, while is misclassified by other methods. We can easily observe that this example has the characteristics of *Unbalanced power relations* and *Authority voice*, and the language expression of the latter is more subtle. Unlike other methods, BERT-PCL can capture the linguistic phenomena of PCL through a discriminative fine-tuning process, and thus detect them correctly. For the second, BERT-PCL and RoBERTa can accurately identify the positive paragraphs, using the sentence representation ability learned by the pre-trained model. The latter two examples are consistently predicted as false negatives and false positives by all methods, respectively. We notice that both paragraphs have been annotated by two human annotators as borderline PCL. Unsurprisingly, these methods also struggle to detect such cases.

As seen in Table 4, only BERT-PCL can correctly determine fine-grained PCL categories of the first two cases, which again illustrates the superiority of our method. It can be noticed that the third example has three PCL sub-categories (i.e., unb., pres., met.) with a certain internal correlation, and the gold label (i.e., merr.) of the fourth example appears too little in the training set. These phenomena increase the difficulty of identifying the two examples, which leads to wrong predicted labels. We believe that identifying multiple related sub-categories simultaneously and controlling the imbalance of positive PCL labels are urgent challenges for Subtask 2.

## 6 Conclusion

In this paper, we propose an advanced BERT-like model and its ensembles to accurately understand and detect patronizing and condescending language. Based on the pre-trained Transformer, we apply two fine-tuning strategies to capture discriminative features from diverse linguistic behaviour and categorical distribution. At SemEval-2022 Task 4, our system achieves 1st in Subtask 1 and 5th in Subtask 2 on the official ranking. Extensive experiments demonstrate the effectiveness and superiority of the proposed system and its strategies.

## Acknowledgements

The research is supported by the Ping An Life Insurance. All the work in this paper are conducted during the SemEval-2022 Competition. Last but not least, we thank the SemEval-2022 organizers for their effort in preparing the challenge, and the reviewers for their insightful and constructive comments.

## References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *SemEval@NAACL-HLT*, pages 54–63. Association for Computational Linguistics.

Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828.

Lilie Chouliaraki. 2006. *The spectatorship of suffering*. Sage.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.

Xiyang Du, Dou Hu, Meizhi Jin, Lianxin Jiang, Yang Mo, and Xiaofeng Shi. 2022. PALI-NLP at SemEval-2022 Task 6: iSarcasmEval- Fine-tuning the Pretrained Model for Detecting Intended Sarcasm. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Howard Giles, Susan Fox, and Elisa Smith. 1993. Patronizing the elderly: Intergenerational evaluations. *Research on Language and Social Interaction*, 26(2):129–149.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *ACL (1)*, pages 328–339. Association for Computational Linguistics.

Dou Hu and Lingwei Wei. 2020. SLK-NER: exploiting second-order lexicon knowledge for chinese NER. In *SEKE*, pages 413–417. KSI Research Inc.

Dou Hu, Lingwei Wei, Wei Zhou, Xiaoyong Huai, Jizhong Han, and Songlin Hu. 2021. A rumor detection approach based on multi-relational propagation tree. *Journal of Computer Research and Development*, 58(7):1395.

Thomas Huckin. 2002. Critical discourse analysis and the discourse of condescension. *Discourse studies in composition*, 155:176.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, pages 1137–1145. Morgan Kaufmann.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *ICLR*. OpenReview.net.

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.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: stochastic gradient descent with warm restarts. In *ICLR (Poster)*. OpenReview.net.

Branka Drljača Margić. 2017. Communication courtesy or condescension? linguistic accommodation of native to non-native speakers of english. *Journal of English as a lingua franca*, 6(1):29–55.

Emilie Morvant, Amaury Habrard, and Stéphane Ayache. 2014. Majority vote of diverse classifiers for late fusion. In *S+SSPR*, volume 8621 of *Lecture Notes in Computer Science*, pages 153–162. Springer.

Sik Hung Ng. 2007. Language-based discrimination: Blatant and subtle forms. *Journal of Language and Social Psychology*, 26(2):106–122.

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. 2020. Social bias frames: Reasoning about social and power implications of language. In *ACL*, pages 5477–5490. Association for Computational Linguistics.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis P. Vlahavas. 2011. On the stratification of multi-label data. In *ECML/PKDD (3)*, volume 6913 of *Lecture Notes in Computer Science*, pages 145–158. Springer.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE 2.0: A continual pre-training framework for language understanding. In *AAAI*, pages 8968–8975. AAAI Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zijian Wang and Christopher Potts. 2019. Talkdown: A corpus for condescension detection in context. In *EMNLP/IJCNLP (1)*, pages 3709–3717. Association for Computational Linguistics.

Lingwei Wei, Dou Hu, Wei Zhou, Xuehai Tang, Xiaodan Zhang, Xin Wang, Jizhong Han, and Songlin Hu. 2020. Hierarchical interaction networks with rethinking mechanism for document-level sentiment analysis. In *ECML/PKDD (3)*, volume 12459 of *Lecture Notes in Computer Science*, pages 633–649. Springer.

Lingwei Wei, Dou Hu, Wei Zhou, Zhaojuan Yue, and Songlin Hu. 2021. Towards propagation uncertainty: Edge-enhanced bayesian graph convolutional networks for rumor detection. In *ACL/IJCNLP (1)*, pages 3845–3854. Association for Computational Linguistics.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *SemEval@NAACL-HLT*, pages 75–86. Association for Computational Linguistics.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. 2021a. Revisiting few-sample BERT fine-tuning. In *ICLR*. OpenReview.net.

Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021b. Mining dual emotion for fake news detection. In *WWW*, pages 3465–3476. ACM / IW3C2.