PclGPT: A Sentiment Large Language Model for Patronizing and Condescending Language Detection

Anonymous ACL submission

Abstract

***Disclaimer**: Samples in this paper may be harmful and cause discomfort!*

Patronizing and Condescending Language (PCL) is a form of harmful communication directed at vulnerable communities. This type of language exacerbates conflicts and confrontations among Internet communities and detrimentally impacts relatively marginalized communities. Traditional pre-trained models exhibit poor detection performance due to the implicit emotional characteristics in the PCL domain, such as hypocrisy and false sympathy. With the rapid development of the Large Language Model (LLM), there is a growing opportunity to utilize extensive emotional semantic features of LLM for sentiment analysis tasks. In this paper, we introduce a comprehensive instruction-tuning framework PclGPT, a new benchmark LLM designed explicitly for Patronizing and Condescending Language. We design the instruction dataset PCL-SFT and build PclGPT-EN/CN by supervised fine-tuning to facilitate cross-language emotion detection. The findings demonstrate that our framework surpass all advanced pre-trained models in classification tasks, including widely employed LLM models like GPT-3.5 and GPT-4. Simultaneously, we confirm PclGPT's substantial capability to detect implicit emotions through fine-grained emotion analysis and fuzzy sample experiments. Our model establishes a crucial basis for further research in PCL and other implicit sentiment analyses.

1 Introduction

Patronizing and Condescending Language (PCL) is a type of toxic speech that specifically targets vulnerable groups. As an important but underexplored branch of toxic speech, timely detection of PCL is an essential means of protecting vulnerable groups from further exclusion and inequality. Compared to traditional tasks such as hate speech, emotional expressions are more subtle and implicit in PCL tasks (e.g., *"Poor kids! Will somebody help them?"*). Although this seemingly sympathetic statement does not contain any offensive words against children, it clearly shows a condescending and discriminatory attitude toward them. It demonstrates an implied disrespect for vulnerable groups and has detrimental implications for the individual being sympathized with. Therefore, further exploration of PCL will have a positive effect on the study of subtle implicit emotions. Conventional supervised models encounter numerous obstacles in this domain, such as the scarcity of annotated data of superior quality, the complexity of resolving contextually PCL emotions, and the unpredictability surrounding the foundation for subjective assessments.

The emergence of the Large Language Model (LLM) has opened up new possibilities for this domain. More world knowledge and a more diverse pre-training corpus make it perform well in various general tasks. Unfortunately, there is insufficient guidance for pre-training knowledge, resulting in its inability to explore its scale effect in specific emotional fields (including PCL tasks). The emergence of prompt engineering has alleviated this problem to a certain extent. However, its lengthy template and redundancy limit the efficiency of LLM. Related LLM research focuses solely on English, and there is a lack of verification work on cross-language detection performance.

To address the above challenges, we concentrate on three main issues: (1) How to effectively design high-quality instruction datasets for subsequent instruction tuning, and compensate for the scarcity of high-quality training data in this domain? (2) How can we build an effective model and migrate the model to perform cross-language tasks, such as Chinese PCL tasks? (3) Patronizing and Condescending Language is primarily filled with implicit emotions, how can the LLM be guided to improve

[1]* Corresponding author.

understanding of those sentiments?

To solve the above problems, we introduce a comprehensive instruction-tuning framework PclGPT for sentiment detection and train our model to explore the LLM's understanding of implicit emotions. In our study, we specifically devise the PCL-SFT instruction dataset by employing the instruction data paradigm to impose additional constraints on both input and output. Subsequently, we develop PclGPT-EN/CN through instruction tuning training, this is the first known type of LLM specifically designed to detect Patronizing and Condescending Language. We further conduct detailed and comprehensive experiments in two key areas: fine-grained emotion classification and the incorporation of implicit fuzzy emotion. Through these experiments, we substantiate the model's proficiency in detecting the implicit emotional nuances within PCL. Our PclGPT framework, coupled with the associated research, represents a substantial contribution, paving the way for advancements in the field of implicit sentiment detection. The main contributions of this paper are summarized as follows:

(1) We construct the first instruction datasets PCL-SFT in detecting Patronizing and Condescending Language and further optimize the instruction rules and data quality.

(2) Based on PCL-SFT, we introduce a comprehensive instruction-tuning framework PclGPT. We first construct PclGPT-EN by using only a small amount of fine-tuning data. Our model surpasses all advanced pre-trained models in classification tasks. We then extend our experiment to the Chinese domain and construct PclGPT-CN, which achieve the same leading results, proving the multilingual transfer capability of our framework.

(3) We demonstrate the value of instruction fine-tuning for implicit emotion detection through fine-grained emotion category detection and adding samples with semantic fuzziness.

2 related works

2.1 PCL Detection Tasks

PCL detection tasks can date back to the work of (Ng, 2007), the author introduced a discriminative definition of condescending language, pointing out that such speech often arises within an imbalanced power dynamic between different groups, and the expression of superior attitudes and discourse on mercy can institutionalize discrimination. (Wong et al., 2014) noted that condescending speech is frequently unconscious, propelled by good intentions, and articulated using embellished language.(Xu, 2022) identified that these unjust treatments of vulnerable groups can exacerbate societal exclusion and inequality, compelling users to exit communities or reduce online participation (Parekh and Patel, 2017). One of the key factors for progress in this area is access to high-quality datasets annotated by experts.(Wang and Potts, 2019) introduced the TalkDown dataset, focusing on PCL in social media, while (Pérez-Almendros et al., 2020) presented the "Don't Patronize Me!" (DPM) dataset, concentrating on how vulnerable groups are portrayed in news reports.(Wang et al., 2023) proposed the pioneering Chinese Condescending Hierarchical Dataset (CCPC).In the realm of detection, transformer-based models are extensively employed for sentiment analysis tasks. For instance,(Pérez-Almendros et al., 2022; Xu, 2022) utilized a modified BERT network for PCL detection tasks, and Lu (Lu et al., 2022) introduced adversarial training to enhance the model's capabilities. While these methodologies represent groundbreaking efforts in detecting PCL, their limitations persist due to insufficient pre-training information, resulting in an incomplete understanding of implicit emotions.

2.2 Domain Fine-tuning LLM

Recently, the emergence of Large Language Models like GPT-3.5(Yang et al., 2023) and GPT-4 (OpenAI, 2023) has led to changes in the entire field of general generated text. However, they are not fully adapted to specific domain tasks. Meanwhile, neither GPT-3.5 nor GPT4 is open-source, which has led to fine-tuning work on open-source LLMs with smaller parameters. Instruction tuning plays a crucial role in shaping LLM's intelligent capabilities (Ouyang et al., 2022; Taori et al., 2023; Chiang et al., 2023). By fine-tuning specific instruction datasets, performance on domain-specific tasks can be significantly improved and aligned with user goals. LLama(Touvron et al., 2023), Vicuna(Chiang et al., 2023), and ChatGLM (Zeng et al., 2022) are all base models that can be used for instruction tuning. At present, the models with instruction tuning and human preference tuning are widely used in the fields of medicine, finance, and law (Singhal et al., 2023; Wu et al., 2023; Cui et al., 2023).

In the field of sentiment detection,(Zhang et al.,

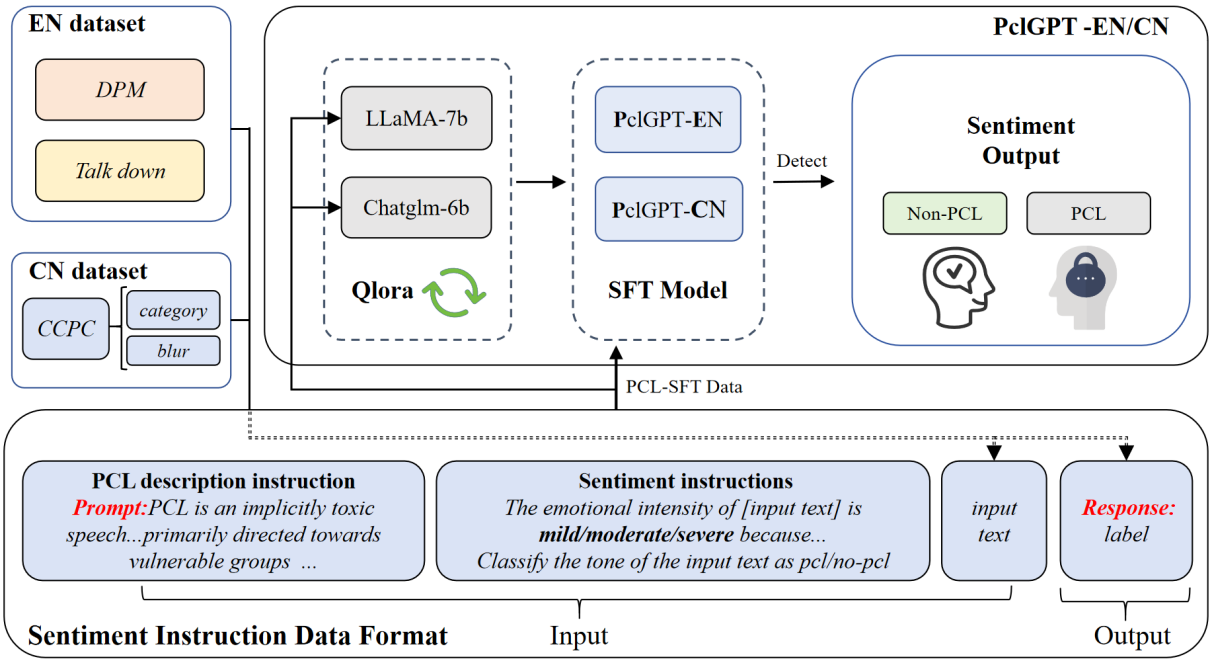

Figure 1: An illustration of the overall PclGPT framework, which encompasses the construction of sentiment instruction datasets and the process of supervised fine-tuning.

2023) introduced instruction tuning to classify emotions in the financial domain; (Nguyen et al., 2023) detected hatred and aggressive emotions by fine-tuning LLM. At present, however, there is currently no instruction tuning method applied to LLMs for tasks related to implicit emotion analysis, such as Patronizing and Condescending Language. The current difficulties include the lack of corpus and the fuzziness of implicit emotional judgment. Thus, proposing a relatively unified fine-tuning paradigm is challenging. Furthermore, effective field research on languages other than English (such as Chinese) is scarce, and this is detrimental to cross-linguistic exploration of sentiment analysis.

As a result, we propose the PclGPT instruction-tuning framework. Our approach is founded on the most popular open-source large language models, such as LLaMA2-7B and ChatGLM3-6B. We fine-tune the model using the PCL-SFT instruction dataset and validate its ability to detect implicit emotions in experiments on fine-grained PCL emotion classification and adding fuzzy emotion samples.

3 PclGPT

Our objective is to elevate the performance of the LLM in targeted implicit emotion classification tasks, such as PCL detection, through the creation of an instruction-tuned large language model. The overall framework of our approach is illustrated in Figure 1. The instruction template we construct uses a combination of descriptions and sentiment instructions to better reflect the implicit semantic features of PCL, then we use instruction tuning to mine the model's ability to discriminate fuzzy emotions.

3.1 Sentiment Instruction Data Format

To better guide LLM, we split the instruction into two parts: ***PCL description instruction*** and ***Sentiment instructions***, as shown in Figure 1.

• **PCL description instruction.** Since PCL is a very subjective emotion category, first we need a complete description of the concept of PCL to guide the model to respond in a standardized format. The description includes the definition and target groups. This part of the content is fixed (e.g., *PCL is a discriminatory and implicitly toxic speech...primarily directed towards vulnerable groups such as the elderly and women...*).

• **Sentiment instructions.** Next, we focus on the significant influence of the intensity of emotional tone on implicit emotions in the process of instruction tuning. Incorporating the emotional intensity labels from the original data, we construct the emotional instructions with a description of the emotional intensity of the input text. The format is specified as (e.g., *The emotional intensity of this*

sentence is mild/moderate/severe because...).

• **Combination.** We combine these two parts and randomly select run instructions from a set of 10 manually created instructions as our input while taking the labels of our original data as the output.

3.2 Instruction Tuning

Instruction tuning utilizes a set of formatted examples in natural language to optimize a pre-trained LLM (Wei et al., 2021). This approach closely aligns with supervised fine-tuning. Initially, we utilize the available English and Chinese datasets to create instructional data to provide supervision signals. Subsequently, we employ the instruction tuning technique to train on the English LLaMA-7B base and the Chinese ChatGLM-6B base separately. The training process employs Qlora fine-tuning, with each model trained using sequence-to-sequence loss. We compare the performance of our model on the test set against advanced pre-trained models. Our approach aims to enhance the precision of identifying condescending emotions and offer guidance for detecting subtle implicit emotions.

• **LLaMA-2-7B**(Touvron et al., 2023) is one of the most advanced English open-source LLM at present. It introduces some optimization measures, including pre-normalization, SwiGLU activation function, and rotation position embedding (RoPE).

• **ChatGLM3-6B**(Du et al., 2022) is an open-source conversational language model that supports Chinese and English bilingualism, based on the General Language Model (GLM) architecture. ChatGLM-6B has been deeply optimized for Chinese question and answer and dialogue, making its performance in the Chinese field more outstanding.

3.3 Implicit Emotion Detection for PCL

Inspired by (Zhang et al., 2023), to evaluate PclGPT's impact on implicit emotions, we test the model on fine-grained emotion classification tasks and compare the results of the model when adding fuzzy implicit PCL examples to the data.

• **Fine-grained sentiment analysis.** Fine-grained sentiment analysis plays an important role in understanding implicit emotions(Tang et al., 2019). Multi-dimensional emotion labels can comprehensively reflect the characteristics of PCL and can observe which category our fine-tuned model has a more significant effect in discriminating more intuitively. The CCPC dataset has more detailed multi-label annotations for each text that is judged to be condescending (specifically, it is labeled as one or more categories of *"Unbalanced Power Relations"*, *"Spectator"*, *"Prejudice Impression"*, *"Appeal"*, and *"Elicit Compassion"*. Therefore, we use CCPC for example and further Split the dataset into five subsets for comparative testing.

• **Add samples with semantic fuzziness.** We conduct additional experiments to assess the detection capabilities of our fine-tuned model for implicit emotions. As a subjective emotion, the ambiguous part of PCL's semantic information is often labeled as intermediate samples during the annotation process. These intermediate samples have more implicit emotions and possess marginal condescending attributes, they will hinder the model's ability to effectively distinguish positive samples. We conduct experiments on the datasets in three situations: without any intermediate samples, with a limited number of intermediate samples, and with all intermediate samples included.

4 PCL-SFT Dataset

4.1 Supervised Fine-tuning

SFT Datasets. We first sorted out all the Chinese and English datasets currently available in the PCL field, as shown in Table 1:

Lan.	Dataset	Source	Scale
EN	DPM	News on Web	10469
	TD	Reddit	68355
CN	CCPC	Weibo,Zhihu	15500

Table 1: Datasets used for instruction tuning. The English parts are used for PclGPT-EN and the Chinese corpus is used for PclGPT-CN.

- **Don't Patronize Me Corpus(DPM).**(Pérez-Almendros et al., 2020) contains 10,469 English paragraphs about potentially vulnerable groups, extracted from the News on the Web (NoW). They are selected from general news reports and labeled hierarchically with numerical labels from 0 to 4, with 0 and 1 being labeled as non-PCL and 2, 3, and 4 as PCL.

- **TalkDown(TD)**(Wang and Potts, 2019). It is a Reddit community data containing 66K En-

Model	Don't Patronize Me			Talk Down		
	P	*R*	*F1*	*P*	*R*	*F1*
BERT	$74.4_{1.1}$	$76.9_{0.7}$	$75.5_{1.0}$	$59.4_{0.1}$	$59.2_{0.3}$	$58.8_{0.8}$
RoBERTa	$74.6_{2.2}$	$79.0_{2.2}$	$76.3_{0.6}$	$64.9_{0.2}$	$63.0_{2.0}$	$59.7_{1.4}$
RoBERTaL	$77.9_{0.2}$	$79.5_{1.9}$	$78.6_{1.0}$	$\mathbf{66.6}_{1.7}$	$62.0_{0.1}$	$59.2_{0.8}$
GPT-3.5	$50.8_{1.2}$	$52.3_{2.4}$	$51.6_{0.7}$	$59.2_{1.2}$	$58.1_{1.4}$	$56.7_{0.3}$
GPT-4	$51.5_{1.2}$	$57.5_{0.8}$	$54.3_{0.7}$	$60.8_{0.2}$	$60.3_{0.4}$	$60.5_{0.4}$
LLaMA-7B	$50.9_{1.3}$	$52.6_{2.5}$	$51.4_{1.2}$	$49.9_{0.8}$	$49.9_{0.4}$	$49.7_{0.4}$
PclGPT **(Ours)**	$\mathbf{78.1}_{0.6}$	$\mathbf{79.8}_{0.8}$	$\mathbf{78.9}_{1.6}$	$66.0_{0.3}$	$\mathbf{65.6}_{0.6}$	$\mathbf{65.3}_{0.6}$

Table 2: Test results of PclGPT and other models on two English datasets. RoBERTaL is RoBERTa-Large.

glish comment/reply pairs. The collected information comes from disadvantaged groups from 2006-2018. One of the replies addressed the span of a specific quote in the comment and contained a word associated with being condescending. The final pair will be marked as one of three categories: PCL, non-PCL, and not sure. We classify the uncertainty as non-PCL.

- **CCPC**(Wang et al., 2023). It contains corpus collected from Chinese social platforms for disadvantaged groups. The dataset now has more than 15k two-level structured annotations. Each sample in the first level has one of two labels: PCL, or non-PCL. At the second level, condescending remarks will be further labeled with one or more of five categories: *"Unbalanced Power Relations"*, *"Spectator"*, *"Prejudice Impression"*, *"Appeal"*, and *"Elicit Compassion"*.

Parameter	Value
Lr	2e-2
Pre Seq Len	128
Training epochs	5
Max Source Len	512
Max Target Len	128
GPUs	V100*4(32G)

Table 3: Instruction tuning basic parameters.

Fine-tune details. For LLaMA2-7B and ChatGLM3-6B, we perform instruction tuning on four V100 GPUs using our PCL-SFT instruction dataset. We employ Qlora to accomplish this procedure and guarantee the consistency of the pertinent training parameters in both Chinese and English. We conduct 5 epochs of tuning and utilized the AdamW optimizer with a learning rate of 2e-2.

Model	CCPC		
	P	*R*	*F1*
BERT	$62.5_{0.2}$	$61.6_{0.7}$	$62.0_{0.5}$
BERT-Multi.	$67.1_{1.7}$	$67.0_{0.9}$	$67.0_{1.2}$
BERT-CN	$68.6_{0.3}$	$69.0_{0.5}$	$68.8_{0.5}$
BERT-CN-wwm	$67.9_{0.7}$	$68.7_{0.1}$	$68.3_{0.2}$
GPT-3.5	$53.1_{2.2}$	$54.2_{1.0}$	$51.2_{2.3}$
GPT-4	$55.4_{0.3}$	$56.3_{0.8}$	$55.7_{0.7}$
Chatglm-6B	$51.9_{0.2}$	$50.2_{1.5}$	$45.2_{1.3}$
PclGPT **(Ours)**	$\mathbf{68.7}_{0.2}$	$\mathbf{72.5}_{1.1}$	$\mathbf{69.8}_{0.4}$

Table 4: Test results of PclGPT and other models on Chinese corpus. BERT-Multi. is multi-language BERT. BERT-CN-wwm is the model using incorporates whole-word masking.

During the training process, we exclusively employ sequence-to-sequence loss for training and map the final generated output to sentiment labels. We set the maximum input text length to 512 tokens to maintain efficiency. Important parameters are recorded in Table 3.

5 Result and Analysis

To evaluate the ability of our model, we conduct sentiment binary classification tasks on two sets of test data: one in English and the other in Chinese. We compare the performance of our PclGPT with a wide range of advanced models. For pre-training models, we use Bert and the advanced variant Roberta. For large language models, we called GPT-3.5 and GPT-4 through the API interface, while using open-source LLMs for direct few-shot testing, such as LLaMA and ChatGLM. To further verify the ability of PclGPT to detect implicit condescending emotion, we carry out detailed tests from two aspects of fine-grained classification and adding fuzzy samples. The main performance metrics of our model include precision, recall, and

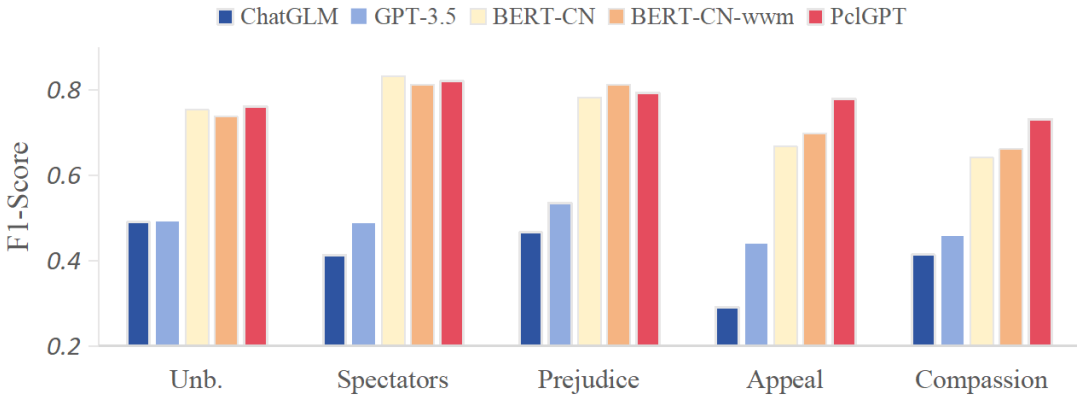

Figure 2: Fine-grained testing across five subcategories of PCL detection:*"Unbalanced Power Relations"*(Unb.), *"Spectators"*, *"Prejudice Impression"*(Prejudice), *"Appeal"*, and *"Elicit Compassion"*(Compassion).

F1 score. The F1 score represents the harmonic mean of precision and recall. We use the macroF1 value in our experiments.

5.1 Overall Performance

Table 2 compares the performance of our PclGPT-EN model on the English test set, while Table 4 displays the results of the PclGPT-CN model on the Chinese test set. The experimental results demonstrate that our PclGPT framework outperforms in both English and Chinese domains. PclGPT's detection capability on the English dataset TalkDown has improved by over 6 percentage points compared with BERT models, while the detection ability for the Chinese dataset CCPC has increased by almost 1 percent. The average performance improvement in detection reaches a notable 17 percent when comparing GPT-3.5 and GPT-4.0 models. This implies that even though the current GPT series models have wider pre-training parameters, they still face difficulties in identifying implicit emotions and lack efficient guidance. Notably, our fine-tuned model consistently outperforms base LLMs such as LLaMA and ChatGLM, showcasing the effectiveness of instruction fine-tuning. In the subsequent sections, we will examine our experimental findings, with a particular emphasis on Fine-grained condescending testing, the addition of semantic ambiguity samples, and case analysis in PCL detection.

5.2 Fine-grained PCL Detection

The test results are shown in Figure 2. It is observed that the effects of GPT and fine-tuned base models are generally low. These models lack guidance similar to the BERT, and it is difficult to detect PCL remarks only through prompt engineering. Our model achieves state-of-the-art(SOTA) results in each subtask category. Looking further at the subcategories, for "Appeal" and "Compassion", the comdescending emotions they contain are the most hidden, and they mostly appear as "false positive" samples. For this type of text, PclGPT has the most obvious improvement effect. This shows that PclGPT can effectively detect implicit speech.

Model	*S-None*	*S-Few*	*S-All*
BERT	67.1(0)	67.2(+0.1)	67.1(-0.6)
ChatGLM	48.1(0)	48.8(+0.7)	48.3(-0.5)
GPT-3.5	64.3(0)	61.3(-3.0)	52.4(-8.9)
GPT-4.0	65.5(0)	63.2(-2.3)	54.5(-8.7)
PclGPT	**67.7(0)**	**71.5(+3.8)**	**72.8(+1.3)**

Table 5: The test results of each model after gradually adding fuzzy samples. The percentage in parentheses indicates the change after addition compared with before addition.

5.3 Add Samples with Semantic Fuzziness

Identifying fuzzy samples, which encompass complex and implicit emotions, is a crucial objective in subjective sentiment analysis. The test results are displayed in Tabel 5. It is evident that when the number of fuzzy intermediate examples increases, both the BERT model and the GPT baseline model experience a decrease in performance. Notably, GPT-3.5 and GPT-4 decline over 8 percentage points, suggesting that they inadequately capture the condescending emotional traits of these fuzzy cases. Our PclGPT is the only model that can effectively detect these fuzzy samples in the S-Few and S-All datasets, which shows that the

Exp.	Original Text	BERT	LLaMA2	GPT4.0	PclGPT	True Label
1	*These poor families face double-digit inflation this year thanks to food , electricity and transport hikes.*	**True**	False	False	**True**	**True**
2	*For some of these male prostitutes, the ' clients ' they picked up on this corner were their only means of survival.*	**True**	False	**True**	**True**	**True**
3	*The fighting raged among the flames: Dazed, starving Jerusalemites wandered lost and distressed through the burning portals.*	False	False	False	**True**	**True**
4	*News <h> Bahamas Gov't denies profiling Haitian migrants*	**False**	True	**False**	True	**False**

Table 6: Sample analysis for PclGPT-EN, the samples were selected from the different sub-categories of PCL.

instruction-tuned model can detect implicit and fuzzy emotions.

5.4 case study for PclGPT-EN

We selected samples of different categories from the test results of the two languages and carried out case tests. The results are detailed in Table 6 and Table 7. Regarding the English section, we conduct a comparative analysis with BERT, LLaMA2, GPT4.0, and PclGPT.

- exp.1 is selected from the "Unbalanced Power Relations" category. This sentence reflects the different identities of the speaking group through the description of poor families and has a discriminatory tendency. GPT4.0 lack the ability to detect descriptive text. BERT and PclGPT successfully identified the text as PCL.

- exp.2 is selected from the "Prejudice Impression" category. This sentence has an unchangeable stereotype of the "male prostitute" group, and the text is the most discriminatory. Except for the LLaMA model, all other models successfully identified the text as PCL.

- exp.3 is selected from the "Elicit Compassion" category. Sympathetic sentences, such as the rhetoric about Jerusalemites in this sentence, often hide the semantics of PCL, it is a sample with semantic fuzziness. BERT and GPT4.0 exhibit subpar performance in detecting intermediate blur examples, whereas PclGPT still demonstrates detection capability by correctly identifying the sample as a True label.

- exp.4 is derived from news reports, which provide an unbiased standpoint, however, PclGPT made an erroneous assessment. This could be attributed to PclGPT's heightened sensitivity towards implicit emotions.

5.5 case study for PclGPT-CN

For the Chinese part, we compare several types of Chinese detection models: BERT-Chinese, ChatGLM3, GPT4.0, and PclGPT for sample analysis.

- exp.1 is taken from the "Unbalanced Power Relations" category. The speaker distinguishes himself from the elderly group in the nursing home, and then expresses his contempt for this group. Except for ChatGLM, all models have detected correct results.

- exp.2 is taken from the "Prejudice Impression" category. The speaker shows the stereotype that single-parent children are difficult to get along with, cannot conceal his contempt for single-parent children, and has a strong condescending mood. All models are correct.

- exp.3 is derived from the "Elicit Compassion" category. The speaker conveys empathy towards the migrant workers by employing

Exp.	Original Text	BERT	Chat-GLM3	GPT4.0	PclGPT	True Label
1	老人既然都送了养老院了，平时肯定是没人管的。 *Since the elderly have been placed in a nursing home, they are undoubtedly left unattended most of the time.*	**True**	False	**True**	**True**	**True**
2	单亲的小孩大概率很难相处。 *Children from single-parent families often face difficulties in getting along with others.*	**True**	**True**	**True**	**True**	**True**
3	农民工挣钱不容易的，确保工资该发就发吧。 *Making a living as a migrant worker is no easy task, let's make sure they receive their rightful wages.*	False	False	False	**True**	**True**
3	脱离群众越来越远讲话越来越轻。 *Growing increasingly distant from the masses, the speeches become more detached.*	**False**	True	True	**False**	**False**

Table 7: Sample analysis for PclGPT-CN.

modal particles. The tone is extremely subtle. Only PclGPT possesses the ability to detect and accurately identify this implicit PCL intonation.

- exp.4 is not a PCL statement. The speaker is a member of the masses, and the statement is not directed at vulnerable groups. PclGPT and BERT can effectively identify this group relationship.

In summary, our PclGPT has greatly enhanced its capacity to detect PCL, particularly in the sympathy subcategory, surpassing other models. This demonstrates that PclGPT has immense potential in implicit emotion detection tasks.

6 Conclusion

In this paper, we propose an implicit emotion detection framework for Patronizing and Condescending Language (PCL) by leveraging the world knowledge and reasoning capabilities of LLM. We construct the first PCL sentiment instruction dataset PCL-SFT. Based on PCL-SFT, we introduce a comprehensive instruction-tuning framework PclGPT. The experimental results of our model exceeded all existing methods, proving the effectiveness of instruction tuning. We also showcase the efficacy of PclGPT for implicit emotion detection by accurately detecting emotions within specific categories and evaluating the inclusion of fuzzy samples. The task has great potential, especially for specific implicit sentiments such as "Elicit Compassion" with significant performance improvements. Our study paves the way for future research on the detection of PCL and other implicit toxic emotions using LLM. At the same time, our research results prove the effectiveness of cross-language research and provide strong support for the protection of vulnerable groups in Chinese and English communities.

7 Limitation

PCL is an implicit and subjective classification of toxic speech. Since the current relevant research is very limited, we need more linguistic foundations to improve the standardized definition of this type of speech. Our current research is still deficient in examining instances of "false positive" occurrences, such as insincere acts of kindness in speech and disingenuous praise towards marginalized communities. Meanwhile, the presence of condescending emotions can be strongly influenced by the speaker's tone and the use of modal particles. These necessitate us to amalgamate our efforts with sarcasm detection to a greater extent.

References

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Junyu Lu, Hao Zhang, Tongyue Zhang, Hongbo Wang, Haohao Zhu, Bo Xu, and Hongfei Lin. 2022. Guts at semeval-2022 task 4: Adversarial training and balancing methods for patronizing and condescending language detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 432–437.

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

Thanh Thi Nguyen, Campbell Wilson, and Janis Dalins. 2023. Fine-tuning llama 2 large language models for detecting online sexual predatory chats and abusive texts. *arXiv preprint arXiv:2308.14683*.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2:13.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Pooja Parekh and Hetal Patel. 2017. Toxic comment tools: A case study. *International Journal of Advanced Research in Computer Science*, 8(5).

Carla Pérez-Almendros, Luis Espinosa Anke, and Steven Schockaert. 2022. Pre-training language models for identifying patronizing and condescending language: an analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3902–3911.

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. *arXiv preprint arXiv:2011.08320*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Feilong Tang, Luoyi Fu, Bin Yao, and Wenchao Xu. 2019. Aspect based fine-grained sentiment analysis for online reviews. *Information Sciences*, 488:190–204.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Hongbo Wang, Mingda Li, Junyu Lu, Liang Yang, Hebin Xia, and Hongfei Lin. 2023. Ccpc: A hierarchical chinese corpus for patronizing and condescending language detection. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 640–652. Springer.

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

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Gloria Wong, Annie O Derthick, EJR David, Anne Saw, and Sumie Okazaki. 2014. The what, the why, and the how: A review of racial microaggressions research in psychology. *Race and social problems*, 6:181–200.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Jinghua Xu. 2022. Xu at semeval-2022 task 4: Prebert neural network methods vs post-bert roberta approach for patronizing and condescending language detection. *arXiv preprint arXiv:2211.06874*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. 2023. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models. *arXiv preprint arXiv:2306.12659*.