CCPC: A Hierarchical Chinese Corpus for Patronizing and Condescending Language Detection

Hongbo Wang[1], Mingda Li[1], Junyu Lu[1], Liang Yang[1], Hebin Xia[1], and Hongfei Lin[1(✉)]

School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China
{dutlaowang,22209217lmd,dutljy,2672054553}@mail.dlut.edu.cn ,
{liang,hflin}@dlut.edu.cn

Abstract. Patronizing and Condescending Language (PCL) is a form of implicitly toxic speech aimed at vulnerable groups with the potential to cause them long-term harm. As an emerging field of toxicity detection, it still lacks high-quality annotated corpora (especially in the Chinese field). Existing PCL datasets lack fine-grained annotation of toxicity level, resulting in a loss of edge information. In this paper, we make the first attempt at fine-grained condescending detection in Chinese. First, we propose CondescendCN Frame, a hierarchical framework for fine-grained condescending detection. On this basis, we introduce CCPC, a hierarchical Chinese corpus for PCL, with 11k structured annotations of social media comments from Sina Weibo and Zhihu. We find that adding toxicity strength (TS) can effectively improve the detection ability of PCL and demonstrate that the trained model still retains decent detection capabilities after being migrated to a larger variety of media data (over 120k).Due to the subjective ambiguity of PCL, more contextual information and subject knowledge expansion are critically required for this field.

Keywords: Patronizing and Condescending Language · Hierarchical Chinese Corpus · Toxic Speech Detection.

1 Introduction

When an entity's language use shows a superior attitude toward others or depicts them in a sympathetic way, we call it Patronizing and Condescending Language(PCL). As an essential subfield of toxic speech, PCL is an open, challenging, and underexploited research area in natural language processing [10,16]. It is a kind of toxic speech aimed primarily at vulnerable communities. Condescending language is more subtle than traditional hate speech or offensive language, which are clearly offensive and easy to detect on the Internet. PCL is often unconscious, motivated by good intentions, and expressed in flowery language[16,6]. It is based on one group's superiority over another and displays an

unbalanced power relationship. These superiorities and compassionate discourses can normalize discrimination and make it less visible [8]. This unfair treatment of vulnerable groups contributes to further exclusion and inequality in society [17], forcing users to leave the community or reduce online participation [16,6], and increasing the risk of depression and suicidal behavior. Thus, PCL detection is a potentially high-impact detection task that can provide theoretical guidance for supporting interventions in online communities [12], assisting linguists in understanding implicit toxicity, and effectively caring for vulnerable groups [15].

Although there has been substantial work on hate speech and offensive language, for example, many researchers use large-scale pre-trained models in deep learning for detection tasks [18,7,19,27,26], and some toxicity models were retrained using the professional corpus for some highly toxic tasks [3], PCL modeling remains very limited. The detection of this language requires special knowledge or familiarity with specific cultural tropes due to their recessive elements [14,9], so one of the critical factors for progress in this area is the need for high-quality datasets annotated by experts to address these subtle detrimental properties. [15] introduce the Talk Down dataset, which is focused on condescending language in social media, [10]introduce the Don ' t Patronize Me! (DPM) dataset, which is focused on the way in which vulnerable communities are described in news stories. In terms of models, [25] introduced adversarial training to improve PCL detecting capability. However, relevant research in the Chinese field is still lagging behind, which is a bottleneck preventing further research. Identifying PCL often seems to require a deep commonsense understanding of human values [24], it requires refinement of the category and intensity of toxicity. However, in relevant PCL datasets, the division of levels and granularity is limited to categories and target groups, and the strength of toxicity is not clearly defined(Tab. 1 compares current datasets with our work).

Table 1. Comparison of current PCL datasets, including PCL categories, Toxicity strength, Target groups, and Context.

Corpus	Language	Source	Size	PCL cate.	Toxicity strength	Target groups	Context
TalkDown [15]	English	Reddit	6510(bal.) 68355(unbal.)				✓
Dont Patronize Me! [10]	English	News on Web	10,469	✓		✓	
CCPC(ours)	Chinese	Weibo,Zhihu	11397	✓	✓	✓	

To fill these gaps, we introduce CondescendCN Frame, the Chinese Internet's first condescending framework. Compared with traditional single classification, it is a more fine-grained hierarchical framework with three levels: (1) Whether toxic, (2) Toxic type, (3) PCL Toxicity strength (Level), PCL category, and

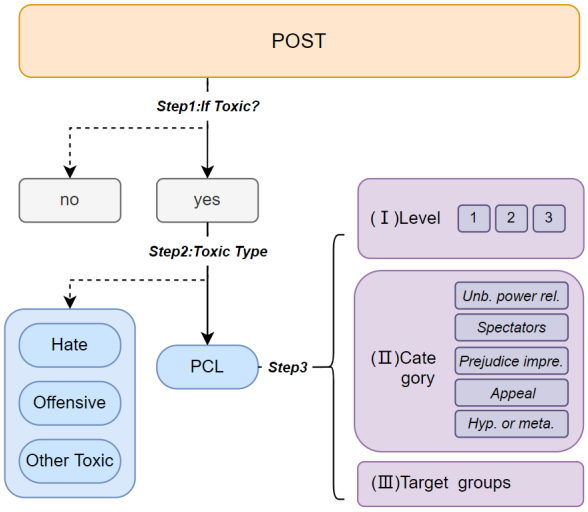

Fig. 1. How to identify condescending language and its type? Here is an example of our CondescendCN framework.

Target groups. Based on the frame(See Fig. 1), we present a fine-grained corpus-CCPC with over 11K comments from two Chinese mainstream social media, Sina Weibo and Zhihu. To verify our dataset, We migrate the data to Weibo for event detection and group identification. Moreover, we present a statistical analysis of CCPC and demonstrate PCL is more likely to target women and children.

2 CondescendCN Frame

Due to the implicit features of condescending texts, it is often difficult to capture the true meaning of condescending semantics from a single point of view with coarse-grained features. We first propose our hierarchical framework for multi-dimensional condescending detection. Then, in multiple pilots, we refine our framework and guidelines by testing small-scale tagging and marginal cases. A methodical interpretation of the free text will increase confidence in machine predictions, and it is of great significance to mitigate deviations in the design and construction of annotations[2,23]. The specific structure of the framework is as follows:

2.1 Toxicity Identification

First, we determine whether the text is toxic. Toxic language is generally defined as rude and uncomfortable remarks [4]. Toxic language detection serves as the foundation for subsequent labeling and is thought of as a binary classification task (toxic or non-toxic).

2.2 Toxicity Type

The second step is to further determine the type of toxicity (hate speech, offensive language, PCL, etc.). Unlike other toxic languages with prominent offensive characteristics, PCL is frequently used for positive contributions due to its subtle toxic expression [11]. Thus, the attack is weak. In this paper, we uniformly classify non-toxic language, hate speech, offensive language, etc. as non-PCL speech, with the remaining classified as PCL.

2.3 PCL Toxicity Strength

In this paper, we creatively use toxicity strength as a category of fine-grained PCL, which will help us better comprehend the toxicity information and reduce our subjective error for edge cases at a fine-grained level [23]. We rank PCL toxicity strength into three levels in increasing order of semantic strength:

1) Weak. Often in the form of "false positive" praising vulnerable groups, such as the qualities worth learning from those living in vulnerable communities. The tone is the most gentle.

2) Middle. Although the author keeps a class distance between himself and the vulnerable groups, the tone will express more sympathy and hypocrisy and will offer shallow opinions to the disadvantaged group from an objective perspective.

3) Strong. The author and the vulnerable group are in completely different classes, and the tone is sharper because of the apparent discriminatory language, superior attitude, and sarcastic semantics toward them.

2.4 PCL Categories

Finally, using the research findings of [10] as a guide, along with the unique characteristics of the Chinese condescending language itself, we propose a detailed linguistic classification of Chinese PCL:

-Unbalanced power relations (unb). Maintain a class and power distance from vulnerable groups, and assist them more as saviors to help them out of difficulty. [20,21].

- Spectators (spe). Without careful consideration, on a whim, spectators offer simple opinions or solutions that cannot solve the problem at its core.

- Prejudice impression (pre). A discriminatory impression with preconceived views on vulnerable groups when offering assistance or advice, but the author's superiority is concealed by friendliness or sympathy. The characteristics of the stereotype are not apparent on the surface.

- Appeal (appe). Calling out on behalf of a group of experts or advocates in the hope that vulnerable groups can change their situation.

- Elicit Sympathy (es). The author may explicitly express compassion and concern for the disadvantaged, or he may use rhetoric, metaphor, etc. to describe the disadvantaged as people in need in order to elicit readers' sympathy.

2.5 PCL Group Detection

PCL mainly targets vulnerable groups. During the corpus collection phase, we categorize the vulnerable groups targeted by the gathered comments for further research. The vulnerable groups include disabled people, women, the elderly, children, commons, single-parent families, and disadvantaged groups(working class, peasant class, etc.).

3 The Corpus

3.1 Data Collection

We collected comments on popular posts from two mainstream Chinese media platforms, Zhihu and Sina Weibo, as our data sources. We limited the scope of our data collection to seven main types of vulnerable groups related to condescending hot topics and events. Then, for each group, we manually compiled a list of keywords and conducted a search within the list's scope.

Table 2. A sample description for the CCPC corpus. The text is divided into levels according to toxicity, toxicity type, toxicity strength, and PCL categories. In our hypothesis, the offensiveness of PCL language is between non-toxic and hate speech.

Exp.	Comment	Toxic?	Toxic Type	Toxicity strength	PCL Categories
1	现在离婚率本来就很高很高了。 *The divorce rate is already very high now.*	no	-	-	-
2	我接触过的单亲家庭出来的孩子，性格多少都不太好，孩子也是受害者啊。 *Children from single-parent families I've met have challenging personalities, and they are also victims.*	yes	pcl	1	pre,es
3	残疾人就业是个严重的问题, 应该给他们更多的岗位。 *The employment of the disabled is a critical issue; more jobs should be created for them.*	yes	pcl	2	unb,appe
4	与其担心别的，农民工朋友倒不如想想如果自己被欠薪了，应该如何合法讨薪才对。 *Instead of worrying about other things, migrant worker friends should consider how to legally obtain their wages if they are not paid.*	yes	pcl	3	unb
5	笑拉了，我就是瞧不起你们小仙女怎么了。 *LOL, what's wrong with me looking down on those fairy girls?*	yes	hate	-	-

To ensure the quality of the corpus, we eliminated posts with fewer than 20 comments and performed additional manual screening. As a result, we collected more than 14K comments on 1082 popular blog posts containing these keywords. We removed duplicate and irrelevant samples (including common fixed tags on Weibo, such as "回复" and "转发微博"), as well as samples with a length of fewer than five characters or without any Chinese character annotations. We retained the emoji in the samples and converted them into the corresponding Chinese text pairs specified on the platform (such as "抱抱*]* "—"*hug]*", "允悲*]*"—"*allow sadly]*")to preserve as much of the emotional semantic information mapped in the emoji as possible [7]. Finally, 11397 comments are retained. A sample of the dataset is shown in Tab. 2.

We carefully considered the ethical and moral implications of data collection. To protect user privacy, all usernames are replaced with a unique "username" token. We ensure that the CCPC corpus is used solely for academic research in this field and make all collected data publicly accessible.

3.2 Data Annotation

Main Annotation. Using our proposed framework as a foundation, we annotated our dataset as follows: We labeled non-toxic speech, hate speech, and offensive language as N-PCL (label 0). For PCL, We introduced **Toxic Strength (TS)**, further labeling based on its condescending toxicity intensity(label 1 to 3). The detailed design idea refers to Section 2.3.

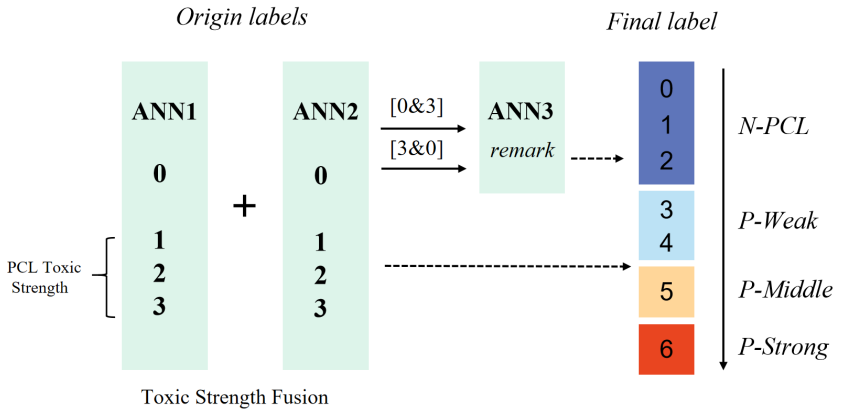

Fig. 2. Toxic Strength Fusion of PCL to produce final label.

Each comment was labeled by two annotators, and a third annotator proofread it for accuracy. To maximize the amount of information captured by the annotations, particularly concerning the more granular evaluation of edge cases [10], we proposed Toxic Strength Fusion, a method that compiled the results of two main annotators and reclassified the labels based on their PCL toxicity intensity.

The final label will be classified into four groups：N-PCL,PCL-Weak,PCL-Middle and PCL-Strong.Fig. 2 illustrates an example of judgment. We conducted two sets of such labeling tasks and finally obtained the dataset by voting. It is important to note that hateful and offensive texts can interfere with PCL toxicity classification, so we did manual proofreading when summing these samples.

Annotator demographics. To ensure the diversity of the annotators, we recruited six annotators (four main markers, two proofreaders) with differing genders, ages, and educational backgrounds(50% women, 50% men; 25±5 ages; three masters, two doctors, one undergraduate). Meanwhile, due to the subtle toxicity unique to the PCL, which is easy to cause differences, we invited two experts in the field of language to guide our team's annotations.

Annotator agreement. We calculated the Kappa Inter-Annotator Agreement (IAA) for binary classification and multi-category labeling. The IAA improved if we omitted all comments labeled as borderline(e.g. Label1/Label2) by at least one annotator [22]. More information is shown in Tab. 3.

Table 3. Inter-annotator agreement table. The left column is the IAA test for PCL detection, and the marginal information is gradually deleted from top to bottom. The table on the right shows the consistency labeling test for the PCL category.

PCL Detection	IAA	PCL Catagory	IAA
All labels	0.62	Unb.power rel.	0.65
		Spectators	0.42
Remove label1	0.64	Pre. impre.	0.59
		Appeal	0.48
Remove label1,2	0.69	Hyp.or meta.	0.71
		Others	0.66

3.3 Data Description

After data collection and labeling, our CCPC corpus contains 11397 comments, including 9251 negative samples (N-PCL) and 2146positive samples (PCL labels are classified as P-Week, P-Middle, and P-Strong as the toxic strength increases), which covers the majority of vulnerable groups in significant Chinese forums (Tab. 4) . Tab. 5 depicts the proportion of three toxicity strengths among vulnerable groups.Tab. 6 provides additional statistics on vulnerable groups. We notice that the condescension rate on the Weibo platform is substantially higher than on Zhihu. In addition, the condescension rate for women and children is considerable on both platforms, with Zhihu having the highest rate for children and Weibo having the highest rate for women. This provides more details on focusing on these groups.Based on these statistics, we will provide an expanded CCPC dataset[1] with a higher data scale (14k) and a broader spectrum of vulnerable populations.

[1] Our dataset and code are available on `https://github.com/dut-laowang/CCPC`

Table 4. Basic statistics of CCPC.

Toxic-Category	Num
N-PCL	9251
P-Weak	1167
P-Middle	439
P-Strong	540
Total	11397

Table 5. PCL toxicity strength statistics for different vulnerable groups.

	P-Weak	P-Middle	P-Strong
Disabled	61.8	16.4	21.8
Women	29.7	29.7	40.6
Elderly	54.6	16.9	28.5
Children	63.7	17.7	18.6
Commons	59.8	17.1	23.1
Single.	62.2	14.7	23.1
Disadv.	57.6	19.7	22.7

Table 6. Statistical Results of CCPC from different Platforms. Platform_p represents samples marked as PCL, whereas prop.(%) represents a percentage.

	Disabled	Women	Elderly	Children	Commons	Single-parents	Disadv. groups	Total
zhihu	838	735	656	858	922	628	815	5452
zhihu_p	66	110	72	177	72	87	123	700
prop.(%)	7.9	15.0	11.0	**20.6**	7.8	13.8	15.1	12.8
weibo	920	760	747	950	864	754	950	5945
weibo_p	167	263	142	323	78	247	226	1446
prop.(%)	18.2	**34.6**	19.0	34.1	9.0	32.8	23.8	**24.3**
Total	1758	1495	1403	1808	1786	1382	1765	11397

4 Experiments

4.1 Baselines

Here we present the primary baseline used in our experiments. The dataset is divided into a ratio of 8:1:1. We limited epoch=15, batch size=32, and set the same random seed. The results are depicted in Tab. 7 and Tab. 8.

BERT. We conducted experiments with PLMs based on BERT and related variants. CCPC was examined using bert-base-uncased[2]($BERT$), bert-base-multilingual-cased[3]($BERT_M$), and bert-base-Chinese[4]($BERT_C$). These PLMs were used as encoders, and we use fully connected layers as classifiers for PCL tasks. We separately evaluated the original label and the label with toxic strength fusion(TS).

BiLSTM. We used a bidirectional LSTM to represent individual words using glove embedding. Our dropout rate is 0.5% for both the LSTM layer and the classification layer. We used precision, recall, and F1 score for these tasks.

[2] `https://huggingface.co/bert-base-uncased`
[3] `https://huggingface.co/bert-base-multilingual-cased`
[4] `https://huggingface.co/bert-base-Chinese`

Table 7. Our work combines toxic strength fusion(TS) to determine whether it is PCL, which is regarded as a binary classification task. The corpus containing TS achieves better F1 results. BiLSTM also received high marks, indicating that LSTM is still effective in the field of short text.

	Input	P	R	F1
$BERT_C$	$\wedge$ *TS*	**0.709**	**0.719**	**0.714**
$BERT_C$		0.682	0.693	0.687
$BERT_M$	$\wedge$ *TS*	0.653	0.646	0.649
$BERT_M$		0.637	0.659	0.643
BERT	$\wedge$ *TS*	0.579	0.590	0.584
BERT		0.586	0.600	0.589
BiLSTM		0.677	0.645	0.656

Next, PCL category detection is viewed as a sentence-level multi-label classification problem, where each paragraph is assigned a subset of PCL category labels.(Tab. 8).

Table 8. Results of categorizing PCL, which is regarded as a multi-label classification task. The categories of PCL include Unbalanced power relations(unb), Spectators(spe), Prejudice impression(pre), Appeal(appe), and Elicit sympathy(es).

	BERT			$BERT_M$			$BERT_C$		
(%)	P	R	F1	P	R	F1	P	R	F1
unb	86.81	98.75	92.40	95.12	92.34	93.71	97.21	94.43	**95.80**
spe	33.33	22.58	26.92	55.56	64.52	59.70	75.01	67.74	**71.19**
pre	63.49	78.43	70.18	72.73	62.75	67.37	73.08	74.51	**73.79**
appe	0.00	0.00	0.00	22.12	24.35	**23.18**	22.22	21.98	22.10
es	17.14	31.58	22.22	34.48	52.63	41.67	38.46	54.63	**45.11**

4.2 PCL Detection for Migration Tasks

We hope that the model we trained can play an active role in detecting PCL on various Chinese mainstream public opinion platforms, which depends on whether our model can obtain accurate condescension rate recognition in a large quantity of external data [15]. We noticed that Weibo is an excellent Chinese platform for evaluating the transferability of models, with the majority of users engaging in community activities with #Keywords.

We confirmed it by comparing condescension rates in different types of vulnerable groups and judging whether the model is accurate for common sense-based recognition.

We chose three distinct categories of disadvantaged-targeted communities for Group recognition: Group A. The community of vulnerable populations.

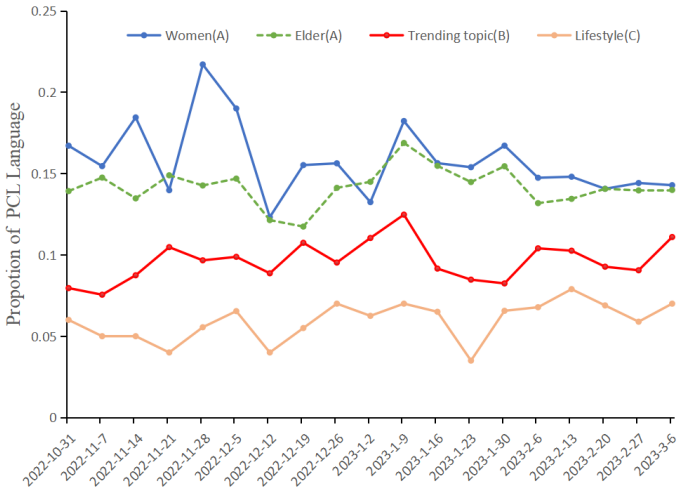

Fig. 3. Condescension rates for different disadvantaged communities. Community A contains *Women*(Blue) and *Elder*(Green); B contains *Trending topics*(Red); C contains *Lifestyle*(Orange). Data were selected weekly from October 2022 to March 2023.

The community is believed to have a high rate of condescension, e.g. *# 妇女 (Women),# 老人 (Elder)*. Group B. Comprehensive community. The condescension rate is at a normal level, e.g. *# 微博热搜 (Trending topics)*. Group C. Non-vulnerable communities. These communities have a low incidence of condescending languages, e.g. *# 生活方式 (Lifestyle),# 娱乐 (Entertainment)*. Between November 2022 and March 2023, we acquired over 120k comments from these communities organized by week. The comparison and validation of the three categories of data are depicted in Fig. 3. We observed higher condescension rates in #Women and #Elder, and the lowest in #Lifestyle. Obviously, models are more inclined to identify PCL in disadvantaged communities. We can also infer that the rate of condescension towards women, a vulnerable group, is much higher than that of males.

5 The ambiguity of PCL

Due to the broad definition of Condescending Language, the judgment of PCL is highly ambiguous. First, it should be determined whether there is a clear class division in condescending discourse, which necessitates more contextual knowledge. It is difficult to discern the identity and class of the speaker based merely on short text sentences. E.g. *I think having a minimal living allowance alone is not enough*, this comment cannot ascertain the class to which 'I' belong or whether the speaker's objectives are directed at vulnerable groups. Second, there is a clear category of 'sympathy' in condescending remarks, but these should be distinguished from those expressing genuine concern, because hypocrisy and sympathy are linguistically similar, but the stated thoughts are not.E.g. *I empathize with you, we are all going through a lot* ,this sentence is not a PCL statement since it does not reflect the superiority complex of different classes

through hypocritical caring, but rather out of concern for the common community.To properly separate these hazy judgments, the model requires more world knowledge and precise definitions.

6 Conclusion and Future Work

PCL language detection is a potentially high-impact detection task for vulnerable communities, which can aid linguists in comprehending implicit toxicity and provide theoretical direction for caring for vulnerable groups. However, current research in the Chinese field lags behind. In this paper, we propose the CondescendCN Frame, the first condescending framework on the Chinese Internet, which divides a comment into a more fine-grained level. On this basis, we construct a Chinese PCL dataset, CCPC, and demonstrate that the addition of toxicity strength(TS) improves PCL detection. The trained model can detect on a larger platform, proving the CCPC's reliability. We conduct sentiment strength analysis, which reveals that condescending language is primarily directed at women and children, who are in critical need of more humane care. Our experiment confirms that PCL detection is still very subjective, and its scientific definition is rather ambiguous. More contextual information and subject knowledge expansion are critically required for these features.This will be the focus of our future research.

References

1. Bell, K.M.: Raising africa?: Celebrity and the rhetoric of the white saviour. PORTAL: Journal of Multidisciplinary International Studies **10**(1), 1–24 (2013)
2. Bussone, A., Stumpf, S., O'Sullivan, D.: The role of explanations on trust and reliance in clinical decision support systems. In: 2015 international conference on healthcare informatics. pp. 160–169. IEEE (2015)
3. Caselli, T., Basile, V., Mitrović, J., Granitzer, M.: Hatebert: Retraining bert for abusive language detection in english. arXiv preprint arXiv:2010.12472 (2020)
4. Dixon, L., Li, J., Sorensen, J., Thain, N., Vasserman, L.: Measuring and mitigating unintended bias in text classification. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. pp. 67–73 (2018)
5. Fortuna, P., da Silva, J.R., Wanner, L., Nunes, S., et al.: A hierarchically-labeled portuguese hate speech dataset. In: Proceedings of the third workshop on abusive language online. pp. 94–104 (2019)
6. Huckin, T.: Textual silence and the discourse of homelessness. Discourse & Society **13**(3), 347–372 (2002)
7. Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P., Mukherjee, A.: Hatexplain: A benchmark dataset for explainable hate speech detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 14867–14875 (2021)
8. Ng, S.H.: Language-based discrimination: Blatant and subtle forms. Journal of Language and Social Psychology **26**(2), 106–122 (2007)
9. Parekh, P., Patel, H.: Toxic comment tools: A case study. International Journal of Advanced Research in Computer Science **8**(5) (2017)
10. Pérez-Almendros, C., Espinosa-Anke, L., Schockaert, S.: Don't patronize me! an annotated dataset with patronizing and condescending language towards vulnerable communities. arXiv preprint arXiv:2011.08320 (2020)

11. Price, I., Gifford-Moore, J., Fleming, J., Musker, S., Roichman, M., Sylvain, G., Thain, N., Dixon, L., Sorensen, J.: Six attributes of unhealthy conversation. arXiv preprint arXiv:2010.07410 (2020)
12. Spertus, E.: Smokey: Automatic recognition of hostile messages. In: Aaai/iaai. pp. 1058–1065 (1997)
13. Straubhaar, R.: The stark reality of the 'white saviour' complex and the need for critical consciousness: A document analysis of the early journals of a freirean educator. Compare: A Journal of Comparative and International Education **45**(3), 381–400 (2015)
14. Van Aken, B., Risch, J., Krestel, R., Löser, A.: Challenges for toxic comment classification: An in-depth error analysis. arXiv preprint arXiv:1809.07572 (2018)
15. Wang, Z., Potts, C.: Talkdown: A corpus for condescension detection in context. arXiv preprint arXiv:1909.11272 (2019)
16. Wong, G., Derthick, A.O., David, E., Saw, A., Okazaki, S.: The what, the why, and the how: A review of racial microaggressions research in psychology. Race and social problems **6**, 181–200 (2014)
17. Xu, J.: Xu at semeval-2022 task 4: Pre-bert neural network methods vs post-bert roberta approach for patronizing and condescending language detection. arXiv preprint arXiv:2211.06874 (2022)
18. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). arXiv preprint arXiv:1903.08983 (2019)
19. Zhou, J., Deng, J., Mi, F., Li, Y., Wang, Y., Huang, M., Jiang, X., Liu, Q., Meng, H.: Towards identifying social bias in dialog systems: Frame, datasets, and benchmarks. arXiv preprint arXiv:2202.08011 (2022)
20. Bell, K.M.: Raising africa?: Celebrity and the rhetoric of the white saviour. PORTAL: Journal of Multidisciplinary International Studies **10**(1), 1–24 (2013)
21. Straubhaar, R.: The stark reality of the 'white saviour' complex and the need for critical consciousness: A document analysis of the early journals of a freirean educator. Compare: A Journal of Comparative and International Education **45**(3), 381–400 (2015)
22. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. biometrics pp. 159–174 (1977)
23. Lu, J., Xu, B., Zhang, X., Min, C., Yang, L., Lin, H.: Facilitating fine-grained detection of chinese toxic language: Hierarchical taxonomy, resources, and benchmarks (2023)
24. Pérez-Almendros, C., Anke, L.E., Schockaert, S.: Pre-training language models for identifying patronizing and condescending language: an analysis. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 3902–3911 (2022)
25. Lu, J., Zhang, H., Zhang, T., Wang, H., Zhu, H., Xu, B., Lin, H.: 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). pp. 432–437 (2022)
26. Min, C., Lin, H., Li, X., Zhao, H., Lu, J., Yang, L., Xu, B.: Finding hate speech with auxiliary emotion detection from self-training multi-label learning perspective. Information Fusion **96**, 214–223 (2023)
27. Lu, J., Lin, H., Zhang, X., Li, Z., Zhang, T., Zong, L., Ma, F., Xu, B.: Hate speech detection via dual contrastive learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2023)