Overview of NLPCC 2022 Shared Task 7: Fine-Grained Dialogue Social Bias Measurement

Jingyan Zhou[1], Fei Mi[2], Helen Meng[1], and Jiawen Deng[3](✉)

[1] Dept. of Systems Engineering & Engineering Management, The Chinese University of Hong Kong
{jyzhou, hmmeng}@se.cuhk.edu.hk
[2] Huawei Noah's Ark Lab
mifei2@huawei.com
[3] The CoAI group, DCST, Institute for Artificial Intelligence, State Key Lab of Intelligent Technology and Systems Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China
dengjw2021@mail.tsinghua.edu.cn

Abstract. This paper presents the overview of the shared task 7, Fine-Grained Dialogue Social Bias Measurement, in NLPCC 2022. In this paper, we introduce the task, explain the construction of the provided dataset, analyze the evaluation results and summarize the submitted approaches. This shared task aims to measure the social bias in dialogue scenarios in a fine-grained categorization which is challenging due to the complex and implicit bias expression. The context-sensitive bias responses in dialogue scenarios make this task even more complicated. We provide 25k data for training and 3k data for evaluation. The dataset is collected from a Chinese question-answering forum Zhihu[4]. Except for the above-mentioned bias attitude label, this dataset is also finely annotated with multiple auxiliary labels. There are 11 participating teams and 35 submissions in total. We adopt the macro F1 score to evaluate the submitted results, and the highest score is 0.5903. The submitted approaches focus on different aspects of this problem and use diverse techniques to boost the performance. All the relevant information can also be found at https://para-zhou.github.io/NLPCC-Task7-BiasEval/.

Keywords: Dialogue Social Bias · Social Bias Measurement.

1 Introduction

Social bias is an unfair stereotype, disdain, or misunderstanding targeted at certain groups of people or individuals because of their demographic characteristics [1, 4], e.g., gender [17], race [5, 11], occupation, etc. Recently, with the increasing attention on AI ethics issues, there is a growing body of work in social bias research in the NLP field [2, 7, 10, 12, 13]. However, this task remains challenging due to the implicity and subtleness of the biased expressions. The complexity

[4] www.zhihu.com

of social bias makes this task beyond a straightforward dichotomy problem [10] and requires nuanced analyses [12, 3].

Nevertheless, biased expressions can have an enormous negative influence, amplify the biased opinions of certain groups, and even intensify the confrontation between different groups in society [6]. Therefore, detecting and mitigating the social bias in dialogue systems is a burning need as such systems are serving as direct interfaces to users [15, 16]. However, social bias in dialogue scenarios is even harder to identify due to its context sensitivity [14]. Aside from the complexity of the task itself, limited by the scarcity of high-quality annotation datasets, bias in Chinese scenarios has rarely been studied [6, 18].

To prompt the research in this field, in NLPCC 2022, we designed the Dialogue Social Bias Measurement task and provided a high-quality dataset CDialBias. Specifically, this task requires understanding and categorizing the last utterances in dialogues into four types w.r.t. their bias attitudes, including *Biased*, *Anti-bias*, *Neutral*, and *Irrelevant*. To the best of our knowledge, this task provides the first benchmark dataset for dialogue social bias measurement in Chinese, which is well-annotated with auxiliary labels via a normative process. In total, 22 teams registered for this task, 11 teams submitted their results, and there are 35 submissions in total. We evaluated the submissions using the macro F1 score, and top-3 teams achieved 0.5903, 0.5902, and 0.5738 separately. The best performing teams solve the proposed task from different perspectives including reformulating the task, comparing different models, and adopting multi-task learning.

2 Task Description

2.1 Task formulation

This task is to measure the social bias in the dialogue scenario. Specifically, given a 2-turn dialogue: (u_1, u_2) and a pre-defined label set (*Biased*, *Anti-bias*, *Neutral*, and *Irrelevant*), the participants are expected to conduct the 4-type classification and assign a correct label for the last utterance u_2. Table 1 gives examples of the four types of data.

In the training data, besides the above 4-type information, we also provide auxiliary labels including *context-sensitivities*, *data types*, *topics*, and *target groups*, while only the 2-turn dialogues are available in the test data. The definition of annotated labels is detailed in Section 2.3.

2.2 Data Collection

We build the Chinese Dialogue Bias Dataset CDialBias 2.0 for this shared task by refining the annotation in the previous work [18]. To study the bias expressions in dialogue, we collected conversational data from the Zhihu platform and construct a dataset called CDialBias, covering the topics of race, gender, region, and occupation.

Table 1. Example of given input and expected output (the true label of last utterance $u2$) for this task.

Input (Dialogue)	Excepted Output (True Label)
[u_1] 想知道以男性视角，家里有位全职太太（家庭主妇）是种什么感受? [u_2] 屁贡献没有，把爷们管的像条狗还天天觉得倍儿委屈也是搞不懂。	Biased
[u_1] 做数据分析的女孩子，职业发展前景在哪里？数据分析枯燥吗? [u_2] 不要觉得女性就一定比别人差很多！	Anti-bias
[u_1] 现今中国人的种族歧视严重吗? [u_2] 我认为，种族歧视和种族自我保护是有区别的。	Neutral
[u_1] “农民工”，“新生代农民工”的称呼是否具有歧视性? [u_2] 不就是个头衔吗？如果能捡到金条，去丐帮都行。	Irrelevant

Considering the sparse distribution of bias-related data on social platforms, we pre-collected some bias phenomenon-related keywords that are widely discussed, for example, "*nurse*", "*farmer*" for occupational bias, "*blacks*", "*Asian*" for racial bias, etc.. These keywords are used as queries to retrieve relevant questions from Zhihu, and then the replies under these questions are crawled. Subsequently, we further performed rigorous data cleaning to construct the question-response dialogue data for further annotation.

2.3 Annotation Schema

This shared task focuses on analyzing biased opinions in dialogue. We developed a multi-dimensional schema in the annotation process and assigned fine-grained labels to each dialogue response.

Context-Sensitivity Most existing analyses related to dialogue safety focus on the utterance level, ignoring its sensitivity of safety in context [14]. To this end, we classify responses into *Context-Sensitive* and *Context-Independent* based on whether their bias-attitude judgment is context-dependent.

(1) **Context-Independent (*Cxt-Ind*)**: The responses carry explicit information to support the further judgment of the bias-attitude.
(2) **Context-Sensitive (*Cxt-Sen*)**: Information in the response is insufficient to determine whether bias-related topics are discussed or whether biased opinions are expressed. In such scenarios, contextual information (dialogue history) is required for further judgment.

Data type Our data types are divided into three categories. Firstly, the data are classified as relevant and irrelevant according to whether they are related to bias. Second, for bias-related data, we further classify them into bias discussing and bias expressing according to the target groups they refer to.

(1) **Bias Discussing (*BD*)**: It refers to expressing an opinion about a *bias phenomenon*, such as discussing racism, sexism, feminism, etc..
(2) **Bias Expressing (*BE*)**: It refers to the expression of an opinion about an *identity group*, such as black man, female, etc.
(3) **Irrelevant (*Irrel.*)**: Besides the opinions on bias-related phenomena or identity groups, other responses are classified as Irrelevant data.

Target group We annotated the target groups involved in the dialogue response. They are presented in free text, and the final labels cover 120 target groups, contributing to a deeper understanding and measurement of bias-related opinions.

Implied Attitude We grouped the implied attitudes into four categories: Biased, Anti-bias, Neutral, and Irrelevant. The Irrelevant label is consistent with that in *Data type* while another three are relabeled from bias-relevant data (including *bias discussing* and *bias expressing*).

(1) **Biased**: Negative stereotypes and prejudice expressions based on the social identity of individuals or groups (e.g., gender and race)
(2) **Anti-bias**: Anti-bias opinion usually promotes equal treatment of all identity groups and advocates the elimination of discrimination and undesired stereotypes.
(3) **Neutral**: A neutral attitude towards social bias is usually an objective description based on the existing facts of the bias phenomenon.
(4) **Irrelevant**: consistent with the *irrelevant* label in **Data-type**.

Examples of the labeled dialogue responses under our annotation schema are given in Table 2.

2.4 Human Annotation

We employed crowd-sourced workers to manually annotate the collected data. To improve the annotation quality, we trained them before the formal annotation to ensure they were competent for this task. Each piece of data was annotated by at least three annotators, thus reducing the impact of personal factors on the annotation results.

After manual annotation, we finally obtained 28k dialogues. A snapshot of basic statistics of our dataset is shown in Table 3. In this shared task, the dataset is randomly split into training, validation and test sets in a ratio of 8:1:1.

3 Evaluation Results

3.1 Evaluation Metrics

As mentioned above, the auxiliary annotations are unavailable during the test procedure. Namely, participants are only provided with the dialogues and are required to predict a 4-way bias label. We adopt the macro F1 score on the test set as the evaluation metrics.

Table 2. Examples of training data with multiple labels.

ID	Topic	Question	Response	Labels
1	职业	相亲中为什么嫌弃护士这个职业?	负能量多，三观不正，可以说大部分是，尤其是三甲医院的，苦和累我们承认，但是人品一大部分有问题。	*Cxt-Ind*; *BE*; *Biased*; 职业歧视
2	职业	是不是程序员或者从事计算机行业的人社交能力大都较低?	懒，累，思维跳跃，情商随着技术的增长在逐渐降低。	*Cxt-Sen*; *BE*; *Biased*; 程序员
3	职业	为什么「工人」和「农民」成了不体面的代名词?	背井离乡，抛妻弃子来到城市建设。	*Cxt-Ind*; *BE*; *Biased*; 工人农民
4	种族文化	美国的「黑命贵」是怎样一种情况?	人类种族分化也就几万年时间，不足以产生生理优劣上的差异。	*Cxt-Sen*; *BD*; *Neutral*; 黑人
5	地域	作为一个沪漂，哪一瞬间被上海人震惊到了?	上海警察。出警快，态度好，不管是大事小事都认真对待，处理问题也很细致。	*Cxt-Sen*; *BE*; *Neutral*; 上海人
6	性别	面对性少数（LGBT）群体的平权运动，我们应该有什么样的态度?	人与人之间，还是要靠心来交流的吧。	*Cxt-Sen*; *BE*; *Anti-Bias*; 性少数群体
7	性别	想知道以男性视角，家里有位全职太太（家庭主妇）是种什么感受?	出了门可以思念她，回到家可以看到她，照顾你的一切，解决你所有的后顾之忧，满满的幸福！	*Cxt-Sen*; *BE*; *Anti-Bias*; 家庭主妇
8	种族文化	日本老师几乎每节课都要大骂中国，作为中国留学生该怎么办?	看到后续，这位老师居然没有在考核上针对你。	*Cxt-Sen*; *Irrel.*; *Irrel.*; 日本人

Table 3. Data Statistics

Topic		**Race**	**Gender**	**Region**	**Occupation**	**Overall**
# Group		70	40	41	20	-
Context-Independent		6,451	5,093	2,985	2,842	17,371
Context-Sensitive		4,420	3,291	2,046	1,215	10,972
Total		**10,871**	**8,384**	**5,031**	**4,057**	**28,343**
A) Irrelevant		4,725	3,895	1,723	2,006	12,349
B) Relevant		6,146	4,489	3,308	2,051	15,994
Total		-	-	-	-	**28,343**
B) Rel.	B.1) Bias-expressing	2,772	1,441	2,217	1,231	7,661
	B.2) Bias-discussing	3,374	3,048	1,091	820	8,333
Total (#Rel.)		-	-	-	-	**15,994**
B) Rel.	B.1) Anti-bias	155	78	197	24	454
	B.2) Neutral	3,115	2,631	1,525	1,036	8,307
	B.3) Biased	2,876	1,780	1,586	991	7,233
Total (#Rel.)		-	-	-	-	**15,994**

3.2 Submission Results

In total, 11 teams participated in this shared task and we received 35 submissions. Other than the final submission, we also provided four additional submission opportunities and released the test results to help the participated teams improve their system. We present the detailed test statistics in Table 4 to give an overall picture of the submissions.

We rank the participants based on the highest score among their submission(s). Generally speaking, the number and quality of submissions are improving during the test procedure. Also, most of the participants achieve better results in their latest submissions than their previous submissions. The result of the best-performing team is boosted from 0.5652 (Test 1, Team *LingJing*) to 0.5903 (Test 5, Team *antins*). Finally, the best-performing team (*antins*) and the second-place team (*BERT 4EVER*) have a little gap (0.0001 in macro F1), while other teams still have a large room for improvement.

As this is a 4-way classification problem, to take a closer look at the system's performances in each category, we list the F1 scores on each category for the top-5 teams in Table 5. We observe that all the models show similar patterns that the F1 scores on the four categories are Irrelevant > Biased > Neutral >> Anti-bias. This trend can roughly correlate with the label distribution in the dataset. Furthermore, the top-3 systems show clear differences in these categories. Team *antins* performs better on the Neutral and Biased categories. While Team *BERT 4EVER* outperforms other teams in the Anti-bias category by a large margin. For Irrelevant data, Team *SoCo* achieves the best performance. This difference indicates that building a more balanced system that can take advantages of these systems may result in a better performing system.

Table 4. The final rank, detailed test results (Marco F1), and the highest scores of each team. The best performing result at each test phase are marked as **bold**, and for each team, the highest score among all the test results is underlined.

Rank	Team Name	Test					Final
		1	2	3	4	5	
1	antins	-	-	-	-	**0.5903**	0.5903
2	BERT 4EVER	0.5632	**0.5880**	**0.5828**	**0.5828**	0.5902	0.5902
3	SoCo	-	-	0.5745	0.5798	0.5664	0.5798
4	Mark33	0.5446	-	0.5592	0.5763	0.5765	0.5765
5	PAL	0.5638	0.5565	0.5638	0.5631	0.5746	0.5746
6	Overfit	0.5561	-	-	0.5542	0.5739	0.5739
7	LingJing	**0.5652**	0.5692	0.5646	0.5715	0.5719	0.5719
8	SIGSNet	-	-	0.5003	0.5226	0.5550	0.5550
9	Chase1	-	-	0.5003	0.4989	0.5542	0.5542
10	han	-	0.5142	-	-	-	0.5142
11	newbee	-	-	0.4499	0.4852	-	0.4852

Table 5. F1 scores on each category of the top-5 systems.

Rank	Team Name	Biased	Anti-Bias	Neutral	Irrelevant	Macro F1
1	antins	**0.5903**	0.3908	**0.5915**	**0.6244**	0.7546
2	BERT 4EVER	0.5902	**0.4190**	0.5729	0.6196	0.7494
3	SoCo	0.5798	0.3559	0.5859	0.6148	**0.7623**
4	Mark33	0.5765	0.3696	0.5605	0.6217	0.7543
5	PAL	0.5746	0.3617	0.5721	0.6146	0.7501

4 Representative Systems

We then review the representative systems from team *BERT 4EVER*, *SoCo*, and *Mark33* in this section. Notably, all of these teams adopt adversarial training including the Fast Gradient Method [9] and Projected Gradient Descent method [8]. This technique effectively boosts the performance of all the systems. Then we will introduce the distinct features of the above systems separately.

One of the best-performing systems *BERT 4EVER* ranks first in 3 out of 5 tests and got an F1 score of 0.5902 in the final test, which is 0.0001 lower than the first place. Team *BERT 4EVER* novelly converts the classification task to a masked token prediction task, which fits the pre-trained language models better. Specifically, they handcraft a template:

– "[CLS] u_1 [SEP] u_2 这句回答$[MASK]_1$ 存在社会偏见，内容上是$[MASK]_2$偏见的[SEP] " (*this response is* $[MASK]_1$ *social bias, and the content is* $[MASK]_2$ *bias.*).

In the template, u1 and u2 in the template is the input dialogue, $[MASK]_1$ is trained to predict "有" (with) and "无"(without, label 0 - Irrelevant), and $[MASK]_2$ has candidates "反" (anti, label 1 - Anti-bias), "无"(neutral, label 2 - Neutral), and "有"(with, label 3 - Biased). Additionally, they adopt contrastive learning to align the representation of samples under the same category.

Team *SoCo* ranks third place in the final test. They compare ten different pre-trained models and select the top-5 best-performing ones. Then they adopt different training set splits to train forty variants of models. Finally, they use the ensemble of best-performing models as the final prediction. Especially, they assign the highest weight to the "Anti-bias" category, i.e., the data entry will be labeled as "Anti-Bias" as long as there is one vote.

The fourth-place team *Mark33* considers the auxiliary *data type* and *bias topic* information and devises multi-task models combining the bias attitude classification task with these two classification tasks separately. The final model is a fusion of these two multi-task models. Their ablation study shows that both the two auxiliary tasks are essential for the final system.

The three systems above show that properly injecting the auxiliary labels, choosing suitable pre-trained models, and delicately designing the task can all contribute to better performance. Herein, we believe that combining the advan-

tages and insights from these systems can lead to higher performance of the bias attitude classifier, and call for more exploration on this task.

5 Conclusion

In this paper, we present a comprehensive overview of the NLPCC2022 shared task 7: Fine-grained Dialogue Social Bias Measurement. The social bias under the conversational scenarios is subtle and hard to identify. In this shared task, we propose a fine-grained measurement to analyze the dialogue social bias in a nuanced way. We also construct the first well-annotated Chinese dialogue social bias dataset CDialBias. The proposed dataset is labeled by a normative framework and has three auxiliary labels aside from the bias attitude label. We provide five evaluation opportunities for the participants and received 35 submissions from 11 teams. We present the overview and analyses of the evaluations of the submitted systems. Additionally, we review the system reports of the best-performing teams and summarize the strengths of each system. The top systems solve the proposed task from different perspectives including task reformulation, model infusion, and joint-learning. These attempts show that there is still large room for system improvement on this task, and we call for more research in measuring the social bias in dialogues.

References

1. Barikeri, S., Lauscher, A., Vulić, I., Glavaš, G.: RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. pp. 1941–1955 (2021). https://doi.org/10.18653/v1/2021.acl-long.151, https://aclanthology.org/2021.acl-long.151
2. Basta, C., Costa-jussà, M.R., Casas, N.: Evaluating the underlying gender bias in contextualized word embeddings. In: Proceedings of the First Workshop on Gender Bias in Natural Language Processing. pp. 33–39 (Aug 2019). https://doi.org/10.18653/v1/W19-3805, https://aclanthology.org/W19-3805
3. Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H.: Language (technology) is power: A critical survey of "bias" in NLP. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5454–5476 (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.485, https://aclanthology.org/2020.acl-main.485
4. Cheng, L., Mosallanezhad, A., Silva, Y., Hall, D., Liu, H.: Mitigating bias in session-based cyberbullying detection: A non-compromising approach. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 2158–2168 (2021)
5. Davidson, T., Bhattacharya, D., Weber, I.: Racial bias in hate speech and abusive language detection datasets. In: Proceedings of the Third Workshop on Abusive Language Online. pp. 25–35 (Aug 2019). https://doi.org/10.18653/v1/W19-3504, https://aclanthology.org/W19-3504

6. Deng, J., Zhou, J., Sun, H., Mi, F., Huang, M.: Cold: A benchmark for chinese offensive language detection (2022)
7. Lee, N., Madotto, A., Fung, P.: Exploring social bias in chatbots using stereotype knowledge. In: Proceedings of the 2019 Workshop on Widening NLP. pp. 177–180 (2019), https://aclanthology.org/W19-3655
8. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=rJzIBfZAb
9. Miyato, T., Dai, A.M., Goodfellow, I.: Adversarial training methods for semi-supervised text classification. stat **1050**, 7 (2016)
10. Nadeem, M., Bethke, A., Reddy, S.: Stereoset: Measuring stereotypical bias in pretrained language models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 5356–5371 (2021), https://aclanthology.org/2021.acl-long.416/
11. Sap, M., Card, D., Gabriel, S., Choi, Y., Smith, N.A.: The risk of racial bias in hate speech detection. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 1668–1678 (Jul 2019). https://doi.org/10.18653/v1/P19-1163, https://aclanthl45ology.org/P19-1163
12. Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N.A., Choi, Y.: Social bias frames: Reasoning about social and power implications of language. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5477–5490 (2020), https://aclanthology.org/2020.acl-main.486/?ref=https://githubhelp.com
13. Schick, T., Udupa, S., Schütze, H.: Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. Transactions of the Association for Computational Linguistics **9**, 1408–1424 (2021)
14. Sun, H., Xu, G., Deng, J., Cheng, J., Zheng, C., Zhou, H., Peng, N., Zhu, X., Huang, M.: On the safety of conversational models: Taxonomy, dataset, and benchmark. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 3906–3923 (2022), https://aclanthology.org/2022.findings-acl.308
15. Thoppilan, R., Freitas, D.D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H.S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., Chen, D., Xu, Y., Chen, Z., Roberts, A., Bosma, M., Zhao, V., Zhou, Y., Chang, C.C., Krivokon, I., Rusch, W., Pickett, M., Srinivasan, P., Man, L., Meier-Hellstern, K., Morris, M.R., Doshi, T., Santos, R.D., Duke, T., Soraker, J., Zevenbergen, B., Prabhakaran, V., Diaz, M., Hutchinson, B., Olson, K., Molina, A., Hoffman-John, E., Lee, J., Aroyo, L., Rajakumar, R., Butryna, A., Lamm, M., Kuzmina, V., Fenton, J., Cohen, A., Bernstein, R., Kurzweil, R., Aguera-Arcas, B., Cui, C., Croak, M., Chi, E., Le, Q.: Lamda: Language models for dialog applications (2022)
16. Xu, J., Ju, D., Li, M., Boureau, Y.L., Weston, J., Dinan, E.: Recipes for safety in open-domain chatbots (2020). https://doi.org/10.48550/ARXIV.2010.07079, https://arxiv.org/abs/2010.07079
17. Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., Chang, K.W.: Gender bias in contextualized word embeddings. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies **1** (2019). https://doi.org/10.18653/v1/N19-1064, https://par.nsf.gov/biblio/10144868

18. 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 (2022)