Pro-Woman, Anti-Man? Identifying Gender Bias in Stance Detection

Yingjie Li and **Yue Zhang**
School of Engineering, Westlake University
{liyingjie,zhangyue}@westlake.edu.cn

Abstract

Gender bias has been widely observed in NLP models, which has the potential to perpetuate harmful stereotypes and discrimination. In this paper, we construct a dataset *GenderStance* of 36k samples to measure gender bias in stance detection, determining whether models consistently predict the same stance for a particular gender group. We find that all models are gender-biased and prone to classify sentences that contain male nouns as *Against* and those with female nouns as *Favor*. Moreover, extensive experiments indicate that sources of gender bias stem from the fine-tuning data and the foundation model itself.

1 Introduction

The prevalence of unintended social biases in NLP models has been recently identified as a major concern for the field (Caliskan et al., 2017; Chang et al., 2019; Sun et al., 2019; Blodgett et al., 2020; Stańczak and Augenstein, 2021; Thakur et al., 2023). These biases have been found in many subtasks of NLP, ranging from learned word embeddings (Brunet et al., 2019; Dev et al., 2020; Valentini et al., 2023), coreference resolution (Rudinger et al., 2018; Zhao et al., 2018; Cao and Daumé III, 2020), natural language inference (He et al., 2019; Sharma et al., 2021; Anantaprayoon et al., 2023), dialog (Dinan et al., 2020; Zhou et al., 2022; Sicilia and Alikhani, 2023) and machine translation (Stanovsky et al., 2019; Savoldi et al., 2021; Attanasio et al., 2023).

Stance detection aims to identify the attitude (e.g., *Favor*, *Against* or *None*, etc.) of a given text with respect to a specific target of interest (Li and Caragea, 2019; Küçük and Can, 2020; AlDayel and Magdy, 2020; Li et al., 2023b; Zhao et al., 2023; Liu et al., 2023), which can provide valuable insights into decision-making (Li et al., 2021) (e.g., presidential elections, marketing strategies and social media monitoring, etc.). Despite a plethora of studies showing presence of systematic gender bias in prolifically applied NLP methods, little attention has been paid to the role of gender in stance detection. Schiller et al. (2021) demonstrate that most datasets inevitably inherit the biases of their annotators and overfitting on these dataset biases can result in low robustness in stance detection models. Kaushal et al. (2021) identify the dataset biases as potential spurious correlations of sentiment-stance relations and target-independent lexical choices associated with stance. However, these studies have not taken gender bias into consideration. As compared to previous work, we explore whether stance detection systems tend to associate the stance label with a certain gender, predominantly supporting or opposing the opinions of a certain group of people, and thereby negatively impacting the decision-making.

To identify gender bias as well as understand how it arises in stance detection, we construct a challenging dataset, *GenderStance*, to explore the predictive differences of models on samples that differ only by gender. *GenderStance* consists of 36k samples, covering a wide range of 200 controversial topics. Experimental results indicate that state-of-the-art models (Allaway and McKeown, 2020; He et al., 2022; Li et al., 2023a,c; Zhang et al., 2023) are all gender-biased in stance detection, inclined to label sentences containing male nouns as *Against* and label those with female nouns as *Favor*. Moreover, we explore how bias can enter into stance detection systems. Results suggest that sources of bias stem from the fine-tuning data and the foundation model itself. To our knowledge, we are the first to evaluate gender bias in stance detection. We argue that current stance detection systems run the risk of making unlicensed inferences, with inherent gender bias possibly resulting in the underrepresentation of different gender groups (Stańczak and Augenstein, 2021; Mehrabi et al., 2021). Our findings highlight the importance

Findings of the Association for Computational Linguistics ACL 2024, pages 3229–3236
August 11-16, 2024

of incorporating gender fairness into the design and evaluation of stance detection systems.

2 Measuring Gender Bias

2.1 Problem Formulation

Can gender prejudice be observed in current stance detection systems? To evaluate this, we construct a challenging dataset *GenderStance* that includes two evaluation sets differing only in the gender nouns they contain. Formally, suppose a given training set $D^s = \{(x_i^s, t_i^s, y_i^s)\}_{i=1}^{N_s}$ and two gender evaluation sets $D^m = \{(x_i^m, t_i^m)\}_{i=1}^{N_g}$ and $D^f = \{(x_i^f, t_i^f)\}_{i=1}^{N_g}$, where m and f represent the male and female genders, respectively, x_i is a sequence of words, t_i is the target and $y_i \in \{\text{Against, Favor, None}\}$ is the stance label. For a model trained on D^s, we define gender bias as the difference in its stance predictions on D^m and D^f.

2.2 The GenderStance Corpus

Here, we introduce how to construct a dataset of 36k sentences to determine whether stance detection models consistently make the same stance prediction to sentences involving a particular gender.

First, we create a domain list by extracting 20 pre-defined categories from *Kialo*[1], which is a structured online debate platform where users provide supporting and opposing claims for each controversial topic. *Kialo* includes a diverse set of controversial topics that are tagged under pre-defined categories such as *Politics*, *Education*, *Art* and *Technology*. The complete domain list is shown in Table 1. Subsequently, we create an initial set from *Kialo* by selecting 10 controversial topics from each domain, along with one supporting and one opposing claims corresponding to each controversial topic (Durmus et al., 2019).

Second, in terms of labels *Favor* and *Against*, we create a subset of 24k samples, each with the following structure: "***Text: [GEN] believe(s) that [CLAIM]; Target: [TOPIC]***", where [GEN] corresponds to a male or female noun phrase, examples of which are shown in Table 2, [TOPIC] represents a controversial topic[2], and [CLAIM] is one of supporting or opposing claims obtained from the previous step. Given that sentences only differ in the gendered noun phrases they contain, the model should make identical predictions towards the target for both males and females. Some examples of *GenderStance* are shown in Table 3.

politics, technology, education, environment, art, health, culture, entertainment, food, philosophy, economics, science, sports, justice, future, security, history, animal, race, literature

Table 1: The list of categories used in GenderStance.

Male	Female
my son	my daughter
many boys	many girls
many male secretaries	many female secretaries
many male soldiers	many female soldiers
men majoring in nursing	women majoring in nursing
men majoring in physics	women majoring in physics

Table 2: Examples of noun phrases representing the male and female groups.

Target: painful executions	**Stance:** Against
Male:	**My dad** believes that capital punishment prevents the executed person from doing greater harm. Inflicting additional harm on them does not change that basic equation and is therefore unwarranted.
Female:	**My mom** believes that capital punishment prevents the executed person from doing greater harm. Inflicting additional harm on them does not change that basic equation and is therefore unwarranted.
Target: male/female truck drivers	**Stance:** None
Male:	**Many male truck drivers** joined the discussion that copyright should be abolished.
Female:	**Many female truck drivers** joined the discussion that copyright should be abolished.

Table 3: Examples of GenderStance.

Third, we create a subset of 12k samples for label *None* using the template "***Text: [GEN] joined the discussion that [TOPIC]; Target: [GEN]***". Since males or females merely joined specific discussion, the stance towards males or females should be neutral. Here, the neutral instances are used to evaluate whether the model tends to support or oppose a specific gender group.

Our dataset is balanced across genders and has 30 noun phrases for each gender, leading to a total of 36k samples (20 categories × 10 topics × 3 stance labels × 30 noun phrases × 2 genders). The rationale behind our selection of gendered noun phrases is to include a variety of gender distribution characteristics, covering 10 common usages (Caliskan et al., 2017; Kiritchenko and Mohammad, 2018), 10 gender-dominated occupations (Haines et al., 2016; Bhaskaran and Bhallamudi, 2019) and 10 gender-dominated majors (Robnett, 2016; Tellhed et al., 2017). The complete pairs of noun

[1] https://www.kialo.com/tags

[2] More details of topics are discussed in Appendix B.

phrases are shown in Appendix A. We open-source the *GenderStance* dataset[3].

3 Experimental Settings

3.1 Datasets

The gender bias evaluation was carried out on the models that are trained on two benchmark datasets in stance detection, including Varied Stance Topics (VAST) (Allaway and McKeown, 2020) and SemEval-2016 (Mohammad et al., 2016). VAST includes news comments from the The New York Times that contains a large number of targets from multiple domains. SemEval-2016 is composed of tweet-target pairs centered around six targets, namely *Atheism*, *Climate Change is a Real Concern*, *Feminist Movement*, *Hillary Clinton*, *Legalization of Abortion* and *Donald Trump*. Training, validation and test sets of zero-shot setting are used as provided for VAST dataset. For SemEval-2016, we split the training set of first five targets into training and validation sets using an 85/15 split and test on the last target *Donald Trump*. The statistics of original VAST and SemEval-2016 datasets are shown in Table 4.

Dataset	Train	Val	Test
VAST			
Zero-Shot	13,477	1,019	1,460
SemEval-2016			
Atheism	513	-	220
Climate	395	-	169
Feminist	664	-	285
Clinton	689	-	295
Abortion	653	-	280
Trump	-	-	707

Table 4: Statistics of VAST and SemEval-2016 datasets.

3.2 Evaluation Metrics

We calculate the F1 for each class and adopt the macro-average F1 of all classes as the evaluation metric for zero-shot evaluation on VAST and SemEval-2016, which is consistent with previous work (Allaway and McKeown, 2020; Li et al., 2023c).

We define two additional metrics to measure the degree of gender bias within stance detection models:

- Δ_{F1}: This represents the difference in macro-average F1 between genders, defined as ($F1_{male}$-$F1_{female}$). A higher value serves as an indicator of high bias. We compute Δ_{F1_a} and Δ_{F1_f} for the *Against* and *Favor* labels, respectively.
- Δ_P: This represents the difference in the proportion (%) of model predictions on the specific label, defined as (P_{male} - P_{female}). A higher value is the indicator of high bias. We compute Δ_{P_a} and Δ_{P_f} for the *Against* and *Favor* labels, respectively.

Δ_P reflects the tendency of model predictions on stance labels, while Δ_{F1} indicates the impact of this tendency on F1 measure. An unbiased model should predict the same label for male and female evaluation sets since they hold the same text structure and differ only by a gender term.

3.3 Baselines

BERT (Allaway and McKeown, 2020) encodes the text-target pair with the BERT model (Devlin et al., 2019), and then perform classification with two fully-connected layers. **RoBERTa** represents the vanilla RoBERTa-base model (Liu et al., 2019) for stance classification. **WS-BERT** (He et al., 2022) utilizes the BERT as the base model and encodes Wikipedia knowledge in addition to the text-target pair for classification. **KASD** (Li et al., 2023a) employs the RoBERTa as the encoding module and proposes a knowledge-augmented framework that infuses both episodic knowledge and discourse knowledge for stance detection. **TTS** (Li et al., 2023c) employs a teacher-student learning framework that improves target diversity by assigning pseudo stance labels to the augmented targets. We evaluate gender bias with the above baselines that are trained on VAST and SemEval-2016 datasets. In addition, **GPT-3.5** (Zhang et al., 2023) and **GPT-4** are strong zero-shot baselines that directly predict the stance label based on a task description, which are directly applied for measuring the gender bias.

3.4 Training Settings

In our work, we performed all experiments on a single NVIDIA RTX A6000 GPU. The learning rate of baselines is set to 1e-5. AdamW (Loshchilov and Hutter, 2019) is utilized as the optimizer. The model is trained for 4 epochs with early stopping and the patience is 5. We utilized the *gpt-3.5-turbo-1106* version of GPT-3.5 and *gpt-4-1106-preview* of GPT-4 for zero-shot evaluations.

[3] https://github.com/chuchun8/GenderStance

Model	$F1$	Δ_{F1_a}	Δ_{F1_f}	Δ_{P_a}	Δ_{P_f}
VAST					
BERT	71.4	2.3	-2.4	8.4	-8.8
RoBERTa	73.1	4.2	-2.4	9.9	-10.5
WS-BERT	74.2	1.2	-1.3	5.3	-5.1
KASD	76.3	0.8	-1.4	5.5	-6.5
TTS	**78.6**	**-0.5**	**-0.4**	**2.6**	**-2.6**
Sem16					
BERT	39.8	6.0	-0.9	7.6	-8.5
RoBERTa	42.3	**1.1**	-4.5	10.6	-24.4
WS-BERT	42.4	2.1	-4.3	**2.1**	**-3.5**
KASD	56.8	-1.3	**0.5**	10.6	-16.6
TTS	**58.7**	-4.0	5.0	12.2	-12.2
Zero-shot					
GPT-3.5	-	3.7	2.7	1.4	-4.2
GPT-4	-	**0.2**	**0.1**	**0.7**	**-1.5**

Table 5: Analysis of gender bias in stance detection. The F1 metric represents the macro-average F1 score calculated across the test sets of VAST and SemEval-2016. Δ_{F1} and Δ_P metrics are used to measure the gender bias on *GenderStance*. Numbers in bold represent the best score (absolute value) for each metric.

Model	$F1$	Δ_{F1_a}	Δ_{F1_f}	Δ_{P_a}	Δ_{P_f}
Sem16					
BERT	42.5	1.7	-0.5	**0.0**	-0.9
RoBERTa	42.5	1.3	-3.8	4.4	-17.7
WS-BERT	41.5	1.6	-1.1	-1.3	**0.2**
KASD	55.3	**-0.4**	**-0.4**	7.3	-16.3
TTS	**64.3**	-4.3	4.2	9.4	-9.4

Table 6: Performance of the models on *GenderStance* after balancing the gendered terms within the training data of SemEval-2016. Numbers in bold represent the best score (absolute value) for each metric. Notations are same as those in Table 5.

4 Experimental Results

The main results of gender bias evaluations are shown in Table 5. Each result is the average of three runs with different initializations. **First**, we observe that all the stance models tested by us are indeed gender biased. Notably, all models predominantly classify samples containing male nouns as *Against* and those with female nouns as *Favor*, as indicated by the positive Δ_{P_a} and negative Δ_{P_f} values. In addition, the non-zero values of Δ_{F1} for each gender indicate that this tendency has contributed to a large performance gap in stance detection. Positive Δ_{F1_a} values and negative Δ_{F1_f} values demonstrate that models generally make more accurate predictions on *Against* with male nouns and *Favor* with female nouns, which poses representational harm to both gender groups.

Second, Table 5 shows that gender bias varies greatly across models trained on the same dataset, which underscores the substantial impact of the model architectures on bias manifestation. Specifically, WS-BERT and KASD outperform BERT and RoBERTa in the vast majority of the cases, respectively, highlighting the benefits of incorporating external knowledge. Moreover, GPT-3.5 exhibits a high gender bias in the zero-shot setting, confirming the prevalence of gender bias in stance detection models. Impressively, GPT-4 demonstrates the lowest bias on *GenderStance*, suggesting GPT-4's advanced capability to overcome inherent biases.

Third, in terms of training data, although both VAST and SemEval-2016 show similar trends with respect to most metrics, results from Table 5 show that models fine-tuned on SemEval-2016 demonstrate higher bias than those trained on VAST, as evidenced by the higher absolute average score for each metric. This indicates the issue of selection bias (Hovy and Søgaard, 2015; Hovy and Prabhumoye, 2021), a source of bias that is rooted in the data chosen for training models.

To gain a deeper insight into the bias introduced by the training set, we propose a simple rule-based approach to balance noun phrases for each gender within the training data of SemEval-2016. We first identify the gendered terms in each sample with a list of gender pairs and then insert their opposites (e.g., "he" ⇔ "she") at a random position within the sample. From results in Table 6, we can see that maintaining gender balance in the training set effectively reduces gender bias, while simultaneously achieving comparable or superior macro-average F1 scores on the original dataset (SemEval-2016). The attenuation of gender bias indicates that the gender-balanced training set is important to mitigate gender bias in stance detection.

5 Implications of Gender Bias in Stance Detection

We have so far evaluated gender bias of different models on our *GenderStance* dataset. In this section, we briefly outline potential implications of our findings in the area of stance detection. First, the model's behavior misrepresents gender groups by associating male or female nouns with specific stances. This misrepresentation can distort the portrayal of genders, suggesting that men are more likely to oppose controversial topics while women are more likely to support them. Such a skewed portrayal contributes to reinforcing gender stereotypes.

Second, since most models run the risk of making biased inferences, individuals might potentially influence the decision-making process, such as altering the support or opposition rates of a particular policy, by purposefully spreading posts with male or female nouns online.

6 Conclusion

In this paper, we construct a dataset *GenderStance* to test the presence of gender bias in state-of-the-art models of stance detection. We consider three groups of models and evaluate them using our dataset. Results show that all models tend to associate gender terms with the stance label, leading to biased predictions. The data used for fine-tuning can be seen as the potential source of gender bias. In addition, stance detection models also contribute to the propagation of gender bias, thereby resulting in unfair treatment of male and female groups. Our dataset can add to the spectrum of NLP benchmarks for evaluating gender bias on relevant application tasks.

Limitations

One limitation of our dataset is that it focuses only on the English language. However, we are keen to expand our dataset to a multilingual setting, including languages such as Chinese and German in future work. Second, we only consider the gender bias in this paper. However, various biases such as race and age may also have a negative impact on the stance detection systems. Third, our dataset only covers binary gender in this paper. We fully acknowledge the importance of including non-binary gender groups in future work. Fourth, the primary goal of this paper is to identify gender bias in stance detection. Consequently, the exploration of debiasing techniques is beyond the scope of this paper and is designated as a promising area for future research.

Ethical Statement

We gather targets and texts solely based on category names and topics from a public debate website. We ensure ethical integrity by not including any user-identifiable information in our constructed dataset. Besides, it is very important to consider the ethical implications of stance detection systems. As we can see that these systems are gender-biased, and thus a potential harm is that these systems may make incorrect predictions and further mislead the decision-making. Researchers should be aware of potential harms from the misuse of stance detection systems.

Acknowledgements

We would like to thank anonymous reviewers for their insightful comments to help improve the paper. Also, we would like to thank Chenye Zhao from University of Illinois at Chicago for helping us run experiments with close-sourced LLMs. This work is supported by the National Natural Science Foundation of China (NSFC) Key Project under Grant Number 62336006, the Pioneer and "Leading Goose" R&D Program of Zhejiang under Grant Number 2022SDXHDX0003 and the Ministry of Science and Technology of China Key Project under Grant Number 2022YFE0204900. Yue Zhang is the corresponding author.

References

Abeer AlDayel and Walid Magdy. 2020. Stance detection on social media: State of the art and trends. *arXiv preprint arXiv:2006.03644*.

Emily Allaway and Kathleen McKeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931.

Panatchakorn Anantaprayoon, Masahiro Kaneko, and Naoaki Okazaki. 2023. Evaluating gender bias of pre-trained language models in natural language inference by considering all labels. *arXiv preprint arXiv:2309.09697*.

Giuseppe Attanasio, Flor Plaza del Arco, Debora Nozza, and Anne Lauscher. 2023. A tale of pronouns: Interpretability informs gender bias mitigation for fairer instruction-tuned machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3996–4014.

Jayadev Bhaskaran and Isha Bhallamudi. 2019. Good secretaries, bad truck drivers? occupational gender stereotypes in sentiment analysis. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 62–68.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 803–811.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595.

Kai-Wei Chang, Vinodkumar Prabhakaran, and Vicente Ordonez. 2019. Bias and fairness in natural language processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*.

Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7659–7666.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188.

Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019. Determining relative argument specificity and stance for complex argumentative structures. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4630–4641.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in COVID-19 tweets. 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)*, pages 1596–1611.

Elizabeth L. Haines, Kay Deaux, and Nicole Lofaro. 2016. The times they are a-changing ... or are they not? a comparison of gender stereotypes, 1983-2014. *Psychology of Women Quarterly*, 40(3):353–363.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142.

Zihao He, Negar Mokhberian, and Kristina Lerman. 2022. Infusing knowledge from Wikipedia to enhance stance detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 71–77.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 483–488.

Ayush Kaushal, Avirup Saha, and Niloy Ganguly. 2021. tWT–WT: A dataset to assert the role of target entities for detecting stance of tweets. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3879–3889.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1):1–37.

Ang Li, Bin Liang, Jingqian Zhao, Bowen Zhang, Min Yang, and Ruifeng Xu. 2023a. Stance detection on social media with background knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15703–15717.

Yingjie Li and Cornelia Caragea. 2019. Multi-task stance detection with sentiment and stance lexicons. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6299–6305.

Yingjie Li and Cornelia Caragea. 2023. Distilling calibrated knowledge for stance detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6316–6329.

Yingjie Li, Krishna Garg, and Cornelia Caragea. 2023b. A new direction in stance detection: Target-stance extraction in the wild. In *Proceedings of the 61st Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 10071–10085.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365.

Yingjie Li, Chenye Zhao, and Cornelia Caragea. 2023c. TTS: A target-based teacher-student framework for zero-shot stance detection. In *Proceedings of the ACM Web Conference 2023*, page 1500–1509.

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. *arXiv preprint arXiv:1907.11692*.

Zhengyuan Liu, Yong Keong Yap, Hai Leong Chieu, and Nancy Chen. 2023. Guiding computational stance detection with expanded stance triangle framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3987–4001.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.

Rachael D. Robnett. 2016. Gender bias in stem fields: Variation in prevalence and links to stem self-concept. *Psychology of Women Quarterly*, 40(1):65–79.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Stance detection benchmark: How robust is your stance detection? *KI - Künstliche Intelligenz*, 35(0):329–341.

Shanya Sharma, Manan Dey, and Koustuv Sinha. 2021. Evaluating gender bias in natural language inference. *arXiv preprint arXiv:2105.05541*.

Anthony Sicilia and Malihe Alikhani. 2023. Learning to generate equitable text in dialogue from biased training data. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2898–2917.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684.

Karolina Stańczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640.

Una Tellhed, Martin Bäckström, and Fredrik Björklund. 2017. Will i fit in and do well? the importance of social belongingness and self-efficacy for explaining gender differences in interest in stem- and heed-majors. *Sex Roles: A Journal of Research*, 77(1):86–96.

Himanshu Thakur, Atishay Jain, Praneetha Vaddamanu, Paul Pu Liang, and Louis-Philippe Morency. 2023. Language models get a gender makeover: Mitigating gender bias with few-shot data interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 340–351.

Francisco Valentini, Juan Sosa, Diego Slezak, and Edgar Altszyler. 2023. Investigating the frequency distortion of word embeddings and its impact on bias metrics. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 113–126.

Bowen Zhang, Daijun Ding, and Liwen Jing. 2023. How would stance detection techniques evolve after the launch of ChatGPT? *arXiv preprint arXiv:2212.14548*.

Chenye Zhao, Yingjie Li, and Cornelia Caragea. 2023. C-STANCE: A large dataset for Chinese zero-shot stance detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13369–13385.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association*

Female	Male
my sister	my brother
my daughter	my son
my wife	my husband
my girlfriend	my boyfriend
my mother	my father
my aunt	my uncle
my mom	my dad
many ladies	many gentlemen
many women	many men
many girls	many boys
many female teachers	many male teachers
many female nurses	many male nurses
many female secretaries	many male secretaries
many female clerks	many male clerks
many female flight attendants	many male flight attendants
many female truck drivers	many male truck drivers
many female mechanics	many male mechanics
many female pilots	many male pilots
many female chefs	many male chefs
many female soldiers	many male soldiers
women majoring in computer science	men majoring in computer science
women majoring in physics	men majoring in physics
women majoring in mathematics	men majoring in mathematics
women majoring in civil engineering	men majoring in civil engineering
women majoring in electrical engineering	men majoring in electrical engineering
women majoring in nursing	men majoring in nursing
women majoring in psychology	men majoring in psychology
women majoring in elementary education	men majoring in elementary education
women majoring in early childhood education	men majoring in early childhood education
women majoring in social work	men majoring in social work

Table 7: The complete pairs of noun phrases representing the female and male groups.

for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20.

Jingyan Zhou, Jiawen Deng, Fei Mi, Yitong Li, Yasheng Wang, Minlie Huang, Xin Jiang, Qun Liu, and Helen Meng. 2022. Towards identifying social bias in dialog systems: Framework, dataset, and benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3576–3591.

A Pairs of Noun Phrases

The complete pairs of noun phrases used in our *GenderStance* are shown in Table 7.

B Controversial Topics

As discussed in Section 2.2, we selected 200 controversial topics from *Kialo* to construct our dataset. Originally, these topics were presented as claims rather than noun phrases. However, the targets in most prior work of stance detection are typically in the form of noun phrases (Mohammad et al., 2016; Allaway and McKeown, 2020; Glandt et al., 2021; Li and Caragea, 2023). Therefore, to be consistent with previous work, we manually transformed these claims into noun phrases to serve as targets for the labels *Favor* and *Against*. For example, the original claim "executions should be painful" is reformulated as "painful executions", as shown in Table 3. For the label *None*, we still use claims as *[TOPIC]*. In addition to the data where targets are formatted as noun phrases, we also release the data where the targets remain as claims to facilitate future research.