2022-12

Identity Term Sampling for Measuring Gender Bias in Training Data

Nasim Sobhani
Technological University Dublin, nasim.x.sobhani@mytudublin.ie

Sarah Jane Delany Prof
Technolgical University Dublin, sarahjane.delany@tudublin.ie

Recommended Citation

Sobhani, N., & Delany, S. J. (2022). Identity Term Sampling for Measuring Gender Bias in Training Data. Springer Nature. DOI: 10.21427/BKM6-RF06

Funder: Science Foundation Ireland

Identity Term Sampling for Measuring Gender Bias in Training Data

Nasim Sobhani* and Sarah Jane Delany

Technological University Dublin, Dublin, Ireland
nasim.x.sobhani@mytudublin.ie
sarahjane.delany@tudublin.ie

Abstract. Predictions from machine learning models can reflect biases in the data on which they are trained. Gender bias has been identified in natural language processing systems such as those used for recruitment. The development of approaches to mitigate gender bias in training data typically need to be able to isolate the effect of gender on the output to see the impact of gender. While it is possible to isolate and identify gender for some types of training data, e.g. CVs in recruitment, for most textual corpora there is no obvious gender label. This paper proposes a general approach to measure bias in textual training data for NLP prediction systems by providing a gender label identified from the textual content of the training data. The approach is compared with the identity term template approach currently in use, also known as Gender Bias Evaluation Datasets (GBETs), which involves the design of synthetic test datasets which isolate gender and are used to probe for gender bias in a dataset. We show that our Identity Term Sampling (ITS) approach is capable of identifying gender bias at least as well as identity term templates and can be used on training data that has no obvious gender label.

Keywords: Machine Learning · Gender Bias · Evaluation.

1 Introduction

Studies have shown gender bias in natural language processing tasks such as machine translation [18], co-reference resolution [23, 25, 17] and abusive and hate speech prediction [6, 14]. Gender bias has also been found in deployed NLP systems. In 2018 Amazon discontinued the use of an AI recruitment tool which showed significant bias against women[1]. These downstream tasks that use machine learning models built on natural language content can reflect biases in the data on which they are trained.

The primary method to measure bias in a downstream task is to measure performance differences across gender as the system's performance should not be

* Corresponding author

[1] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

influenced by gender. This requires a way to isolate gender in the test instances which are used to measure performance. This is typically done by using synthetic test data that is appropriate for the task at hand. This test data is designed through the use of templates which can be filled in with content relevant for the task and duplicated for different gender identities. As an example, in an abusive content prediction task in work by [14], the template sentence "You are a $<$ *adjective* $>$ $<$ *identity term* $>$" generated a number of test instances labelled for the classification task (abusive and non-abusive) and identified for gender. $<$ *adjective* $>$ was replaced with adjectives such as disgusting, filthy, nasty for abusive instances and adjectives such as lovely, excellent, incredible for non-abusive instances, while $<$ *identity* $>$ was replaced with common gender identity pairs such as man/woman, boy/girl. This generated gender-swapped labelled test instances that were used to measure the difference in performance across genders.

There are some challenges with these template approaches. The artificial nature of the generated text does not reflect the true distribution and content of the task data. The templates have to be designed specifically for the downstream task and are not general across tasks. In addition the actual performance of these generated test datasets on the downstream task has been shown to be poor.

As an alternative to synthetic test data this paper proposes an approach to a more confident measure of gender bias by selecting appropriate test data from the original datasets and identifying their gender to allow the measurement of task performance across genders. Our approach, which we call Identity Term Sampling (ITS), is compared with the identity term template approach on the task of abusive content detection. We also apply it to a text classification task where the data is not typically expected to have gender bias and we show no significant gender bias evident.

The rest of this paper is structured as follows, section 2 discusses related work in measuring gender bias in natural language tasks, section 3 explains our ITS approach, section 4 details the evaluation of our approach and the results and findings are discussed in section 5.

2 Related Work

Natural Language Processing (NLP) models and systems are trained on human generated text content and they can reflect existing biases in the data when used in downstream applications [6, 14]. In addition to the training data itself, word embeddings which are distributed representations that are generated from large corpora of natural language and are used to represent words and sentences, can reflect and sometimes even amplify certain characteristics of the data including gender stereotypes [2, 3, 26].

As a first step towards reducing bias in an NLP system, we need to identify and measure any bias that might exist. Over the last few years a lot of research has been conducted to identify and measure bias in the training data [6, 26, 11] and in embeddings that might be used to represent the training data [2,

24]. An effective technique for evaluating bias in training data, which is known as gender-swapping, involves replacing female/male definitional words by their equivalent male/female definitional words in the test set and comparing the overall performance of the system. The difference between the original test set and the gender-swapped results illustrates the system's fairness [11].

Another technique to evaluate gender bias is generating a synthetic test set with test instances that isolate gender. This approach is called Gender Bias Evaluation Testsets (GBETs) by [21], and has been used to evaluate bias in a variety of different NLP tasks including sentiment analysis [10], abusive language detection [6, 14] and coreference resolution [25, 17].

GBETs can be generated in different ways depending on the NLP task to be tackled. For instance, a GBET for coreference resolution named GAP [23] is a human labeled ambiguous pronoun-name pairs corpus mined from Wikipedia. Similarly, to analyse gender bias in coreference resolution [7] constructed a dataset which is also scraped from Wikipedia, OpenSubtitle and Reddit comments. The template approach described above is also used to generate GBETs and involves creating sentence templates, that include gender identification words, appropriate for the downstream task. Pairs of sentences are generated from the template, one for each gender, and differences in the performance of the NLP system between the generated test sentences with a male and female gender identity facilitate the measurement of gender bias in the dataset. This gender identity template approach has been used in variety of different NLP tasks including sentiment analysis [10], abusive language detection [6, 14] and coreference resolution [25, 17].

More recently StereoSet [12] and CrowS-Pairs [13] GBETS have been proposed to evaluate bias in language models. These GBETs are crowd-sourced, template based which are created and annotated by crowdsourcing to measure bias in different domains. Each example consists of a pair of stereotype and anti stereotype sentences in case of CrowS-pairs. However, StereoSet contains of triplets of sentences with each instance corresponding to a stereotypical, anti-stereotypical or a meaningless association. An additional study presents a large GBET dataset called HOLISTICBIAS for measuring bias. This dataset is assembled by using a set of demographic descriptor terms in a set of bias measurement templates and can be used to test bias in language models [19].

There are a variety of measures used to detect gender bias in NLP methods [20]. Most of the recent work on evaluating gender bias in NLP systems use variations on Hardt et al.'s work on equalised odds and equal opportunity [9]. These measures are group measures and use the gender distributions in the training data rather than the democratic parity measure which insists on equal outcomes for both genders regardless of prevalence or ground truth.

There has also been a lot of work in identifying gender bias in word embeddings which have become a common form of representation of textual content in NLP systems. The existence of gender stereotypes in pre-trained word embeddings has been shown by [24, 2] and in contextualized word embeddings including ELMO by [15, 1]. The Word Embedding Association Test (WEAT) [3] has also

been proposed to measure model bias inside word embeddings through the difference in the strength of association concepts.

3 Approach

As the extent of gender bias in a natural language system is evident by the task performance differences across genders, the test data used to measure performance needs to include gender. The first step in the proposed approach, which we call Identity Term Sampling (ITS), is to identify the gender of instances in the training dataset in order to identify appropriate test instances which can used to measure performance in the downstream task. Our approach then randomly selects the gendered test instances from the training data to be used to estimate the gender bias.

The gender identification step in Identity Term Sampling is based on the frequency of gender identity words in a data instance. ITS can assign gender to those instances that contain at least one gender identity term. The gender identity terms we use are those terms from a list of gender definitional pairs proposed in work by [2] and are given in Table 3b. These ten gender pairs were found by crowdsourcing to be the most frequent words used to define gender among a list of gender definitional and stereotype gender association words. For each instance in our datasets the frequency of male and female identity terms that occur in the text content is counted. The gender assigned to the data instance is the gender with the larger frequency of identity terms. Data instances with equal numbers of male and female gender identity terms are not identified with a gender as there was no obvious gender.

As an initial validation of the ITS approach we compared the gender identified by ITS against the actual gender on the BiasBios dataset [5], a dataset of 397,340 biographies across 28 different occupations. The ITS technique successfully identified 91.8% of the biographies correctly with only 4.1% misidentified and just over 4% were identified as no obvious gender.

To explore the gender identification approach we applied it to a number of datasets of user generated content which are used for text classification tasks. These datasets include two Twitter datasets used for the identification of abusive content and a review dataset used for sentiment analysis or opinion prediction. Twitter datasets used for abusive content detection are highly likely to exhibit bias and are used in other bias identification work [14, 4]. A hotel review dataset is less likely to exhibit gender bias in the training data.

The **Hate Speech** dataset [22] is a collection of almost 17K tweets consisting of 3,383 samples of sexist content, 1,972 samples of racist content and 11,559 neutral samples.

The dataset is transformed to a binary classification problem by labelling the sexist and racist samples as "abusive" class and neutral samples as "non abusive" class.

The **Abusive Tweets** dataset is a large scale crowd-sourced dataset, collected by [8]. The size of the dataset is just under 100k tweets and it is annotated

with four labels: *hateful*, *abusive*, *spam* and *none*. By combining the *none* and *spam* instances into a "non-abusive" class, and the hateful and abusive instances to an "abusive" class, we transform the dataset to a binary classification task, similar to the Hate Speech dataset.

The **Hotel Reviews** dataset has been scraped from booking.com and made available in Kaggle[2]. The dataset contains almost 515,0000 reviews and scores for 1493 luxury hotels across Europe. The classification task is to predict whether a textual review is a good or a bad review (i.e. a satisfied or unsatisfied customer). Each review in the dataset has a rating between 2.5 and 10 where higher is better in terms of satisfaction. The reviews were split into two classes: "unsatisfied" for reviews with a rating of less than 5, and "satisfied" for those with a rating of 5 or higher. The original dataset is highly imbalanced with 95% of the reviews in the "satisfied" class.

Table 1 shows the overall size and the per class and per gender distribution of data for the three datasets.

Table 1: Class distribution, gender identified data percentage and overall size for each dataset

Dataset	**Class**	**Class%**	**Gender identified** F(%)	M(%)	**Size**
Hate Speech	Abusive	31.4	3.6	1.6	16K
	Non Abusive	68.6	1.9	3.3	
Abusive Tweets	Abusive	32.1	2.0	2.9	100K
	Non Abusive	67.9	2.2	4.5	
Hotel Review	Unsatisfied	4.3	0.1	0.2	515K
	Satisfied	95.7	1.4	1.7	

To illustrate the effect of gender identification, each data instance is categorised into one of four groups. Data instances that do not have any of the gender identity words in them are categorised as No-Gender (NG). Data instances which contains equal numbers of male and female identity terms are categorised as Equal-Gender (EG). The other two categories are Positive Gender (PG) and Strongly Positive Gender (SPG) and use the proportion of male and female identity terms. The data instance is identified as the gender with the higher proportion of identity terms. If the proportion is between 50% and 75% the data instance is categorised as Positive Gender, and if it is 75% or higher, it is categorised as Strongly Positive Gender.

Table 2 describes the results of gender identification on the datasets, showing the size proportion of each dataset with gender identified and the proportion of the data of each category with gender identified. It is evident that most of the gendered data in all datasets is categorized as Strongly Positive indicating that typically over 75% of the definitional words in the gendered data are for one specific gender. As a result of applying the proposed method, 11% of Hate Speech data and almost 12% of the Abusive Tweets data are gender identified.

[2] https://www.kaggle.com/datasets/jiashenliu/515k-hotel-reviews-data-in-europe

The Hotel Reviews dataset has significantly less gendered instances with only 3.6% with gender identified.

Table 2: The results of identifying gender in the datasets, showing the size and proportion of each dataset with gender and the proportion of the gendered data of each category: EG equal gender, PG positive gender, SPG strongly positive gender.

				Percentage of gender identified data			
Dataset	**NG(%)**	**Gender Data**	**EG(%)**	**Female**		**Male**	
				PG(%)	SPG(%)	PG(%)	SPG(%)
Hate Speech	89.0	1758(11.0%)	4.6	0.7	49.9	0.6	44.3
Abusive Tweets	88.1	11914(11.9%)	3.1	0.8	34.5	1.0	60.6
Hotel Review	96.4	18771(3.6%)	3.8	1.1	41.6	1.3	52.2

4 Evaluation

The aim of the evaluation is to measure gender bias using our ITS approach for creating test instances identified with gender which are necessary for measuring the difference in task performance across genders. We compare this with using the synthetic test instances generated using the identity term template approach.

The evaluation uses the text classification tasks of abusive content detection on the Hate Speech and Abusive Tweets datasets described in section 3 above. We also include an evaluation on the Hotel Review dataset where the expectation of gender bias in the data is less. Due to the highly imbalanced class distribution in the Hotel Review dataset, a subset of the data was sampled. A equal distribution of both classes that did not have the gender identified was sampled from the dataset in the dataset and this was added to the test data to give a subset of just under 60,000 instances with a class distribution of 63%/37% for "satisfied"/"unsatisfied".

For classification a simpler version of the deep neural network model used by [5] is used. Our model consists of an embedding layer as an input layer using Word2Vec embedding, followed by a bidirectional long short term memory (BiLSTM) layer to encode the input sentence, a dropout layer and a linear output layer with cross entropy to compute the loss which comes after a dense layer with Relu activation. All hyper parameters were tuned on a 20% split of the training data.

Gender bias is measured using the test instances that have gender identified. As can be seen from Table 2 the number of examples with gender identified in the datasets is limited. As we need gendered examples in both training and test splits, we split the datasets 70%/30% into stratified training and holdout testing splits. The full holdout test split is used to measure overall task performance.

For our proposed ITS approach the test split is used to measure gender bias. We repeat this process five times and report average performance measures across the five holdout splits.

Table 3: The templates, identity term pairs and the nouns & attributes

(a)

Templates
You are a < *adjective* > < *identity term* > < *verb* > < *identity term* >
Being a < *identity term* > is < *adjective* >
I am a < *identity term* >
I hate < *identity term (plural form)* >

(b)

Identity Pairs
woman/man girl/boy
female/male she/he
daughter/son her/his
herself/himself gal/guy
mother/father Mary/John

(c)

Nouns & Adjectives & Verbs	Target
disgusting, filthy, nasty, rotten, horrible, terrible awful,stupid, moronic, dumb, ugly,repulsive, vile idiotic,shitty, fucked, kill, murder, hate, destroy	Abusive
great, fun, nice, neat, happy, good, best, fantastic wonderful, lovely, excellent, incredible, friendly gracious, kind, caring, hug, like, love, respect	Non-Abusive

For the identity term template approach the test instances which are generated from the templates are used to measure gender bias. Following work by [14] the templates we used are given in Table 3a. Table 3b lists the identity term pairs we used to give sets of paired gendered test data. These pairs are the same as those we used for the identification of gender in our ITS approach. Table 3c shows the nouns and adjectives used to fill the templates.

The identity term template approach generated 1480 synthetic test samples in total, 740 pairs with equal sets of male and female instances and equal distribution across the "abusive" and "non-abusive" classes. The distribution of the test instances for our ITS approach varied slightly for each holdout split. Table 4 shows the percentage of the dataset that was used as test data and the female and male distribution of the test data per class for both ITS and identity term template approach across the three datasets. This shows that the amount of gendered test data varies regardless of approach while the template approach generates a standard set of synthetic test data.

5 Results & Discussion

Task performance is measured using average class accuracy due to the imbalanced class distributions in all datasets as evident in Table 1. We measure gender bias using True Positive Rate Gap (TPR_{gap}) [16] which is an equality of opportunity measure and measures the differences in the gender specific true positive rates and is defined in Equation 1.

$$TPR_{gap} = \mid TPR_{male} - TPR_{female} \mid \tag{1}$$

Table 4: Percentage of the dataset used as gendered (G) test data and the distribution of gendered test data for Identity Term Sampling (ITS) and Identity Term Template (ITT) across the five holdout splits

Dataset	**Class**	**Identity Term Sampling**			**Identity Term Template**		
		G(%)	F(%)	M(%)	G(%)	F(%)	M(%)
Hate Speech	Abusive	$1.6\pm4 \times 10^{-4}$	$69\pm3 \times 10^{-2}$	$31\pm3 \times 10^{-2}$	4.6	50	50
	Non Abusive	$1.5\pm3 \times 10^{-4}$	36 ± 10^{-2}	64 ± 10^{-2}	4.6	50	50
Abusive Tweets	Abusive	$1.5\pm4 \times 10^{-4}$	41 ± 10^{-2}	59 ± 10^{-2}	0.7	50	50
	Non Abusive	$2.0 \pm3 \times 10^{-4}$	33 ± 10^{-4}	67 ± 10^{-4}	0.7	50	50
Hotel Review	Unsatisfied	0.1 ± 10^{-4}	$40\pm5 \times 10^{-3}$	$60\pm7 \times 10^{-3}$	-	-	-
	Satisfied	0.9 ± 10^{-4}	$45\pm9 \times 10^{-3}$	$55\pm8 \times 10^{-3}$	-	-	-

The results of measuring gender bias using both the identity term template approach (labelled ITT) and our new Identity Term Sampling (ITS) approach for the Hate Speech and Abusive Tweets datasets are displayed in Figures 1a to 1c. Each figure gives results for a single dataset and the left hand y-axis is classification performance and the right hand axis is the TPR_{gap} which reflects the gender bias. Each figure gives the performance on the test data for each class and for each gender. The True Positive Gap TPR_{gap} for each class is also displayed on the graph.

Across the Hate Speech and Abusive Tweets datasets (Figures 1a & 1b) where some level of gender bias may be expected, the TPR_{gap} is higher for our proposed ITS method than for the template method. It is significantly higher in the Hate Speech dataset. This shows that our proposed method is identifying gender bias at least as well as the template approach which uses synthetic data. It also suggests that the use of test data that is aligned with the original data as it is extracted from it, may be a more confident way of identifying gender bias in the data.

Looking at the gender level classification results on both datasets to identify where this gap comes from, the pattern is the same across both datasets. The accuracy on the female data is lower than the male data for the "non abusive" class. This indicates that examples of non-abusive content that are identified as female (i.e. more likely to be about women) are classified incorrectly as "abusive" more often that examples of non-abusive content that are identified as male, i.e about men. And the reverse happens in the "abusive" class, examples of abusive content that are identified as female are more often classified correctly as "abusive" than examples of abusive content that are identified as male. This suggests that the model built on this training data is demonstrating gender bias by treating gender differently. This pattern is extremely evident in the Hate Speech dataset.

Figure 1c shows the results of the Hotel Review dataset. As it is difficult to generate appropriate identity term templates that will be adequately representative for this domain, we do not include figures for the identity term template approach. As can be seen from the figure, the ITS gender gap for this dataset is very small. This is not surprising as we would not generally expect there to be significant gender bias in user generated hotel reviews. However, it is worth

Table 5: Accuracy per class, average class accuracy (ACA) on the gendered test data for identity term template (ITT) & ITS approaches and ACA for each dataset.

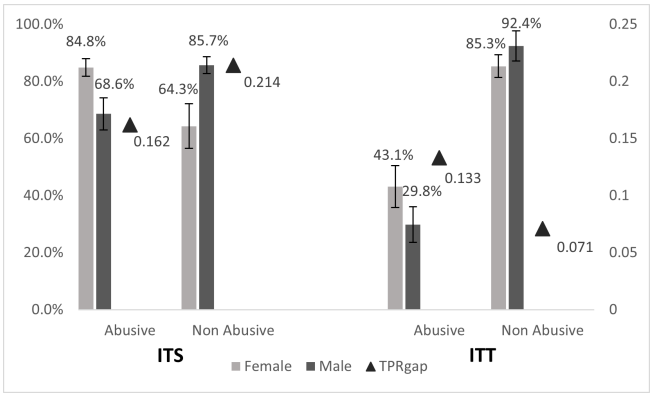
Dataset	**Class**	**Class Accuracy**		**Gender identified test ACA**		**Overall testset**
		ITS(%)	ITT(%)	ITS(%)	ITT(%)	**ACA(%)**
Hate Speech	Abusive	79.8±0.06	40.0±0.14	78.9±0.08	64.6±0.11	82.2±0.02
	Non Abusive	78.0±0.09	89.3±0.09			
Abusive Tweets	Abusive	84.2±0.01	39.0±0.23	88.4±0.01	65.3±0.17	91.4±0.007
	Non Abusive	92.5±0.02	97.6±0.03			
Hotel Review	Unsatisfied	64.8±0.16	-	75.7±0.05	-	84.4±0.06
	Satisfied	86.6±0.06	-			

noting that the ITS TPR_{gap} for the "unsatisfied" class in the Hotel Reviews is higher than the TPR_{gap} for the template based approach for the "non abusive" class in the Abusive Tweets dataset. This suggests that there may be some element of gender bias in this dataset, specifically in the "unsatisfied" class - the pattern is similar to that identified in the other two datasets. The examples of "unsatisfied" content which are identified as female (i.e. about women) are more slightly more often classified correctly as "unsatisfied" than reviews that are identified as male (i.e. about men).

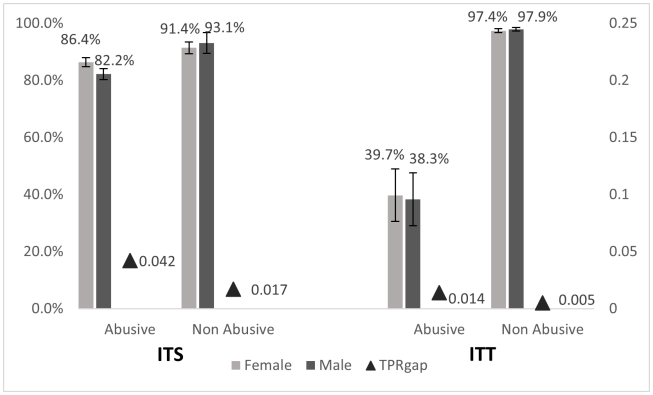
The classification results on the holdout test data and on the gendered test data for each dataset across the five holdout splits is shown in Table 5. The last column in the table shows the average class accuracy (ACA) on the full test data, averaged with the standard deviation across all five holdout splits. This shows how well the model can perform at the task of abusive content prediction with the ACA on the Abusive Tweets dataset higher at 91% than the Hate Speech at 82%. The gender-identified ACA columns show the performance of the model on just the test data with gender identified for both the ITS and identify term template (ITT) approaches. Across the two abusive content datasets the proposed ITS approach achieves significantly better performance on the gendered test data than the template approach. This is not surprising as the ITS test data is sampled directly from the original training data. However, this suggests that the templates used to measure gender bias are not reflective of the data as the model is unable to classify them well. The class accuracy columns in the table show the average class accuracy with standard deviation results for the test data with gender identified. In both abusive content datasets, the ITT approach has a very has very poor classification performance on the abusive class with less than 50% accuracy in both cases and a high standard deviation, suggesting that the template sentences generated for the abusive content do not reflect at all the actual abusive content in the datasets. The use of the original data which the proposed ITS approach achieves a significantly better performance on the abusive class suggesting better test data.

6 Conclusions & Future Work

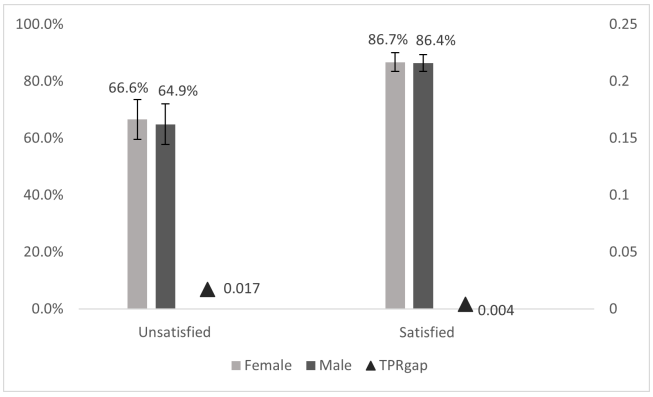
In this work, we propose an Identity Term Sampling technique to overcome one of the challenges faced in evaluating bias in training data which is the absence of



(a) Hate Speech



(b) Abusive Tweets



(c) Hotel Review

Fig. 1: Accuracy and TPR_{gap} for Identity Term Sampling (ITS) and Identity Term Template (ITT)

gender in existing datasets. The proposed method addresses the challenges and the limitations of using GBETs by automatically identifying gender for some instances in a dataset and using these to evaluate the gender bias. We evaluated the performance of ITS on an abusive content classification task using datasets which are likely to contain gender bias and a sentiment analysis task using a dataset which is less likely to contain gender bias.

Our experiment results show ITS can identify gender bias at least as well as existing template based approaches. Classification results on the gendered test data used to measure gender bias show that template based approaches do not generate test data that is appropriate for the task at hand while ITS uses test data that is better aligned to the task. While the gender identification performed in this work might be considered naive, we suggest that this approach has some promise as a more confident mechanism of measuring gender bias through automatic identification of gender. Future work will consider including more focused methods of identifying gender in text instances.

Although ITS has shown promising results in this work, it should be mentioned it might be challenging to use ITS on some types of natural language datasets. User generated content including movie and book reviews potentially can contain a wide range of gender identity words and it may be challenging to identify a single gender. More focus on refining the identification of gender in ITS may help in this respect.

Our evaluation of ITS focused on using the dataset itself for the evaluation of gender bias without applying any data augmentation techniques often used in this domain. In future work we will consider the impact of applying gender swapping as a data augmentation technique on the test instances that are generated by the ITS approach giving additional test data and equal distribution of test data. Finally, it has been observed that a wide range of research into gender bias predominantly focuses on distinguishing two genders, male and female, neglecting the fluidity and continuity of gender as a variable [20]. Future work will consider extending the ITS approach to non binary genders and also include gender-neutral linguistic norms such as 'they' in English.

Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

1. Basta, C., et al.: Evaluating the underlying gender bias in contextualized word embeddings. In: Proc of the 1st Workshop on Gender Bias in NLP. ACL (2019)
2. Bolukbasi, T., et al.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in NeurIPS (2016)

3. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017)
4. Davidson, T., et al.: Racial bias in hate speech and abusive language detection datasets. In: Proc of the 3rd Workshop on Abusive Language Online. ACL (2019)
5. De-Arteaga, M., et al.: Bias in bios: A case study of semantic representation bias in a high-stakes setting. In: Procs of FAT* (2019)
6. Dixon, L., et al.: Measuring and mitigating unintended bias in text classification. In: Proc of the 2018 AAAI/ACM Conf on AIES. AIES '18, ACM (2018)
7. Emami, A., et al.: The KnowRef coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution. In: Proc of ACL (2019)
8. Founta, A.M., et al.: Large scale crowdsourcing and characterization of twitter abusive behavior. In: Twelfth Int AAAI Conf on Web and Social Media (2018)
9. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29**, 3315–3323 (2016)
10. Kiritchenko, S., Mohammad, S.: Examining gender and race bias in 200 sentiment analysis systems. In: Procs of Conf on Lexical & Computational Semantics (2018)
11. Lu, K., et al.: Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday. Springer (2020)
12. Nadeem, M., Bethke, A., Reddy, S.: StereoSet: Measuring stereotypical bias in pretrained language models. In: Proc of ACL and the 11th IJCNLP). ACL (2021)
13. Nangia, N., Vania, C., Bhalerao, R., Bowman, S.R.: CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In: EMNLP (2020)
14. Park, J.H., Shin, J., Fung, P.: Reducing gender bias in abusive language detection. In: Proc of EMNLP. ACL (2018)
15. Peters, M.E., et al.: Deep contextualized word representations. In: Proc of the NAACL. ACL (2018)
16. Prost, F., Thain, N., Bolukbasi, T.: Debiasing embeddings for reduced gender bias in text classification. In: Proc of the 1st Workshop on Gender Bias in NLP (2019)
17. Rudinger, R., et al.: Social bias in elicited natural language inferences. In: Proc of the First ACL Workshop on Ethics in NLP. ACLs (2017)
18. Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., Turchi, M.: Gender bias in machine translation. Transactions of the ACL **9**, 845–874 (2021)
19. Smith, E.M., et al.: " I'm sorry to hear that": finding bias in language models with a holistic descriptor dataset. arXiv preprint arXiv:2205.09209 (2022)
20. Stanczak, K., Augenstein, I.: A survey on gender bias in natural language processing. arXiv preprint arXiv:2112.14168 (2021)
21. Sun, T., et al.: Mitigating gender bias in natural language processing: Literature review. In: Proc of the ACL. ACL (2019)
22. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In: Procs NAACL Student Workshop (2016)
23. Webster, K., Recasens, M., Axelrod, V., Baldridge, J.: Mind the GAP: A balanced corpus of gendered ambiguous pronouns. Transactions of the ACL (2018)
24. Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., Chang, K.W.: Gender bias in contextualized word embeddings. In: Proc of the NAACL). ACL (2019)
25. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Gender bias in coreference resolution: Evaluation and debiasing methods. In: Proc of the NAACL (2018)
26. Zhao, J., et al.: Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In: Proc of the EMNLP (2017)