Towards Fairer NLP Models: Handling Gender Bias In Classification Tasks

Nasim Sobhani and Sarah Jane Delany
SFI Centre for Research Training in Machine Learning
Technological University Dublin
nasim.x.sobhani@mytudublin.ie, sarahjane.delany@tudublin.ie

Abstract

Measuring and mitigating gender bias in natural language processing (NLP) systems is crucial to ensure fair and ethical AI. However, a key challenge is the lack of explicit gender information in many textual datasets. This paper proposes two techniques, Identity Term Sampling (ITS) and Identity Term Pattern Extraction (ITPE), as alternatives to template-based approaches for measuring gender bias in text data. These approaches identify test data for measuring gender bias in the dataset itself and can be used to measure gender bias on any NLP classifier. We demonstrate the use of these approaches for measuring gender bias across various NLP classification tasks, including hate speech detection, fake news identification, and sentiment analysis. Additionally, we show how these techniques can benefit gender bias mitigation, proposing a variant of Counterfactual Data Augmentation (CDA), called Gender-Selective CDA (GS-CDA), which reduces the amount of data augmentation required in training data while effectively mitigating gender bias and maintaining overall classification performance.

1 Introduction

In recent years, there has been a significant growth in research analyzing biases present in natural language processing (NLP) systems and models. This includes studies on biases present in embedding spaces, which are representations of words and sentences generated from large text data (Bolukbasi et al., 2016; Caliskan et al., 2017; Gonen and Goldberg, 2019; Zhao et al., 2017; May et al., 2019)and in large language models (Wan et al., 2023; Kotek et al., 2023).

Researchers have investigated how these biases manifest in NLP systems across a range of tasks, coreference resolution (Rudinger et al., 2017; Zhao et al., 2018), machine translation (Vanmassenhove et al., 2021; Savoldi et al., 2021; Stanovsky et al., 2019), sentiment analysis (Kiritchenko and Mohammad, 2018), and hate speech/toxicity detection (Park et al., 2018; Dixon et al., 2018), among others. As NLP models are trained on human-generated text data, they can acquire and propagate societal biases present in that data when deployed in real-world applications, leading to concerns about discriminating outputs (Park et al., 2018).

Machine learning models can be deliberately designed with a specific bias aligned with their intended purpose. For example, a toxic comment detector is meant to be biased toward giving higher scores to actual toxic comments over non-toxic ones. However, such models are not intended to discriminate based on attributes like gender that might be evident in comments. If a model exhibits this behavior by scoring comments differently due to gender references, it is considered an unintended and undesirable bias. While the bias towards accurately identifying toxic content is the intended goal, any bias that leads to unfair treatment or discrimination based on attributes such as gender is regarded as an unintended bias that needs to be addressed (Dixon et al., 2018). Biased algorithmic outcomes from AI systems can negatively impact users, creating a feedback loop that amplifies existing biases (Mehrabi et al., 2021). These harmful effects can impact different groups based on the nature of the bias, such as women facing discrimination from gender biases, minorities affected by racial biases, or specific age groups impacted by age-related biases. Evaluating and mitigating these unintended biases is crucial for developing trustworthy, fair, and ethical AI systems.

Bias Statement.In textual classification tasks, gender bias refers to the presence of systematic errors or unfairness in predictions related to gender within the text data. Our key concern is the potential allocational harms arising from such systematic gender biases in NLP models, where the systems may disproportionately misclassify or make inaccu-

Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 167–178

rate predictions for text associated with a particular gender group (Blodgett et al., 2016; Barocas et al., 2017). For instance, a sentiment analysis model might demonstrate gender bias by associating certain emotions or sentiments more strongly with one gender, regardless of the context (Jentzsch and Turan, 2022). Hate speech detection models can also display gender biases towards specific identity terms due to factors like uneven distribution in datasets and excessive use of certain identity terms in hate speech sentences. For instance, terms such as "women" and "feminism" may often be associated with sexist comments in benchmark datasets, leading to incorrect generalisations by the model (Park et al., 2018; Mozafari et al., 2020). This could lead to unfair censorship or moderation applied disproportionately to one gender. Similarly, biased fake news detectors may struggle more to identify misinformation targeting or involving females versus males. Such gender disparities in NLP system performance can propagate societal biases and enable discriminatory downstream impacts. Our normative stance is that an ideal NLP system should perform equally well regardless of the gender mentioned or associated with the input text. Significant differences in accuracy across genders in core classification tasks is an undesirable outcome that can enable allocational harms through unfair allocation of negative consequences like censorship, spread of misinformation, or mischaracterisation.

A primary method for identifying gender bias in an NLP system is to measure whether the performance differs across genders. However, one of the main challenges in many textual corpora is the absence of explicit gender identification.

Gender Bias Evaluation Testsets (GBETs), named by (Sun et al., 2019) have been employed to address this challenge. GBETs facilitate gender identification by creating synthetic test sets that isolate specific groups of individuals. This enables the evaluation of bias across various natural language processing (NLP) tasks. There are three types of GBETs (Stanczak and Augenstein, 2021), template-based datasets, natural language-based datasets, and datasets generated for probing language models. The template approach involves creating sentence templates with words related to gender and the specific task being evaluated. From each template sentence individual sentences are generated, one for each gender. The performance of the NLP system is then compared across the groups of this synthetic test data, one group for each gender, allowing for the measurement of gender bias. This gender identity template approach has been used (across binary genders) for various NLP tasks, including abusive language detection (Dixon et al., 2018; Park et al., 2018), sentiment analysis (Kiritchenko and Mohammad, 2018), and coreference resolution (Zhao et al., 2018; Rudinger et al., 2017). Additionally, the gender identity template has been extended to include non-binary genders (Sobhani et al., 2023).

While template-based approaches offer a way to create gender bias evaluation datasets, they face certain limitations. The artificially generated text may not accurately represent the true distribution and content of real-world data for the target task. Additionally, the templates need to be carefully designed for each specific downstream task, lacking generalisability across different NLP applications. Furthermore, studies have demonstrated that the performance of these synthetically generated test datasets on the intended downstream tasks is often poor.

In this work, we propose two techniques to identify gender in natural language text to facilitate measuring gender bias in NLP systems, aiming to overcome the limitations of template-based approaches. The first technique, Identity Term Sampling (ITS), is a knowledge-light approach built upon the work by (Sobhani and Delany, 2022) which we further extend in this study. The second technique, Identity Term Pattern Extraction (ITPE), is a more knowledge-intensive alternative that we propose to address the shortcomings of ITS. Both techniques involve selecting the test set used to measure gender bias in the NLP model from the main dataset itself, ensuring that the test data aligns with the training dataset for the target task and is not synthetically produced like template data. By leveraging the dataset itself, these techniques enable a more reliable and representative evaluation of gender bias within the NLP model's intended domain and data characteristics.

We apply these new techniques, ITS and ITPE, to measure gender bias across a diverse range of natural language processing classification tasks involving textual data about people. Such tasks, including hate speech detection, fake news identification, and sentiment analysis, are more likely to exhibit gender bias due to the presence of personal references and mentions within the text.

In addition, we use the ITPE approach in a proposed variant of Counterfactual Data Augmenta-

tion (CDA)(Lu et al., 2020), which we call Gender-Selective CDA (GS-CDA). This variant selectively applies CDA only to the gender-identified instances in the training set, using our proposed ITPE technique. We demonstrate that GS-CDA effectively reduces gender bias gaps (in some cases more than CDA itself) while maintaining overall classification performance with the significant benefit of reducing the computational overhead of augmenting the entire training data.

2 Approach

To address the challenge of the lack of gender identification for evaluating gender bias in NLP models, we propose two distinct techniques: Identity Term Sampling (ITS) which is a knowledge-light approach, and Identity Term Pattern Extraction (ITPE), a more knowledge-intensive approach. These techniques aim to determine whether the natural language text is talking about a person and to identify the gender of that person by leveraging gender identity terms and associated patterns within the text. By applying these techniques to datasets that may be used to train models for downstream classification tasks, a section of the dataset, with gender identified, can be used as test data to measure the gender bias of the model built on that training data.

Identity Term Sampling (ITS) uses the frequency of gender identity terms in a data instance to identify the gender in a sample of text that could be about a person. Table 1 presents the list of gender identity terms used by ITS. The basis of this is a list of gendered nouns from (Hoyle et al., 2019) augmented by additions pronouns and nouns such as "her/his/him," "herself/himself," "guy/gal," "male/female," and "dad/mum/mom."

ITS can assign gender to those data instances in a dataset that contains at least one gender identity term. In each data instance, the frequency of male and female identity terms listed in Table 1 as well as words ending with "man/men/woman/women" is counted within the text content. 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 gender as there was no obvious gender. ITS is quite a naive approach and does not provide a large number of gender-assigned examples. Therefore, we explored a knowledge-intensive approach to identify more gender-assigned instances in the datasets.

Male		Female	
Singular	Plural	Singular	Plural
man	men	woman	women
boy	boys	girl	girls
father	fathers	mother	mothers
son	sons	daughter	daughters
brother	brothers	sister	sisters
husband	husbands	wife	wives
uncle	uncles	aunt	aunts
nephew	nephews	niece	nieces
emperor	emperors	empress	empresses
king	kings	queen	queens
prince	princes	princess	princesses
duke	dukes	duchess	duchesses
lord	lords	lady	ladies
knight	knights	dame	dames
waiter	waiters	waitress	waitresses
actor	actors	actress	actresses
god	gods	goddess	goddesses
policeman	policemen	policewoman	policewomen
postman	postmen	postwoman	postwomen
hero	heroes	heroine	heroines
wizard	wizards	witch	witches
steward	stewards	stewardess	stewardesses
guy	guys	gal	gals
male	males	female	females
dad	dads	mum/mom	mums/moms
he	–	she	–
his/him	–	her/hers	–

Table 1: Seed words concepts

Identity Term Pattern Extraction (ITPE) is our proposed more knowledge-intensive approach which leverages a comprehensive set of part-of-speech (POS) patterns that contain gender identity terms.

The algorithm splits the data instance into individual sentences and parses each sentence to look for the POS patterns listed in Table 2. When a pattern is found, it is checked against the gender identity terms in Table 1 and the sentence is assigned the gender of the matched identity term. The approach works through the pattern list in the order stated. Once a gendered match is found, the instance has a gender identity.

In cases where there are multiple occurrences of the matched pattern, the algorithm counts the frequency of male and female gender identity terms within the data instance. The gender with the higher cumulative frequency across these patterns is then assigned as the label for that instance. In cases where the data instance contains multiple sentences, the algorithm determines the overall gender label for that data instance by selecting the majority gender across all sentences.

To illustrate how ITPE and ITS operate in practice, we can examine the sentence:

Order	POS Pattern	Examples
1	subject	he, she, my mother, that guy
2	pronoun-noun	his cookbook, his name, her choice, her face
3	adjective-noun	male oppression, stupid man, female announcer, female character
4	noun-noun	boy scout, boy teams, women comedian, woman commentator
5	pronoun-verb	he did, he thinks, she changed, she thought
6	proposition-pronoun	to him, for him, about her, to her
7	verb-pronoun	tell him, reassuring him, loves her, find her
8	determiner-noun	the man, that boy, a girl, this woman
9	pronoun-adjective-noun	his real name, her first mate

Table 2: POS patterns used for ITPE with examples

"Despite facing criticism from some men in the industry, the pioneering female CEO confidently presented her innovative strategy to the board, earning praise from her colleagues for her bold vision."

ITPE would first identify the subject "the pioneering female CEO". This matches the subject pattern (Order 1 in Table 2), and "female" is a gender-specific term. Consequently, ITPE would immediately label this sentence as female and terminate the process. In contrast, ITS would count the frequency of gender identity terms from Table 1. In this sentence, ITS would count the female terms "female" and "her" (which appear three times), and the male term "men". With five female terms and one male term, ITS would assign a female gender label to this sentence. This example demonstrates how both techniques successfully identify the gender in the text, through different mechanisms.

2.1 Evaluation

The performance of the ITS and ITPE techniques is evaluated on six natural language datasets to assess their accuracy in identifying gender. The selected datasets are all related to people and include the gender (male or female) of the person in the text. These datasets, described in Table 3, include:

BiasBios (De-Arteaga et al., 2019), a dataset of 397,340 biographies across 28 different occupations each with gender identified as male/female.

Wizard of Wikipedia (Dinan et al., 2018), consisting of conversations between two people discussing a topic related to Wikipedia biographies. It contains approximately 11K conversations annotated with "ABOUT" labels regarding man/woman/non-binary (Dinan et al., 2020). For validating our technique, we only used the dataset instances related to man/woman.

WikiBias (Wan et al., 2023) is a collection of approximately 11K personal biography datasets scraped from Wikipedia, including demographic and biographic information (Sun and Peng, 2021).

The gender subset of the **StereoSet** dataset (Nadeem et al., 2021), consisting of 378 data instances manually labeled as male/female.

CryanSets dataset (Soundararajan et al., 2023) is generated using ChatGPT from lexicons of gender-coded words from gender-coded lexicons. It includes gendered language that captures and reflects stereotypical characteristics or traits of a particular gender. From the datasets mentioned in this paper, we combined the Cryan dataset sets 1, 2, and 3, resulting in a combined dataset of approximately 8K instances including male and female labels.

Jigsaw, Unintended Bias in Toxicity Classification, a dataset from Kaggle[1] which contains comments where each comment is accompanied by a toxicity label. A subset of comments have been labeled with values ranging from 0 to 1, representing the extent of various identity attributes (such as male, female, ethnicity, etc) in the comment. For our purposes, we only consider the subset of data with male/female values greater than 0.5, resulting in approximately 63K data instances which include male and female labels.

Dataset	Gender Distribution(%)		Size
	F	M	#instances
BiasBios	46.2	53.8	396616
Wizard	19.7	80.3	9481
Wikibias	46.1	53.9	11452
StereoSet	49.5	50.5	378
CryanSets	49.7	50.3	7894
Jigsaw	59.0	41.0	63454

Table 3: Characteristics of datasets used to evaluate ITS and ITPE

ITS and ITPE were run on each of these datasets and those data instances that were successfully assigned gender were identified. Performance was

[1] https://kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification

Dataset	ITPE —Precision(%)—				ITS —Precision(%)—				Overlap
	Overall	F	M	GI%	Overall	F	M	GI%	%
BiasBios	99.6	99.9	99.3	96.3	99.9	99.9	99.9	95.5	98.7
Wizard	94.8	89.4	96.8	50.0	95.4	90.4	97.0	45.1	87.2
Wikibias	99.0	99.4	98.6	84.1	95.8	97.7	94.2	79.7	94.4
StereoSet	94.4	97.0	92.1	95.2	95.2	96.2	94.0	82.0	85.6
CryanSets	99.4	99.4	99.3	84.7	99.5	99.4	99.6	74.2	79.8
Jigsaw	81.8	93.8	71.9	83.9	89.0	96.9	81.2	77.1	74.2

Table 4: Performance of ITS and ITPE Gender Identification Techniques. Overall Precision, F% (Female Precision), M% (Male Precision), amount of Gender-Identified(GI) data using ITPE and ITS, Overlap between ITPE and ITS

evaluated by measuring precision, the percentage of those data instances with gender identified, that had the gender correctly assigned. Since these techniques are designed to identify gender in a subset of instances in the dataset (which can subsequently be used to measure gender bias) we are only concerned with the accuracy of the instances extracted by the techniques and not necessarily all instances in the dataset.

Table 4 shows the results of applying ITPE and ITS on these datasets which includes the overall precision and the precision of female and male instances extracted from each dataset. It also gives the percentage of data instances from each dataset, labeled GI%, that were identified with a gender.

When comparing the overall precision, and the precision of male and female gender identification between the ITPE and ITS approaches, we observe that the differences in precision are relatively small, with both approaches demonstrating high precision in accurately identifying gender in textual datasets.

Looking at the numbers, we can observe that for all datasets, the *GI%* column has higher percentages for ITPE as compared to ITS. By using NLP techniques the ITPE technique is able to identify a larger amount of data instances with gender information than the ITS approach.

Generally, both ITPE and ITS successfully identify the gender over 80% of instances except in the Wizard dataset. This lower percentage could be attributed to the fact that the Wizard dataset contains more names than gendered pronouns or other explicit gender references. Since ITPE and ITS primarily rely on identifying gendered words and pronouns, they struggle to determine the gender for instances that do not include any such gender-specific terms.

The *Overlap* column in Table 4 provides insights into the intersection between the data instances gender identified by ITPE and ITS techniques. This overlap is measured using the Jaccard index, a measure of similarity between two sets. A higher Jaccard Index indicates a greater overlap between the data instances identified by both techniques. Our examination reveals that the overlap between ITPE and ITS is high but the techniques do differ in what they identify.

Both ITPE and ITS exhibit high precision in accurately identifying the gender of data instances across various datasets, with ITPE achieving a higher percentage of gender-labeled dataset instances compared to ITS.

3 Measuring Bias in an NLP Task

To evaluate the efficacy of the proposed ITPE technique in measuring bias for various NLP tasks, we apply it to identify gender on several datasets that do not initially provide gender identification. We then use this to measure gender bias on a number of different types of NLP classification tasks, including hate speech detection, fake news identification, and sentiment analysis. We focus on using ITPE as it generally can identify gender for a larger number of data instances in a dataset.

We use three hate speech datasets, two fake news, and a sentiment analysis dataset. Table 5 gives the characteristics of each dataset used including size and class distribution.

The **HateSpeech** dataset (Waseem and Hovy, 2016) 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 into a binary classification problem by labeling the sexist and racist samples as the "offensive" class and neutral samples as the "non-offensive" class.

Task	Dataset	Class	Class (%)	Gender-Identified (%) F (%)	Gender-Identified (%) M (%)	Size
Hate Speech Detection	HS (W& H)	Offensive	31	25.2	9.5	17K
		Non-offensive	69	11.5	4.3	
	HS (Davidson)	Offensive	83	12.0	9.0	24K
		Non-offensive	17	5.1	11.4	
	SBIC	Offensive	53	18.0	18.0	35K
		Non-offensive	47	9.0	14.0	
Fake News Identification	WELFake	Real	51	7.0	14.2	71K
		Fake	49	2.0	6.3	
	FakeNews (Kaggle)	Real	48	1.2	6.0	44K
		Fake	52	9.0	19.0	
Sentiment Analysis	MOJI	Positive	69	5.0	5.0	2M
		Negative	31	4.0	4.0	

Table 5: Class distribution, percentage of gender-identified data, and overall size for each dataset

The **HateSpeech and Offensive** dataset (Davidson et al., 2019) is a collection of almost 24k tweets. The majority of tweets are considered to be offensive language (77%), almost 17% are labeled as non-offensive and only almost 6% of the tweets are flagged as hate speech samples. By assigning the "offensive" class label to samples exhibiting hate speech and offensive, and the "non-offensive" label to non-offensive samples, we convert the dataset into a binary classification problem.

The **Social Bias Inference Corpus (SBIC)** dataset (Sap et al., 2020) over 44K posts collected from various sources of potentially biased online content including Twitter, Reddit, and hate sites. Each post is annotated by crowdsourcing workers on Amazon Mechanical Turk, with different annotations per post. For classification in this study, we selected the data with offensive and non-offensive categories as the target labels.

The Word Embedding over Linguistic Features for Fake News Detection **(WELFake)** dataset (Verma et al., 2021) consists of about 71K news articles with 35K real and about 37K fake news from popular news datasets. The dataset includes the title and body of the news, for the purpose of gender identification we only used the title.

The second **FakeNews** dataset is a Kaggle dataset (Lifferth, 2018) consisting of about 44K instances, each labeled as reliable or unreliable. Each article in the dataset is provided with both a title and body text. However, for the purpose of gender bias evaluation and classification, we only use the title.

The **MOJI** dataset (Blodgett et al., 2016) contains over 2M tweets that are used for sentiment analysis, categorising them as either positive or negative. Additionally, the dataset provides details regarding the type of English used in the tweets, which is a sensitive attribute in fairness-aware methods. This attribute distinguishes between African-American English (AAE) and Standard-American English(SAE).

For the classification tasks, we use a pre-trained BERT model (Devlin et al., 2019) from the Hugging Face library (Wolf et al., 2020). The datasets are split into stratified training and holdout testing splits, with an 80/20 ratio. The hyperparameters of the model are tuned on a 20% split of the training data for each dataset. The full holdout test split is used to measure the overall task performance (accuracy) of the models. To evaluate classification performance, we use average class accuracy (ACA).

Measures for evaluating gender bias in NLP systems are often built upon the work of Hardt et al. (2016) on equal opportunity and equalized odds. These measures utilize the gender distributions in the training data, rather than insisting on equal outcomes for both genders regardless of the ground truth prevalence (democratic parity). Equality of opportunity considers where the predictions are independent of gender but conditional on the ground truth or positive outcome in the training data. In this work, we adapt the TPR_{gap} measure used by (Prost et al., 2019), which measures the difference in the True Positive Rates across genders classification task, to a more general measure $Class_{gap}$ to quantify disparities in a model's performance across genders. For a given class c, the $Class_{gap}$ is defined as Equation 1.

$$Class_{gap}(c) = TPR_{c,female} - TPR_{c,male} \quad (1)$$

Where $TPR_{c,g}$ is the True Positive Rate for class c and gender g,

A positive value for $Class_{gap}$ indicates a bias

Data	Classgap		Class ACC(%)		ACA	Template-based ACA
	Off	Non-Off	Off	Non-Off	(%)	(%)
HS (W& H)	0.093	-0.086	85.5	80.1	82.7	68.6
HS (Davidson)	0.020	-0.083	97.8	88.2	93.0	73.0
SBIC	0.033	-0.109	83.9	77.7	80.8	78.5

(a) Hate Speech Detection

Data	Classgap		Class ACC(%)		ACA
	Real	Fake	Real	Fake	(%)
WELFake	0.010	-0.047	97.8	91.2	96.1
Fakenews	0.011	-0.005	95.8	99.3	97.5

(b) Fake News Identification

Data	Classgap		Class ACC(%)		ACA
	Pos	Neg	Pos	Neg	(%)
Moji	0.0001	0.009	90.1	73.9	82.0

(c) Sentiment Analysis

Table 6: Classification and Bias results: Class gap, accuracy per class, average class accuracy (ACA) on the test data

towards females, the model performs better in predicting that class for female instances. Conversely, a negative value indicated bias towards males and better performance for male instances. Values close to zero represent little bias.

We measure bias using the subset of data that is gender identified in the hold-out test set. As the dataset is randomly split into train and test sets, to ensure the robustness of our evaluation and obtain a reliable estimate of the model's performance and gender bias, we repeat the splitting process three times and report the average results.

The *Gender-Identified* column in Table 5 shows the amount of female and male data that is gender-identified using the ITPE technique. The hate speech datasets, which would include more gender-specific words than other areas, tend to have a higher proportion of data identified as female than male. On the other hand, the fake news datasets have less data identified as female and more identified as male. This is not very surprising if we consider the domains. It is worth noting that for the MOJI dataset, although the percentages of 5% for the positive sentiment class and 4% for the negative sentiment class per gender may seem low, the dataset is quite large, and these percentages represent a substantial number of instances available for bias evaluation.

Table 6 presents the classification performance and gender bias results for the hate speech detection 6a, fake news identification 6b, and sentiment analysis 6c tasks. Results include the gender bias $Class_{gap}$ metric and class accuracy for each class, and the overall average class accuracy (ACA). Additionally, for the Hatespeech datasets, we report the average class accuracy (ACA) obtained using the template-based technique for comparison.

Looking at the $Class_{gap}$ results for hate speech in Table 6a the positive value in the offensive class means that the model correctly classifies female instances as abusive more than males, and the negative value in the non-offensive class, means it is incorrectly classifying female examples as abusive. This demonstrates a bias against females, as female instances are classified as offensive more frequently than instances involving males even those female instances that are not actually offensive.

Additionally, we compared our proposed approach using gender-identified instances from the original data (ITPE approach) with a template-based synthetic test set generation method. The template-based approach, following the work by (Park et al., 2018), was applied specifically to the hate speech dataset, as it is more accessible for this type of dataset compared to others. For the hate speech dataset, the template-based approach generated 1480 synthetic test samples in total, with 740 pairs of male and female instances equally distributed across the "offensive" and "non-offensive" classes. The average class accuracy (ACA) for the template-based test set is reported in the *Template-based ACA* column of Table 6a. When comparing template-based ACA with the ACA of our ITPE approach, we observe that for all datasets, the template-based approach exhibits very poor classification performance. This suggests that the generated template sentences do not accurately reflect the actual content present in the datasets.

Table 6b presents the results for the fake news detection task. The bias demonstrated here is the opposite effect of the hate speech. The positive values are for the real class and the negative values

are for the fake class, indicating that the model tends to perform better at identifying fake news for male instances compared to female instances and is inclined to consider real news as fake more for the male instances. The level of bias is significantly smaller though than the bias in the hate speech.

Table 6c shows the results for the sentiment analysis task on the MOJI dataset. There is very little bias shown in this dataset, but the differences suggest that the model has a slight tendency to predict more female instances as having negative sentiment as compared to the male instances.

As the MOJI dataset had labels for the type of English, African American English (AAE) and Standard American English (SAE), we had the opportunity to explore potential gender gap differences between a subset of AAE and SAE, to see if any disparities emerged when considering the racial characteristics present in language expression. We focused on the $Class_{gap}$ within each subset. The results are presented in Table 7.

There is little bias in the AAE subset with the $Class_{gap}$ values showing a minimal difference between male and female instances. However, in the SAE dataset, there is more bias shown with the positive sentiment $Class_{gap}$ exhibiting a positive value, and the negative sentiment $Class_{gap}$ with negative value. Essentially, this suggests that for Standard American English, the model tended to classify more male-written text as negative sentiment and female-written text as positive sentiment. In contrast, such distinctions were not observed in the African American English subset.

Subset	Classgap		Class ACC(%)		ACA
	Pos	Neg	Pos	Neg	(%)
AAE	-0.0005	0.005	94.2	78.2	86.2
SAE	0.021	-0.041	86.0	69.5	77.7

Table 7: Gender Bias Analysis for a subset of African American English (AAE) and Standard American English (SAE)

In general, the results reveal more pronounced gender bias in the hate speech detection task compared to fake news identification and sentiment analysis which may not be surprising due to the nature of the task. Hate speech models exhibit substantial class gender gaps, indicating biases in classifying offensive content based on gender mentions. In contrast, fake news detection models show relatively smaller gender gaps, while sentiment analysis exhibits negligible bias. However, upon examining individual groups of African American English and Standard American English in the sentiment analysis task, gender bias is observed within the Standard American English texts.

4 Using ITPE in Bias Mitigation

We have seen in the previous section that the models for hate speech detection exhibit gender bias. Mitigating bias in machine learning models is a critical challenge to ensure fairness and prevent discrimination against protected groups. Strategies employed for bias mitigation can be categorized into three main approaches: pre-processing, in-processing (during training), and post-processing (Ravfogel et al., 2020; Han et al., 2022). Pre-processing techniques adjust the training dataset prior to model training to achieve balanced representations across protected groups such as gender and race. A common approach is resampling the training set, such that the number of instances within each protected group is equal. One popular pre-processing technique for mitigating gender bias is Counterfactual Data Augmentation (CDA) (Lu et al., 2020). CDA augments the training data with gender-swapped examples, building upon basic gender word swapping (e.g., "he" to "she") while addressing key limitations. It handles co-references to maintain grammatical consistency, swapping gendered words that co-refer to proper nouns (e.g., "Queen Elizabeth" to "King Elizabeth"). CDA offers a systematic approach to augmenting the data with counterfactual examples, providing a comprehensive solution to reduce gender bias encoding.

In-processing or during-training approaches introduce constraints into the model optimization process. A widely adopted method is adversarial training, which jointly trains a discriminator to recover protected attributes from the model's representations and the main model to make accurate predictions while preventing the discriminator from determining the protected attributes (Zhang et al., 2018; Elazar and Goldberg, 2018).

While adversarial training has been shown to reduce bias in machine learning models (Zhang et al., 2018; Han et al., 2021), one of its key limitations is that it requires having access to sensitive attribute labels (e.g. gender, race) during the training process. The need for annotated sensitive attribute data can be restrictive, as such labels may not always be available in the data.

If we consider a task like hate speech identification and the datasets used in the previous section the sensitive attribute, gender, is not identified in

Dataset	Class	Original%				CDA%				GS-CDA%			
		Gap	Class	ACA	TSize	Gap	Class	ACA	TSize	Gap	Class	ACA	TSize
HS (W&H)	Off	0.093	85.5	82.7	13K	0.072	81.8	81.1	26K	0.039	83.2	82.3	16K
	Non-off	-0.086	80.1			-0.050	80.5			-0.060	81.3		
HS(Davidson)	Off	0.020	97.8	93.0	20K	0.024	97.6	92.8	38K	0.021	97.7	92.7	24K
	Non-off	-0.083	88.2			-0.075	88.0			-0.053	87.7		
SBIC	Off	0.033	83.9	80.8	28K	0.011	84.4	80.5	56K	0.017	84.1	80.8	36K
	Non-off	-0.109	77.7			-0.068	76.6			-0.032	77.6		

Table 8: Comparison of before and after applying Counterfactual Data Augmentation (CDA) and Gender-Selective Counterfactual Data Augmentation (GS-CDA) Bias Mitigation Techniques for Hate Speech Detection.Classification and Bias results: Class gap, Accuracy per class, average class accuracy (ACA) on the test data, and Training Size(TSize) per each dataset

the data preventing using adversarial training to mitigate the bias in these models. So, a pre-processing technique such as CDA can be used to reduce this bias. One of the limitations of CDA is that it significantly increases the size of the training data, as it augments the training data with gender-swapped versions.

We propose a variant on CDA called Gender-Selective Counterfactual Data Augmentation (GS-CDA) where CDA is selectively applied only to the data instances in the training set that were identified as containing gender information using the ITPE technique.

To evaluate how useful GS-CDA is in bias mitigation, we use the same approach discussed in Section 2. The results of classification and gender bias after applying the CDA and GS-CDA to training data are shown in Table 8.

Comparing the original classification and gender bias results on hate speech datasets in Table 8 with the results after applying bias mitigation techniques we observed a notable reduction in gender bias gaps.

Compared to the original models, applying CDA during training data augmentation leads to a reduction in gender bias gaps. Notably, CDA lowers the offensive $Class_{gap}$ from 0.093 to 0.072 on the HateSpeech(W&H) dataset and the non-offensive $Class_{gap}$ from 0.109 to 0.068 on the SBIC dataset. However, the classification accuracy (ACA) remains almost the same. The GS-CDA variant demonstrates even more promising results. GS-CDA achieves further reductions in gender bias gaps, outperforming both the original models and the full CDA approach. On the HateSpeech(W&H) dataset, GS-CDA lowers the offensive $Class_{gap}$ to 0.039 and the non-offensive $Class_{gap}$ to 0.060, while on SBIC, the non-offensive gap is reduced to 0.032. Remarkably, GS-CDA maintains comparable or slightly improved ACA compared to the original models. These findings suggest that selectively augmenting gender-identified instances is an effective strategy for mitigating bias while preserving overall classification performance.

The *Tsize* columns in the table show the number of training instances for the original datasets before any mitigation, as well as the training set size after applying the mitigation techniques. As can be observed, the training set size for CDA is almost twice as large as the original dataset size. However, the training set size for GS-CDA is significantly smaller than that of CDA, adding only around 20% to the original dataset size. GS-CDA offers an additional benefit over the full CDA approach by avoiding the computational expense associated with doubling the training data size, as is the case with CDA.

5 Conclusion

This paper addresses the challenge of measuring and mitigating gender bias in NLP systems by proposing ITS and ITPE as techniques for identifying gender information in textual data, which can be used as an alternative to template-based approaches for measuring gender bias. Through the evaluation on multiple datasets, we demonstrate the techniques performance in accurately assigning gender labels. By applying ITPE, we demonstrated measuring gender bias in various NLP classification tasks, including hate speech detection, fake news identification, and sentiment analysis. We showed that these techniques facilitate measuring gender bias in a wide variety of NLP classification tasks, which offers significant benefits over the existing template technique which has only been used for hate speech detection.

However, it is important to acknowledge the limitations of our techniques. One limitation is the inability to recognize names. This is primarily because names vary significantly across different cultures and regions, and many libraries do not adequately support some names including Irish, Asian,

and other ethnic groups. Additionally, some names are unisex, making gender identification based on names alone tricky and often inaccurate. Another important limitation is that this approach only considers binary gender, which excludes non-binary and other gender identities.

In addition, we have used the ITPE technique to mitigate observed gender bias by introducing Gender-Selective Counterfactual Data Augmentation (GS-CDA), a variant of the popular CDA approach. GS-CDA selectively augments only the gender-identified instances during training, leveraging ITPE's capabilities. Our results show that GS-CDA effectively reduces gender bias gaps while maintaining overall classification performance, outperforming the conventional CDA approach and using less augmented data.

The proposed techniques, ITPE and GS-CDA, offer practical alternatives to template-based methods for measuring and mitigating gender bias in NLP systems. By addressing the limitations of template techniques and efficiently augmenting training data, these approaches pave the way for fairer and more ethical AI systems. As future work, these techniques will be extended to other protected attributes and applied to a broader range of NLP tasks to promote algorithmic fairness and responsible AI development. In addition, they will be extended to include non-binary and transgender individuals, emphasizing the importance of addressing the full spectrum of gender identities in NLP research. While our proposed methods have shown effectiveness in certain NLP tasks, it will be very intriguing to see how these methodologies generalize across different languages and cultures and perform in more diverse or complex datasets.

Acknowledgments

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 license to any Author Accepted Manuscript version arising from this submission.

References

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual conference of the special interest group for computing, information and society*, page 1. Philadelphia, PA, USA.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357.

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

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Maria De-Arteaga et al. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. ACM.

Jacob Devlin et al. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 ACL: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. ACL.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-dimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Lucas Dixon et al. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AIES*, AIES '18, page 67–73, New York, NY, USA. ACM.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the ACL: Human Language Technologies*, pages 609–614, Minneapolis, Minnesota. ACL.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021. Diverse adversaries for mitigating bias in training. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2760–2765, Online. Association for Computational Linguistics.

Xudong Han et al. 2022. FairLib: A unified framework for assessing and improving fairness. In *Proceedings of the 2022 Conference on EMNLP: System Demonstrations*, pages 60–71, Abu Dhabi, UAE. ACL.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323.

Alexander Miserlis Hoyle et al. 2019. Unsupervised discovery of gendered language through latent-variable modeling. In *Proceedings of the of the ACL*, pages 1706–1716, Florence, Italy. ACL.

Sophie Jentzsch and Cigdem Turan. 2022. Gender bias in BERT - measuring and analysing biases through sentiment rating in a realistic downstream classification task. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 184–199, Seattle, Washington. Association for Computational Linguistics.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24.

William Lifferth. 2018. Fake news.

Kaiji Lu et al. 2020. Gender bias in neural natural language processing. *Logic, language, and security: essays dedicated to Andre Scedrov on the occasion of his 65th birthday*, pages 189–202.

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Ninareh Mehrabi et al. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861.

Moin Nadeem et al. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the ACL and Conference on NLP (Volume 1: Long Papers)*, pages 5356–5371, Online. ACL.

Ji Ho Park et al. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on EMNLP*, pages 2799–2804, Brussels, Belgium. ACL.

Flavien Prost, Nithum Thain, and Tolga Bolukbasi. 2019. Debiasing embeddings for reduced gender bias in text classification. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 69–75, Florence, Italy. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

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.

Nasim Sobhani and Sarah Jane Delany. 2022. Identity term sampling for measuring gender bias in training data. In *Irish Conference on Artificial Intelligence and Cognitive Science*, pages 226–238. Springer.

Nasim Sobhani, Kinshuk Sengupta, and Sarah Jane Delany. 2023. Measuring gender bias in natural language processing: Incorporating gender-neutral linguistic forms for non-binary gender identities in abusive speech detection. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1121–1131.

Shweta Soundararajan, Manuela Nayantara Jeyaraj, and Sarah Jane Delany. 2023. Using chatgpt to generate gendered language. In *2023 31st Irish Conference on*

Artificial Intelligence and Cognitive Science (AICS), pages 1–8. IEEE.

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

Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. *arXiv preprint arXiv:1906.00591*.

Jiao Sun and Nanyun Peng. 2021. Men are elected, women are married: Events gender bias on Wikipedia. In *Proceedings of the 59th Annual Meeting of the ACL and the 11th International Joint Conference on NLP*, pages 350–360, Online. ACL.

Tony Sun et al. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the ACL*, pages 1630–1640. ACL.

Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. 2021. Neutral rewriter: A rule-based and neural approach to automatic rewriting into gender-neutral alternatives. *arXiv preprint arXiv:2109.06105*.

Pawan Kumar Verma, Prateek Agrawal, Ivone Amorim, and Radu Prodan. 2021. Welfake: word embedding over linguistic features for fake news detection. *IEEE Transactions on Computational Social Systems*, 8(4):881–893.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "kelly is a warm person, joseph is a role model": Gender biases in LLM-generated reference letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748, Singapore. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Thomas Wolf et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 EMNLP*, pages 38–45, Online. ACL.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

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

Jieyu Zhao et al. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on EMNLP*, pages 2979–2989, Copenhagen, Denmark. ACL.