Using ChatGPT to generate Gendered Language

1st Shweta Soundararajan
School of Computer Science
Technological University Dublin
Dublin, Ireland
shweta.x.soundararajan@mytudublin.ie

2nd Manuela Nayantara Jeyaraj
School of Computer Science
Technological University Dublin
Dublin, Ireland
manuela.n.jeyaraj@mytudublin.ie

3rd Sarah Jane Delany
School of Computer Science
Technological University Dublin
Dublin, Ireland
sarahjane.delany@tudublin.ie

***Abstract*—Gendered language is the use of words that denote an individual's gender. This can be explicit where the gender is evident in the actual word used, e.g. mother, she, man, but it can also be implicit where social roles or behaviours can signal an individual's gender - for example, expectations that women display communal traits (e.g., affectionate, caring, gentle) and men display agentic traits (e.g., assertive, competitive, decisive). The use of gendered language in NLP systems can perpetuate gender stereotypes and bias. This paper proposes an approach to generating gendered language datasets using ChatGPT which will provide data for data-driven approaches for gender stereotype detection and gender bias mitigation. The approach focuses on generating implicit gendered language that captures and reflects stereotypical characteristics or traits of a particular gender. This is done by engineering prompts to ChatGPT that use gender-coded words from gender-coded lexicons. The evaluation of the datasets generated shows good instances of English-language gendered sentences that can be identified as those that are consistent with gender stereotypes and those that are contradictory. The generated data also shows strong gender bias.**

***Index Terms*—natural language processing, machine learning, large language models, ChatGPT, gendered language, prompt engineering, zero-shot prompting**

I. Introduction

Gendered language is a phenomenon that has long existed in human communication [1]. In NLP, gendered language refers to the use of language that explicitly or implicitly indicates the gender of a person, animal, or object [2]–[4]. While gendered language can be useful in certain contexts, it can also perpetuate gender stereotypes and bias [5], [6]. Gendered wording in job advertisements has been shown to exist and sustain gender inequality [7]–[9]. The use of gendered language has promoted gender stereotypes in various situations, for e.g., in dialogues within Bollywood movies [10], biographical pages of notable people [11].

Gendered language occurs in real world interactions but also in textual content used in and by NLP systems. The availability of gendered language datasets that exhibit gender stereotypes and gender bias is a step towards alleviating gender stereotyping or bias in NLP systems.

This paper proposes an approach to generating corpora of gendered language using pre-trained Large Language Models (LLMs), specifically ChatGPT. The ability to generate text samples that resemble human-labelled data has been improved by the advent of these models [12]. LLMs have been used for data generation previously in the medical field for medical report generation and medical recommendations [13], for generating clinical text data [14], [15] and for training data augmentation in low-resource data scenarios [14], [16]. They have also been used for generating training data for fine-tuning multilingual models [17].

The proposed approach generates examples of gendered language using lexicons of gender-coded words. Gaucher et al. [7] define these masculine-coded and feminine-coded words, such as those associated with gender stereotypes, as gendered wording. People often change how they refer to or speak about others depending on who they are [18], [19], and, in particular, based on the gender identity of the individual being spoken about [20], [21].

The focus of the proposed approach is on using adjectives which have been shown to reflect stereotypical characteristics or traits of a gender [22]–[27]. Adjectives which describe women have been shown to differ from those used to describe men in numerous situations such as in movie dialogues [10], job advertisements [7]–[9], recommendation letters for faculty positions [28], fashion magazines [24], [26], Wikipedia articles [11], fictional stories [25] and children's picture books [29].

While there are some datasets available that have been created for stereotype detection [30]–[34], the focus of these datasets was on crowdsourcing to collect examples of stereotypes rather than creating gendered language that may or may not include gender stereotypes. The majority of these datasets [30], [31], [34] feature examples that go beyond gender stereotypes. Furthermore, the instances associated with gender stereotypes in these datasets cover all the categories of gender stereotypes outlined by Deaux & Lewis [35] and Chiril et al. [33], though the specific type of gender stereotype remains unidentifiable. Our work centers on adjectives that can give rise to gender stereotypes, specifically in two of these cases, namely personality traits and physical appearance.

Two lexicons of gender-coded words were used in the generation of the gendered language; the first was created by Gaucher in work that examines job descriptions [7] and a more recent lexicon where adjectives were scraped from Wikipedia and annotated with gender using crowdsourcing [32]. This lexicon is not publicly available but a portion of it was provided by the authors.

979-8-3503-6021-9/23/$31.00 ©2023 IEEE

The prompts provided to ChatGPT requested sentences about people that include the gendered word from the lexicon. Sentences are labelled as *consistent with gender stereotypes* if the person in the sentence is the same gender as the gendered word used and labelled *contradictory to gender stereotypes* otherwise.

A dataset was created for each lexicon, where an instance in the dataset was a generated sentence including at least one word from the lexicon. Classification models were built on the datasets and evaluated to see how accurately the gender stereotypes could be identified using state of the art text classification approaches. Multiple datasets were generated from each lexicon and evaluated to ensure stability in classification performance.

We found that both lexicons were able to generate examples of gendered language that exhibit gender stereotypes and gender bias. Sentences that are consistent with gender stereotypes were distinguished from those contradictory to gender stereotypes with good accuracy. The gender bias of the generated data was measured and showed considerable bias.

The rest of the paper is structured as follows. Section II discusses how gender is depicted in textual content and existing work on identifying gendered language and creating gender lexicons. Section III outlines our approach and describes the datasets generated and section IV discusses the evaluation undertaken. The paper concludes in section V identifying some future work on how these datasets will be used.

II. Related Work

Gendered language can take many forms from the use of gendered pronouns (e.g., "he" or "she") to gendered titles (e.g., "Mr." or "Ms."). In some languages, gendered language is even more pervasive, with gendered nouns and adjectives that must agree with the gender of the subject or object. A categorisation of gender in language technology, distilled from Ackerman [36], Cao & Daumé III [37], and Bartl and Leavy [38], is that gender can be categorised as either **linguistic**, the expression of gender within grammar and language or **social**, which relates to the cultural and social roles, behaviour and identity of individuals in society. Fig. 1 depicts this categorisation.

According to Cao & Daumé III [37] linguistic gender includes grammatical gender, referential gender and lexical gender. Grammatical gender is a system of noun classification in many languages, although not English, where nouns are categorised based on agreement between the nouns and their dependents. In languages with grammatical gender, nouns are typically either masculine, feminine, or neuter, with each gender having its own set of rules for determining the appropriate gender for a particular noun. Referential gender in language refers to the gender of a person, animal, or object that is being referred to in the text. Third-person pronouns and gendered titles are the most obvious examples in English. Lexical gender is a semantic property of a word, which is assigned a gender based on its inherent characteristics or meaning. For example, in English, the noun "man" is typically associated with the masculine gender while the noun "woman" is associated with the feminine gender.

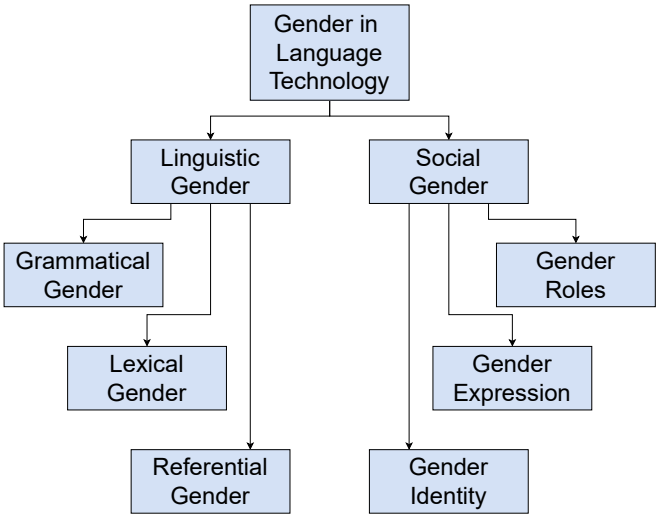

Fig. 1. Categorisation of Gender in Language Technology distilled from Ackerman [36], Cao & Daumé III [37], and Bartl and Leavy [38]

Social gender, which refers to the cultural and social aspects associated with gender in a particular society or community, includes concepts such as gender identity, gender expression, and gender roles [36]. Gender identity refers to an individual's internal sense of their gender, which may or may not correspond to their biological sex [39]. Gender expression refers to the way an individual presents their gender to the world through their appearance, behavior, and mannerisms [40]. Whereas gender roles are societal expectations and norms associated with gender, including behaviors, attitudes, and activities that are considered appropriate for men and women [41]. Gender roles are often learned from a young age through socialisation and can vary significantly across cultures and time periods. For example, in some cultures, men may be expected to be breadwinners while women are expected to take care of the home and children [42].

The way gender is expressed in language can reflect or reinforce social gender norms and stereotypes, but it can also challenge or subvert them [43]. The referential identity of an individual may or may not match their social gender. Social gender has become a prominent consideration when it comes to inferring a subject or entity's gender, or properly addressing a subject or entity so as to align the real-world gender identity to the intended language used in text.

All these scenarios can give rise to gendered language in text content. Lexical gender and referential gender are explicit in that the gender is evident in the words used which is useful for detecting gender in content. Social gender is more implicit. Work on identifying gender bias [44]–[46] and mitigating gender bias [44], [47]–[49] focuses more on linguistic gender, using the pronouns and gender identity terms, words that carry lexical gender.

Gender lexicons are available which include pre-compiled

lists of words representing social characteristics and behaviors that distinguish men and women [50], [51]. The Personal Attributes Questionnaire (PAQ) [52] and Bem Sex Role Inventory (BSRI) [50] are two of the earliest gender lexicons. These gender lexicons are based on college students' self-reported characteristics, which are measured through questionnaires on their self-assessment and self-valuation of feminine and masculine characteristics.

Gaucher et al. [7], developed a list of gendered words from lists of agentic and communal words (e.g., individualistic, competitive, committed, supportive) and masculine and feminine trait words (e.g., ambitious, assertive, compassionate, understanding) from previous psychology studies [50], [53]–[55]. The Gender Decoder for Job Ads tool [56] extended Gaucher et al.'s [7] lexicon and used it to provide an overall judgment on the gender-coding of job advertisements by determining the ratio between masculine-coded and feminine-coded words.

The traditional use of gendered word inventories or gender lexicons such as the BSRI, to assess gender roles and self-perceptions as feminine or masculine has been a widely accepted standard for over 40 years [57]. However, the BSRI lexicon scores have been reassessed [58], [59] and these studies reported that the femininity scores have reduced and women are less likely to endorse traditionally feminine characteristics as representative of themselves. This suggests that societal gender norms should be updated regarding stereotypical characteristics of masculinity and femininity. Responding to this, Cryan et al. [32] proposed a new gender lexicon created by scraping Wikipedia for lists of candidate words and using crowdsourcing for annotation.

Several research efforts have produced datasets for stereotype detection [30]–[34]. These studies involved creating datasets by extracting content from various sources or using crowdsourcing to gather instances of stereotypes, rather than generating gendered language that may or may not contain gender stereotypes. Most of these datasets [30], [31], [34] also include examples that extend beyond gender stereotypes such as race, ethnicity, age, occupation, and various social categories. Additionally, when examining instances related to gender stereotypes in these datasets, it is important to note that they encompass all types of gender stereotypes as described by Deaux & Lewis [35] and Chiril et al. [33], including personality traits, domestic behaviors, occupations, and physical appearance, but the type of gender stereotype is not identifiable. Our work specifically focuses on adjectives that can contribute to gender stereotypes in two of these cases: personality traits and physical appearance. Furthermore, some of these efforts [30], [31] concentrate on generating datasets to detect gender bias in language models rather than solely identifying gendered language within textual content. Diverging from approaches that rely on crowdsourcing for dataset creation, our methodology involves generating synthetic datasets utilizing pre-trained language models, thereby eliminating the need for crowdsourcing. These datasets are crafted with the purpose of facilitating gender stereotype detection in textual content and aiding bias mitigation.

III. Approach

The proposed approach to producing examples of gendered language uses ChatGPT to generate sentences that are about individuals and that contain adjectives that are gender-coded. The adjectives used were selected from two lexicons which we call Gaucher [7] and Cryan [32], the more recent lexicon. Cryan's lexicon is not publicly available but the researchers who generated it provided a version to the authors for use.

Gaucher's lexicon contains 82 gender-coded words that are either labelled as masculine or feminine. While the words listed in Gaucher's lexicon are not all adjectives, the words are stemmed and can be used as adjectives, verbs or nouns. For instance, the word "aggress*" can be used as an adjective (aggressive) or as a noun (aggression). A limitation of Gaucher's lexicon is the size, which is small having been manually collected from lists of words, including traits and behaviors associated with males and females, that were published in previous studies. The version of Cryan's lexicon that we received contained 2903 male-coded and 2504 female-coded words with a gender score associated with each word.

As Cryan's lexicon contained words that were not adjectives, preprocessing was done on the lexicon to filter out the non-adjectives[1] leaving 1845 masculine and 1675 feminine adjectives.

The candidate words in Cryan's lexicon were scraped from Wikipedia. The most commonly used adjectives were annotated with a gender score as male or female by crowdsourcing and these labelled words were used to predict the gender scores for the remaining words. The version of the lexicon provided appeared to be incomplete, as the most commonly used words reported in the paper were missing. We extracted the most commonly used adjectives that were visible in the word cloud diagram in Cryan et al.'s [32] paper. The words extracted from the word cloud diagram included 299 gendered adjectives, of which 152 were masculine and 147 were feminine. Examples include *beautiful*, *emotional*, *glamorous* and *dependent* for females and *muscular*, *courageous* and *aggressive* for males.

The gender score of each word was based on the extent to which a word is associated with a particular gender. The gender score ranges from +20 (male) to -20 (female), with higher gender scores indicating that the word is more associated with males and lower gender scores indicating that the word is more associated with females. We focused on the adjectives that were more strongly gender coded by using those that were outside 1 standard deviation of the mean gender score. It should be noted that gender scores were not available for the most commonly used words, those extracted from the word cloud diagram. Nevertheless, as they were the most commonly occurring adjectives, most with obvious gender coding, we

[1]To identify adjectives we used the Free Dictionary API (`https://dictionaryapi.dev/`) to get the Parts of Speech of the words; for the words for which the POS wasn't available at the API, the POS was obtained manually from Wiktionary, and only the words with adjective as their POS were retained.

included these words in the lexicon. Table I specifies the details of the size and distribution of both lexicons used.

TABLE I
SIZE AND DISTRIBUTION OF LEXICONS USED WHERE M INDICATES MALE-CODED WORDS AND F INDICATES FEMALE-CODED WORDS.

Lexicon	#M	#F	Size
Gaucher	42 (51.22%)	40 (48.78%)	82
Cryan	702 (53.71%)	605 (46.29%)	1307

Fig. 2 shows the pipeline for generating the sentences based on the lexicon words using ChatGPT. We used OpenAI's ChatGPT (gpt-3.5-turbo) through the API to generate the sentences. Zero-shot prompting[2] [60] of ChatGPT was used for data generation - several studies have indicated that ChatGPT exhibits impressive performance in data generation tasks when utilized with zero-shot prompting [15], [16]. We chose zero-shot prompting over one-shot and few-shot prompting [60] because the latter methods require examples for each prompt, which can be data-intensive. In contrast, zero-shot prompting can accommodate a broader range of prompts without requiring specific examples. Additionally, we aimed to avoid having the model generate sentences based on the examples or templates provided in the prompt. Instead, we wanted the generated sentences to be distinct and unique. We instructed ChatGPT to generate sentences by passing our instruction prompt along with gendered words from the lexicon to the ChatGPT API. In order to ensure accurate and effective responses from ChatGPT, it is important to frame the instruction prompt in a manner that ChatGPT understands and can respond to correctly [60], [61]. The instruction prompt should be concise and to the point, providing sufficient information for ChatGPT to comprehend the intent. Similar to work done by Tang et al. [15], prompt engineering was performed by constructing several instruction prompts and selecting the final prompt by reviewing the generated sentences. Examples of prompt refinements that were made include inputting lexicon words in batches of 20 into ChatGPT - this was more effective as ChatGPT produced longer and more complex sentences and included more than one adjective when a limited number of lexicon words were included in the prompt, setting sentence length, instructing to include all the types of English sentences and tenses, and setting the context style. An example of the prompt used to generate sentences about females is shown in Table II. Sentences about males were generated in the same way.

We asked ChatGPT to generate sentences about males and females using both masculine and feminine adjectives. Examples of generated sentences are given in Table III where MM refers to sentences about males with masculine terms, FF refers to sentences about females with feminine terms, MF refers to sentences about males with feminine terms, and FM refers to sentences about females with masculine terms.

[2]Zero-shot prompting is a technique that involves providing a task description to a language model instead of direct supervision or training data, enabling the model to perform a specific task. The task description is typically in the form of a prompt or question, which guides the model in generating the desired output.

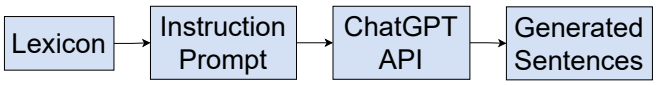

Fig. 2. Pipeline for data generation using ChatGPT

TABLE II
INSTRUCTION PROMPT PROVIDED TO CHATGPT TO GENERATE SENTENCES ABOUT FEMALES WHICH INCLUDE THE ADJECTIVES WORD_1, WORD_2, ETC. [X] IS REPLACED BY THE NUMBER OF GENDER-CODED WORDS PASSED TO CHATGPT WITH THE INSTRUCTION PROMPT.

Generate [x] sentences in the style of context from newspapers, magazines, children's books, job advertisements, story books, etc. using all the words listed below. Simple sentences, compound sentences, complex sentences, and compound-complex sentences must all be included in the list of sentences. Each and every sentence must necessarily be about a female, females, woman, women, girl, or girls. Any tenses and any parts of speech can be used in the sentences. The sentences can use pronouns, nouns, the name of a person, etc. to refer to the female, females, woman, women, girl, or girls being discussed in the sentences. All of the sentences must use one or more of the words mentioned below as an adjective or noun to depict the characteristic or traits of the female, females, woman, women, girl, or girls being discussed in the sentence. More than 15 words must be used in each sentence. word_1, word_2, word_3, ...word_n

TABLE III
EXAMPLES OF CHATGPT GENERATED SENTENCES.

Examples	Label
The **daring** boy jumped off the high dive with a **fearless** attitude. The **bearded** man spoke in a blunt manner, highlighting his **authoritative** stance on the current political scene. The **ruthless** businessman would stop at nothing to beat out his competition and succeed.	MM
The **caring** mother hugged her daughter tight, knowing she needed the comfort after a long day. The **attractive** woman walked confidently down the street, turning heads with her beauty. The **emotional** woman couldn't help but tear up at the beautiful sight of the sunset.	FF
The **motherly** man gently hugged his son and whispered words of encouragement into his ear. The **pretty little** boy had a contagious smile that brightened up the room. The **beautiful** man, who was also quite **fragile**, tried his best to be strong while dealing with his emotional struggles	MF .
The high school girl was more **aggressive** on the basketball court than any of her male teammates. The **irreverent** woman had a **sarcastic** sense of humor and often made jokes that left people laughing. The **headstrong** woman refused to let her unfortunate circumstances define her, which was a major display of her resilience.	FM

IV. EVALUATION

In order to evaluate the data generated by ChatGPT, we tested whether the generated sentences could be identified as consistent with gender stereotypes or not in two ways - by human annotators and using a supervised machine learning approach.

The generated sentences about males using masculine adjectives and females using feminine adjectives were labelled as *consistent with gender stereotypes*. Those about males using feminine adjectives and females using masculine adjectives were labelled as *contradictory to gender stereotypes*. Table IV gives these labelling details while the size and data distribution of the datasets generated by ChatGPT are given in Table V. Three datasets (set 1, 2, & 3) were generated from each lexicon for evaluation purposes. Those generated from the Gaucher lexicon had the same distribution across the labels. However, the datasets generated using Cryan's lexicon differed slightly in class distribution. This was due to the model's tendency to occasionally generate an inconsistent number of output sentences [62], [63], even though we added instructions in the prompt to generate a specified number of sentences.

TABLE IV
LABELLING OF SENTENCES GENERATED BY CHATGPT.

Categories of Sentences	Label
Sentences about males with masculine terms (MM) Sentences about females with feminine terms (FF)	Consistent with Gender Stereotypes (S)
Sentences about males with feminine terms (MF) Sentences about females with masculine terms (FM)	Contradictory to Gender Stereotypes (NS)

To perform a sanity check on the labels assigned to the sentences, four human annotators (two males and two females) were tasked with labelling a random selection of 50 generated sentences. We ensured a balance between samples that are consistent with gender stereotypes and those that are contradictory to gender stereotypes. We provided the human annotators with a clear definition of gendered language, supplemented by a few illustrative examples of both labels. The inter-annotator agreement between each annotator and the labels given was measured using Cohen's Kappa. The average of the Cohen's Kappa scores across all annotators was 0.80 (with individual scores of 0.84, 0.80, 0.72 and 0.84). This confirms that there was substantial agreement between the annotators and our labels. Table VI shows examples of sentences that received different levels of agreement among the annotators. This includes cases of complete agreement, where all annotators correctly labelled the sentences, and cases of partial agreement, where at least one annotator incorrectly labelled the sentences. Every sentence has at least one annotator who assigned the correct label. So, there are no sentences for which all the annotators assigned the wrong label.

We then calculated the classification accuracy at distinguishing sentences that are consistent with and contradictory to gender stereotypes in each dataset using the pre-trained transformer model BERT. We used stratified 5-fold cross validation with the training data further split 80% : 20% for hyperparameter tuning. For each fold, the 80% of the training data was used to fine-tune the BERT model and the remaining 20% was used to identify the optimal hyperparameters. Finally, sentence classification is performed on the testing data with the optimal hyperparameters for each fold.

As an estimate of the consistency of the generated data, multiple datasets were generated from the same lexicon and the stability of the classification performance was measured across all versions generated. We followed the same procedure for all the datasets generated as ChatGPT gives different results each time for the same prompt.

LLMs have been shown to have gender bias [30], [31], [45] which is evident if the behaviour of the system is different for men than for women. The primary method to measure gender bias is to measure performance differences across gender as the system's performance should not be influenced by gender. Gender bias in classification systems can be measured using the True Positive Rate Gap (TPR_{gap}) [64]. This is an equality of opportunity measure which measures the differences in the gender specific true positive rates. It is defined in (1) where TPR is the *True Positive Rate*.

$$\text{TPR}_{\text{gap}} = \mid \text{TPR}_{\text{male}} - \text{TPR}_{\text{female}} \mid \tag{1}$$

Table VII shows the classification accuracy and the gender bias of the datasets generated by ChatGPT. Each of these is shown with a measure of standard deviation across the different generated datasets for the same lexicon. The accuracy results in Table VII show that it is possible to identify stereotypes on both datasets generated from the two lexicons. However the Cryan lexicon outperforms the Gaucher lexicon. This discrepancy could be attributed to the relatively smaller size of the Gaucher lexicon. The accuracy is similar across both classes (consistent with gender stereotypes and contradictory to gender stereotypes), although the accuracy for identifying the consistent with gender stereotypes is slightly lower than that of the contradictory with stereotype across both lexicons. It is significant that both datasets show a notable gender bias, with the datasets derived from Gaucher exhibiting a significantly higher bias.

Analyzing the incorrect predictions of the model showed some of the sentences were incorrectly predicted as the gender-coded words were associated with an animal or an object rather than a person. Although the prompt instruction requested sentences about people, infrequently sentences were generated about objects. This happened when the adjective wasn't suitable to describe a character trait for a person. Examples of this include *atmospheric*, *archaeological* and *hierarchical*. In addition, when the sentences contained both male-coded and female-coded words the sentences may be incorrectly classified by the model. The first example of partial agreement in Table VI is an example of this. This is an issue that the annotators also had when they were labelling.

V. CONCLUSION & FUTURE WORK

In this paper we propose an approach to generating gendered language datasets using ChatGPT. The generation of the datasets is driven by lexicons that include words which capture

TABLE V
DATA DISTRIBUTION OF THE DATASETS OF GENERATED SENTENCES.

Dataset	#Consistent with gender stereotypes			#Contradictory to gender stereotypes			Size
	MM	FF	Total	MF	FM	Total	
Gaucher	42 (51.22%)	40 (48.78%)	82 (50%)	40 (48.78%)	42 (51.22%)	82 (50%)	164
Cryan (set 1)	707 (53.89%)	605 (46.11%)	1312 (49.87%)	615 (46.63%)	704 (53.37%)	1319 (50.13%)	2631
Cryan (set 2)	714 (54.13%)	605 (45.87%)	1319 (50.11%)	611 (46.53%)	702 (53.47%)	1313 (49.89%)	2632
Cryan (set 3)	702 (53.55%)	609 (46.45%)	1311 (49.83%)	618 (46.82%)	702 (53.18%)	1320 (50.17%)	2631

TABLE VI
ANNOTATOR AGREEMENT LEVELS FOR SENTENCES: EXAMPLES OF COMPLETE AND PARTIAL AGREEMENT.

Level of agreement	Examples
Complete	The bubbly girl couldn't contain her joyful laughter during the party. (FF) The emotional man broke down in tears as he read a heartfelt letter from his son abroad. (MF) The unafraid male firefighter ran into the burning building to rescue those trapped inside. (MM)
Partial	Although seen as submissive, the female CEO proved to be a powerful force in the boardroom. (FF) Despite facing rude remarks, the headstrong girl remained independent in her actions. (FM) The feeble man struggled to lift the heavy boxes as the other men watched on. (MF)

TABLE VII
ACCURACY AND GENDER BIAS PERFORMANCE FOR DATASETS GENERATED FROM THE TWO LEXICONS. THE ACCURACY AND GENDER BIAS FIGURES SHOW STANDARD DEVIATION ACROSS MULTIPLE DATASETS GENERATED WITH THE SAME PROMPTS.

Dataset	Gaucher	Cryan
Overall Accuracy (%)	67.9 ± 6.5	77.2 ± 2.7
TPR of S (%)	67.1	76.2
TPR of NS (%)	68.7	78.1
TPR_{gap} in S (%)	10.7 ± 7.8	7.3 ± 5.6
TPR_{gap} in NS (%)	17.1 ± 11.6	8.7 ± 7.2

the characteristics or traits of a particular gender. The prompt to ChatGPT requests generated sentences that are about people and include specific gender-coded words from the lexicons used.

We show that the datasets produced include good instances of English natural language sentences. Sentences were labelled as consistent with gender stereotypes where the gender of the person was consistent with the gender of the gendered word used in the generation and labelled as contradictory to gender stereotypes otherwise. Labels given by human annotators for a random sample of generated sentences correlated highly with these labels.

Using a state of the art text classification model, the instances that are consistent with gender stereotypes were distinguished from those that are more contradictory to gender stereotyping with good accuracy. All generated datasets show significant gender bias in the classification task which supports the view that LLMs capture bias from the training data on which they are trained [65]–[68].

One limitation in our work is that we focused on the binary genders of male and female when generating the datasets. This decision was influenced by the lack of gender lexicon resources available for other genders. Future work will look at generating more inclusive gendered data, at how to extend this work and generate gendered data for non-binary and trans genders. This may be done by exploring the generation of non-binary lexicons which can be used in a similar way to generate non-binary gendered data.

The datasets we created are of moderate size, limited by the sizes of the lexicons used to generate sentences. In addition there was a certain amount of noise in the lexicons - where certain adjectives caused ChatGPT to generate sentences about objects rather than people. Future work will look at cleaning and extending these gender lexicons to provide a much larger dataset.

The availability of gendered language datasets will assist in work on detecting gender stereotypes and mitigating bias in textual content. These datasets are publicly available in our repository (`https://doi.org/10.21427/AYM7-SF29`). We intend to use these datasets for implicit gendered language detection tasks.

ACKNOWLEDGMENT

This work was funded by Technological University Dublin through the TU Dublin Scholarship – Presidents Award.

REFERENCES

[1] C. L. Sidner, "Focusing for interpretation of pronouns," *American Journal of Computational Linguistics*, vol. 7, no. 4, pp. 217–231, 1981.
[2] R. Thomson, T. Murachver, and J. Green, "Where is the gender in gendered language?" *Psychological Science*, vol. 12, no. 2, pp. 171–175, 2001.
[3] R. S. Bigler and C. Leaper, "Gendered language: Psychological principles, evolving practices, and inclusive policies," *Policy Insights from the Behavioral and Brain Sciences*, vol. 2, no. 1, pp. 187–194, 2015.
[4] F. Hamidi, M. K. Scheuerman, and S. M. Branham, "Gender recognition or gender reductionism? the social implications of embedded gender recognition systems," in *Proceedings of the 2018 chi conference on human factors in computing systems*, 2018, pp. 1–13.
[5] M. Bucholtz and K. Hall, "Language and identity," *A companion to linguistic anthropology*, vol. 1, pp. 369–394, 2004.
[6] C. Leaper and R. S. Bigler, "Gendered language and sexist thought." *Monographs of the Society for Research in Child Development*, 2004.
[7] D. Gaucher, J. Friesen, and A. C. Kay, "Evidence that gendered wording in job advertisements exists and sustains gender inequality." *Journal of personality and social psychology*, vol. 101, no. 1, p. 109, 2011.

[8] S. Tang, X. Zhang, J. Cryan, M. J. Metzger, H. Zheng, and B. Y. Zhao, "Gender bias in the job market: A longitudinal analysis," *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, no. CSCW, pp. 1–19, 2017.

[9] R. E. Tokarz and T. Mesfin, "Stereotyping ourselves: gendered language use in management and instruction library job advertisements," *Journal of Library Administration*, vol. 61, no. 3, pp. 301–311, 2021.

[10] N. Madaan, S. Mehta, T. Agrawaal, V. Malhotra, A. Aggarwal, Y. Gupta, and M. Saxena, "Analyze, detect and remove gender stereotyping from bollywood movies," in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 92–105.

[11] C. Wagner, D. Garcia, M. Jadidi, and M. Strohmaier, "It's a man's wikipedia? assessing gender inequality in an online encyclopedia," in *Proceedings of the international AAAI conference on web and social media*, vol. 9, no. 1, 2015, pp. 454–463.

[12] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He *et al.*, "A comprehensive survey on pretrained foundation models: A history from bert to chatgpt," *arXiv preprint arXiv:2302.09419*, 2023.

[13] S. Wang, Z. Zhao, X. Ouyang, Q. Wang, and D. Shen, "Chatcad: Interactive computer-aided diagnosis on medical image using large language models," *arXiv preprint arXiv:2302.07257*, 2023.

[14] H. Dai, Z. Liu, W. Liao, X. Huang, Z. Wu, L. Zhao, W. Liu, N. Liu, S. Li, D. Zhu *et al.*, "Chataug: Leveraging ChatGPT for text data augmentation," *arXiv preprint arXiv:2302.13007*, 2023.

[15] R. Tang, X. Han, X. Jiang, and X. Hu, "Does synthetic data generation of llms help clinical text mining?" *arXiv preprint arXiv:2303.04360*, 2023.

[16] S. Ubani, S. O. Polat, and R. Nielsen, "Zeroshotdataaug: Generating and augmenting training data with chatgpt," *arXiv preprint arXiv:2304.14334*, 2023.

[17] A. Michail, S. Konstantinou, and S. Clematide, "Uzh_clyp at semeval-2023 task 9: Head-first fine-tuning and chatgpt data generation for cross-lingual learning in tweet intimacy prediction," *arXiv preprint arXiv:2303.01194*, 2023.

[18] J. R. Rickford, F. McNair-Knox *et al.*, "Addressee-and topic-influenced style shift: A quantitative sociolinguistic study," *Sociolinguistic perspectives on register*, vol. 235, p. 276, 1994.

[19] D. Hymes, "Ways of speaking," *Explorations in the*, 2009.

[20] R. Lakoff, "Language and woman's place," *Language in society*, vol. 2, no. 1, pp. 45–79, 1973.

[21] P. Eckert and S. McConnell-Ginet, "Communities of practice: Where language, gender, and power all live," in *Locating power: Proceedings of the second Berkeley women and language conference*, vol. 1. Berkeley, CA: Berkeley University, 1992, pp. 89–99.

[22] C. Hoffman and M. A. Tchir, "Interpersonal verbs and dispositional adjectives: The psychology of causality embodied in language." *Journal of personality and social psychology*, vol. 58, no. 5, p. 765, 1990.

[23] A. Maass, "Linguistic intergroup bias: Stereotype perpetuation through language," in *Advances in experimental social psychology*. Elsevier, 1999, vol. 31, pp. 79–121.

[24] S. Arvidsson, "A gender based adjectival study of women's and men's magazines," 2009.

[25] E. Fast, T. Vachovsky, and M. S. Bernstein, "Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community," in *Tenth International AAAI Conference on Web and Social Media*, 2016.

[26] A. Morelius, "The use of adjectives in contemporary fashion magazines: A gender based study," 2018.

[27] N. Ellemers, "Gender stereotypes," *Annual review of psychology*, vol. 69, pp. 275–298, 2018.

[28] J. M. Madera, M. R. Hebl, and R. C. Martin, "Gender and letters of recommendation for academia: agentic and communal differences." *Journal of Applied Psychology*, vol. 94, no. 6, p. 1591, 2009.

[29] J. A. Williams Jr, J. Vernon, M. C. Williams, and K. Malecha, "Sex role socialization in picture books: An update," *Sociology Department, Faculty Publications*, p. 8, 1987.

[30] M. Nadeem, A. Bethke, and S. Reddy, "Stereoset: Measuring stereotypical bias in pretrained language models," *arXiv preprint arXiv:2004.09456*, 2020.

[31] N. Nangia, C. Vania, R. Bhalerao, and S. R. Bowman, "Crows-pairs: A challenge dataset for measuring social biases in masked language models," *arXiv preprint arXiv:2010.00133*, 2020.

[32] J. Cryan, S. Tang, X. Zhang, M. Metzger, H. Zheng, and B. Y. Zhao, "Detecting gender stereotypes: lexicon vs. supervised learning methods," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–11.

[33] P. Chiril, F. Benamara, and V. Moriceau, ""be nice to your wife! the restaurants are closed": Can gender stereotype detection improve sexism classification?" in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 2833–2844.

[34] R. Pujari, E. Oveson, P. Kulkarni, and E. Nouri, "Reinforcement guided multi-task learning framework for low-resource stereotype detection," *arXiv preprint arXiv:2203.14349*, 2022.

[35] K. Deaux and L. L. Lewis, "Structure of gender stereotypes: Interrelationships among components and gender label." *Journal of personality and Social Psychology*, vol. 46, no. 5, p. 991, 1984.

[36] L. M. Ackerman, "Syntactic and cognitive issues in investigating gendered coreference," *Glossa*, 2019.

[37] Y. T. Cao and H. Daumé III, "Toward gender-inclusive coreference resolution," *arXiv preprint arXiv:1910.13913*, 2019.

[38] M. Bartl and S. Leavy, "Inferring gender: A scalable methodology for gender detection with online lexical databases," in *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, 2022, pp. 47–58.

[39] J. Sunderland and L. Litosseliti, "Gender identity and discourse analysis," *Gender identity and discourse analysis*, pp. 1–343, 2002.

[40] D. L. Rubin and K. L. Greene, "Effects of biological and psychological gender, age cohort, and interviewer gender on attitudes toward gender-inclusive/exclusive language," *Sex Roles*, vol. 24, pp. 391–412, 1991.

[41] U. Gabriel, P. Gygax, O. Sarrasin, A. Garnham, and J. Oakhill, "Au pairs are rarely male: Norms on the gender perception of role names across english, french, and german," *Behavior research methods*, vol. 40, no. 1, pp. 206–212, 2008.

[42] J. Brutt-Griffler and S. Kim, "In their own voices: Development of english as a gender-neutral language: Does learning english promote gender equity among asian international students?" *English Today*, vol. 34, no. 1, pp. 12–19, 2018.

[43] M. Hellinger and H. Bußmann, "Gender across languages: The linguistic representation of women and men," *Gender across languages*, pp. 1–26, 2015.

[44] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," *Advances in neural information processing systems*, vol. 29, 2016.

[45] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Men also like shopping: Reducing gender bias amplification using corpus-level constraints," *arXiv preprint arXiv:1707.09457*, 2017.

[46] J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K.-W. Chang, "Gender bias in contextualized word embeddings," *arXiv preprint arXiv:1904.03310*, 2019.

[47] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Gender bias in coreference resolution: Evaluation and debiasing methods," *arXiv preprint arXiv:1804.06876*, 2018.

[48] J. Zhao, Y. Zhou, Z. Li, W. Wang, and K.-W. Chang, "Learning gender-neutral word embeddings," *arXiv preprint arXiv:1809.01496*, 2018.

[49] M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, and A. T. Kalai, "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*, 2019, pp. 120–128.

[50] S. L. Bem, "The measurement of psychological androgyny." *Journal of consulting and clinical psychology*, vol. 42, no. 2, p. 155, 1974.

[51] P. Rosenkrantz, S. Vogel, H. Bee, I. Broverman, and D. M. Broverman, "Sex-role stereotypes and self-concepts in college students." *Journal of consulting and clinical psychology*, vol. 32, no. 3, p. 287, 1968.

[52] J. T. Spence, R. Helmreich, and J. Stapp, "Personal attributes questionnaire," *Developmental Psychology*, 1974.

[53] C. Hoffman and N. Hurst, "Gender stereotypes: Perception or rationalization?" *Journal of personality and social psychology*, vol. 58, no. 2, p. 197, 1990.

[54] L. A. Rudman and S. E. Kilianski, "Implicit and explicit attitudes toward female authority," *Personality and social psychology bulletin*, vol. 26, no. 11, pp. 1315–1328, 2000.

[55] J. A. Bartz and J. E. Lydon, "Close relationships and the working self-concept: Implicit and explicit effects of priming attachment on agency and communion," *Personality and Social Psychology Bulletin*, vol. 30, no. 11, pp. 1389–1401, 2004.

[56] K. Matfield, "Gender decoder for jobs ads," 2014.

[57] M. L. Dean and C. C. Tate, "Extending the legacy of sandra bem: Psychological androgyny as a touchstone conceptual advance for the study of gender in psychological science," *Sex Roles*, vol. 76, pp. 643–654, 2017.
[58] J. M. Twenge, "Changes in masculine and feminine traits over time: A meta-analysis," *Sex roles*, vol. 36, pp. 305–325, 1997.
[59] K. Donnelly and J. M. Twenge, "Masculine and feminine traits on the bem sex-role inventory, 1993–2012: A cross-temporal meta-analysis," *Sex roles*, vol. 76, pp. 556–565, 2017.
[60] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
[61] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.
[62] Q. Chen, H. Sun, H. Liu, Y. Jiang, T. Ran, X. Jin, X. Xiao, Z. Lin, Z. Niu, and H. Chen, "A comprehensive benchmark study on biomedical text generation and mining with chatgpt," *bioRxiv*, pp. 2023–04, 2023.
[63] P. P. Ray, "Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," *Internet of Things and Cyber-Physical Systems*, 2023.
[64] F. Prost, N. Thain, and T. Bolukbasi, "Debiasing embeddings for reduced gender bias in text classification," *arXiv preprint arXiv:1908.02810*, 2019.
[65] S. Singh, "Is chatgpt biased? a review," 2023.
[66] S. Ghosh and A. Caliskan, "Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages," *arXiv preprint arXiv:2305.10510*, 2023.
[67] E. A. Van Dis, J. Bollen, W. Zuidema, R. van Rooij, and C. L. Bockting, "Chatgpt: five priorities for research," *Nature*, vol. 614, no. 7947, pp. 224–226, 2023.
[68] S. Shahriar and K. Hayawi, "Let's have a chat! a conversation with chatgpt: Technology, applications, and limitations," *arXiv preprint arXiv:2302.13817*, 2023.