GenBiT: measure and mitigate gender bias in language datasets

Authors:
Kinshuk Sengupta, Global Data Science and Analytics , kisengup@microsoft.com
Rana Maher, Global Data Science and Analytics, rana.maher@microsoft.com
Declan Groves, Global AI, degroves@microsoft.com
Chantal Olieman, Global AI, chantal.olieman@microsoft.com

Keywords: *gender bias, responsible ai, ai for good, machine learning, language modeling, text analytics*

Abstract

Natural Language Processing (NLP) systems have shown incredible results while solving business problems as part of many automated solutions. However, it has become clear that many NLP systems suffer from various biases often inherited from the data on which these systems are trained. The prejudice is exhibited at multiple levels spilling from how individuals generate, collect, and label the information leveraged into datasets. Datasets, features, and rules in machine learning algorithms absorb and often magnify such biases present in datasets. Therefore, it becomes essential to measure preferences at the data level to prevent unfair model outcomes. The paper introduces GenBiT, a tool to measure gender bias. The model is designed based on word co-occurrence statistical methods. In addition to measuring bias, a novel approach for mitigating gender bias is introduced based on contextual data augmentation powered by language models combined with random sampling, sentence classification, and filtering on targeted gendered pieces of data to eliminate unintended gender bias in multilingual training data. Our experiments demonstrate that this ensembled mitigation approach can ensure historical gender biases are reduced in conversational parallel multilingual datasets. This facilitates fairer machine learning model training over the augmented datasets to improve fairness and inclusiveness across a range of potential model applications.

1. Introduction

Recently, there has been a growing awareness of bias in machine learning (ML) models and artificial intelligence (AI) systems [1]. The existence of discrimination in AI systems has become one of the top industry concerns, often reflecting the historical biases inherent in human decisions. Building ethical and socially responsible AI starts with awareness of the potentially harmful effects and applying appropriate methods to protect groups impacted negatively.

Numerous studies have investigated the possible implication of data and algorithmic bias in real-world contextual learning models and the influence on fairness and customer decision-

making [2]. Some well-documented cases of harmful biases occurring in ML models include facial recognition technologies [3] and predictive analytics that are increasingly being leveraged to make hiring decisions [4], to prompt financial investigations, and to determine health care risk assessments [5].

Every state of AI development and deployment can introduce biases; conscious and unconscious bias introduced by the AI scientists and engineers in the decisions they have made in AI system design, in the learning algorithms themselves, or biases in the data on which the algorithms are trained and tested [6].

Identifying bias in the trained model is challenging and sometimes impossible as such algorithms are not always inspectable. However, identifying data bias will help system architects make more ethical design judgments in their models and enable AI feature teams to ensure that models trained on or evaluated on this data are not biased. Therefore, tackling bias at the data level is arguably more impactful than addressing it later in the AI lifecycle. Natural language processing (NLP) models that leverage text corpora are potentially more prone to potential societal biases entrenched in the data. At the same time, the systems learn inapt correlations between the final decisions and sensitive attributes such as race and gender [7].

Gender bias, the focus of our work, is the preference or prejudice towards one gender over another. Several teams across Microsoft have identified gender bias as an important topic that can reinforce existing societal biases. Previous research work from Kate Crawford [8] shows how gender bias dominates artificial intelligence models for most tasks. Many of these tasks rely on historical data and have inadequate or unfair representation of the female gender. That has led to gender-based harms in many previously mentioned scenarios, such as hiring decisions in engineering disciplines being heavily skewed against female candidates [4].

Despite the increasing awareness of the consequences of gender bias in ML and the potential of resulting harm to customers and end-users, practical solutions for quantifying and mitigating gender bias are still limited. In NLP tasks, where we are training on large volumes of corpora, the lack of available solutions is particularly apparent when considering languages beyond English. Our work aims to define, quantify, and mitigate this unintended bias to improve fairness across a range of potential model applications.

The main contributions of this paper are summarized as follows:

1. We introduce a measurement framework for gender bias in datasets. The method leveraged helps determine if gender is uniformly distributed across data by measuring the strength of association between pre-defined gendered words and other words in the corpus via co-occurrence statistics and applies to many languages.
2. This paper further proposes handling grammatical gender in language datasets using sentence masking and post-facto lemmatization techniques.
3. In addition to measuring gender bias, we present a bias mitigation approach to evaluate and mitigate the gender bias in data through a robust and efficient manner powered by language modeling techniques that is also easily scalable to multiple languages. We provide reliable, more balanced datasets that can be used to test/train AI systems. The proposed approach is generic and can also be applied to other bias categories to promote fairness and inclusiveness.

2. GenBiT Framework

This section discusses the GenBiT[1] (Gender bias in text toolkit) approach to computing gender bias scores for datasets based on co-occurrence counts of words in the dataset with a given list of pre-defined gender definition words provided as reference. The gendered word list (which we refer to as the 'gender definition lexicon') is a manually curated dictionary created for five languages (EN=English, DE=German, FR=French, RU=Russian, ES=Spanish). A detailed statistical formulation of the approach is described in the following section.

2.1. Metric Calculation Approach

We derive our approach from that of Bordia & Bowman (2019) [9], who collect a word co-occurrence matrix across the tokenized input data and, from these counts, calculate conditional probabilities via the maximum likelihood method (MLE) denoted as,

$$P(w|g) = \frac{count(w,g)/\sum_i count(w_i,g)}{count(g)/\sum_i count(w_i)}$$

For each word in the corpus, *w*, the above formula calculates the probability of it co-occurring with any male-gendered or female-gendered word, *g,* from the gender definition lexicon. Co-occurrence counts between *w* and *g* are collected if *w* and *g* occur with a pre-defined context window of length *c.* To add greater importance to words that appear in proximity to gender definition words, we apply back-off weighting so that each co-occurrence count is multiplied by the discount value $0.95^{distance(w,g)}$. To avoid non-zero counts, add *1/N* smoothing, with *N* being the number of unique tokens in the dataset. Probabilities are returned as log values to prevent overflow.

To quantify the bias measurement scores, we choose two key critical metrics for bias assessment. The method leverages co-occurrence frequency counts and conditional probability as described above, and iterative benchmarking was performed to validate the final metrics used in the framework.

- Average absolute bias score: $avg(abs(count(w|g_m)/count(w|g_f)))$
- Average absolute bias conditional score: $avg(abs(P(w|g_m)/P(w|g_f)))$

The algorithmic implementation accepts a list of strings; the length of each element of the list is not a constraint. However, each component of the list may represent text from an entire file, a single paragraph, or a single sentence. The non-empty list constraint is applied i.e., a list should contain at least one element, and that the members of the list are python string types.

2.2. Gender Definition Lexicons

The gender definition lexicon consists of words that represent concepts that can only ever be attributed to a person of a specific gender; they represent words that are associated with gender by their definition. Entries are unambiguous and consist primarily of nouns or pronouns that refer to a person of a specific gender.

In the current version, we have a male lexicon and female lexicon, with an approximate correspondence between each; in general, for every male-gendered form, there is one (or more) female gendered form(s), and vice versa. The current version does not account for non-gender, but this is work currently underway. The English lexicon gender definition pairs were based on work published by Bordia & Bowman (2019) [9]. Authors further extended them manually to include additional concepts that fell under the broad definition, including the most common gendered professions as taken from US census information (e.g., waiter/waitress, businessman/businesswoman). The English lexicon was reviewed by members of Microsoft's

[1] https://aka.ms/genbit

Aether Bias and Fairness Natural Language working group.

The lexica for other languages were created by linguists, using the English lexicon as a guideline but working from a concept definition to avoid any inherent bias working from English. The lexica were further extended to account for all gender and linguistic case variations of entries as well as language/market specific synonyms, which for lead to significantly increases in lexicon sizes, particularly for morphologically rich languages like German and Russian (cf. Table 1). The process also included professions that typically carry gender markers in languages exhibiting grammatical gender (e.g., *engineer* in English has different male/female forms in most other languages).

Language	male	female
English (en)	230	231
Italian (it)	244	389
German (de)	482	583
Spanish (es)	236	365
French (fr)	234	299
Russian (ru)	679	1406

Table 1: Gender definition lexicon sizes

2.3. Masked Lemmatization for Grammatical genders

Grammatical gender is a linguistic characteristic of many languages where nouns are assigned a particular gender and require grammatical agreement with any words associated with these nouns, such as adjectives, articles, verbs, and pronouns. Our initial implementation of GenBiT did not cater explicitly to grammatical gender or lexical variants. There were few serious challenges in model performance during rapid experimentations and benchmarking, where metric scores obtained were inflated due to grammatical gender.

In English, words that refer to a gendered subject only ever have one form, e.g., "*The waiter was polite*"/"*The waitress was polite*.". In contrast, these words would have multiple forms in many other languages, e.g., French: "*Le serveur étail poli*"/"*Le serveuse était polie*.". If we have <*serveur*, *serveuses*> in our gender definition lexicon, then we would want to treat both forms of "*polite*" (<*poli*, *polie*>) as the same word to accurately capture any bias here.

The GenBiT core algorithm was modified using masking and lemmatization to resolve challenges due to grammatical gender and additional word forms. We chose the technique based on results from multi-stage benchmarking conducted using multilingual datasets. An optimal approach to handle grammatical gender identified was to apply lemmatization to 'normalize' gender variants and mask gender definition words. The modification of masking the gender definition words was carried out to prevent them from being at risk of alteration (e.g., words like "*waiter*" and "*waitress*" would both be lemmatized to the root form "*waiter*"). The implementation is followed by treating stopwords, punctuation treatments, and other noise reduction. Once a sentence is lemmatized, the original surface forms are re-constructed based on the original sentence, with masked gender definition words. Lemmatization has the additional benefit of reducing data sparseness for all languages, as word variants, e.g., '*run*,' '*running*,' '*runs,*' etc., are all normalized to the same root form, e.g., '*run*.'

Next, a thorough benchmarking was performed to evaluate the impact of the algorithmic implementation on multiple language datasets, as discussed in later sections of this paper. A high-level process flow is described in figure 1

In addition to lemmatization, we adjusted pronouns included in our gender definition word lexica. For example, the pronoun '*it*' in English is commonly used to refer to inanimate objects, whereas '*he*' and '*she*' refer to people. In many other languages, the masculine and feminine pronouns can refer to either people or too masculine/feminine inanimate objects (e.g.,

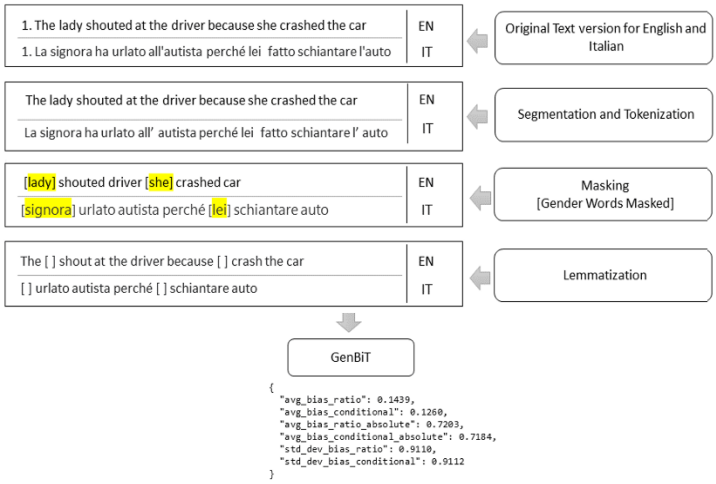

Figure 1: Implementation pipeline GenBiT

'*table*,' '*chair*' etc., carry gender markers in many languages). To identify biased gender associations, we only wish to concentrate on words associated with animate objects (i.e., people). Without a co-reference resolution component, we have no way of disambiguating pronoun referencing; therefore, we removed all ambiguous pronouns from all the non-English lists.

The purpose of conducting benchmarking and A/B testing was to find the correlation for the proposed hypothesis(H_0), i.e., whether the handling of grammatical gender delivers minimal variation and more stable scores across different languages. The results and detailed analysis from our benchmarking experiments on the GenBiT metrics are provided in Section 4.1

3. Bias Mitigation Framework

Once we can identify and measure bias in datasets, the next decisive stage is mitigating that bias. This section introduces a novel approach for mitigating and eliminating unintended gender bias in multilingual datasets.

Previous related work on gender bias mitigation at the data level uses gender-swapping techniques and counterfactual data augmentation. The methods use sentences augmented via a targeted approach that requires dependency parsers and exhaustive lists of finite gendered nouns [10], [11]. Due to the need for such resources, this approach is not readily applicable to other languages, particularly languages with rich morphological inflections.

3.1 An Ensemble Approach

In our work, we apply an ensemble approach, where we combine targeted random over-sampling together with data augmentation using contextual word embeddings as provided by mBERT[2], a multilingual bi-directional transformer-based language representation model. mBERT provides sentence representations for 104 languages (model configuration: 12-layer, 768-hidden, 12-heads, 110M parameters).

Part of our solution makes use of a simple rule-based sentence classifier (or tagger) to initially label sentences with a gender class; 'male,' 'female,' 'mixed,' or 'no gender/neutral.' Sentences classified as 'male' contain male words from the gender definition lexicon (and no female tokens/words). In contrast, conversely, sentences classified as 'female' contain female words from the gender definition lexicon and no male tokens/words. Additionally, we label sentences that contain both male and female gender definition words as 'mixed gender' and sentences that do not contain any words from the lexicon as 'no gender/neutral' sentences.

During the process, we target the sub-sample of the dataset that belongs to the minority gender class, as identified using our sentence classifier, and then apply randomized over-sampling followed by data augmentation using mBERT to members of that sub-sample. The newly created synthetic instances are then added to the original dataset to create a more balanced dataset. The

[2] bert/multilingual.md at master · google-research/bert · GitHub

method avoids the potential problem of model over-fitting due to straightforward over-sampling, particularly for datasets that exhibit a small number of samples for the minor class. Different techniques like Wordnet[3] were evaluated for word and synonym replacement. Finally, mBERT was chosen over others due to its contextualized bidirectional representation in many languages. However, sequential augmentation was implemented to allow users to select multiple augmentation techniques.

1. **Text Processing:** Process and tokenize input text into sentences, paragraphs, or deal with it as a continuous text based on data nature and defining the context window according to the nature of data.
2. **Sentence Tagger/Classifier:** Classify data according to the defined gender bias type. This classifier is used to evaluate the distribution of biased groups based on gendered words in the dataset. In our case, we classify the input text into four classes: Male, Female, Mixed Gender, and No Gender.

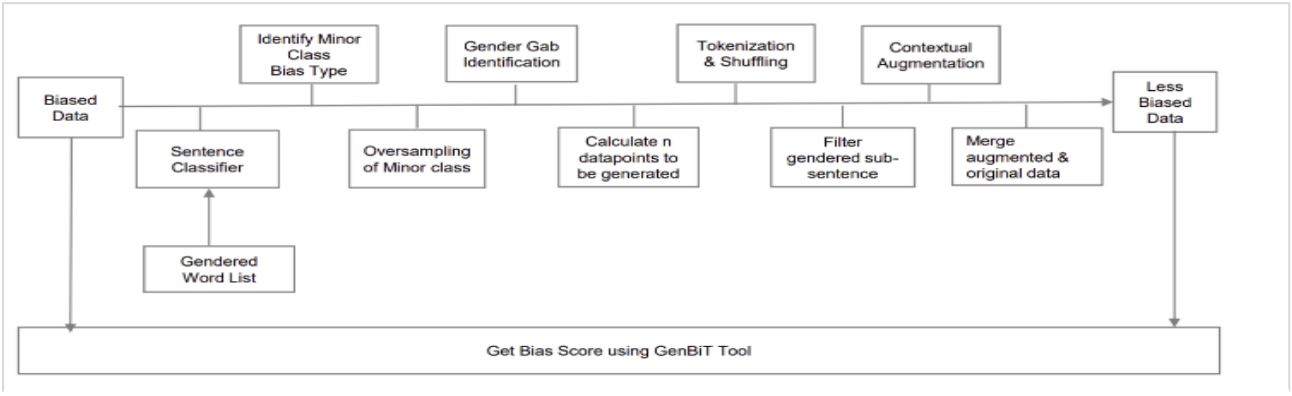

Figure 2: Mitigation approach procedure step by step to mitigate bias in the dataset

We monitor the 'gender gap' (the gender class imbalance) during the process and iterate until we have closed this gap significantly. We use both our sentence classifier and GenBiT metrics to measure the balance and bias of the resulting dataset. The approach is quite generic for any language and not dependent on the bias (gender, race, religion). The algorithmic flow is illustrated in Figure 2, and the various steps are further described in Section 3.2.

3.2 Mitigation Algorithm

As mentioned previously, the first component of the mitigation technique is based on implementing a sentence classifier/tagger focus mitigation on the more biased content of the text. The classifier was used for the evaluation of the distribution of biased and unbiased groups across the dataset.

The second component of the mitigation approach is generating contextualized data that balances the training dataset and reduces unintended gender bias. Now, to minimize the scalability issues for long sentences and large corpora, several tasks have been added before running the augmentation task using mBERT as described below:

1. **Identify required Data Generation**: Identify *n* data points needed from each sample in the minor class by using the classifier result as an input to calculate the gap compared to the class of interest, in our case, male/female-biased class.

[3] WordNet Interface (nltk.org)

2. **Random Sampling**: Leverage a random over-sampler to boost the original samples of the minor class with a percent range from 10% to 20% to avoid overfitting problems that may result from the replication of actual instances, especially if the sample size of the minor class is too small.
3. **Shuffling**: Use of Albumentation package leveraging python utilities for computer vision and NLP to shuffle each sample into sub-sentences and form a new version from each instance [12].
4. **Filter Gendered Sentences:** Target gendered substances within each sentence as candidates for the application of augmentation (this is somewhat equivalent to the notion of phrase rewriting)
5. **Contextual augmentation:** using contextual word embedding for sentence augmentation by applying mBERT.

The last three steps are repeated *n* times according to the defined gap based on the classifier explained above. Finally, the augmented data samples generated using mBERT are merged with the original data to create a more balanced dataset. Below is an example of the sentence transformation.

Original Text:

An awful film! They've taken the taste of a first famous female Renaissance painter and mangled it beyond recognition. My complaint is not that they've taken liberties with the facts; if the story were good, it must have been up against some real stinkers to be nominated for the Golden Globe.

Shuffled & Augmented Text:

They've taken the story of the first famous German grave painter and mangled it beyond recognition. An awful film! My complaint is not that they've taken liberties with the facts; if the story were good, it must have been up against some real stinkers to be nominated for the Golden Globe.

Validation and testing took place on multiple multilingual datasets of different domains and sizes, which we discuss in Section 4.2.

4. Experiments and Results

As part of the verification process, we set up the experimental testbeds using WinoMT, Winogender, and Curated-WinoMT datasets for benchmarking and analysis. The Wino-MT dataset [13] is a gender-balanced dataset that comprises at least two gender terms per sentence apart from pronouns such as *he/him* or *she/her*. Other datasets like Winogender suffer from a lack of appropriate gender terms in the text, leading to poor metric calculation and unreliable scores. Hence, to study the absolute impact of gender term co-occurrence, we manually prepared the datasets with varied size and sentence contexts, imitating a balanced representation of male/female-gendered sentences and neutral sentences. Further, the GenBiT metric is evaluated using the sampled datasets of different record sizes and shapes—the estimated scores from experiments as illustrated in Table 2 and Table 3 in the ***Appendix 1***.

4.1 GenBiT Metric Results

Based on the modified algorithmic implementation, we finalized a range of scores post-multi-level rapid experiments. An improvement was observed, which supported our hypothesis(H_0), as shown in Table 2.

Language	Score Range	Data Size	Bias % Indicator (Moderate-high)
EN	0.30-1.0+	>400 Samples	> 0.30
IT	0.50-1.5+	>400 Samples	> 1.00
DE	0.60-2.4+	>200 Samples	> 0.60
ES	0.60-2.5+	>400 Samples	> 0.60
FR	0.50-1.3+	>200 Samples	> 0.60
RU	0.80-2.3+	>400 Samples	> 1.10+

Table 2: GenBiT Score range for biased dataset.
Source: aka.ms/genbitdocs

The score range depicts the degree of biases present in the datasets, indicating that the higher

the score, the more biased the datasets are. The bias percentage indicates the median point score for each language and aids in estimating the degree of bias in datasets ranging from moderate to high.

Table 3 (**Appendix *1***) signifies scores obtained on constraint experiments performed on manually curated datasets using the WinoMT schema. The results demonstrate that as we increase the bias in the datasets, there is a positive correlation on the score range while maintaining stable scores. Lastly, Table 4 (***Appendix 1***) represents the evaluation of GenBiT metrics on the Winogender schema. The Winogender base file consists of 720 records. Next, to study the impact in the score's ranges, the base datasets were inflated through random sampling sentence augmentation techniques to calculate the reflections of average bias condition absolute score vs. percentage of male/female-gendered definition word.

4.2 Bias Mitigation and Impact analysis

The experiments below show the impact of our mitigation technique. The experiments involved testing the approach across different datasets and measuring the distribution of the various gender classes based on tagging sentences as 'male,' 'female,' 'mixed gender,' or 'no gender/gender neutral')**.**

For these experiments, we made use of the News Commentary and IMDB datasets[4]. The size of the News Commentary dataset[5] and IMDB is about 24k observations for each dataset. For the News Commentary dataset, the text length can go up to 70 words per each data entry. While the text length for a single review in IMDB dataset can go up to 500 words. Note that each data sample can consist of multiple sentences.

News Commentary Dataset: Experiments results depicted in Table 5 illustrates the impact of the mitigation technique on different sizes of the dataset across English, French, and Italian. The output of the sentence classifier specifically for male-classified and female-classified sentences improved significantly after mitigation on different corpora sizes as in Table 5, with minimal impact on mixed and gender-neutral samples, indicating a more balanced dataset.

The sentence classifier helped to identify the weight of the less biased class and resulted in a more targeted accurate data augmentation to the class of interest. Table 4.2.3 presents the impact of our approach using mBERT on the bias scores along with the percentage of male-gendered words; the two metrics indicate whether the percentage of bias in the dataset is significant or not. An improvement noted in the bias scores post-mitigation technique was applied to the datasets. A notable improvement for the more biased languages like the Italian version of the News Commentary dataset exhibits more bias than English and French. After applying our mitigation technique, it further resulted in a higher bias score to achieve a 29% drop compared to the 4% drop in the bias and 17% drop for French. Hence, we observed a correlation between the percentage of bias in the data and the impact of mitigation. We also explored the effect of mitigation on other metrics like the number of words and frequency of gender words. An improvement in the scores ranges observed that aligns with the progress of the bias score. Some languages displayed a more female bias for specific corpora, like French, which exhibits female bias for the News Commentary dataset across different data sizes.

As part of the mitigation study, we explored the impact of using the resulting mitigated balanced

[4] https://www.imdb.com/interfaces/

[5] https://opus.nlpl.eu/News-Commentary.php

News Commentary		Before Mitigation						After Mitigation						
Lang	**sample size**	**Female**	**Male**	**Mixed**	**Neut.**	**%male words**	**bias score**	**Female**	**Male**	**Mixed**	**Neut.**	**%male words**	**bias score**	**rel. change**
EN	6k	139	483	41	11337	**82%**	**1.27**	424	486	47	11387	**58%**	**1.26**	**-0.8%**
	12k	220	972	70	22733	**79%**	**1.33**	856	974	80	22837	**58%**	**1.30**	**-2.3%**
	24k	52	217	15	5716	**81%**	**1.58**	175	217	17	5756	**56%**	**1.52**	**-3.8%**
FR	6k	544	194	16	11246	**31%**	**1.33**	544	463	20	11323	**39%**	**1.17**	**-12%**
	12k	1096	377	31	2249	**31%**	**1.45**	1097	907	34	22676	**40%**	**1.26**	**-13%**
	24k	261	96	8	5653	**31%**	**1.67**	261	205	14	5685	**40%**	**1.39**	**-17%**
IT	6k	154	3203	89	8554	**95%**	**1.84**	2325	3256	253	9215	**65%**	**1.33**	**-28%**
	12k	267	6377	160	17191	**95%**	**1.94**	4502	6512	470	18621	**65%**	**1.41**	**-27%**
	24k	79	1601	36	4284	**95%**	**2.05**	1039	1638	112	4733	**65%**	**1.43**	**-30%**

Table 5: Sentence gender classifier distributions (female, male, mixed, neutral), % of male gender definition words and bias scores before and after mitigation on News Commentary dataset across different languages and sample sizes.

data created using our approach on the performance of machine learning classifiers. The results demonstrated a significant improvement in the performance of a classifier model like Support Vector Machine (SVM). The classifier is tested on the augmented version of the News Commentary dataset for English language and then fed into data loader to train the model. Comparison of the f1-score took place between training the original dataset and training the augmented version through leveraging the similarity matrix for the male versus female classes and comparing results pre-post mitigation where the f1-score increased from 0.70 to 0.87.

IMDB Dataset: Experiments shown in Table 6 demonstrates the impact of our technique, especially the contextual augmentation part on datasets with long sentences. Similarly, like News Commentary, experiments showed an improvement in the results of sentence classifier, a drop in the bias scores and percentage of male definition words, as shown in Table 5. Our bias metric is quite sensitive to bias, so a reduction of even 0.1 could be considered significant, this was concluded based on the increase in the bias score that occurred when we doubled the size of the dataset from 6k to 12k observations. For example, News Commentary showed an increase by 0.1 in bias score for French and Italian while maintaining the same percent of male words for the larger dataset. Experiments showed that the more augmentation, the better the impact on the bias score. Thus, in the algorithm, we enabled the option for tuning to allow a more significant number of gendered sentences to be augmented based on the computing environment. It further facilitates running large language models like mBERT on Azure ML artifacts or using multiple GPUs on Azure.

IMDB		Before Mitigation				After Mitigation				
Lang	**sample size**	**Female**	**Male**	**%male words**	**bias score**	**Female**	**Male**	**%male words**	**bias score**	**rel. change**
EN	24K	1095	2795	0.60	**0.51**	2586	2807	0.54	**0.77**	**-51%**
DE	24K	617	1625	0.31	**1.45**	1216	1626	0.59	**0.50**	**-61%**
IT	24k	543	12607	0.84	**0.89**	108	12663	0.76	**0.86**	**+3%**

Table 6: Sentence gender classifier distributions for female vs male and bias scores before and after mitigation on IMDB dataset across different languages.

5. Conclusions and Next Steps

The paper provides details around the implementation of how practitioners can detect gender bias in datasets for multiple different languages. The work addresses the challenge of grammatical gender in non-English languages by designing a generalizable solution scalable across other languages.

The interpretation of the score is highly dependent on two key factors: a) percentage of male or female gendered definition words contained in the corpus, and b) average bias conditional absolute score. Both the data points prove to be most helpful in delivering an understandable and actionable metric.

The benchmarking across a manually curated multilingual dataset demonstrated the stability and robustness of the GenBit metrics. We artificially introduced gender bias in the dataset via targeted oversampling. We observed corresponding bias reflected in the calculated metrics correlated with the increased occurrence of male-gendered definition words present in the text.

The paper proposed an ensemble mitigation approach using the combined techniques of data augmentation and contextual word embeddings using language models, which generally showed better improvements (i.e., reduction) in the gender bias score across different corpora and languages. The approach is generic for any bias type and can adequately handle both male-directed and female-directed gender bias. Further improvements can be made by allowing mBERT to augment a larger volume of sentences.

The flexible sequential augmentation approach can act as an extension to allow the user to choose either a single or multiple ways of augmentation to take advantage of resources like WordNet for a word replacement and mBert.

As a future direction for this study, we are pursuing the adoption of GenBiT based data sampling that enables the selection of more gender-balanced datasets in a compliant fashion when training and fine-tuning significant representation models provided by the Turing ELR (Enterprise Language Representation) team.

Further, it would be possible to extend our measurement and mitigation approaches to other categories of bias and compare the impact relative to gender bias. Also, a comparative analysis can be beneficial to compare our method with different existing approaches like gender-swapping, especially across languages other than English.

6. Data & Code

A thorough benchmarking activity was carried out as part of our research activities to study the correlation of score ranges across different datasets. The study helped us examine how gender bias could influence the overall ML task using multilingual parallel datasets (Winogender-schema, WinoMT, Curated-WinoMT, IMDB, News Commentary, TedTalks, and few others). Users could access the documentation for the API and reference results from the benchmarking via https://aka.ms/genbit and https://aka.ms/genbitdocs. The documentation provides pointers to sample datasets (WinoMT-curated) as reference for quick evaluation and experimentation with a Python Notebook of assessment and replication of the work described in this paper.

References

[1] D. Hovy and S. Spruit, "The social impact of natural language processing.," *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,* vol. 2, no. Short Papers, 2016.

[2] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys,* vol. 56, no. 6, pp. 1-35, 2021.

[3] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in PMLR 81:77-91,* 2018.

[4] S. Leavy, G. Meaney, K. Wade and D. Greene, "Mitigating Gender Bias in Machine Learning Data Sets. In: Boratto L., Faralli S., Marras M., Stilo G. (eds) Bias and Social Aspects in Search and Recommendation. BIAS 2020.," *Communications in Computer and Information Science,* vol. 1245.

[5] D. Leslie, A. Mazumder, A. Peppin, K. W. M and A. Hagerty, "Does “AI” stand for augmenting inequality in the era of covid-19 healthcare? BMJ 2021; 372 :n304 doi:10.1136/bmj.n304," *BMJ,* vol. 372:n304, 2021.

[6] Blodgett, "Language (Technology) is Power: A Critical Survey of “Bias” in NLP," *University of Massachusetts Amherst,* 2020.

[7] J. Zhao, S. Mukherjee, S. Hosseini, K. Chang and A. & Awadallah, "Gender Bias in Multilingual Embeddings and Cross-Lingual Transfer," *ACL,* 2020.

[8] K. Crawford, "Artificial intelligence white guy problem. The New York Times," 2016.

[9] B. Shikha and B. S. R, "Identifying and Reducing Gender Bias in Word-Level Language Models," *Association for Computational Linguistics,* pp. 7--15, 2019.

[10] K. Lu, P. Mardziel, F. Wu, P. Amancharla and D. A, "Gender Bias in Neural Natural Language Processing," *In: Nigam V. et al. (eds) Logic, Language, and Security. Lecture Notes in Computer Science,* vol. 12300, 2020.

[11] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief and J. e. a. & Zhao, "Mitigating Gender Bias in Natural Language Processing: Literature Review," *Proceedings Of The 57Th Annual Meeting Of The Association For Computational Linguistics,* 2019.

[12] A. Buslaev, V. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin and K. AA, "Albumentations: Fast and Flexible Image Augmentations," *Information,* vol. 11, no. 125, 2020.

[13] G. Stanovsky, A. Smith and L. Zettlemoyer, "Evaluating Gender Bias in Machine Translation," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* p. 1679–1684, 2019.

Appendix 1

Freq cutoff	Num words considered	Freq. of male-gendered words	Freq. of female-gendered words	Percentage of male gendered words	Avg bias ratio absolute	Avg conditional bias absolute	# Of samples	Remarks
8.515202	78	191	187	0.505291	0.1607549	0.1629159	300	Balanced
8.27938	88	219	215	0.504608	0.1814063	0.1822537	350	Balanced
8.227701	112	296	291	0.504259	0.175754	0.176689	500	Balanced
8.241381	113	296	295	0.500846	0.194678	0.194864	750	+ 100 **Neutral Sentence**
8.5645	114	300	390	0.434783	0.328916	0.308218	850	**+100 female-gendered sentences**
8.976574	129	300	486	0.381679	0.536571	0.439762	950	**+100 female-gendered sentence**
8.976574	135	502	391	0.56215	0.342696	0.309756	1050	**+200** male-gendered sentences
8.976574	149	620	391	0.613254	0.511666	0.43284	1150	**+100** male gendered sentences
8.976574	157	681	391	0.635261	0.583734	0.478745	1200	**+50** male gendered sentences

Table 3. Benchmarking of Genbit on Curated-WinoMT dataset

Freq cutoff	Num words considered	Freq. of male gendered words	Freq. of female-gendered words	Percentage of male gendered words	Avg bias ratio absolute	Avg conditional bias absolute	% Of artificial bias infused	language
1599.135298	70	35461	15600	0.6944830	0.96804	0.216496	100	EN
9584.811785	70	212761	93600	0.6944780	0.96900	0.210316	120	EN
63464.77343	74	1489321	655200	0.6944772	0.97788	0.234303	140	EN
17.2	40	224	41	0.845283	2.13985	1.886913	100	ES
17.21378094	39	230	45	0.836363	2.20611	1.968895	200	ES

Table 4: Experiment results from Wino-gender-schema-master dataset through the introduction of artificial bias in the dataset