Curation of Benchmark Templates for Measuring Gender Bias in Named Entity Recognition Models

Ana Cimitan[1], Ana Alves Pinto[2], Michaela Geierhos[3]
University of the Bundeswehr Munich[1,3], ZITiS[2]
acimitan@computer.org[1], ana.alvespinto@zitis.bund.de[2], michaela.geierhos@unibw.de[3]

Abstract

Named Entity Recognition (NER) constitutes a popular machine learning technique that empowers several natural language processing applications. As with other machine learning applications, NER models have been shown to be susceptible to gender bias. The latter is often assessed using benchmark datasets, which in turn are curated specifically for a given Natural Language Processing (NLP) task. In this work, we investigate the robustness of benchmark templates to detect gender bias and propose a novel method to improve the curation of such datasets. The method, based on masked token prediction, aims to filter out benchmark templates with a higher probability of detecting gender bias in NER models. We tested the method for English and German, using the corresponding fine-tuned BERT base model (cased) as the NER model. The gender gaps detected with templates classified as appropriate by the method were statistically larger than those detected with inappropriate templates. The results were similar for both languages and support the use of the proposed method in the curation of templates designed to detect gender bias.

Keywords: BERT, masked token prediction, gender gap

1. Introduction

Assessment of gender bias in language models (LMs) has been performed in several previous studies through the use of benchmark datasets, with bias being detected at the model prediction level (Zhao et al., 2018; Rudinger et al., 2018; Webster et al., 2018; Kiritchenko and Mohammad, 2018; Stanovsky et al., 2019; Escudé Font and Costa-jussà, 2019; Costa-jussà et al., 2020; Mehrabi et al., 2020). Despite of the large number of such benchmark datasets, their curation remains a challenge, as they are specifically tailored to test a particular use case scenario, e.g., they support a particular NLP task and/or are of a particular language register. Thus, there is no 'one size fits all' approach, and curating benchmark datasets can still be a time-consuming and resource-intensive task. The evaluation of gender bias is particularly important in the context of NER, as the latter is widely used to detect, among other things, person entities. Previous research has shown that NER models are sensitive to the context in which entities are embedded, a result with implications also for bias assessment using benchmark datasets. For NER, especially Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) have been shown to achieve excellent performance on common NER datasets, such as OntoNotes 5.0 and CoNLL-2003 (Li et al., 2022). However, despite these remarkable results, two adversarial attack studies showed that BERT models for NER are sensitive to the context of entities (Simoncini and Spanakis, 2021; Dirkson et al., 2022). In both studies, adversarial attacks were successfully performed, causing the model to incorrectly predict the named entities after manipulating the context of the entities at the character, word, or sentence level. In light of this research, it is important to examine, in benchmark templates used to assess gender bias in BERT models, the context in which entities are immersed, as well as to investigate *How to evaluate whether templates are appropriate for bias assessment?*.

The experimental work presented here provides, to the best of our knowledge, one of the first investigations into the robustness of benchmark datasets for NER and on how such datasets can be hardened for bias detection. We propose a method for filtering benchmark data, which we will here call the *'Fill-Mask Templates Filter'*. The method aims to identify benchmark templates that have a higher probability of detecting gender bias in BERT-based NER models, thus enabling more resource-efficient curation of benchmark datasets through low-cost pre-filtering of the templates. The Fill-Mask Templates filter proposed here is based on masked token prediction and tested on language-specific BERT (`BERT-base-cased`; Devlin et al. (2019)) models, fine-tuned for NER on WikiNEuRal (Tedeschi et al., 2021) for English and German. These languages were chosen because the gender context plays a different role in these two languages (in German, for example, gender is implicit in other context words), and the research group has linguistic expertise in both English and German.

LREC-COLING 2024, pages 4238–4246
20-25 May, 2024.

The effectiveness of the proposed Fill-Mask Templates Filter is evaluated by computing the false negative rate (FNR) during person-type entity detection and by assessing the gender gap based on all 411,000 unique permutations of 300 different templates. Finally, we examine the templates for lexical complexity and linguistic features and compare these for the templates that passed/failed the Fill-Mask Templates Filter.

The paper is organized as follows: In the following section, we present related work, which is divided into two parts. The first part deals with the existing research on the evaluation of bias in BERT models based on predicted token probability, while the second part deals with work analyzing BERTs ability to identify entities by their contextual information. In the third section we describe the Fill-Mask Templates Filter, which forms the methodology used here to filter benchmark data. The experimental implementation, described in Section 4, starts with the specification of pre-trained base models and their fine-tuning for NER. It then proceeds to the curation of the benchmark data and concludes with the computation of the gender gaps. The results are presented in the fifth section, and in Section 6 we discuss, in face of the results obtained, the robustness and reliability of the data-related benchmark approaches for bias testing of LMs. In the seventh section we draw the conclusions of the study and provide an outlook for future work. Finally, in Section 8, we outline the limitations of this work, and in Section 9 we address the ethical implications.

2. Related Work

Given the lack of published research on the robustness of benchmark templates for bias assessment, this section provides a brief overview of the work worth mentioning in the context of the current study. In the following, we focus on previous research that addresses two aspects that inspired our work: (i) bias research by means of token probability measured by fill-in-the-[blank/mask/gap] style templates. Here, we limit the review to studies that include BERT; and (ii) context-based NER in BERT, with special attention to the identification of person-type entities by their contextual information.

2.1. Bias Measurement

Log Probability Bias Score (LPBS). Kurita et al. (2019) measured gender bias based on the masked token prediction task. They proposed the LPBS to quantify the amount of bias in BERT embeddings by querying the underlying pre-trained model. Using self-curated sentences as templates, e.g. `[TARGET] is a [ATTRIBUTE]`, they estimated the association between each of two *targets* (e.g. he/she) and one *attribute* (e.g. programmer). The difference between the calculated associations for two targets he/she represents the LPBS and can be used to estimate gender bias.

Discovery of Correlations (DisCo). Webster et al. (2020) introduced DisCo, an intrinsic analysis to detect and measure gendered correlations in pre-trained contextual representations. DisCo is based on a set of templates, such as `[PERSON] studied [BLANK] at college.`, containing two predefined slots. While the `[PERSON]` slot is filled with elements taken from lists of gender-related words (first names or gendered nouns), the `[BLANK]` slot is used as a placeholder, for which a model under test is asked to generate any vocabulary item. Thus, an indication of bias is obtained if the predictions differ in gender.

Context Association Test (CAT). Nadeem et al. (2021) proposed CAT, a test that assesses stereotypical biases in pre-trained language models. Together with the associated crowdsourced StereoSet dataset, consisting of 16,995 CATs, the test can assess a model's preference for ranking stereotypical contexts higher than anti-stereotypical contexts. The StereoSet contains two types of CATs: intrasentence and intersentence. An intrasentence CATs consists of a fill-in-the-blank sentence template and a set of three attributes: a stereotypical, an anti-stereotypical, and an unrelated attribute. Biases can be estimated by determining which attribute has the highest probability of filling the gap. In intersentence CATs a sample consists of a set of sentences. A context sentence related to a target group and several attribute sentences: a stereotypical, an anti-stereotypical and an unrelated sentence. The bias can be assessed based on which attribute sentence is more likely to be a successor to the context sentence. The StereoSet dataset contains CATs to test four forms of stereotype bias: gender, occupation, race, and religion.

In the context of this work, another method for bias assessment has been proposed (Nangia et al., 2020; Kaneko and Bollegala, 2022). Both approaches are based on the pseudo-log-likelihood (PLL) (Wang and Cho, 2019; Salazar et al., 2020) score, where a sentence is iterated with each token being masked one by one, and the log likelihoods of all masked tokens are computed and summed up. Both approaches use data sets with sentence pairs containing a stereotype and an anti-stereotype sentence (Nangia et al., 2020; Kaneko and Bollegala, 2022). Although bias is estimated by applying masking, we do not investigate these approaches further here because they do not make use of fill-in-the-gap templates.

2.2. Named Entity Recognition

Recently, Davody et al. (2022) described how pretrained BERT model can be used for NER by giving the model some contextual information about the entity to be labeled. The authors introduced a cloze-style approach that can be used as a standalone, zero- or few-shot classifier for NER.The last approach proposed by Davody et al. (2022), the few-shot approach, will not be discussed here, as it was not crucial for our experimental design.
The basic method of Davody et al. (2022) takes advantage of a low-cost classification using the pretrained LM by giving the model a cloze-style query, for example,

> *'I would like to visit Munich next week. Munich is a **[MASK]**.'*
> *(Davody et al., 2022)*

While the first part of the prompt – *'I would like to visit Munich next week.'* – provides the contextual information about the entity, the second part – *'Munich is a **[MASK]**.'* – serves as a predefined template. In this example, the following 5 tokens were predicted by BERT (large): `city`, `success`, `democracy`, `capital`, `dream`, with `city` having the highest probability value (0.43) of all predicted tokens. Using a representative word list defined for each entity (Person, Organization, City, Ordinal and Date), the prediction can then be correctly mapped to an entity type. Since `city` is part of the representative word list of the entity Location[1], *'Munich'* would be correctly mapped to the entity type Location. Therefore, the trained model predicts the probabilities for all tokens contained in the representative word lists and the entity type containing the word with the highest predicted probability is selected. Davody et al. (2022) evaluated this approach among several LM architectures and the experiments showed that BERT in particular can identify person-type entities based on context alone with a high performance (F-score of 80 %).

3. Method

The results provided by Davody et al. (2022) raise the question: *When does the context of an entity in a sentence appear as sufficiently personal to the model?* Or: *When is a sentence too simple?* Inspired by their work, we propose the Fill-Mask Templates Filter, a method based on masked token prediction that allows filtering for benchmark templates that carry a high level of person-related context and thus may be unsuitable for detecting gender bias in LMs.

How the process works. Given a potential benchmark sentence template, the first step of the Fill-Mask Templates Filter method is to replace the demographic identifier with a [MASK] token. This sentence is then fed into the `fill-mask` pipeline[2], which retrieves the token with the highest probability for the masked demographic identifier. Based on this token, the Fill-Mask Templates Filter classifies the template context as person-related (PR) or non-person-related (NPR) using representative words. The flowchart in Figure 1 describes the steps taken during the implementation of the Fill-Mask Templates Filter to categorize a template.

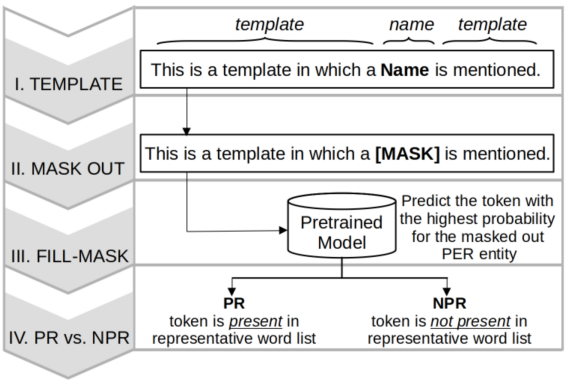

Figure 1: Diagram of the implementation steps performed to classify a template as PR vs. NPR by the Fill-Mask Templates Filter.

A representative list of words. Similar to Davody et al. (2022), who defined representative word lists for each entity type to perform NER entity mapping, we defined two language-specific representative word lists that aim to linguistically reflect the high-level concept of the entity person. Davody et al. (2022) suggested a number of possible strategies for generating representative word lists. The authors also emphasized that the approaches can be combined, which we did in our experiment. For the following implementation, we took advantage of two curation strategies suggested by Davody et al. (2022): (i) the use of appropriate prompts to curate the representative word list using the most likely mask filler for each entity, and (ii) the curation of the representative word list by a domain expert.
According to strategy (ii), we built on the representative word list[3] for the person-type entity designed by Davody et al. (2022). When following strategy (i), we extended the word lists with a list of pro-

[1] location, city, country, region, area, province, state, town (Davody et al., 2022)

[2] https://huggingface.co/docs/transformers/main_classes/pipeline#transformers.FillMaskPipeline

[3] person, man, woman, boy, girl, human, someone, kid (Davody et al., 2022)

nouns[4], since we observed the words – he/she [EN]; er/sie [GER] – as the most likely person-related mask fillers predicted by the models.

Categorization of PR vs. NPR. Based on whether the most likely mask filler predicted by the `fill-mask` pipeline for a given template is included in the representative word list or not, the template is classified as carrying person-related context. Thus, the classification can be summarized as follows:

- PR - most likely mask filler is *present* in representative word list
- NPR - most likely mask filler is *not present* in representative word list

4. Implementation

The flowchart in Figure 2 describes the three experimental phases of the study: (i) model training, definition of (ii) benchmark templates, and (iii) gender gap calculation.

(i) Model Training. The first phase consisted of fine-tuning the two language-specific BERT models to the NER task. In order to achieve equivalent experimental conditions for the two languages tested – English and German – a dataset was chosen that is available for both languages. Thus, both monolingual `BERT-base-cased` models were fine-tuned for NER with the corresponding WikiNEuRal (Tedeschi et al., 2021) training dataset split. For more detailed information on the model training and the statistics of each dataset split, see the Model Training in Appendix.

(ii) Benchmark Templates. The benchmark templates used to assess gender bias in NER models typically consist of two main elements: a person mention that the NER model should identify as a person-type entity, and a sentence context in which the person mention is embedded in (Mehrabi et al., 2020). We will refer to them here as *(a) first name lists* and *(b) sentence templates*.

(a) First Name Lists. The English and German first name lists were obtained from two public resources that track first name statistics in the United States and Germany, respectively. For the United States, we used the 'Baby Names from Social Security Card Applications'[5] dataset, which contains all names from Social Security card applications after 1879. For Germany, we retrieved the data from the 'Statistics of the Most Common First Names for Each Birth Year'[6]. From both datasets, we derived the 500 most popular first names from 2008 to 2021. This is the largest time range common to both datasets, with the highest number of first names registered. Additional features such as year of birth, assigned gender at birth (AGAB), and the popularity of the name, represented by the frequency of the first name, were also taken into account. We merged the first names of these fourteen years according to AGAB for each country and reduced them to the unique first name mentions while preserving the popularity/frequency ranking. In order to get first name lists of equal length for both languages, we cut the name lists of all four [AGAB, country] groups to the size of the shortest of these lists, which was 685 first names.

(b) Sentence Templates. To select the *(b) sentence templates*, we ran the Fill-Mask Templates Filter and created two sets of templates for each language – person-related (PR) and non-person-related (NPR) templates. The textual data for the sentence templates were extracted from the WikiNEuRal (Tedeschi et al., 2021) test data splits to keep in the same domain of the training data. Furthermore, for both languages, the test-dataset splits were reduced to sentences with exactly one person-type annotation[7], which is referenced in a gender-neutral way to ensure correct syntax and semantics of the sentence. For the Fill-Mask Templates Filter implementation, entities annotated as person were replaced with the [MASK] token. Each of the [EN/GER, PR/NPR] groups contains 75 templates randomly selected from the language-specific PR and NPR bins, for 150 templates per language and 300 templates in total.

(iii) Gender Gap Calculation. Finally, in the last phase of the study, we calculated the gender gap for each template. First, we composed language-related benchmark templates by inserting 685 first names of each AGAB group from the *(a) first name lists* into each of the 300 *(b) sentence templates*. This means 685 variations of a composed template (first name embedded in sentence template) for each AGAB group and thus 1,370 variations per template in total. We then determined for each template and each group, female assigned at birth (F_{emale}AAB) and male assigned at birth (M_{ale}AAB),

[4] 1st person, 2nd person, 3rd person [cased and uncased]; for both language-specific lists, see the Representative Word Lists in Appendix

[5] https://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-national-data

[6] https://www.beliebte-vornamen.de/22932-wissenschaftliche-quelle.htm

[7] WikiNEuRal is annotated according to the IOB2 tagging scheme. According to this scheme, a person entity can consist of several person-type tags: at least one B-PER (beginning of entity) and several I-PER (is inside entity). Since in the course of this paper we perform the bias testing with first names only, person entities consisting of multiple PER tags were reduced to a [MASK] token before executing the Fill-Mask Templates Filter.

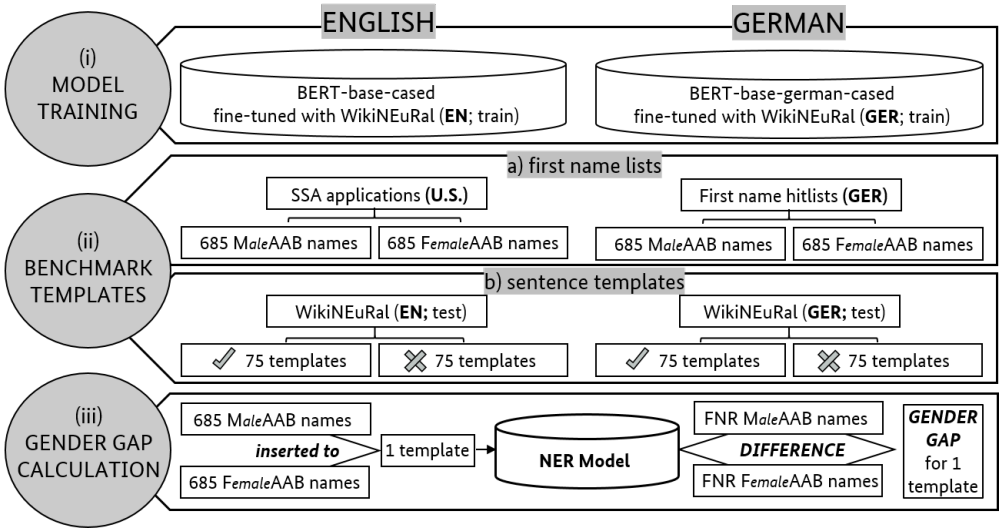

Figure 2: Diagram of the experimental steps performed to assess gender bias in BERT fine-tuned for NER with PR and NPR templates.

how many first name mentions were not identified as person-type entities by the NER model – FNR, as well as the corresponding type of bias (F_{emale}AAB bias, if the FNR was larger for this group, or M_{ale}AAB bias otherwise).

5. Results

PR vs. NPR Templates. As shown in Figure 3, for both languages, English and German, the gender gap (in percentage values) obtained with templates classified as NPR by the Fill-Mask Templates Filter was greater than that obtained with templates classified as PR. A Mann-Whitney U test, performed separately for each language, showed that the difference between the template groups was significant at the p = 0.05 level. According to these results, an existing gender bias in the trained BERT model is more likely to be detected when scoring is done with previously classified NPR templates. In other words, these results suggest that NPR templates are better at detecting potential gender bias in a BERT NER model.

Lexical Complexity. To investigate possible reasons for the difference in results between PR and NPR templates, we evaluated several linguistic features for the two template groups (cf. Table 1). In addition to raw text features such as word and sentence length, morpho-syntactic and syntactic features were also evaluated. For example, lexical density, calculated as the ratio of content words (verbs, nouns, adjectives, and adverbs) to the total number of tokens in the template. As for syntactic features, we evaluated the depth of the parse tree, calculated as the longest path from the root of the dependency tree to a leaf. The Flesch Reading Ease score (Flesch, 1979) is a commonly used measure of the readability of a text, with higher values indicating an easier to understand text. All scores were calculated separately for each sentence and reported as mean (M) and median (Mdn) values. The set of features considered in Table 1 captures different aspects of sentence complexity. Overall, with the exception of lexical density, the computed values indicate a slight tendency towards longer sentences, greater parse tree depth and lower readability for the NPR templates compared to the PR templates. This means that NPR templates, which have a higher probability of being detected, have a higher linguistic complexity. The largest difference occurred in the readability score.

Bias Direction. For 271 out of 300 templates, a higher F_{emale}AAB false negative rate was obtained, while 20 had higher M_{ale}AAB false negative rate values, and 9 templates had similar false negative rate values for M_{ale}AAB and F_{emale}AAB (i.e., no gender gap). The highest M_{ale}AAB-related gender gap achieved was about 10 % (gray square in Figure 3, for German). Thus, the 300 templates and 411,000 template permutations tested here do not provide a clear evidence of a specific bias direction, but rather a tendency of both BERT NER models to perform better in identifying M_{ale}AAB first names, namely in 90.33 % of the test cases. The gender gap analysis presented here refers only to the two models trained in the course of the experimental implementation and serves as a proof-of-concept for the introduced Fill-Mask Templates Filter.

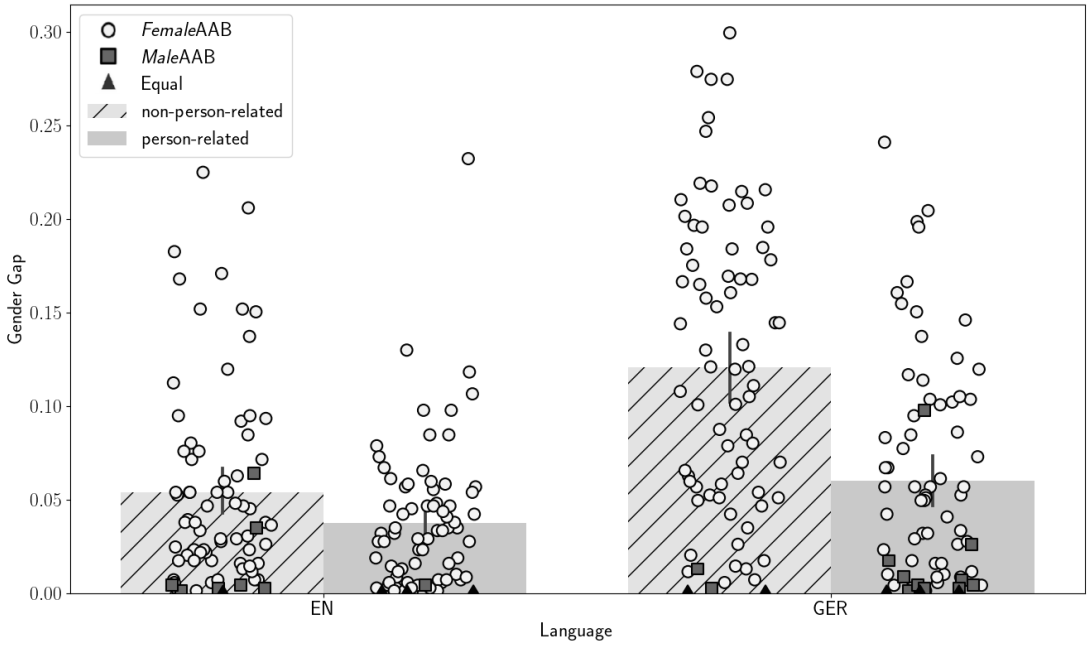

Figure 3: Gender gaps (module of the difference between FNR for F_{emale}AAB and M_{ale}AAB groups in %) for PR and NPR templates. Circles/squares indicate F_{emale}AAB/M_{ale}AAB bias respectively. Black triangles indicate no gender gap.

Language	EN				GER			
Classification	PR		NPR		PR		NPR	
Summary Statistics	M	Mdn	M	Mdn	M	Mdn	M	Mdn
Word Length	4.11	4.18	**4.22**	**4.19**	4.97	4.87	**5.11**	**5**
Sentence Length	14.89	13	**19.23**	**17**	12.05	11	**15.69**	**13**
Lexical Density	0.38	**0.39**	0.38	0.38	**0.38**	0.38	0.37	**0.39**
Max. Depth of Parse Tree	4.4	4	**5.25**	**5**	4.28	4	**4.81**	4
Flesch Reading Ease	56.32	61.34	**36.34**	**44.08**	58.25	64.5	**38.05**	**55.5**

Table 1: Assessment of linguistic features.
PR=person-related; NPR=non-person-related; M=mean; Mdn=median

6. Discussion

A method, called here Fill-Mask Templates Filter, is proposed to filter out sentence templates to be used in the assessment of gender bias in BERT-based NER models. A larger gender gap was found when sentences classified as NPR by the method were used in a person detection task compared to sentences classified as PR. These results support previous findings on the vulnerability of BERT models to entity context Simoncini and Spanakis (2021) and Dirkson et al. (2022), and support the need to curate benchmark templates for bias with particular attention to their suitability for detecting bias. Furthermore, the results presented also highlight the weaknesses of bias evaluation using benchmark data in LMs. A major drawback of this approach, which is also highlighted by the results of this study, is that the results can be difficult to interpret if the direction of the gender gap is ambiguous. In addition, benchmark datasets for bias detection should be curated to cover the actual demographic distribution of the use case under consideration, otherwise no bias will be detected. Demographics, represented by demographic identifiers such as first names, remain a challenging factor to reflect in linguistic expression. However, in a black-box scenario, without further knowledge of or access to a NER model, bias assessment using benchmark datasets may be the only option available to justify the use of this approach. Thus, additional research is needed to better understand data-based bias evaluation of LMs, e.g., regarding appropriate benchmark data size, appropriate demographic reflection, and interpretation of ambiguous results.

7. Conclusion and Future Work

The assessment of gender bias is particularly important in the context of NER, since it is widely used to detect, among other things, person entities. Therefore, we have paid special attention to this task. However, despite the promising results for NER, the question of the generalizability of the method presented here remains unanswered. Further research is needed to evaluate the extent to which these results can be transferred to other model architectures and other NLP tasks.

In addition, this study provided an initial assessment of the quality of the benchmark data, which we consider to be a first step in future research on how to evaluate such templates for their ability to detect bias in LMs. In the work presented here, this evaluation was based on the most likely mask filler predicted by the model. A natural extension of this work is to extend the analysis to the n-most likely mask fillers, i.e., the top-5 or top-10. One possible shortcoming that could be associated with such an extension from just one parameter (the *one* most likely mask filler) to an entire parameter list (a ranked *list* of mask fillers) is the dilution of the results. For example, if a person-related word is ranked 10th in the list of the most likely mask fillers with a very low probability, the template is still classified as a PR template. A complementary metric, e.g., the probability scores of the proposed masked fillers, can be additionally considered in this case.

In order to develop a complete picture of which mask fillers can be considered as person-related, additional work is needed to examine more person-related words. The results provide initial indications that the gender gap tends to be smaller when the most likely mask filler is a person's first or last name, or even an artistic pseudonym. This was also observed for other person-related words, such as a person's job title. The results of these observations suggest that a wordlist-based approach has its weaknesses due to the large variety of person-specific words, and that at least its extension is an important topic for future research.

8. Limitations

This study was limited by the lack of resources that would allow it to adequately reflect the non-binary community. Two statistical resources were used to curate the *(a) first name lists*, both of which capture the binary assigned sex at birth AGAB. Thus, this study, being limited to these two groups, F_{emale}AAB and M_{ale}AAB, lacks the representation of a non-binary community. An additional uncontrolled factor related to the two statistical sources mentioned is the possibility that they do not reflect the actual demographic distributions in the two countries in a coherent and comparable way. For the German list, the data were curated from a representative sample of birth reports throughout the country. Thus, the sex assignment can be considered as AGAB. The English list was compiled from statistics on Social Security card applications for births that occurred in the United States. Therefore, we assume that the sex assignment of this data set is also AGAB. Third, the scope of this study was limited to the type of bias analyzed. Although the proposed method could be applied to template selection for other types of social bias benchmarking, we evaluated its applicability to gender bias detection. A natural extension of this work would be to analyze its generalizability to other social biases.

Ethics Statement

The propagation of gender bias from LMs to the real world is an undisputed possibility that can have negative consequences for humans. The method described here can contribute to the development and application of fairer algorithms by allowing NER models to be benchmarked for gender bias more efficiently. Given this, the authors see no ethical concerns with the research presented in this paper. However, there is one aspect that we would like to clarify regarding the gender groups considered here. In the course of evaluating the gender gap, we focused on a representation of the F_{emale}AAB and M_{ale}AAB groups that restricts gender to a binary model. As mentioned in Limitations, the experimental setup for the evaluation of the Fill-Mask Templates Filter relies on country-specific statistical resources about infants' first names. To our knowledge, there are no comparable resources that cover both countries and support a non-binary model for gender classification. This limitation hinders an accurate reflection of the non-binary community in the gap analysis.

Acknowledgements

The work described in this paper is performed in the H2020 project STARLIGHT ("Sustainable Autonomy and Resilience for LEAs using AI against High priority Threats"). This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 101021797.

9. Bibliographical References

Marta R. Costa-jussà, Pau Li Lin, and Cristina España-Bonet. 2020. GeBioToolkit: Automatic extraction of gender-balanced multilingual corpus of Wikipedia biographies. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4081–4088, Marseille, France. European Language Resources Association.

Ali Davody, David Ifeoluwa Adelani, Thomas Kleinbauer, and Dietrich Klakow. 2022. Token is a mask: Few-shot named entity recognition with pre-trained language models. In *International Conference on Text, Speech and Dialogue*.

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

Anne Dirkson, Suzan Verberne, and Wessel Kraaij. 2022. Breaking bert: Understanding its vulnerabilities for named entity recognition through adversarial attack.

Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy. Association for Computational Linguistics.

R. Flesch. 1979. *How to Write Plain English: A Book for Lawyers and Consumers*. Harper & Row.

Masahiro Kaneko and Danushka Bollegala. 2022. Unmasking the mask – evaluating social biases in masked language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):11954–11962.

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.

Lars Kjeldgaard and Lukas Nielsen. 2021. Nerda. GitHub. Available at `https://github.com/ebanalyse/NERDA`.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, and Aram Galstyan. 2020. Man is to person as woman is to location: Measuring gender bias in named entity recognition. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, HT '20, page 231–232, New York, NY, USA. Association for Computing Machinery.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

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

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Walter Simoncini and Gerasimos Spanakis. 2021. SeqAttack: On adversarial attacks for named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 308–318, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *ArXiv*, abs/2010.06032.

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.

A. Appendix

A.1. Representative Words List

The English words highlighted in *italic* are collected by Davody et al. (2022) as a representative word list for the person entity type. Their German translations have also been added to the German list.

English: I, me, my, mine, myself, you, your, yours, yourself, he, him, his, himself, she, her, hers, herself, *person, man, woman, boy, girl, human, someone, kid*

German: ich, mich, mir, mein, meiner, du, dich, dir, dein, deiner, sie, ihr, ihrer, er, ihn, ihm, seiner, *Person, Mann, Frau, Junge, Mädchen, jemand, Kind*

A.2. WikiNEuRal Dataset Statistics

Split	Art.	Sent.	Tok.	μLen.	μNEs
EN	50k	116k	2.73M	23.53	1.67
GER	50k	124k	2.19M	17.66	1.42

Table 2: WikiNEuRal dataset statistics (Tedeschi et al., 2021).

Split	PER	ORG	LOC	MISC	Other
EN	51k	31k	67k	45k	2.4M
GER	60k	32k	59k	25k	1.87M

Table 3: WikiNEuRal entities statistics (Tedeschi et al., 2021).

A.3. Model Training

As a pre-trained model, `bert-base-cased` and `bert-base-german-cased` were used. For the implementation of the NER task we used the NERDA framework (Kjeldgaard and Nielsen, 2021).