Lexical Gender Made Simple: A Scalable Methodology for Gender Detection with Online Lexical Databases

Anonymous ACL submission

Abstract

The evaluation of gender bias in Natural Language Processing relies on the use of gendered expressions, such as pronouns and words with lexical gender. Up until this point, researchers have manually compiled lists that record lexical gender for individual words. However, manual compilation leads to static information if lists are not periodically updated and categorization requires value judgements by annotators and researchers. Moreover, words that are not covered by the list fall out of the range of analysis. To address these issues, we devised a dictionary-based method to automatically detect lexical gender that can provide a dynamic, up-to-date analysis with high coverage. Our approach reaches 90% accuracy in determining the lexical gender of words retrieved randomly from a Wikipedia sample, and when testing on a manually compiled list that the method aims to replace.

1 Introduction

Within the field of Natural Language Processing (NLP) there is a growing body of research on gender bias in trained models as well as on allocational and representational harms caused by the deployment of these models. There have moreover been increasing calls for early and thorough data description and curation in order to gain insights into how, for instance, gender stereotyping or quality of service bias is propagated from data into an NLP model. What both these strands of research on gender bias have in common, is their reliance on words related to gender.

In English, gendered words most commonly include pronouns (*he*, *she*, *they*, etc.), and also words that carry lexical gender, such as *boyfriend*, *policewoman* or *prince*. Previous works on gender bias in NLP have mostly used manually compiled lists of words carrying lexical gender to for example mitigate gender stereotyping through data augmentation (Lu et al., 2020), assess trans-exclusionary bias in co-reference annotations (Cao and Daumé III, 2020), or evaluate gender inequalities in Wikipedia article titles (Falenska and Çetinoğlu, 2021). Such manually curated lists, however are limited in their coverage of terms that contain lexical gender and can become outdated if not maintained.

To address this issue, we present a scalable algorithmic method to determine lexical gender by querying a word's dictionary definitions for a small subset of definitively gendered words. Our method allows for high-coverage, instantaneous detection of words carrying lexical gender, which eliminates the need to manually compile and maintain static lists of gendered words. This not only facilitates the extension of previous work on gender bias in NLP, but can also be used for a more detailed analysis on the representation of gender in large-scale language datasets used to train large language models like BERT (Devlin et al., 2019) or GPT-2 (Radford et al., 2019).

By combining the gender labels obtained from Merriam Webster Online (Merriam-Webster, 2021) and WordNet® (Princeton University, 2010), our method reaches an accuracy of 90% in determining the lexical gender of words in a random sample of 150 Wikipedia articles. Using only labels obtained from querying Merriam Webster, the method also reaches 90% accuracy on a list of words carrying lexical gender adapted from previous research. The code for the algorithm along with evaluation methods and datasets will be available upon publication.

In the following sections we outline the conceptions of linguistic gender used in this research and subsequently present an overview of research on gender in NLP that relies on curated lists of gendered words. Section 3 gives a detailed overview of the algorithm and Section 4 introduces the datasets used to assess our gender detection algorithm. We present quantitative and qualitative results in Section 5 and discuss limitations as well as avenues for future development.

2 Background

When dealing with the category of gender in the context of computational linguistics, it is important to make a distinction between the social category of gender and gender in a linguistic sense. While social gender relates to the complex property, performance and experience of one's own and others' gender within society (Ackerman, 2019), linguistic gender describes the expression of gender within grammar and language. In English, linguistic gender mainly encompasses ways to express gender as female, male or gender-indefinite (Fuertes-Olivera, 2007), while social gender, as an extra-linguistic category, includes a more fluid view of gender aside from male and female categories. This includes transgender, genderqueer and other non-binary experiences and expressions of gender (Darwin, 2017). Therefore, as Bucholtz (1999) and Cao and Daumé III (2020) point out, there is no "one-to-one" mapping between social and linguistic gender. However, they are influenced by each other and subject to changing norms in society (Fuertes-Olivera, 2007).

Since this research explicitly focuses on lexical gender in English, which is a linguistic category, we give an overview of linguistic gender in English in Section 2.1. Section 2.2 explores the role lexical gender information plays in different areas of research on gender bias in NLP, which simultaneously present possible areas of application for our method of lexical gender detection.

2.1 Linguistic gender in English

The taxonomy of linguistic gender in this work builds upon the approach developed by (Cao and Daumé III, 2020) and incorporates work by Corbett (1991), Hellinger and Bussmann (2003) and Fuertes-Olivera (2007).

Within linguistic gender, Cao and Daumé III (2020) differentiate between grammatical, referential, and lexical gender. **Grammatical gender** refers to the distinction of noun classes based on agreement between nouns and their dependants. English, as a natural or notional gender language (McConnell-Ginet, 2013), does not have grammatical gender, but it has referential as well as lexical gender. **Referential gender**, as the name suggests, is used to refer to the social gender of a specified extra-linguistic entity. Thus, it "relates linguistic expressions to extra-linguistic reality, typically identifying referents as 'female', 'male', or 'gender-indefinite.' " (Cao and Daumé III, 2020). In English, pronouns fall under the category of referential gender. **Lexical gender**, which we focus on in this work, is non-referential but a semantic property of a given linguistic unit, which can be either masculine, feminine[1] or gender-indefinite/gender-neutral. Ackerman (2019) calls these words "definitionally gendered". Words that carry lexical gender can require semantic agreement in related forms, such as, for instance, using the pronoun *his* in connection with the word *stuntman* in the sentence 'Every *stuntman* needs to rehearse *his* stunts.' (Fuertes-Olivera, 2007). In English, lexical gender is usually not morphologically marked. Exceptions to this rule include the suffixes *-man* to denote masculine gender, such as in *policeman*, or *-ess* to denote feminine gender, such as in *waitress*. It should moreover be noted that lexical gender is exclusively a linguistic property. However, words that carry lexical gender can be used to express referential gender if a concrete referent is specified (Cao and Daumé III, 2020).

2.2 Lexical gender in gender bias research

The evaluation and mitigation of gender biases in NLP datasets and models is reliant on referential expressions of gender, such as pronouns and, but also words that carry lexical gender. These pieces of research vary in application, as well as the number of gendered expressions considered, which start at two up to around 120 words. Most works assess binary differences between male and female gender. However, an emergent strand of NLP research is also concerned with non-binary gender expressions (Cao and Daumé III, 2020) and creating gender-neutral datasets and systems (Vanmassenhove et al., 2021). The following considers example use-cases of lexicons of terms carrying lexical gender. These simultaneously represent a variety of applications for our lexical gender detection algorithm.

Dataset evaluation The most straight forward form of using gendered words is to assess the distribution of gendered words in a corpus. Zhao et al. (2019) counted *he*/*she* pronouns in the One Billion Word Benchmark (Chelba et al., 2013) to show male-skew in the training data for the ELMo language model (Peters et al., 2018), which is the primary focus of their analysis. This analysis ad-

[1] We use the terms *masculine* and *feminine* instead of *male* and *female* here in order to underline the purely linguistic, i.e. semantic, property of lexical gender

dressed calls for better data evaluation (Bender et al., 2021; Rogers, 2021) prior to or alongside with model bias analyses.

Retrieval for analysis Limited-scope lists of word that carry lexical gender were used by Caliskan et al. (2017) to retrieve Word2Vec embeddings (Mikolov et al., 2013) and perform the Word Embedding Association Test (WEAT). This test measured stereotyping by calculating implicit associations between eight *male/female* word pairs and words related to maths or science and arts. Guo and Caliskan (2021) used an adapted version of the WEAT, the CEAT, to asses intersectional biases in contextualized word embeddings (ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), OpenAI GPT (Radford et al., 2019; Brown et al., 2020)). Another use-case in which gendered words were used for retrieval is research by Falenska and Çetinoğlu (2021), who assessed gender bias in Wikipedia articles. As a first step, they filtered the article titles for a limited number of words that carry lexical gender.

Creation of synthetic evaluation data In sentence-based analyses of gender-bias, lists of words with lexical gender can also be used to fill placeholders in sentence templates and thus create synthetic sentences with different gendered entities. For example, Kiritchenko and Mohammad (2018) created the Equity Evaluation Corpus (EEC) to analyze gender stereotyping in sentiment analysis systems which inspired the creation of the Bias Evaluation Corpus with Professions (BEC-Pro), that was used to analyze associations between gendered entities and professions in BERT (Bartl et al., 2020). Similarly, Sheng et al. (2019) used the word pair *the man/the woman* as fillers within sentence-start prompts for open-ended natural language generation (NLG) and the subsequent analysis of gender biases in the generated sentences.

In a rare instance of research on non-binary representations of gender in NLP, (Cao and Daumé III, 2020) used gendered lists of words to find and hide lexical gender in the GAP dataset (Webster et al., 2018). The dataset created in this way was used to measure gender and trans-exclusionary biases in coreference resolution performed by both humans and machine-learning models.

Data manipulation Extensive lists of gendered words were used in the context of Counterfactual Data Augmentation (CDA), which replaces words with masculine lexical gender with their feminine variants and vice versa in a corpus. This is done in order to create training or fine-tuning data for gender bias mitigation. For instance, Lu et al. (2020) "hand-picked" gender pairs to swap in CDA and Maudslay et al. (2019) added first names to the list of words to be swapped.

Another kind of data manipulation, this time aiming not for the opposite but for neutral gender, was performed by Vanmassenhove et al. (2021). They used lists of unnecessarily gendered job titles (e.g. *mailman/mailwoman*), unnecessarily gendered feminine forms (e.g. *actress*), and generic uses of the suffix *-man* (such as in *freshman*) in the extended version of their *Neutral Rewriter*, which re-writes sentences with explicit mentions of gender into their gender-neutral variants (*mail carrier*, *actor* and *first-year student*).

3 Method: Automatic Detection of Lexical Gender

The main goal of this work is to produce a dynamic, high coverage, scalable method to determine the lexical gender of a target word, to replace previously used manually compiled lists. For this purpose, we leveraged the fact that if a word has lexical gender, its definition includes words from a small set of definitively gendered words carrying the same lexical gender. In the following, we describe the main algorithm setup, additional parameters and heuristics, as well as a method to combine lexical gender labels from different databases.

3.1 Algorithm construction

The method we outline utilises the increasing availability of machine readable established dictionaries such as Merriam Webster Online (Merriam-Webster, 2021) and the lexical database WordNet (Princeton University, 2010) to identify gendered terms. The following is an example of how lexical gender is captured within Merriam-Webster's (2021) definitions of *nun* and *monk* in (1) and (2) :

(1) *nun*: a woman belonging to a religious order

(2) *monk*: a man who is a member of a religious order and lives in a monastery

Both definitions mention the lexical gender of the referent through a gendered word, in this case *man* and *woman*. Initial analyses showed that gendered words are more likely to occur at the beginning of a definition and definitions often used

the words *female/male* or *woman/man* to specify lexical gender. In identifying gendered terms therefore, we considered the presence and amount of up to eight definitively gendered words, such as *male/female*, *man/woman* etc., in the target word's definitions to draw inferences about its lexical gender.

For retrieval of the definitions, we accessed WordNet through the Natural Language Toolkit (NLTK) API (Bird et al., 2009) and Merriam Webster Online (Merriam-Webster, 2021) through HTTP requests. Additionally, we applied a rationale for combining lexical gender labels of the two databases, which will be discussed in 3.3.

Once the definitions for a given target word were retrieved, the process of obtaining lexical gender was the same for both Merriam Webster and WordNet. We determined whether a word has masculine, feminine or neutral lexical gender, by counting occurrences of a number of word pairs which have clearly defined feminine or masculine lexical gender, such as the pairs *female/male* and *woman/man*. If the combined definition texts contain more masculine than feminine terms, the word was labelled with masculine lexical gender, and vice versa. If the same number of masculine and feminine words was found within a set of definitions, which includes the case in which none of the pre-defined gendered terms can be found, the word was labelled with neutral lexical gender.

3.2 Parameters

We additionally used three variable parameters to limit the number of definitions and word tokens queried, as well as the number of definitively gendered words to use for the query.

Number of definitions d We limited the number of definitions, because definitions that occur early on have a higher likelihood of describing a more general sense of the word, while later definitions relate to very specific word senses. Therefore, we retrieved only the first d definitions that the dictionary lists for the word. In the initial experiments, the default value for d was determined to be $d = 5$.

Number of tokens t We also experimented with limiting the number of tokens within a given definition to see whether definitively gendered terms were more likely to be mentioned earlier in a given definition. The definitions were tokenized using NLTK (Bird et al., 2009). We took the first t tokens of each definition. Regarding the number of tokens in a definition, we tested the algorithm with $t = 5$ and $t = 10$ in our experiments and find $t = 10$ to produce optimal results.

Number of gendered word pairs w The word pairs used during experiments are listed in Table 1. The first two word pairs, *woman/man* and *female/male*, as well as the pair *girl/boy*, are most commonly used to describe the gender of a person or animal, while the rest of the words describes gendered family relations. The latter were chosen in order to account for cases in which the lexical gender of a person is described in relation to another person by using family terms, which is, for example, the case for the definition of *baroness* in Merriam Webster: "the wife or widow of a baron" (Merriam-Webster, 2021).

We found that limiting the number of gendered pairs to $w = 5$ provides the best results. Moreover, if the target word is part of the definitively gendered pairs or their plural forms it was automatically classified with the respective lexical gender.

3.3 Combination of lexical gender labels

In order to settle conflicts between the two dictionary-based lexical gender labels and potentially increase algorithm performance, we joined the lexical gender labels of both dictionaries into a combined label. In cases in which a word cannot be found in one dictionary, the other dictionary label was used. If no label can be found for either, a 'neutral' label was given.

In order to determine the best combination method for cases in which both labels were found, we analyzed label conflicts for our two test datasets, which are described in detail in Section 4. The conflicts are shown in Table 2. We differentiated two types of conflict: 1. feminine vs. masculine label, and 2. gendered (fem or masc) vs. neutral label conflict. Interestingly, Table 2 shows no feminine vs. masculine conflicts for neither the gold standard nor the small Wiki150 dataset. All conflicts were due to one dictionary providing a gendered and the other a neutral label. Most of these conflicts have gendered true labels in the gold standard data while in the small Wiki150 dataset, most label conflicts have the true label of 'neutral'. Since we assumed the sampled data from Wikipedia to emulate naturally occurring data most closely, we resolved these conflicts as neutral in the combined label.

	$w = 2$							
	$w = 5$							
	$w = 8$							
feminine	woman	female	wife	daughter	mother	girl	sister	aunt
masculine	man	male	husband	son	father	boy	brother	uncle

Table 1: Words carrying explicit lexical gender; w = number of pairs used for experiments

dataset	**gold standard**						**Wiki150-sample**					
***n* conflicts out of instances**	21 out of 119 (18.1%)						63 out of 150 (29.3%)					
type of conflict	fem vs. masc			fem/masc vs. neutral			fem vs. masc			fem/masc vs. neutral		
	0			21			0			63		
true label	masc	fem	neut	masc	fem	neut	masc	fem	neut	masc	fem	neut
	0	0	0	12	8	1	0	0	0	7	1	55

Table 2: Conflicts between lexical gender labels obtained from Merriam Webster and WordNet

3.4 Morphological Heuristics

Aside from the lexical database method described above, we additionally applied two morphological heuristics and one heuristic relating to punctuation. Morphological heuristics were applied before querying the dictionaries, while the punctuation-related heuristic was applied when a word cannot be found in the dictionary.

We classified words containing the suffixes *-man* and *-boy* or *-woman* and *-girl* into masculine and feminine lexical gender, respectively. Regular expressions were used in order to ensure that words with the suffix *-woman*, which includes *-man*, were not classified as masculine, but as feminine.

In order to account for differing uses of punctuation within terms, different forms of words are examined if a term contains punctuation characters and is not contained within a dictionary. For example, the word *land-lady*, spelled with a dash, is not contained in WordNet, while *landlady* is. Therefore, if a word cannot be detected, we check for possible punctuation in a phrase, remove it and try again with the resulting word. This also applies to the case when non-detection is caused by a whitespace character.

4 Data

We used two test datasets to evaluate and run the algorithm. The first dataset, which we called *gold standard* hereafter, contains nouns that have a clear lexical gender and were mainly sourced from previous research on gender bias. The second dataset contains 150 randomly sampled Wikipedia articles, which we used to extract gendered nouns. The following describes both datasets in detail. An overview of overlap between the two datasets can be found in Table 6 in the Appendix.

4.1 Gold Standard

In order to gain insights into the performance of the dictionary-based algorithm for lexical gender retrieval, we compiled a list of words that have an almost unambiguous lexical gender, which acts as the *gold standard*. This gold standard was developed based on a lexical gender list by (Cao and Daumé III, 2020) with the addition of more words retrieved from online lists for learners of English[2][3][4]. They were then filtered for explicitness of lexical gender, which means that for example, the pair *actor/actress* would not be considered since the word *actor* is nowadays used for both male and female referents. We moreover added neutral gender replacements for word pairs for which such an alternative exists. And example would be the triplet *headmaster*-MASC, *headmistress*-FEM, *headteacher*-NEUT. The final list is comprised of 48 masculine, 48 feminine and 23 neutral words. We provide the full gold standard list in Table 5 in the Appendix.

[2] www.vocabularypage.com/2017/03/gender-specific-nouns.html

[3] 7esl.com/gender-of-nouns/

[4] learnhatkey.com/what-is-gender-in-english-grammar/

4.2 Wikipedia Sample

This research aims at providing a flexible, scalable, and high-coverage method for lexical gender detection. Therefore we additionally tested the approach on more naturalistic data, namely a random sample of 150 articles from English Wikipedia obtained through the wikipedia python library[5]. We will abbreviate the sample corpus as *Wiki150* hereafter.

The articles were then cleaned and tokenized into sentences using NLTK (Bird et al., 2009), which were subsequently processed with spacy to obtain part-of-speech (POS) tags for each word. All singular and plural nouns (POS-tags: NN, NNS) were then extracted and analyzed for lexical gender. Words that were identified as nouns but contain special characters due to cleaning and tokenization errors were dropped. This method provided us with 4,187 nouns, as illustrated under Wiki150 dataset in Table 3.

In order to test the performance of the algorithm, the instances of the Wiki150 dataset needed true labels. A corpus size of 4,187 instances, however, was beyond the scope of this research to manually label and represents the kind of corpus size that we aim to label automatically. We therefore filtered Wiki150 for nouns that were labelled as either *masculine* or *feminine* by Merriam Webster Online or WordNet. Like this, we specifically target gendered nouns and obtain a corpus similar to the gold standard corpus, but sourced from naturally occurring text. The resulting corpus *Wiki150-sample* was subsequently labelled for 'true' lexical gender by the researchers (Cohen's $\kappa \approx 0.96$). There were four instances in total that caused annotator disagreement due to word disambiguation issues, which were *fellow*, *master*, *ram* and *suitor*. For reasons of simplicity, it was decided to exclude these from the final evaluation. We discuss the issue of word sense disambiguation in the context of this research further in Section 5.3. The specifications of the Wiki150-sample dataset can be found in Table 3.

In line with previous research on gender bias in Wikipedia (Wagner et al., 2015; Falenska and Çetinoğlu, 2021), which found an over-representation of male entities in the encyclopedia, Table 3 shows that there are approximately twice as much mentions of distinct entities with masculine lexical gender in our small Wikipedia sample than there of entities with feminine lexical gender.

[5] https://pypi.org/project/wikipedia/

5 Results and Discussion

5.1 Quantitative results

An overview of algorithm performance on the gold standard dataset and the reduced Wiki150 sample can be found in Table 4. We report the weighted average of precision, recall, and F1-measure due to unbalanced classes in our test data.

Table 4 shows that overall, our method reached an accuracy of 70% or higher in each experiment configuration. However, determining the best experiment settings was is challenging due to varying performance our two test datasets. Our best performing approach on the gold standard in terms of accuracy queries only Merriam Webster (90%), while the best performance on the Wiki150 sample utilised a combination of sources (90%).

This difference in performance between the two test datasets is not surprising given their respective label distributions, which are displayed in Table 3. There are more neutral nouns in the Wiki150 sample, while the gold standard contains more gendered instances. Since the combined approach performed 'conflict resolution' for the most part by assigning neutral labels, its performance was higher on the small test set.

This dynamic can also be observed in Figure 1, which shows confusion matrices for the combined approach on both the gold standard dataset (1a) and the Wiki150-sample (1b). Figure 1a shows that on the gold standard, the combined classifier mislabelled eight feminine and 16 masculine instances as neutral, but did not mislabel any of the neutral instances as either masculine or feminine. In contrast, both these classification mistakes can be found on the Wiki150 sample (Figure 1b).

Another issue, which only occurred when testing on the gold standard dataset, concerns words that could not be found. The first is *single person*, which we chose as the gender-neutral alternative for *bachelor*/*spinster*. The fact that it was not found could be due to the fact that *single person* is more of a composite phrase than a joined expression. Moreover, single people are often described using the adjective *single* in a predicative way, such as in the sentence 'He is single.', instead of 'He is a single person.' The other word that could not be found is *child-in-law*, which is the gender-neutral variant of *son-in-law* and *daughter-in-law*. Here, the issue could be frequency of use, since *child-in-law* is less established than its gender-specific variants.

	gold (N=119)	Wiki150-sample (N=146)			Wiki150 dataset (N=4187)		
POS	**NN**	**NN**	**NNS**	**comb.**	**NN**	**NNS**	**comb.**
masc	47	36	22	58	36	19	55
fem	47	21	8	29	25	7	32
neut	22	39	20	59	2732	1285	4017
NF	-	-	-	-	71	11	82
all	116	96	50	146	2865	1322	4187

Table 3: Composition of evaluation corpora for lexical gender detection algorithm. **NF** = not found
Note: for *Wiki150 full*, combined predicted labels were used, because no gold labels exist for this dataset

	gold standard (N=119)				Wiki150-sample (N=146)			
measure	**P**	**R**	**F1**	**Acc**	**P**	**R**	**F1**	**Acc**
WordNet	0.91	0.82	0.85	**0.82**	0.73	0.70	0.66	**0.70**
Merriam Webster	0.94	0.9	0.91	**0.9**	0.82	0.79	0.78	**0.79**
Combined	0.9	0.78	0.81	**0.78**	0.9	0.9	0.9	**0.9**

Table 4: Quantitative results for lexical gender detection of gold standard and Wiki150-sample

5.2 Qualitative results

In the following we will go into more detail on specific classification errors that occur due to outdated and gender-exclusive definitions in the lexical databases or due to historically close associations of words to a single gender.

Some misclassifications of masculine terms as neutral can be traced back to outdated definitions representing a male-as-norm viewpoint (Fuertes-Olivera, 2007). As an example, consider the definition for the word *businessman* in WordNet (Princeton University, 2010) in (3).

(3) *businessman* - a person engaged in commercial or industrial business (especially an owner or executive)

Even though *businessman* contains the masculine suffix *-man*, its definition is generic. This is most likely due to the fact that *businessman* was once used for business people of all genders. However, since feminine or neutral equivalents (*business woman, business person*) are widely used nowadays, we see the current WordNet definition in need of an update.

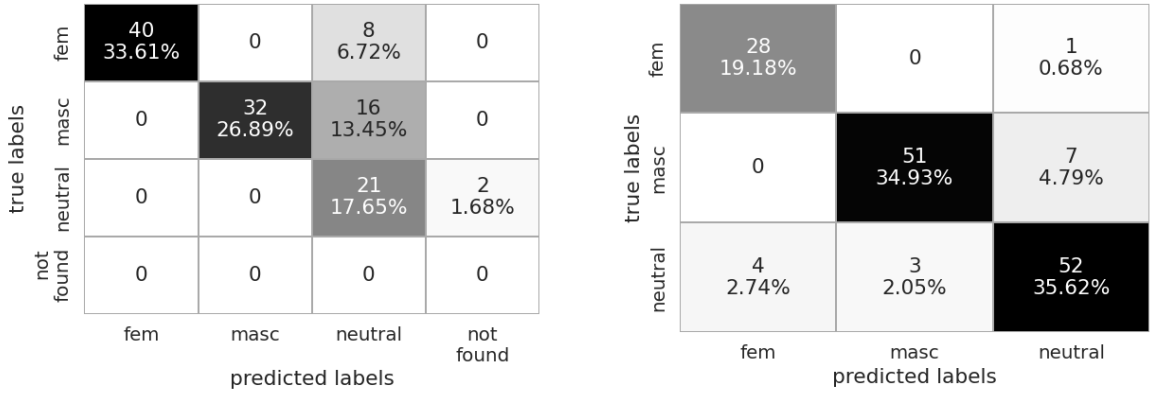
Conversely, outdated definitions can also cause misclassifications of neutral terms as masculine, such as for the word *crew* in WordNet. We show the first and fourth definitions in Example (4), in order to illustrate how the masculine label was obtained.

(4) *crew*

1. the men and women who man a vehicle (ship, aircraft, etc.)
4. the team of men manning a racing shell

In the first definition, the words *men and women* are used to describe the crew of any vehicle. However, in the fourth definition, which describes the crew of a racing shell (a type of rowing boat), only the word *men* is used, causing the lexical gender label to be masculine since the definitions taken together contain more masculine than feminine words. However, the fourth definition could also have been worded like the first definition, or worded using the word *people*, since racing shells can be crewed by people of any gender.

Another, however foreseeable classification error occurred for the words *dowry*, *pregnancy*, and *contraceptives*, which all were classified as feminine by Merriam Webster (Merriam-Webster, 2021), even though they have neutral lexical gender. This error was caused since these terms are closely associated with female social gender. For example, the most prevalent contraceptive in Europe in 2019 was the birth control pill (Statista, 2020), which is currently only widely available for people with a



(a) gold standard

(b) Wiki150-sample

Figure 1: Confusion matrices for combined labels
words that were not found in (a): *single person*, *child-in-law*

female reproductive system. However, contraceptives can reference any form of pregnancy prevention and should therefore have a neutral definition. Moreover, the fact that the definition for *pregnancy* includes specific references to female gender is a form of trans-exclusionary bias, since people with a uterus who do not identify as female can still get pregnant.

5.3 Limitations and Future Developments

We have selected dictionaries to obtain the lexical gender of a word, because they represent a relatively objective resource that is expected to list neutral and non-stereotypical definitions of words. However, as shown in Section 5.2, dictionaries are after all a human-curated resource and as such can still carry human biases and outdated definitions, which in turn lead to biased or outdated results. In order to (at least partially) mitigate this bias, we plan on including more dictionaries into an updated version of our algorithm and thus use a voting mechanism with a more diverse set of lexical gender predictions.

Another limitation of the present work concerns word sense disambiguation, since whether or not a word contains lexical gender depends on its sense in context. As an example, the word *ram*, can either mean a male sheep or an instrument to apply brute force in order to open something, among others. In the sense of a male sheep, the lexical gender of *ram* is clearly masculine while in the sense of the brute-force instrument, it is neutral. Differences in the lexical gender of word senses can also be caused by semantic shifts, such as for the word *master*, which traditionally refers to a man who is in control of e.g. servants or a household. However, in an academic context its meaning has shifted and now refers to an academic degree, or more broadly to a person of undefined gender who has reached a high level of skill in a given discipline. In this work we excluded four words from the evaluation on the Wiki150 dataset sample due to annotator disagreement caused by word sense disambiguation issues. Therefore, future work will integrate word sense disambiguation within the algorithm.

6 Conclusion

We have presented a method to automatically determine the lexical gender of a given word by querying its dictionary definitions. The performance of the algorithm on a gold standard dataset of gendered nouns based on related literature, as well as set of nouns sampled from a set of 150 randomly selected Wikipedia articles, reached up to 90% accuracy. Previous research on gender bias in NLP used manually compiled lists of gendered words for data evaluation, retrieval, manipulation and the synthetic creation of data. In contrast, our method is scalable and has a high, dynamic coverage, which gives it a variety of applications within past and future research on gender bias in NLP. These include e.g. the assessment of gender representations in large-scale corpora, the retrieval of gendered words for which gender-neutral replacements need to be found, as well as determining whether male-centric language such as epicene *he* is used in coreference resolution clusters.

References

Lauren M Ackerman. 2019. Syntactic and cognitive issues in investigating gendered coreference. *Glossa*.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. Conference Proceedings.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Mary Bucholtz. 1999. Gender. *Journal of Linguistic Anthropology*, 9(1/2):80–83.

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

Yang Trista Cao and Hal Daumé III. 2020. Toward Gender-Inclusive Coreference Resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Ciprian Chelba, Tomás Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, and Phillipp Koehn. 2013. One billion word benchmark for measuring progress in statistical language modeling. *CoRR*, abs/1312.3005.

Greville G. Corbett. 1991. *Gender*. Cambridge University Press.

Helana Darwin. 2017. Doing gender beyond the binary: A virtual ethnography. *Symbolic Interaction*, 40(3):317–334.

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.

Agnieszka Falenska and Özlem Çetinoğlu. 2021. Assessing Gender Bias in Wikipedia: Inequalities in Article Titles. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 75–85, Online. Association for Computational Linguistics.

Pedro A. Fuertes-Olivera. 2007. A corpus-based view of lexical gender in written Business English. *English for Specific Purposes*, 26(2):219–234.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 122–133.

Marlis Hellinger and Hadumod Bussmann. 2003. *Gender across languages: the linguistic representation of women and men*, volume 11. J. Benjamins, Amsterdam;Philadelphia;.

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.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*, pages 189–202. Springer.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5267–5275.

Sally McConnell-Ginet. 2013. Gender and its relation to sex: The myth of 'natural' gender. In Greville G. Corbett, editor, *The Expression of Gender*, pages 3–38. De Gruyter Mouton.

Merriam-Webster. 2021. Merriam-webster.com.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Princeton University. 2010. About WordNet.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Anna Rogers. 2021. Changing the World by Changing the Data. *arXiv:2105.13947 [cs]*. ArXiv: 2105.13947.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Statista. 2020. Contraceptive use of women in Europe by method.

Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. 2021. NeuTral Rewriter: A rule-based and neural approach to automatic rewriting into gender neutral alternatives. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8940–8948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *Ninth international AAAI conference on web and social media*.

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.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. 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 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Appendix

category	masculine	feminine	neutral
misc	man	woman	person
	male	female	
	boy	girl	child
	boyfriend	girlfriend	partner
	gentleman	lady	
	groom	bride	
	bachelor	spinster	single person
	lad	lass	
	manservant	maidservant	servant
	steward	stewardess	attendant
	wizard	witch	
occupation	policeman	policewoman	police officer
	fireman	firewoman	fire fighter
	headmaster	headmistress	head teacher
	landlord	landlady	renter
	milkman	milkmaid	
	salesman	saleswoman	salesperson
	chairman	chairwoman	chairperson
	businessman	businesswoman	business person
religion	monk	nun	
	friar	nun	
family	father	mother	parent
	dad	mum	
	dad	mom	
	son	daughter	child
	daddy	mummy	
	daddy	mommy	
	brother	sister	sibling
	uncle	aunt	
	grandfather	grandmother	grandparent
	grandson	granddaughter	grandchild
	husband	wife	spouse
	father-in-law	mother-in-law	parent-in-law
	nephew	niece	
	son-in-law	daughter-in-law	child-in-law
	stepfather	stepmother	stepparent
	widower	widow	
title	duke	duchess	
	baron	baroness	
	count	countess	
	earl	countess	
	czar	czarina	
	king	queen	
	prince	princess	
	signor	signora	
	sir	madam	
	viscount	viscountess	
	Mr.	Mrs.	Mx.

Table 5: Masculine, feminine and neutral nouns of the gold standard dataset

dataset	gold standard	overlap	Wiki150-sample
masc	Mr., baron, boyfriend, count, czar, dad, duke, earl, father-in-law, fireman, headmaster, lad, landlord, manservant, milkman, policeman, salesman, signor, sir, son-in-law, stepfather, steward, viscount, widower, wizard	bachelor, boy, brother, businessman, chairman, daddy, father, friar, gentleman, grandfather, grandson, groom, husband, king, male, man, monk, nephew, prince, son, uncle	baseman, bull, dude, emperor, freeman, freshman, knight, layman, nobleman, ombudsman, papa, patriarch, ram, spokesman, stableman, statesman
fem	Mrs., baroness, businesswoman, chairwoman, countess, czarina, daughter-in-law, duchess, female, firewoman, granddaughter, grandmother, headmistress, landlady, lass, madam, maidservant, milkmaid, mom, mother-in-law, mum, mummy, nun, policewoman, saleswoman, signora, stepmother, stewardess, viscountess, widow, witch	aunt, bride, daughter, girl, girlfriend, lady, mommy, mother, niece, princess, queen, sister, spinster, wife, woman	actress, barmaid, gal, hen, hind, maid
neutral	Mx., attendant, business person, chairperson, child, child-in-law, fire fighter, grandchild, grandparent, head teacher, parent, parent-in-law, partner, person, police officer, renter, salesperson, servant, sibling, single person, spouse, stepparent		baggage, ball, bass, bird, blade, blood, breast, costume, court, crew, dean, dowry, dress, ed, fellow, honor, honour, honours, horse, liver, lizard, marksmanship, master, member, mill, name, neighbor, nurse, parity, polygyny, pop, pregnancy, rake, rating, relation, relief, specimen, suitor, sweetheart, transformation, womanhood, youth

Table 6: Overlap of words with feminine, masculine and neutral lexical gender between gold standard corpus and Wiki150-sample