Proposed Taxonomy for Gender Bias in Text; A Filtering Methodology for the Gender Generalization Subtype

Yasmeen Hitti* and **Eunbee Jang*** and **Ines Moreno*** and **Carolyne Pelletier***
{hittiyas, jangeunb, morenoin, pelletic}@mila.quebec
Mila, Université de Montréal / 6666 St-Urbain, Montreal, QC H2S 3H1

Abstract

The purpose of this paper is to present an empirical study on gender bias in text. Current research in this field is focused on detecting and correcting for gender bias in existing machine learning models rather than approaching the issue at the dataset level. The underlying motivation is to create a dataset which could enable machines to learn to differentiate bias writing from non-bias writing. A taxonomy is proposed for structural and contextual gender biases which can manifest themselves in text. A methodology is proposed to fetch one type of structural gender bias, Gender Generalization. We explore the IMDB movie review dataset and 9 different corpora from Project Gutenberg. By filtering out irrelevant sentences, the remaining pool of candidate sentences are sent for human validation. A total of 6123 judgments are made on 1627 sentences and after a quality check on randomly selected sentences we obtain an accuracy of 75%. Out of the 1627 sentences, 808 sentence were labeled as Gender Generalizations. The inter-rater reliability amongst labelers was of 61.14%.

1 Introduction

The feminist movement which debuted in the late 1960s was a response to gender discourses that had been problematic and often biased (Messerschmidt et al., 2018). Ever since, more emphasis has been guided towards outlining these issues in societal roles, sports, media, religion, culture, medicine, and education. Haines et al. (2016) have shown, despite time, from the 1980s to 2014, that the perception of gender roles has remained stable for men and women, hence the presence of biased sociocultural expectations to this day. Research concerning gender equality has been focused on quantitative data analysis and has resulted in empirical evidence of inequalities in different sectors. Examples include school enrollments and job employments, all which have failed to provide the source responsible for these inequalities (Unterhalter, 2015). Although the root of these imbalances remain ambiguous, it is known that social norms have greatly influenced and reinforced inconsistencies while referring to specific genders (Robeyns, 2007).

Language is known to reflect and influence society in its perception of the world. For these reasons there has been constant effort to promote bias-free and non-sexist writing to empower the fairness movement. However, to our knowledge, no quantitative study on gender bias in text at the dataset level has been done. In the era of Machine Learning (ML), gender biases are translated from sourced data to existing algorithms that may reflect and amplify existing cultural prejudices and inequalities (Sweeney, 2013) by replicating human behavior and perpetuating bias. Thus, there is a need to approach this issue in a ML context in the hope that it will help raise awareness and minimize discrimination at the human-level. To do so, rather than removing gender bias in current ML models we want to create a dataset with which to train a model to detect and help correct gender bias in written form. In the long run, our dataset would ideally be extended to encompass all types of bias such as race, religion, sexual orientation, etc.

1.1 Contributions

- Provide a high-level definition of gender bias in text
- Present an approach to find one of the subtypes of gender bias, Gender Generalization
- Provide a small labeled dataset for Gender Generalization bias.

*equal contribution

Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing, pages 8–17
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

2 Related Work

Current ML research has identified gender bias in various models, each with its own evaluation and debiasing methods. In Natural Language Processing (NLP), gender bias has been studied in word embeddings, coreference resolution and recently, in datasets. Previous work on gender bias in writing has been addressed by linguists with the creation of inclusive writing. In the field of gender studies, gender gaps have been explored through social contexts.

2.1 Word Embedding

In NLP, word embeddings have become a powerful means of word representations. Bolukbasi et al. (2016) first experimented with gender in word embeddings and found that the presence of gender stereotypes were highly consistent in popularly used word representation packages such as Glove and word2vec (Bolukbasi et al., 2016). To better understand the gender bias subspace, gender specific words were investigated to compare their distances with respect to other words in the vector space (Bolukbasi et al., 2016). It is claimed that the unequal distances measured are due to the corpora on which an embedding has been trained on and reflect the usage of language which contain cultural stereotypes (Garg et al., 2018).

Hard debiasing was developed following the findings on gender bias in word embeddings. This method introduced by Bolukbasi et al. (2016) has the main objective of debiasing word embeddings while preserving their properties in the embedding space. To achieve the debiasing of an algorithm, the assumption was that a group of words needed to be neutralized to ensure that gender neutral words were not affected in the gender subspace of the embedding. Following this work, Zhao et al. (2018b) have approached the problem differently and have uptaken the task of training on debiased word embeddings from scratch by introducing gendered words as seed words. Furthermore, Gonen and Goldberg (2019) has shown with clustering that debiased word embeddings still contain biases and concluded that the existing bias removal techniques are insufficient, and should not be trusted for providing gender-neutral modeling.

2.2 Coreference Resolution

Coreference resolution is a task aimed at pairing a phrase to its referring entity. In the context of this paper, we are interested in pairing pronouns with their referring entities. Recent studies by Rudinger et al. (2018) suggest, however, that state-of-the-art coreference resolvers are gender biased due in part to the biased data they have been trained on. For example, OntoNotes 5.0, a dataset used in the training of coreference systems, contains gender imbalances (Zhao et al., 2018a). One such example of these imbalances are in the frequency of gendered mentions related to job titles: ”Male gendered mentions are more than twice as likely to contain a job title as female mentions”. Zhao et al. (2018a) showed that coreference systems are gender biased in this same context of job occupations since they link pronouns to occupations dominated by the gender of the pronoun more accurately than occupations not dominated by the gender of the pronoun.

When coreference resolution decisions are used to process text in automatic systems, any bias present in these decisions will be passed on to downstream applications. This is something that we must keep in mind as we rely on coreference resolution in our filtering system in the later section.

2.3 Datasets

In the past few years, the ML community has created new text datasets with respect to gender discrimination and have focused on hate speech, stereotypes and relatedness to gender ambiguous pronouns. Twitter posts have been the preferred source of investigation when it comes to understanding and capturing human bias although this may only focus on one type of gender bias. The Equity Evaluation Corpus is a dataset of 8,640 English sentences with a race or gendered word and evaluates the sentiment towards these sentences. The measurement of sentiment was achieved by training on the SemEval-2018 Tweets (Kiritchenko and Mohammad, 2018). Abusive language datasets have also been based off of tweets and identify sexist and racist language (Waseem, 2016). GAP is a dataset focused on sentences which have references to entities; this dataset is composed of sentences with proper nouns and ambiguous gendered pronouns (Webster et al., 2018).

2.4 Gender Bias in Writing

2.4.1 Inclusive Writing

Gender-neutral writing was developped to avoid sexism and generic mental images for gender roles (Corbett, 1990). Guidelines for inclusive writing were created following surveys and offer insights on different biases including gender related biases (Schwartz, 1995). A study by Vainapel et al. (2015) demonstrated that male-inflected terms in a survey have affected the responses of women leading to lower task value beliefs. Motivations behind the utilization of gender inclusive writing is to disrupt the current educational system which is tailored for masculinized vocational professions (Ray et al., 2018).

2.4.2 Gender Gap

The gender gap in writing has resurfaced multiple times through history. The meaning of gender was studied by Simone de Beauvoir and was defined as something which is prescribed by society with preferences towards men (Cameron, 2005). This societal role of the toy and media culture has influenced the writing of boys and girls at schools and has related boys to violence and girls to subordinate roles (Newkirk, 2000). The online writing of of women and men on Wikipedia has also been unequal as most editors have been males thus creating a gender gap in their content (Graells-Garrido et al., 2015).

3 Proposed Gender bias Taxonomy

As most work in the ML community related to gender bias has been focused on debiasing existing algorithms, the creation of a dataset will enable to tackle the issue at its root and allow for observation of its impact on different ML models.

The first step to the data creation is to quantify the qualitative definition of gender bias. Thus, a gender bias taxonomy is proposed after consulting language and gender experts. We define gender bias in text as the use of words or syntactic constructs that connote or imply an inclination or prejudice against one gender. Gender bias can manifest itself structurally, contextually or both. Moreover, there can be different intensities of biases which can be subtle or explicit.

3.1 Structural Bias

Under our definition, structural gender bias occurs when bias can be traced down from a specific grammatical construction. This includes looking up of any syntactic patterns or keywords that enforce gender assumptions in a gender neutral setting. This type of bias can be analyzed through popularly used text processing techniques used in NLP.

3.1.1 Gender Generalization

The first subtype of structural bias, that we refer to as Gender Generalization, appears when a gender-neutral term is syntactically referred to by a gender-exclusive pronoun, therefore, making an assumption of gender. Gender-exclusive pronouns include: *he, his, him, himself, she, her, hers and herself.*

- "**A programmer** must always carry **his** laptop with **him**." - gives a fact about an arbitrary programmer and assumes a man to be the programmer by referring to "*he*".

- "**A teacher** should always care about **her** students." - gives a fact about an arbitrary teacher and assumes a woman to be the teacher by referring to "*she*".

Counter example:

- "**A boy** will always want to play with **his** ball." - although representing a stereotype, it is not assuming the gender for a gender neutral word since the word boy (gendered - male) is linked to a male pronoun. Thus, it is not Gender Generalization bias.

3.1.2 Explicit Marking of Sex

A second subtype of structural bias appears with the use of gender-exclusive keywords when referring to an unknown gender-neutral entity or group.

- "**Policemen** work hard to protect our city." - the use of "*policemen*" instead of "*police officers*" directly excludes all women that could also hold that position.

- "The role of **a seamstress** in the workforce is undervalued." - the usage of a gender-marked title for women for a job that can be done by both sexes is biased unless referring only to the female counterpart.

3.2 Contextual Bias

On the other hand, contextual gender bias does not have a rule-based definition. It requires the learning of the association between gender-marked keywords and contextual knowledge. Unlike structural bias, this type of bias cannot be observed through grammatical structure but requires contextual background information and human perception.

3.2.1 Societal Stereotype

Societal stereotypes showcase traditional gender roles that reflects social norms. The assumption of roles predetermines how one gender is perceived in the mentioned context.

- "Senators need their **wives** to **support** them throughout their campaign." - the word "*wife*" is depicted as a supporting figure when we do not know the gender of the senator and the supporting figure can be a male partner, a husband.
- "The event was **kid-friendly** for all the **mothers** working in the company." - assumes women as the principal caretakers of children by using the word "*mothers*" instead of using "*parent*" that would encompass possibly more workers.

3.2.2 Behavioural Stereotype

Behavioural stereotypes contain attributes and traits used to describe a specific person or gender. This bias assumes the behaviour of a person from their gender.

- "**All boys** are **aggressive**." - misrepresentation of all boys as aggressive.
- "**Mary** must **love dolls** because **all girls like playing with them**." - assumes that dolls are only liked by girls.

4 Empirical Pilot Study

Two different surveys were deployed, first to better understand if the proposed definition of gender bias was well accepted and second to decide whether categorical or binary labeling should be used when presenting sentences to human labelers. The definition survey was distributed to individuals from the field of sociolinguistics, linguistics, and gender studies. The second survey on categorical and binary labeling was deployed on Mechanical Turk[1] and to the same gender and language experts for the definition survey.

4.1 Definition Survey

The survey form was designed to be shared with individuals who had some relatedness to the topic in a research context. The motivation was to start a dialogue across disciplines to observe if some sort of consensus could be achieved and to recognize potential factors influencing the bias towards gender. The questions asked were short answers, long answers and multiple choices.

Questions:

1. Do you think gender bias is influenced by demographics (gender, age, geographic location, professional status., etc...)? Please justify your answer.
2. Where is it most likely to find gender bias? (work place, home, legal system, academia, media and other)
3. Do you think there are subtypes of gender bias? (yes/no)
4. If yes, which are the subtypes of gender bias?
5. Our current understanding of gender bias in text is : *Gender bias in text is the use of words/syntactic constructs that connote or imply an inclination or prejudice against one gender. It can be structural (when the construction of sentences show patterns that are closely tied to the presence of gender bias) or contextual (when the tone, the words use or simply the context of a sentence shows gender bias).* Do you agree with this definition?
6. Would you add/remove something to/from the previous definition?
7. Do you have any comments/feedback?
8. Having a well-labeled dataset is key for the success of our project. In the future, would you be willing to help label a subset of sentences as gender biased or non-gender biased?

4.2 Data Presentation Survey

A data presentation survey was sent out to the same group of people and was also launched on a crowdsourcing platform, Mechanical Turk. The survey had two sections of 10 questions; the first section contained categorical labeling with all of

[1] https://www.mturk.com/

the potential types of gender bias in text and the second section was binary labeling confirming if a sentence was gender biased or not. At the end of each survey, optional feedback was collected from the participants to ask for their preference and clarity on labeling format. The sentences chosen to be presented to the participants were selected from various journal sources which had been web-scraped previously.

4.3 Deductions

Both surveys provided insightful information for the data collection. The responses from the definition survey included:

- 90% agreed that gender bias is influenced by demographics.
- Respondents had consensus that gender bias can be found in academia, households, media, legal systems, sport coverage, literature and in medical treatments.
- 100% agreed that there are different subtypes of gender bias in writing.
- The top three subtypes identified were stereotypes with 100% agreement, Gender Generalizations with 90% agreement and abusive language with 80% agreement.

A total of 44 participants responded to the data presentation survey and 77.3% preferred binary labeling versus categorical labeling. A good takeaway from this survey was that the presentation of all subtypes of gender bias for categorical labeling may complicate understanding of different definitions we present for future labelers to be able to identify every types of biases. Following both surveys, we decided to focus on extracting one subtype of biases at a time.

5 Methodology

In the previous section, we define different types of biases that can occur which can induce both explicit and implicit biases. In this paper, we focus on one of the structural biases, Gender Generalization, that can be analyzed through observing the syntactic structure of text. Under our definition, Gender Generalization occurs when a gendered pronoun is linked to a gender-neutral term in a gender-free context.

5.1 Corpora Selection

The frequency of Gender Generalizations in texts are unknown and for this reason different types of writing styles were considered for exploration. The biggest challenge in corpus selection was finding sources which talked about human individuals in a general way rather than specific individuals. Our starting point was the IMDB dataset (Maas et al., 2011), followed by multiple corpora from Project Gutenberg [2]. This selection provided a range of writings from the 1800s to modern colloquial English. The texts from Project Gutenberg used for the experiment were: Business Hints for Men and Women, Magna Carta, The Federalists Papers, The Constitution of the United States of America: Analysis and Interpretation, The Common Law, Langstroth on the Hive and the Honey-Bee: A Bee Keeper's Manual, Scouting For Girls: Official Handbook of the Girl Scouts, Boy Scouts Handbook and Practical Mind-Reading. These texts were chosen on the belief that we could capture Gender Generalization sentences; this selection includes guidelines, law and instructions.

5.2 Preprocessing

All texts were preprocessed in order to pass on to the filters and labelers. All texts were split into sentences, no punctuation was stripped and letter cases remained in their original form for integrity purposes. For the IMDB dataset, HTML tags were removed and text was decoded from unicode matching the closest ASCII characters to handle any special symbols present in the text. All text from Project Gutenberg came in a text format in UTF-8 encoding. All document formatting of indentations, blank spaces and quotation marks were removed.

5.3 Design of Filters

The objective behind gathering Gender Generalization sentences is to start constructing a dataset of gender biased sentences with a subtype of bias that is easy to recognize structurally. To gather text data that falls into this category of bias, we have decided to filter sentences based on their syntactic structure. The strategy was to find all the links between expressions that refer to the same entity in text and observe their property with respect to

[2] https://www.gutenberg.org/

the gender they are associated with. Following our definition, the main characteristics of Gender Generalization bias is the existence of a link between a gendered pronoun to any human entity that is not tied to any gender.

Identifying gender-free mentions was challenging since they appear in diverse forms and are closely connected with their context in which they appear, making it necessary for human validation. The filters were used as tools to reduce the scope of the labeling pool, which was sent to the labelers for human judgment.

The filters were applied to every sentence and if any sentence did not meet one of the criteria, it was removed from the potential pool of Gender Generalization candidates. The order of filters applied were as such: *coreference resolution, verification of gendered pronoun, human-name removal, gendered-term removal, and pronoun-link.* The coreference resolution was achieved using AllenNLP and the other filters were dependent on the NLTK library.

5.3.1 Coreference Resolution Filter

Coreference resolution was chosen as a filter for fetching Gender Generalizations as it is by definition identifying different mentions referring to the same entity. AllenNLP's[3] pre-trained model was used to gather coreference clusters. This model implements the current state-of-the-art end-to-end neural coreference resolution by Lee et al. (2017) which is trained on a biased word embedding (Bolukbasi et al., 2016). The models utilizes GloVe and Turian embeddings (Pennington et al., 2014) which result in preferred resolution for gendered pronouns. While the accuracy of coreference resolvers given the gender of the pronoun may differ, it did not affect our coreference resolution filter since we were simply interested in using the resolver to indicate the presence of an antecedent linked to a pronoun. As such, the accuracy of the resolver was of diminished concern.

5.3.2 Gendered Pronoun Filter

After acquiring the information of coreference relationships, we filtered out sentences which we know confidently are not human related. Generalization of gender by definition assumes a particular pronoun to be assigned to a person entity with an unknown gender. Such datapoints were traced down by checking the existence of gendered pronouns in text using simple list manipulations. The gendered pronouns in our list included: *he, him, his, himself, she, her, hers, herself.*

5.3.3 Human Name Filter

While sentences containing human names can be biased, they were not identified as a Gender Generalization. This type of bias requires gender-free context and having a specific person referenced to a gendered pronoun enforces gender in the text as seen in the example below.

- "**Jason** must not abandon the place where **he** was brought up." - The pronoun "*he*" is used because it refers to Jason who is a male.
- "**A politician** must not abandon the place where **he** was brought up." - Exhibiting gender bias because the pronoun "*he*" was used when "*a politician*" is a gender-free term.

To make our system recognize human names, we utilized Named Entity Recognition (NER) from Natural Language Toolkit (NLTK). For every mention in a coreference cluster, we checked if NER classifies the mention as a person-type category when tokenized sentences were fed into the system; identified clusters resulted in the removal of sentences.

5.3.4 Gendered Term Filter

Gendered terms are the words which exhibit specific gender and confirm a person's gender without needing context . For example, the term '*sister*' always refers to female sibling and is always associated with female pronouns whereas '*brother*' refers to male sibling with male pronouns. These types of terms in the coreference relationship were discarded for Gender Generalization bias text mining. Since there is no such system that detected gender assignments of human words, we explored the Lesk algorithm from NLTK which performs Word Sense Disambiguation (WSD) using WordNet. WordNet is a lexical database for the English language and it provides access to dictionary definitions along with related synonyms. The Lesk algorithm utilizes sense-labeled corpus to identify word senses in context using definition overlap.

Our approach was to acquire the adequate word sense of mentions in the coreference cluster given sentences as a context for WSD. The Morphy algorithm in WordNet was then utilized; it uses a

[3] `https://allennlp.org/`

combination of inflectional ending rules and exception list to find the base form of the word of interest. When the base forms were attained, we looked up the definitions associated with their synsets (word sense token). If the definitions contained any gendered terms in table 1 , the sentence was removed.

Type	Male Term	Female Term
Base Term	male man boy	female woman girl
Pronoun	he him his himself	she her hers herself
Family Term	husband father son brother grandfather grandson uncle nephew	wife mother daughter sister grandmother granddaughter aunt niece

Table 1: Gendered terms used in the filter.

Below in Table 2 are some example words that have passed through the definitions of human nouns that we have obtained.

Word	Definition	Gendered?
landlord	a landowner who leases to others	No
landlady	a landlord who is a **woman**	Yes
gentleman	a **man** of refinement	Yes
lady	a polite name for any **woman**	Yes
actor	a theatrical performer	No
actress	a **female** actor	Yes

Table 2: Example definitions provided by WordNet.

5.3.5 Pronoun Link Filter

The pronoun link filter detected any coreference clusters that are linked with just pronouns. Our definition of structural Gender Generalization requires at least one gender-neutral human entity in each datapoint. If a cluster contained only pronoun links, the original mention happened in the scope outside of the sentence which was considered. Thus, these sentences were removed from the labeling pool before they were sent to the human labelers.

5.4 Crowdsourcing

The labeling task was designed and implemented on the crowdsourcing platform Figure Eight[4] (previously known as CrowdFlower). The questionnaire form was created based off of a template for categorical labeling of data provided by the crowdsourcing platform. The categories presented to the labelers were "*Gender Generalization*" , "*not a Gender Generalization*" and "*problematic sentence*". The third option was added as a choice for labelers to indicate when a sentence did not have any mention of a human entity, if the sentence was not grammatical and if the sentence was wrongly picked up by our filters.

Labelers were presented with 10 sentences per page and a limit was set to 100 judgements per labeler. Each page of the task contained a random number of golden sentences to ensure the quality of labelers. The golden set is a set of 20 sentences which were labeled by gender and language experts. The golden sentences were used as a mechanism to filter good labelers from bad labelers. The labelers had to label correctly 80% of the golden sentences presented to them in order for their results to be taken into account. Each sentence needed three trusted judgments at a minimum before obtaining the final label.

To ensure better quality of the data, additional measures were taken to ensure labelers were taking the time to understand the proposed definition of Gender Generalization. Level 2 contributors who were endorsed as experienced, higher accuracy contributors on Figure Eight were chosen to participate in the task. This provides us with a set of labelers that were more experienced. Each time a new page of 10 sentences were presented, the labelers had to spend a minimum of 120 seconds on each page. Equally, the Google translate option on Figure Eight was disabled for labelers while participating in this task in order to preserve the context of the sentences presented to them.

6 Results

Once the 15,000 datapoints from IMDB train set were split into sentences, the dataset contained 180,119 sentences. The 9 Project Gutenberg corpora yielded a total of 55,966 sentences. The

[4] https://www.figure-eight.com/

search space of IMDB was reduced to 7876 candidate sentences for the labeling pool, representing 4.4% of the original set used. The search space of all Project Gutenberg ebooks was reduced to 1627 sentences, representing 2.7% of the original data. It is important to note that the quality of the pretrained models used in the filters can impact the sentences retained.

As a preliminary test to validate the quality of sentences filtered from IMDB, randomly chosen 1000 sentences were sent for labeling. It was observed that sentences provided from movie reviews were person specific and they contained information about specific movie characters, actors or directors rather than displaying gender assumptions towards gender-neutral human entities. This introduced too much noise in the data and the quality of the filtration was altered accordingly. Thus, true Gender Generalization sentences were less likely to be found even after going through human validation due to vast noise in the data. This suggests that finding adequate data sources for Gender Generalization is important and confirms our hypothesis that good source for Gender Generalization is dependent on the style of writing.

Corpora from Project Gutenberg on the other hand contained sentences that can be applied to general population, making them more relevant to Gender Generalization bias. We present our result on label quality in the later section. Furthermore, it is observed that the amount of Gender Generalization candidate sentences represented a small fraction of each corpus explored from Project Gutenberg.

As seen in table 3, the search space for Gender Generalization was greatly reduced when the filtering approach was undertaken. This allowed for only the relevant sentences to be validated by human labelers. Reducing the search space helps human labelers to focus on one type of syntactic structure, which can directly impact the quality of final labels. Finally, 808 out of 1627 filtered sentences were accepted as Gender Generalization bias which accounts for 49.7% of filtered data across our corpora and 819 are labeled as not Gender Generalization bias.

Source	S	$\frac{\Sigma C}{\Sigma S}$	$\frac{\Sigma T}{\Sigma C}$
Boy Scouts Handbook	6330	3.6%	61.2%
Business Hints for Men and Women	2162	4.0%	63.6%
The Common Law	6101	5.8%	53.8%
The Constitution of the United States of America	21920	2.2%	44.3%
The Federalists Papers	5981	1.5%	27.0%
the Hive and the Honey-Bee: A Bee Keeper's Manual	4430	4.0%	36.5%
Magna Carta	407	4.4%	55.5%
Official Handbook of the Girl Scouts	7687	1.9%	51.7%
Practical Mind-Reading	948	4.5%	55.8%
All Corpora	55,966	2.9%	49.7%
Mean	-	2.9%	55.4%
Standard deviation	-	1.4%	11.9%

Table 3: Candidate and Gender Generalization sentences by source - {S: total number of sentences in each corpus, C: sentences remaining after filtration, T: sentences identified as true Gender Generalization bias}

6.1 Quality of Judgments

A total of 6123 judgements were made on the potential Gender Generalization candidates (a set of 1627 sentences). Out of the total amount of judgements, 4881 were trusted and accepted as final labels; these judgments represent 79.7% out of the total judgments. Each sentence was validated three times by the labelers who maintained a minimum accuracy of 80% on the golden sentences. 1242 judgements were untrusted, meaning the labelers who did not maintain an accuracy of 80% of the golden sentences were not accounted for in the final labeling; these judgments represent 20.3% of the total judgements.

Full agreement of labels only happened for 637 out of 1627 presented sentences. The remaining 990 sentences had an agreement of 66.7% which means 2 out of 3 labels were in accordance per data point. The inter-rater reliability for the full set of sentences was of 61.14%. Consequently, we decided to investigate a random subset of sentences to evaluate the quality and to better understand the low level of agreement. A total of 108 sentences, 12 sentences from each corpus, were randomly chosen from the final labeled pool to test the quality of labels assigned to each sentence. An f1 score of 73.9% was achieved with 75% accuracy. The percentage of correctly labeled Gender Generalization sentences were 70.4% and correctly labeled not Gender Generalization bias was 79.6% respectively. Sentences falsely classified as true Gender Generalization bias exhibited gender bias that did not fall into the Gender Generalization category. Moreover, sentences which should have been filtered out that remained in the labeling pool also created confusion, suggesting that improving the quality of the filters could impact the quality of the final labels. On the other hand, falsely classified as not Gender Generalization bias sentences tend to be in longer length and contained multiple pronouns linked to different human entities. This suggests that the labeler's judgment is altered when longer attention span is required. Following this, a minimum and maximum time allocation for labeling can be studied in the future as Cooley et al. (2018) observes that predefined social attributions may affect human perception and consequently may affect our labeling.

7 Conclusion and Future Work

In this paper, we propose a gender bias taxonomy as well as a means for capturing Gender Generalization sentences. The purpose of capturing these sentences is to build a dataset so that we can train a ML classifier to identify gender bias writing as well as to see the impact of clean dataset on different ML models. In future work, we hope to propose a method to capture the other types of gender bias in text that we identified in our taxonomy. Capturing qualitative bias is a challenging task and there is a need for designing systems in order to better understand bias. The approach we took was based off the proposed definition that was translated into a fetching mechanism which can aid human validation. With an initial set of 55,966 sentences, the search space was filtered down to 1627 candidates of which 808 were labeled as Gender Generalization. The presence of Gender Generalizations in text was small and represented below 5% of each corpus explored.

Our method suggests that there is a small search space for sentences with Gender Generalizations. Future work to increase the number of fetched sentences an quality of labeling are:

- Explore different state-of-art models for filters
- Upgrade to an automatized filtering and classification mechanism to enhance the quality and quantity of the labeling pool.
- Explore different data presentation for labeling (ie. longer response time, highlighting parts of sentences, etc)
- Create different methodologies to look for different types of gender bias in text.
- Create a full dataset of different gender biases in text.

Acknowledgements

We would like to acknowledge the guidance of our mentors at Mila: Kris Sankaran, Dmitriy Serdyuk, and Francis Grégoire. Also, thank you to Professor Deborah Cameron and Professor Sally McConnell-Ginet for taking the time to speak with us via email and Skype. Lastly, we'd like to acknowledge the aiforsocialgood.ca summer lab, where this project was originally created.

References

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

Deborah Cameron. 2005. Language, gender, and sexuality: Current issues and new directions. *Applied linguistics*, 26(4):482–502.

Erin Cooley, Hannah Winslow, Andrew Vojt, Jonathan Shein, and Jennifer Ho. 2018. Bias at the intersection of identity: Conflicting social stereotypes of gender and race augment the perceived femininity and interpersonal warmth of smiling black women. *Journal of experimental social psychology*, 74:43–49.

Maryann Z Corbett. 1990. Clearing the air: some thoughts on gender-neutral writing. *IEEE Transactions on Professional Communication*, 33(1):2–6.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*.

Eduardo Graells-Garrido, Mounia Lalmas, and Filippo Menczer. 2015. First women, second sex: Gender bias in wikipedia. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 165–174. ACM.

Elizabeth L Haines, Kay Deaux, and Nicole Lofaro. 2016. The times they are a-changing... or are they not? a comparison of gender stereotypes, 1983–2014. *Psychology of Women Quarterly*, 40(3):353–363.

Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*.

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 142–150. Association for Computational Linguistics.

James W Messerschmidt, Michael A Messner, Raewyn Connell, and Patricia Yancey Martin. 2018. *Gender reckonings: New social theory and research*. NYU Press.

Thomas Newkirk. 2000. Misreading masculinity: Speculations on the great gender gap in writing. *Language Arts*, 77(4):294–300.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Sarah M Ray, Ovidio Galvan, and Jill Zarestky. 2018. Gender-inclusive educational programs for workforce development. *Adult Learning*, 29(3):94–103.

IAM Robeyns. 2007. When will society be gender just?

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. *CoRR*, abs/1804.09301.

Marilyn Schwartz. 1995. *Guidelines for Bias-Free Writing*. ERIC.

Latanya Sweeney. 2013. Discrimination in online ad delivery. *arXiv preprint arXiv:1301.6822*.

ES Unterhalter. 2015. Measuring gender inequality and equality in education. In *Proceedings of workshop hosted by UNGEI*. United Nation Girls' Initiative (UNGEI).

Sigal Vainapel, Opher Y Shamir, Yulie Tenenbaum, and Gadi Gilam. 2015. The dark side of gendered language: The masculine-generic form as a cause for self-report bias. *Psychological assessment*, 27(4):1513.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

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, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*.