Survey: Bias in NLP

Niteesh Mallela and **Pushpak Bhattacharyya**
Department of Computer Science and Engineering
Indian Institute of Technology Bombay
{niteesh, pb}@cse.iitb.ac.in

Abstract

In this paper, we present a comprehensive study of societal biases that result from the application of standard NLP tasks, focusing on how data and techniques contribute to biases and the progress towards detecting them. We look at different metrics for measurement of bias. We analyze how social biases with respect to different demographics get reflected on corpora of major entertainment source like movies and multilingual social media posts.

1 Problem Definition

Bias is the presence of any prejudice or favouring toward a person or a group based on their innate or acquired features when it comes to decision-making. Consequently, a biased algorithm is one whose conclusions are weighted in favour of a specific demographic. Bias is also exhibited in multiple components of a Natural Language Processing (NLP) system including the training data, resources, pretrained models (e.g. word embeddings) (Bolukbasi et al., 2016), and algorithms themselves. A very popular example of bias in machine translation: "He is a nurse. She is a doctor" was translated to Hungarian and back to English. The round-trip translation resulted in "She is a nurse. He is a doctor" (Sun et al., 2019a), depicting representation bias. Bias is ubiquitous, being exhibited in Caption Generation, Speech Recognition, Sentiment Analysis, Language Model, Word Embedding and so on.

It has been observed that the Entertainment industries like Bollywood, Hollywood are also riddled with social biases (Khadilkar et al., 2021a) as their content reflects social norms or beliefs in some form. (Chattarjee, 2016)(Khan and Taylor, 2018) In this work we will explain various methodologies to detect these biases in Hollywood movies[1] along with techniques to detect the demographic groups against which the bias has occurred by identifying the biased dialogue turns and also introduce a new dataset for social bias detection in Hindi along with novel multilingual training framework which helps to get better performance compared to baselines.

2 Motivation

Movie biases can be introduced by the screenwriter's own bias in addition to the story's demands. As the material may spark controversy, annoyance, and financial loss, movie production companies prefer to confirm that any bias included in a script is a result of the story's demand. Recall that the substance of films like The Last Temptation of Christ, The Birth of a Nation, and textit has caused controversy. As a result, the production companies take care to screen[2] out potentially harmful speech during the original scripting stage. Before production, the scripts go through several versions to check their content. This process is heavily dependent on human intervention, including decisions and attempts to control it.

A recent Bollywood movie on acid attack, Chhapak, was inspired from a true story of an acid attack survivor who set up an NGO and was a recipient of the International Women of Courage award. Her biopic and her initiative of *Stop Acid Sale* when released, triggered regulatory legislation that made it difficult to buy certain types of acids without legal authorization. Devising NLP methods to identify how popular entertainment influences society will be a worthy future research challenge.

An AI-supported solution to detect the bi-

[1] `https://imsdb.com/`

[2] `https://en.wikipedia.org/wiki/List_of_banned_films`

ases existing in the screenplay at the authoring stage is urgently needed as Deep Learning (DL) models approach human-level accuracy in a variety of tasks. This can reduce human labour and speed up the entire scripting process. Although data is essential to DL models, there is currently no dataset that can be used to uncover biases in the domain of movie scripts.

People in India, as a multilingual country, like to express themselves in their native speech. Because social media is such a strong communication tool, any prejudice in these messages can have disastrous implications. The most important job in combating such operations is to detect bias. Hindi is the world's third most widely spoken language. Given the prevalence of Hindi content on social media platforms, early detection of biased texts in languages such as Hindi is very critical.

3 Bias in NLP

Bias is a fraught and complex term with partially overlapping, or even competing, definitions ((Campolo et al., 2017)). In sociology, bias is a prejudice in favor or against a person, group or community that is considered to be unfair. In a similar context machine learning models may make predictions that are skewed towards certain groups of people.In the field of computer vision, some face recognition algorithms fail to detect faces of black users or labeling black people as "gorillas"((Crawford, 2017)). In the field of audio processing, it is found that voice-dictation systems recognize a voice from a male more accurately than that from a female (Tatman, 2017). If we take deep NLP,word embeddings and related language models are widely trained on large databases from the Internet and may encode stereotyped biased knowledge (Garrido-Muñoz et al., 2021).

As models and datasets become increasingly large and complex, it is critical to detect the biases and evaluate the fairness of models according to multiple definitions of bias and mitigate them in learned representations. We should develop techniques that empower everyone in NLP to combat bias, that is, the "unjust, unfair, or prejudicial treatment of people related to race, age, sexual orientation, religion, gender, and other characteristics historically associated with discrimination and marginalization, when and where they manifest in algorithmic systems or algorithmically aided decision-making" ((Kai-Wei et al., 2019)). Ultimately, "people who are the most marginalized, people who would benefit the most from such technology, are also the ones who are more likely to be systematically excluded from this technology" because of bias of the machine learning models ((Kai-Wei et al., 2019)).

In this section we will will a thorough survey of the methods available for detecting various types of biases in NLP pipeline. We will also discuss the different metrics used by the researchers to quantify the bias detected by the system along with the mitigation methods to alleviate their impacts.

4 Detection of Bias

In this section, we discuss methodologies to detect the biases across variety of NLP tasks along with the work around for many of them.As word embeddings are one of the main building block of the neural NLP models, we will also focus on the detection of embedding biases and the algorithms to debias them.

4.1 Text Representations

Word embeddings have become an important component in many NLP models and are widely used for a vast range of downstream tasks. However, these word representations have been proven to reflect social biases (e.g. race and gender) that naturally occur in the data used to train them ((Caliskan et al., 2017); (Garg et al., 2017)). For example, the vector for the adjective honorable would be close to the vector for man, whereas the vector for submissive would be closer to woman. These stereotypes are automatically learned by the embedding algorithm and could be problematic if the embedding is then used for sensitive applications such as search rankings, product recommendations, or translations. An important direction of research is to develop algorithms to debias the word embeddings. In (Alipourfard et al., 2018) authors noticed that while using state-of-the-art word embeddings in word analogy tests, "man" would be mapped to "computer programmer" and "woman" would be mapped to "homemaker." This bias toward woman triggered the authors to propose a method to debias word embeddings by proposing a method that respects the embeddings for gender-specific words but debiases embeddings for gender-neutral words by following these steps:

1. Identify gender subspace. Identifying a direction of the embedding that captures the bias (Bolukbasi et al., 2016).

2. Hard debiasing or soft debiasing:

   (a) Hard debiasing (neutralize and equalize). Neutralize puts away the gender subspace from gender-neutral words and makes sure that all the gender-neutral words are removed and zeroed out in the gender subspace ((Bolukbasi et al., 2016)). Equalize makes gender-neutral words to be equidistant from the equality set of gendered words (Bolukbasi et al., 2016).

   (b) Soft bias correction. Tries to move as little as possible to retain its similarity to the original embedding as much as possible, while reducing the gender bias. This trade-off is controlled by a parameter (Bolukbasi et al., 2016).

Following on the footsteps of these authors, other future work attempted to tackle this problem (Zhao et al., 2018b) by generating a gender-neutral version of (Glove called GN-Glove) that tries to retain gender information in some of the word embedding's learned dimensions, while ensuring that other dimensions are free from this gender effect. This approach primarily relies on Glove as its base model with gender as the protected attribute. However, a recent paper (Gonen and Goldberg, 2019) argues against these debiasing techniques and states that many recent works on debiasing word embeddings have been superficial, that those techniques just hide the bias and don't actually remove it. Many of the debaising works related to word embedding have focused on binary labels(e.g. male/female); but most real-world demographics attributes like race, religion are not binary. Many of these demographic attributes have more than 2 sub-categories. (Manzini et al., 2019a) paper has described a technique to debias the word embeddings for multiclass demographic attributes. A recent work (Brunet et al., 2018) took a new direction and proposed a preprocessing method for the discovery of the problematic documents in the training corpus that have biases in them, and tried to debias the system by perturbing or removing these documents efficiently from the training corpus. Most debiasing techniques, however, concentrate on post-processing pre-trained word embeddings. In a very recent work (Zhao et al., 2019a), authors target bias in ELMo's contextualized word vectors and attempt to analyze and mitigate the observed bias in the embeddings. They show that the corpus used for training of ELMo has a significant gender skew, with male entities being nearly three times more common than female entities. This automatically leads to gender bias in these pre-trained contextualized embeddings. In (May et al., 2019a) authors extend the research in detecting bias in word embedding techniques to that of sentence embedding. They try to generalize bias-measuring techniques, such as using the Word Embedding Association Test (WEAT (Caliskan et al., 2017)) in the context of sentence encoders by introducing their new sentence encoding bias-measuring techniques, and the Sentence Encoder Association Test (SEAT). They used state-of-the-art sentence encoding techniques, such as CBoW, GPT, ELMo, and BERT, and find that although there was varying evidence of human-like bias in sentence encoders using SEAT, more recent methods like BERT are more immune to biases. That being said, they are not claiming that these models are bias-free, but state that more sophisticated bias discovery techniques may be used in these cases, thereby encouraging more future work in this area.

4.2 Natural Language Understanding Tasks

Coreference resolution is the task to identify all the entity mentions in a given text which correspond to the same entity. (Webster et al., 2018) defines and measures biases through a disparity in correctly resolving pronoun-name relationships for the male and female genders using Gendered Ambiguous Pronouns (GAP) dataset ((Webster et al., 2018)). The Maybe Ambiguous Pronoun (MAP) dataset ((Cao and Daumé III, 2020)) expands GAP to go beyond binary genders with a broader dataset. Both WinoBias ((Zhao et al., 2018a)) and Winogender ((Rudinger et al., 2018)) generate Winograd schema style datasets to investigate occupational gender stereotypes. Additionally, (Lu et al., 2018) create simple sentence templates to evaluate biases using the ratio of accurate pronoun resolution for stereotypical vs non-stereotypical occupational associations. While all of the above can potentially cover additional demographics and undesired associations, it is important to question which is more applicable to investigate harms faced by a group.

Natural Language Inference determines the directional relationship between two sentences, as to whether the second sentence (hypothesis) is entailed, contradicted, or neutral to the first sentence (premise). (Dev et al., 2019) demonstrates how the task captures and mirrors stereotypical associations (with binary gender, religion, etc) learned by text representations. Their bias measure consists of a dataset with sentence pairs: one sentence with an explicit demographic attribute (e.g., gender), and the other with implicit, stereotypical associations (e.g., occupations). Bias is measured as the accuracy of models in identifying that all sentences have no directional relation, i.e., classified having the 'neutral' label. Since an overall score is calculated for bias over a set of templates, a variety of templates can be independently assessed together to evaluate fairness of NLI model outcomes across multiple demographic groups, thus not restricting measurements to a single stereotype.

Sentiment or language polarity analysis of text is useful for understanding consumer perception from reviews, tweets, etc. However, this task has been demonstrated to be stereotypically influenced by demographic characteristics such as race and gender ((Kiritchenko and Mohammad, 2018)), age ((Diaz et al., 2018)) and names of individuals ((Prabhakaran et al., 2019)). Existing works keep sentence templates constant between samples and change the assumed demographic attribute of the person in a sentence (e.g.,through changing names). This ideally should not change the sentiment classification of the sample— changes in sentiment indicate the existence of stereotypical associations. Since evaluation hinges on this contrast in classification across groups, bias against a group is also measured in comparison to another. (Elazar and Goldberg, 2018) has show the protected demographic information like gender, age leaks into intermediate representation of neural networks trained on text data for an emoji-based sentiment detection task. They also have suggested an adversarial method to decrease the leakage of protected attribute while performing the sentiment detection task.

Question Answering models perform reading comprehension tasks and also propagate stereotypical associations from underlying language representations, as demonstrated through UnQuover ((Li et al., 2020)). In this work, biases exhibited by QA systems are measured using constructed sentence templates containing limited direct demographic information (e.g., names) accompanied by underspecified questions containing no related demographic information. The setup is such that all sub-categories of a demographic attribute (e.g., religion: Christian, Buddhist, etc) should be equally predicted as the answer. A statistically significant, higher value for one sub-category is interpreted as bias. This gives us the understanding of comparative biases across several demographic dimension values and is a closer reflection of the complexities of real-world biases.

(De-Arteaga et al., 2019) set up a measure for evaluating bias in text classification where the task is to predict a person's occupation given their biography.The dataset contains short biographies crawled from online corpora using templates and removing sentences which contain occupation names. Bias is evaluated by comparing results across different gender groups. (Zhao et al., 2020) extend the original dataset to Spanish, French, and German. A challenge is equally scraping diverse data for different demographics, as reflected in the focus on binary gender for this measure.

Toxic language ranges from more explicitly offensive forms (e.g., vulgar insults) to more subtle forms (e.g., microaggressions). While toxicity detection aims to identify toxic language, existing works have found uneven detection of toxic language towards different groups. (Prabhakaran et al., 2019) show that there are varying levels of toxicity towards different names. (Dixon et al., 2018) analyze biases in a toxicity classification model through the Wikipedia Talk Pages dataset as well as through a templated test set. Jigsaw ((Jigsaw, 2019)) contains comments from the Civil Comments platform labeled with six types of toxicity (e.g., toxic, obscene, etc) and identity attributes (e.g., white, woman, etc).

Hate speech detection is the task of identifying abusive language that is specifically directed towards a particular group. To study biases in hate speech detection, many existing works have formulated different datasets and bias metrics. (Davidson et al., 2017) and (Founta et al., 2018) annotate Twitter datasets for hate speech detection. (Blodgett et al., 2016) provide a corpus of demographically-aligned text with geolocated messages based on Twitter. (Sap et al., 2019a); (Xia et al., 2020) use those datasets to show racial biases through a higher false positive rate for AAE, while (Davidson et al., 2019a) use the dataset of (Blodgett et al., 2016) for racial bias evaluation by comparing probabilities of tweets from different social groups being predicted as hate speech.

4.3 Natural Language Generation Tasks

Autocomplete generation is the task of having a language model generate continuations from a prompt. (Sheng et al., 2019) and (Huang et al., 2020) both curate sets of prompts containing different demographic groups to prompt for inequalities in generated text. The former uses a regard metric to measure social perception towards groups, and the latter uses distributional differences in sentiment scores. Whereas these two works manually curate prompt sets, (Dhamala et al., 2021) extract the beginnings of Wikipedia articles to collect the BOLD dataset of prompts about various demographic groups. The authors then use several metrics (sentiment, toxicity, regard, etc) to measure biases in generated text. There are also works that extract existing prompts and augment the prompt set with manual annotations. For example, (Groenwold et al., 2020) use extracted African American English prompts to create a parallel set of White-Aligned English Twitter prompts and compare the sentiment of generated texts. While manually constructed prompts allow for more targeted evaluations, automatically extracted prompts allow for more comprehensive and syntactically-varied evaluations.

For **machine translation**, the English WinoMT dataset ((Stanovsky et al., 2019)) is a widely used dataset for quantifying gender biases. By concatenating examples from Winogender ((Rudinger et al., 2018)) and WinoBias ((Zhao et al., 2018a)), the authors create a challenge set to assess translations of stereotypical and nonstereotypical occupations for gendered coreference associations. There are also extensions of WinoMT for different languages ((Kocmi et al., 2020)) and datasets collected through mining ((Webster and Gonen, 2020)). Bias metrics for translation typically rely on translation accuracy. A challenge for translation bias measures is obtaining correct translations in several languages, which is perhaps simpler for manually constructed prompts with similar syntax.

Dialogue generation is similar to autocomplete generation in that both require the model to generate a text continuation given some prompt. The differences lie in the use contexts—dialogue generation is used for specific tasks (e.g., patient help) within some domain (e.g., healthcare). (Liu et al., 2019) construct a Twitter based dataset with parallel context pairs between different groups, and (Liu et al., 2020) rely on extracted conversation and movie datasets to evaluate gender biases. Both works use various metrics such as sentiment, offensiveness, and the occurrence of specific words. (Dinan et al., 2019) present an example of a bias measure that uses a crowdsourced dataset (LIGHT from (Urbanek et al., 2019)) to evaluate gender biases—in this case, through the percentage of gendered words. There are many possible bias metrics for this open-ended task and limited examination on trade-offs between different metrics.

5 Metrics for measurement of Bias

5.1 Implicit Association Test (IAT)

In psychology, the Implicit Association Test (IAT) is used to measure subconscious gender bias in humans, which can be quantified as the difference in time and accuracy for humans to categorize words as relating to two concepts they find similar versus two concepts they find different ((Greenwald et al., 1998); (Caliskan et al., 2017)). For example, subjects are much quicker if they are told to label insects as unpleasant and flowers as pleasant than if they are asked to label these objects in reverse. The fact that a pairing is faster is taken to indicate that the task is more easy, and therefore that the two subjects are linked in their mind subconsciously. Similarly, to measure subconscious associations of genders with arts and sciences, participants are asked to categorize words as pertaining to (males or the sciences) or (females or the arts) (Nosek et al., 2009). The participants are then asked to categorize words as pertaining to (males or the arts) or (females or the sciences). If participants answered faster and more accurately in the former setting, it indicates that humans subconsciously associate males with the sciences and females with the arts.

5.2 Word Embedding Association Test (WEAT)

(Caliskan et al., 2017) propose the Word Embedding Association Test (WEAT) as a way to examine the associations in word embeddings between concepts captured in the Implicit Association Test (IAT) intended to assess implicit stereotypes held by test subjects, such as unconsciously associating stereotypically black names with words consistent with black stereotypes. It is considered that this measure is analogous to reaction time in the IAT, since the shorter time implies a semantic 'nearness' ((Mcdonald and Lowe, 1998)).

1. **Adopting Psychological Tests WEAT :** The null hypothesis is that there is no difference between the two sets of target words in terms of their relative similarity to the two sets of attribute words.In formal terms, let X and Y be two sets of target words of equal size, and A,B the two sets of attribute words

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B) \tag{1}$$

where each addend is the difference between the mean of cosine similarities of the respective attributes:

$$s(w, A, B) = \text{mean}_{a \in A} \cos(w, a) - \text{mean}_{b \in B} \cos(w, b) \tag{2}$$

In other words, s(w,A,B) measures the association of the word w with the attribute, and s(X,Y,A,B) measures the differential association of the two sets of target words with the attribute. To compute the significance of the association between (A, B) and (X, Y) a permutation test on s(X, Y, A, B) is used.

$$p = \Pr\left[s\left(X_i, Y_i, A, B\right) > s(X, Y, A, B)\right] \tag{3}$$

where the probability is computed over the space of partitions (X_i, Y_i) of $X \cup Y$ so that X_i and Y_i are of equal size. The effect size is defined to be

$$d = \frac{\text{mean}_{x \in X}\, s(x, A, B) - \text{mean}_{y \in Y}\, s(y, A, B)}{\text{std\_ dev}_{w \in X \cup Y}\, s(w, A, B)} \tag{4}$$

The idea is that the more positive the value given by WEAT, the more the target X will be related to attribute A and target Y to attribute B. On the other hand, the more negative the value, the more target X will be related to attribute B and target Y to attribute A. Commonly these values are between +/-0.5 and +/-2. The ideal score is 0. [3]

5.3 Sentence Embedding Association Test (SEAT)

SEAT, as proposed by(May et al., 2019b), compares sets of sentences, rather than sets of words, by applying WEAT to the vector representation of a sentence. Because SEAT operates on fixed-sized vectors and some encoders produce variable-length vector

[3] https://www.kdnuggets.com/2020/08/word-embedding-fairness-evaluation.html

sequences, SEAT uses pooling as needed to aggregate outputs into a fixed-sized vector. We can view WEAT as a special case of SEAT in which the sentence is a single word. In fact, the original WEAT tests have been run on the Universal Sentence Encoder ((Cer et al., 2018)).

To extend a word-level test to sentence contexts, each word is slotted into each of several semantically bleached sentence templates such as "This is $< word >$.", "$< word >$ is here.", "This will $< word >$.", and "$< word >$ are things.". This design tries to focus on the associations a sentence encoder makes with a given term rather than those it happens to make with the contexts of that term that are prevalent in the training data; a similar design was used in a recent sentiment analysis evaluation corpus stratified by race and gender ((Kiritchenko and Mohammad, 2018)).

5.4 Mean Average Cosine Similarity (MAC)

WEAT as proposed by (Caliskan et al., 2017) provides a geometric interpretation of the distance between two sets of target words and two sets of attribute words. The mean average cosine similarity (MAC) uses the intuition behind WEAT and applies this notion to a multiclass domain as proposed by (Manzini et al., 2019a). Instead of comparing the associations of one target set T_1 and an attribute set A_1, to the association of T_2 and A_2, MAC considers the association of one target set T_1 to all attribute sets A at one time.

The MAC metric is computed by calculating the mean over the cosine distances between an element t in a target set T to each element in an attribute set A, as seen below equation, in which the cosine distance is defined as $cos_{distance}(t, a) = 1 - cos(t, a)$. This is repeated for all elements in T to all attribute sets. The MAC then describes the average cosine distance between each target set and all attribute sets.

$$s_{MAC}\left(t, A_j\right) = \frac{1}{\left|A_j\right|} \sum_{a \in A_j} \cos_{\text{distance}}(t, a) \quad (5)$$

5.5 Metrics to Detect Bias in NLG Tasks

Language generation tasks often involve stochastic generation of open-ended and lengthy texts, traits that are not directly compatible with traditional algorithmic bias definitions (e.g.,equalized odds, equal opportunity, demographic parity ((Dwork et al., 2011)). Because of the difficulty in defining metrics, existing works define bias loosely as demographic inequality and use intermediate proxy metrics to comparatively measure bias. Examples include:

1. **Regard Ratio:** negative-neutral-positive regard score ratios of text generated from bias-inducing prompts. ((Sheng et al., 2019))

2. **Sentiment Ratio:** negative-neutral-positive sentiment score ratios of text generated from African American English (AAE) versus White-Aligned English (WAE) prompts. ((Groenwold et al., 2020))

3. **Individual and Group Fairness through Sentiment:** comparisons of the sentiment distributions of generated text across demographics and prompts. ((Huang et al., 2020))

4. **Gendered Word Co-occurrence Score:** mean and standard deviations of the absolute log ratio of probabilities: $P(word|femaleterms)$ to $P(word|maleterms)$ across all words in generated text. ((Bordia and Bowman, 2019))

6 Bias in Movie Scripts

Despite the fact that there have been many efforts to identify social biases in texts ((Sap et al., 2017; Kagan et al., 2020; Garcia et al., 2014; Xu et al., 2019)), less attention has been paid to doing so in the area of entertainment. Various studies demonstrate the gender disparity and stereotypes in popular culture Fast et al. (2016), Ramakrishna et al. (2015) and Ramakrishna et al. (2017). rely on linguistic characteristics to measure age, racial, and gender disparities in literature and film Khadilkar et al. (2021b). Unlike other earlier efforts, this one compares gender prejudice and other subtle biases between the two major film industries of Bollywood and Hollywood using Cloze tests and WEAT measurements.

Despite the fact that there have been many efforts to identify social biases in texts, less attention has been paid to doing so in the area of entertainment. Various studies demonstrate the gender disparity and stereotypes in popular culture. rely on linguistic characteristics to measure age, racial, and gender disparities in literature and film. Unlike other earlier efforts, this one compares gender prejudice and other subtle biases between the two major film industries of Bollywood and Hollywood using Cloze tests and WEAT measurements.

7 Multilingual Bias Detection

The presence of social bias in language representations is mainly due to the undesired and skewed associations within the training data. Considering the increasing societal impact of NLP applications, studying these undesired relationships is the scientific endeavour ((Bender and Friedman, 2018; Crawford, 2017)). The initial works to tackle this issue aimed at measuring and mitigating gender biases from word embeddings ((Bolukbasi et al., 2016; Caliskan et al., 2017; Zhao et al., 2017; Garg et al., 2017; Sun et al., 2019b)). Additionally, There have been multiple works to detect race, religion bias in word embedding ((Manzini et al., 2019b)). Many follow-up works ((May et al., 2019c; Zhao et al., 2019b; Kurita et al., 2019)) have also focused on contextualised language representation like BERT.

More recently, many datasets ((Nadeem et al., 2021; Nangia et al., 2020; Sap et al., 2020)) have been created to measure social biases like gender, race, profession, religion, age, etc. Blodgett et al. (2021) has reported that these datasets lack clear definitions and have ambiguities and inconsistencies in annotations. Researchers have also investigated these biases through various NLP tasks like machine translation ((Stanovsky et al., 2019; Savoldi et al., 2021)), question answering((Li et al., 2020)), coreference resolution((Webster et al., 2018)).

There have been a lot of notable efforts towards detection of data bias in hate speech and offensive languages ((Waseem and Hovy, 2016; Davidson et al., 2019b; Sap et al., 2019b; Mozafari et al., 2020)). Borkan et al. (2019) has discuss the presence of unintended bias in hate speech detection models for identity terms like islam, lesbian, bisexual, etc. Recent studies have also investigated the usefulness of counter-factual data augmentation ((Dixon et al., 2018; Nozza et al., 2019; de Vassimon Manela et al., 2021)) to reduce the effect of unintended bias in these tasks.

However, most of the research in bias detection, mitigation are in English language and have focused on western culture. Few recent works have explored the issue of social bias in languages such as Arabic, Italian, Spanish, French, and Korean ((Lauscher et al., 2020; Sanguinetti et al., 2020; Zhou et al., 2019; Kurpicz-Briki, 2020; Moon et al., 2020)). There are very few research works towards tackling this challenge on Indian context. Pujari et al. (2019) explore bianry gender bias in Hindi languages and Gupta et al. (2021) investigate gender bias in Hindi-English machine translation using different fairness metrics. Sambasivan et al. (2021) analyze and discuss multiple dimensions of algorithmic fairness in India. Through a detailed qualitative study, the authors suggest seven potential dimension of algorithmic unfairness in India such as, Caste, Gender, Religion, Ability, Class, Sexual Orientation, Ethnicity.

Kumar et al. (2021) released a multilingual dataset in four languages like Hindi, Bangla, Meitei, and Indian English. The dataset has social media comments which are mostly code-mixed with English and annotated for labels like gender bias, religion bias, class bias, and ethnic bias. In this paper, we majorly focus on the political bias, personal attacks, religion bias, and other biases like race, gender, etc. in Hindi language.

8 Summary

In this paper we have summarized all the previous works related to different biases in variety of NLP tasks like Natural language understanding and generation tasks and also discussed about various metrics for measurement of bias. Some works related to multilingual bias were discussed.

References

Nazanin Alipourfard, Peter G. Fennell, and Kristina Lerman. 2018. Can you trust the trend: Discovering simpson's paradoxes in social data. *CoRR*, abs/1801.04385.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of african-american english. *CoRR*, abs/1608.08868.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping norwegian salmon: an inventory of pitfalls in fairness benchmark datasets. 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 1004–1015.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 491–500, New York, NY, USA. Association for Computing Machinery.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard S. Zemel. 2018. Understanding the origins of bias in word embeddings. *CoRR*, abs/1810.03611.

Aylin Caliskan, Joanna Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356:183–186.

Alex Campolo, Madelyn Rose Sanfilippo, Meredith Whittaker, and Kate Crawford. 2017. *AI Now 2017 Report*. AI Now Institute at New York University.

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.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Sushmita Chattarjee. 2016. 'english vinglish' and bollywood: what is 'new' about the 'new woman'?

Kate Crawford. 2017. The trouble with bias.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019a. Racial bias in hate speech and abusive language detection datasets. *CoRR*, abs/1905.12516.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019b. Racial bias in hate speech and abusive language detection datasets. *CoRR*, abs/1905.12516.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *CoRR*, abs/1703.04009.

Maria De-Arteaga, Alexey Romanov, Hanna M. Wallach, Jennifer T. Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Cem Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. *CoRR*, abs/1901.09451.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2232–2242, Online. Association for Computational Linguistics.

Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. 2019. On measuring and mitigating biased inferences of word embeddings. *CoRR*, abs/1908.09369.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: dataset and metrics for measuring biases in open-ended language generation. *CoRR*, abs/2101.11718.

Mark Diaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. *Addressing Age-Related Bias in Sentiment Analysis*, page 1–14. Association for Computing Machinery, New York, NY, USA.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2019. Queens are powerful too: Mitigating gender bias in dialogue generation. *CoRR*, abs/1911.03842.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2011. Fairness through awareness. *CoRR*, abs/1104.3913.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. *CoRR*, abs/1808.06640.

Ethan Fast, Tina Vachovsky, and Michael S. Bernstein. 2016. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior.

David Garcia, Ingmar Weber, Venkata Rama, and Kiran Garimella. 2014. Gender asymmetries in reality and fiction: The bechdel test of social media. In *In International AAAI Conference on Weblogs and Social*.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2017. Word embeddings quantify 100 years of gender and ethnic stereotypes. *CoRR*, abs/1711.08412.

Ismael Garrido-Muñoz , Arturo Montejo-Ráez , Fernando Martínez-Santiago , and L. Alfonso Ureña-López . 2021. A survey on bias in deep nlp. *Applied Sciences*, 11(7).

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. *CoRR*, abs/1903.03862.

Anthony G. Greenwald, Debbie E. Mcghee, and Jordan L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464–1480.

Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. 2020. Investigating african-american vernacular english in transformer-based text generation. *CoRR*, abs/2010.02510.

Gauri Gupta, Krithika Ramesh, and Sanjay Singh. 2021. Evaluating gender bias in hindi-english machine translation. *CoRR*, abs/2106.08680.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83, Online. Association for Computational Linguistics.

Jigsaw. 2019. Jigsaw unintended bias in toxicity classification.

Dima Kagan, Thomas Chesney, and Michael Fire. 2020. Using data science to understand the film industry's gender gap. *Palgrave Communications*, 6(1):92.

Chang Kai-Wei, Ordonez Vicente, Mitchell Margaret, and Prabhakaran Vinodkumar. 2019. Bias and fairness in natural language processing.

Kunal Khadilkar, Ashiqur R. KhudaBukhsh, and Tom M. Mitchell. 2021a. Gender bias, social bias and representation: 70 years of b\$^h\$ollywood. *CoRR*, abs/2102.09103.

Kunal Khadilkar, Ashiqur R. KhudaBukhsh, and Tom M. Mitchell. 2021b. Gender bias, social bias and representation: 70 years of b^{h}ollywood.

Subuhi Khan and Laramie Taylor. 2018. Gender policing in mainstream hindi cinema: A decade of central female characters in top-grossing bollywood movies. *International Journal of Communication*, 12(0).

Svetlana Kiritchenko and Saif M. Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *CoRR*, abs/1805.04508.

Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020. Gender coreference and bias evaluation at WMT 2020. *CoRR*, abs/2010.06018.

Ritesh Kumar, Enakshi Nandi, Laishram Niranjana Devi, Shyam Ratan, Siddharth Singh, Akash Bhagat, and Yogesh Dawer. 2021. The comma dataset V0.2: annotating aggression and bias in multilingual social media discourse. *CoRR*, abs/2111.10390.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations.

Mascha Kurpicz-Briki. 2020. Cultural differences in bias? origin and gender bias in pre-trained german and french word embeddings.

Anne Lauscher, Rafik Takieddin, Simone Paolo Ponzetto, and Goran Glavaš. 2020. AraWEAT: Multidimensional analysis of biases in Arabic word embeddings. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 192–199, Barcelona, Spain (Online). Association for Computational Linguistics.

Tao Li, Tushar Khot, Daniel Khashabi, Ashish Sabharwal, and Vivek Srikumar. 2020. Unqovering stereotyping biases via underspecified questions. *CoRR*, abs/2010.02428.

Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2019. Does gender matter? towards fairness in dialogue systems. *CoRR*, abs/1910.10486.

Haochen Liu, Wentao Wang, Yiqi Wang, Hui Liu, Zitao Liu, and Jiliang Tang. 2020. Mitigating gender bias for neural dialogue generation with adversarial learning. *CoRR*, abs/2009.13028.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018. Gender bias in neural natural language processing. *CoRR*, abs/1807.11714.

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W. Black. 2019a. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *CoRR*, abs/1904.04047.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019b. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in 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 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019a. On measuring social biases in sentence encoders. 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 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019b. On measuring social biases in sentence encoders. *CoRR*, abs/1903.10561.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019c. On measuring social biases in sentence encoders. 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 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Scott Mcdonald and Will Lowe. 1998. Modelling functional priming and the associative boost. In *Proceedings of the 20th Annual Meeting of the Cognitive Science Society*, pages 675–680. Erlbaum.

Jihyung Moon, Won-Ik Cho, and Junbum Lee. 2020. Beep! korean corpus of online news comments for toxic speech detection. *CoRR*, abs/2005.12503.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *CoRR*, abs/2008.06460.

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. *arXiv preprint arXiv:2010.00133*.

B. A. Nosek, F. L. Smyth, N. Sriram, N. M. Lindner, T. Devos, A. Ayala, Y. Bar-Anan, R. Bergh, H. Cai, and K. et al. Gonsalkorale. 2009. National differences in gender-science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences*, 106(26):10593–10597.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended bias in misogyny detection. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 149–155.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. *CoRR*, abs/1910.04210.

Arun K. Pujari, Ansh Mittal, Anshuman Padhi, Anshul Jain, Mukesh Jadon, and Vikas Kumar. 2019. Debiasing gender biased hindi words with word-embedding. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, ACAI 2019, page 450–456, New York, NY, USA. Association for Computing Machinery.

Anil Ramakrishna, Nikolaos Malandrakis, Elizabeth Staruk, and Shrikanth Narayanan. 2015. A quantitative analysis of gender differences in movies using psycholinguistic normatives. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1996–2001, Lisbon, Portugal. Association for Computational Linguistics.

Anil Ramakrishna, Victor R. Martínez, Nikolaos Malandrakis, Karan Singla, and Shrikanth Narayanan. 2017. Linguistic analysis of differences in portrayal of movie characters. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1669–1678, Vancouver, Canada. Association for Computational Linguistics.

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

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining algorithmic fairness in india and beyond. *CoRR*, abs/2101.09995.

Manuela Sanguinetti, Gloria Comandini, Elisa Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, and Irene Russo. 2020. Haspeede 2 @ evalita2020: Overview of the evalita 2020 hate speech detection task.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019a. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019b. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. *ACL*.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334, Copenhagen, Denmark. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *CoRR*, abs/2104.06001.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *CoRR*, abs/1909.01326.

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.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019a. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019b. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. *CoRR*, abs/1903.03094.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Kellie Webster and Hila Gonen. 2020. Automatically identifying gender bias in machine translation using perturbations.

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(0):605–617.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. *CoRR*, abs/2005.12246.

Huimin Xu, Zhang Zhang, Lingfei Wu, and Cheng-Jun Wang. 2019. The cinderella complex: Word embeddings reveal gender stereotypes in movies and books. *PLOS ONE*, 14(11):1–18.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. *CoRR*, abs/2005.00699.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019a. Gender bias in contextualized word embeddings. *CoRR*, abs/1904.03310.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019b. Gender bias in contextualized word embeddings.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. *CoRR*, abs/1804.06876.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. *CoRR*, abs/1809.01496.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. *CoRR*, abs/1909.02224.