How Biased Are We?
Automated Detection of Gendered Language

Ananya
Department of Computer Science
University of California, Irvine
aananya@uci.edu

Sameer Singh
Department of Computer Science
University of California, Irvine
sameer@uci.edu

Abstract

Detecting the different ways in which gender bias is encoded in language is an important and challenging task, however, it currently requires a manual description of how *gendered language* is expressed. Apart from requiring considerable effort for each language/domain/author, such an analysis does not provide a quantifiable measure of this bias, nor is it identified at the sentential level. Based on the intuition that gendered language is indicative of the gender of its mentions, we automatically annotate data to train a classifier to predict the gender of any mention from its context. Such a classifier is then used on unseen text to predict the gender of its mentions, the confidence of which indicates the level of bias in the text. We present preliminary implementation and results in this abstract.

1 Introduction

Gendered language is the use of words and phrases that discriminate the gender. Examples of gendered language can be found in the use of stereotypes e.g. linking women to homemakers and men to programmers (Bolukbasi et al., 2016) or when pronouns, adverbs, adjectives, noun are used carelessly, e.g. when the masculine pronoun "he" is used to refer to both sexes or when the masculine or feminine pronoun is used exclusively to define roles by sex e.g. using "her" when talking about a nurse. Due to its importance in understanding societal phenomena, detecting gendered language has been an area of active interest. Bias in language has been studied across different fields like teaching evaluations by students (Centra and Gaubatz, 2000), high school textbooks (Otlowski, 2003; Gharbavi and Mousavi, 2012; Hamid et al., 2008; Macaulay and Brice, 1997), Wikipedia edits (Recasens et al., 2013), media content (Ali et al., 2010; Len-Ríos et al., 2005; Smith, 1997) and sports journalism (Eastman and Billings, 2000; Tyler Eastman, 2001; Kinnick, 1998; Fu et al., 2016). These approaches to detect bias are either domain-specific, or rely on techniques such as counting male and female occurrences, or require manual annotations, construction of keywords and lexicons, carrying out surveys etc., or focus on whether the corpus on a whole is biased towards a particular gender. With the large scale and variety of text readily available for analysis, there is a crucial need for robust, automated, and domain-independent methods to detect and quantify the gender bias in language.

In this paper, we present a framework that requires minimal supervision and uses deep learning to detect bias at mention level. We are interested in two kinds of gender biases here: **factual bias**, and **stereotypical bias**. We define factual bias as the bias occurring in text due to objective events that cannot be reported in a gender-neutral way. For illustration, the sentence "Queen *Elizabeth* was born in 1533." Stereotypical bias, however, depends on the particular linguistic construction (lexical and syntactic), and thus reflects on the author (and their perception of the readers).

In this paper, we propose a pipeline for statistical modeling of bias and provides a concrete way to quantify it. In particular, we automatically train a classifier for sentence-level gender bias detection based solely on NER-tagged text and quantitatively analyze the *gendered*-ness of any sentence using this model. The main advantages of our pipeline and method are: (1) *Flexibility*, in that it can be applied across different domains with minimal manual intervention, (2) *Sentence-level* detection, as opposed to article- or corpus-level analysis in most of the previous work, providing more granularity, and (3) *Quantitative Measure* of

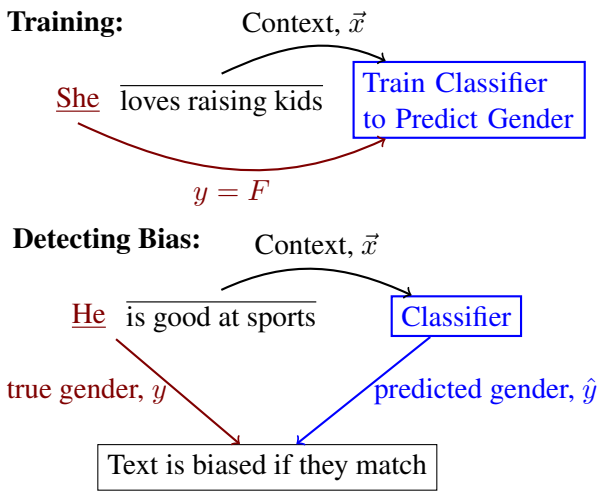

Figure 1: **Overview:** Using an unlabeled corpus, we train an accurate classifier to predict the gender for each mention given the context (i.e. with mention hidden). On each mention in the target sentences, we check whether the predicted gender matches the true prediction, with level of agreement indicating the *gendered*-ness of the text.

the extent of "gendered"-ness of a sentence, allowing large-scale, detailed analyses and comparisons. We present a concrete instance of this framework for news articles, using a bi-directional LSTM, and show examples of gendered and gender-neutral occurrences automatically identified by our approach.

2 Gender Bias Detector

By definition, gendered language is the use of words and phrases that discriminate[1] the gender, or, in other words, the gender of the mentioned people should be easy to predict from context if the text is gendered, and otherwise not. Humans learn to detect gender-biased language in a given context based on a lifetime experience of reading and observing society, and learning the types of events and language that are unique to each gender. When we observe such events and language occurring with the gender they are unique to, we detect it as a use of biased language.

In this paper, we use this intuition to train an accurate classifier that can predict the gender of individual mentions based on their context, and use it to quantify how discriminative a piece of text is in determining the gender of the mentioned people. As illustrated in Figure 1, a large corpus of text is used to train the classifier (with minimal manual supervision), the predictions of which can be used to quantify the *gendered*-ness of any other sentence from the same domain.

[1] in the machine learning sense of the word

Classifier	Accuracy
Random	67.4
Support Vector Machines	69.0
Logistic Regression	69.4
Our LSTM model	73.8
Human estimate[2]	77.5

Table 1: **Accuracy of Gender Prediction** on held-out sentences spanning 11 years. [2]Based on a small subset of 100 sentences, labeled by one user.

Here we focus on news articles as the domain in which we train the classifier to predict the gender of named entities and pronouns from the context they appear in. Using a subset of NYT containing nearly a million sentences tagged with NER (as released by Napoles et al. (2012)), we classify each named person mention and pronoun as male or female, the former using a combination of the Genderize API, US-census data, and phonetic features (only confident predictions, > 0.75, were retained), resulting in 2.2M male and 1.1M female references. For porting this pipeline to a new domain, we would require an appropriate mention detector.

We use two LSTMs (Gers et al., 2000) to encode the context of each mention (one for the text before the mention, and one for after), with a sigmoid layer on the concatenation of the final hidden states to predict the gender of the mention. We train the classifier using the data described with binary cross-entropy loss and Adam optimization algorithm, including dropout and early stopping for regularization.

3 Preliminary Results

Here we describe the accuracy of our classifier in detecting gender from context, and example sentences we estimate as biased.

Evaluation of Gender Detection Before using the classifier to estimate the gender bias, it is crucial to ensure it is an accurate proxy of a reader in predicting gender from text. On evaluation data, the classifier achieves an accuracy of 73.8% (Table 1 presents the accuracy of different classifiers). This is impressive given that news articles mostly contain factual descriptions that should be gender neutral. In order to evaluate the difficulty of the

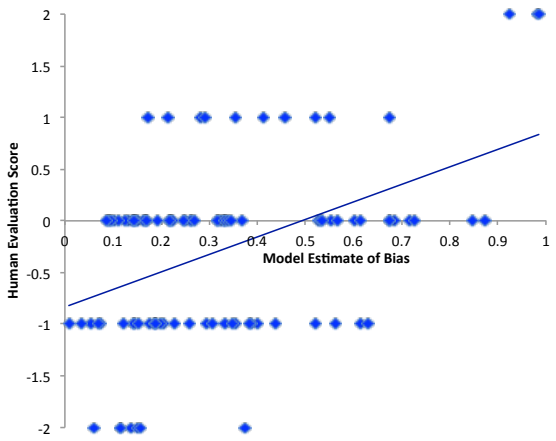

Figure 2: **User-Model Correlation** computed between a human's guess on whether a mention is female (higher score indicates higher confidence that it is female, and lower indicates male) and our model's estimate of feminine bias (as noted by the probability of the female gender by our classifier). The correlation coefficient is 0.21 here.

task, we selected a sample of 100 sentences, and presented them to a user whilst hiding the mention itself. This user was able to correctly predict the gender only 74% of the time, suggesting that our classifier may be reasonably encoding the context to accurately predict gender from it. Although further evaluation is needed, it is clear that the classifier is able to reasonably encode the context to accurately predict gender from it.

Detecting Gendered Language We use the classifier's probability of the *correct* gender as an estimation of how gendered the language is: a high probability indicates the gender is heavily reflected in the context. To demonstrate that our model exhibits such a behavior, in Figure 2 we show that the user's confidence in how *female* the context is correlates with the predicted probability from our classifier. We present examples of detected bias in Table 2; the first two show contexts for which a high level of bias is estimated, while the classifier has very low confidence for the third example, indicating gender-neutral use of language.

4 Discussion and Future Work

We presented an implementation and preliminary evaluation of an unsupervised gender bias detector, however there remain a number of shortcomings that provide exciting avenues for future work.

Female Bias: *Dress by ■ has a lace bra top.*
Male Bias: *Most Sunnis voted for a coalition called Iraqiya , led by ■, a secular Shiite who served as an interim Prime Minister in 2004.*
Gender-Neutral: *"It's giving ordinary citizens a whole new power they never had before," said ■, author of the book 'The Virtual Community.'*

Table 2: **Examples:** Sentences for which our detector predicts a high level of gendered language usage for the mentions (the ■'s), along with one gender-neutral usage.

In this work, we do not delineate between factual information and the intentional use of the stereotype as gendered language. Since news primarily consists of factual information, the predictions of a model trained on news articles mostly capture the factual/societal gender bias, not so much the linguistic/stereotype ones. It is thus important to note that this measure of bias is only a slight judgment of the author (insofar as they select the facts to present), but instead a reflection on our society.

We will characterize the types of bias we are unable to capture, by identifying the assumptions behind this pipeline and our approach. For instance, our automated pipeline relies on working POS and NER taggers, and uses the Genderize API, phonetics cues, and Census lists to determine genders of mentions, which may introduce systematic, domain-specific errors into the annotations. As another example, it is worth noting that since the current English language use is mostly limited to binary gender identities (both in grammar and in usage), we have treated gender as binary in this work, however we do recognize genderqueer and/or non-binary identities.

Last, we analyze only news articles here, which differ substantially from most other domains since news articles mostly contain facts, and not personal opinions/views. On the other hand, fiction, for example, explicitly exhibits writer's opinions and views about how different genders are perceived in society, and what the expected norm is. In news, one would find more factual bias; while in fiction, we might find more stereotypical bias. It would thus be interesting to study multiple domains, to not only estimate the bias in each, but also to investigate how it is expressed across domains.

References

Omar Ali, Ilias N Flaounas, Tijl De Bie, Nick Mosdell, Justin Lewis, and Nello Cristianini. 2010. Automating news content analysis: An application to gender bias and readability. In *WAPA*. pages 36–43.

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.

John A Centra and Noreen B Gaubatz. 2000. Is there gender bias in student evaluations of teaching? *The Journal of Higher Education* 71(1):17–33.

Susan Tyler Eastman and Andrew C Billings. 2000. Sportscasting and sports reporting the power of gender bias. *Journal of Sport & Social Issues* 24(2):192–213.

Liye Fu, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Tie-breaker: Using language models to quantify gender bias in sports journalism. *arXiv preprint arXiv:1607.03895* .

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 2000. Learning to forget: Continual prediction with lstm. *Neural computation* 12(10):2451–2471.

Abdullah Gharbavi and Seyyed Ahmad Mousavi. 2012. The application of functional linguistics in exposing gender bias in iranian high school english textbooks. *English Language and Literature Studies* 2(1):85.

Bahiyah DatoHj Abdul Hamid, Mohd Subakir Mohd Yasin, Kesumawati Abu Bakar, Yuen Chee Keong, and Azhar Jalaluddin. 2008. Linguistic sexism and gender role stereotyping in malaysian english language textbooks. *GEMA Online® Journal of Language Studies* 8(2).

Katherine N Kinnick. 1998. Gender bias in newspaper profiles of 1996 olympic athletes: A content analysis of five major dailies. *Women's Studies in Communication* 21(2):212–237.

María E Len-Ríos, Shelly Rodgers, Esther Thorson, and Doyle Yoon. 2005. Representation of women in news and photos: Comparing content to perceptions. *Journal of Communication* 55(1):152–168.

Monica Macaulay and Colleen Brice. 1997. Don't touch my projectile: Gender bias and stereotyping in syntactic examples. *Language* pages 798–825.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*. Association for Computational Linguistics, Stroudsburg, PA, USA, AKBC-WEKEX '12, pages 95–100. http://dl.acm.org/citation.cfm?id=2391200.2391218.

Marcus Otlowski. 2003. Ethnic diversity and gender bias in efl textbooks. *Asian EFL Journal* 5(2):1–15.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *ACL (1)*. pages 1650–1659.

Kevin B Smith. 1997. When all's fair: Signs of parity in media coverage of female candidates. *Political Communication* 14(1):71–82.

Andrew C. Billings Susan Tyler Eastman. 2001. Biased voices of sports: Racial and gender stereotyping in college basketball announcing. *Howard Journal of Communication* 12(4):183–201.