Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning

Susan Leavy
University College Dublin
Dublin, Ireland
susan.leavy@ucd.ie

ABSTRACT

Artificial intelligence is increasingly influencing the opinions and behavior of people in everyday life. However, the over-representation of men in the design of these technologies could quietly undo decades of advances in gender equality. Over centuries, humans developed critical theory to inform decisions and avoid basing them solely on personal experience. However, machine intelligence learns primarily from observing data that it is presented with. While a machine's ability to process large volumes of data may address this in part, if that data is laden with stereotypical concepts of gender, the resulting application of the technology will perpetuate this bias. While some recent studies sought to remove bias from learned algorithms they largely ignore decades of research on how gender ideology is embedded in language. Awareness of this research and incorporating it into approaches to machine learning from text would help prevent the generation of biased algorithms. Leading thinkers in the emerging field addressing bias in artificial intelligence are also primarily female, suggesting that those who are potentially affected by bias are more likely to see, understand and attempt to resolve it. Gender balance in machine learning is therefore crucial to prevent algorithms from perpetuating gender ideologies that disadvantage women.

KEYWORDS

Gender Bias, Machine Learning, Text Analytics

1 INTRODUCTION

There is a growing awareness of the effects of bias in machine learning. For instance, in a system used by judges to set parole, the evaluation of the likelihood of offending was found to be biased again black defendants [2]. Facial recognition software embedded in most smart phones also works best for those who are white and male [6]. Scoring systems, fueled by potentially biased algorithms, are increasingly being used to make decisions about people's lives in relation to finance, jobs and insurance [10]. Kate Crawford aptly captured the ultimate cause of the prevalence of gender bias in artificial intelligence; "Like all technologies before it, artificial intelligence will reflect the values of its creators" [11]. Societal values that are biased against women can be deeply embedded in the way language is used and preventing machine learning algorithms trained on text from perpetuating bias requires an understanding of how gender ideology is manifested in language.

Developers of artificial intelligence are overwhelmingly male. Those who have recognized and are seeking to address this issue are overwhelmingly female (Kate Crawford, Fei-Fei Li and Joy Buolamwini to name but a few). It follows that to avoid gender biased algorithms influencing decisions in our society, diversity in the area of machine learning is essential. The benefits of diversity in the workplace are well documented and largely stem from the inclusion of a range of critical perspectives. Diversity in the development of machine learning technologies could accelerate solutions to the issue of gender bias by improved assessment of training data, incorporation of concepts of fairness in algorithms [17] and the assessment of the potential impact of gender bias in the context of the intended use of the technology.

There have been attempts to address gender bias in machine learning through the review of learned gender-based associations and modification of the algorithms to exclude stereotypes [5]. However, there is little consideration of the decades of research that exist on the relationship between gender ideology and language. Incorporating gender theory, in particular feminist linguistic theory, into the approach to machine learning from textual data may prevent learning of gender bias and avoid the need to modify the algorithms.

2 GENDER BIAS IN LANGUAGE

Many of the debates in artificial intelligence on the topic of gender bias mirror those related to gender equality in society since the 1960s. It is important that computer scientists look to such debates so that negative consequences for women due to gender bias are not repeated. Feminist studies from the 1960s analyzed how women were often represented as passive, emotional and irrational in literature [22] and how the media presented idealized portrayals of femininity[12]. In the later part of the 20th century feminist theorists questioned the active role of language in the perpetuation of gender ideologies in society [7]. These seminal works identified ways in which gender ideology is embedded in language and how this can influence people's conceptions of women and expectations of behavior associated with gender. These gender ideologies are still embedded in text sources and result in machine learning algorithms learning stereotypical concepts of gender [5].

To ascertain the importance of addressing gender bias in machine learning, a lot can be learned from experiments in the 1970s

GE'18, May 28, 2018, Gothenburg, Sweden

ACM ISBN 978-1-4503-5738-8/18/05…$15.00

showing its damaging effects [4, 20]. These studies prompted the development of guidelines to avoid the use of gender biased or sexist language [9, 21]. For example, the publisher McGraw-Hill adopted editorial guidelines to avoid sexist language [26]. It would be unfortunate to have to wait until gender biased machine learning algorithms repeat the injustices of the past before action preventing gender bias is taken.

3 LEARNING BIAS FROM TEXT

Work within the field of stylistics on gender and language has identified recurring linguistic features of language that are attributable to gender bias [23]. This work lends itself to a computational approach to identifying gender bias and could be used to remove it from training data for a machine learning algorithm. The following demonstrates how an abstract concepts such a gender bias can be operationalized into measurable features of text that can be computationally identified. This connection of theoretical and critical perspectives on language to the feature extraction stage of machine learning is the key to addressing bias in artificial intelligence.

3.1 Naming

Gender bias can be recognized in terms used to describe groupings of men and women. For instance a father is often described as a 'family man' with no commonly used equivalent such as 'family woman' [28]. Terms such as 'single mum', 'working mother', 'career woman' and 'mother' commonly used in the media also reveals social preconceptions of women [23]. Occupational terms used in relation to women were found to be often pre-modified by a gender specification such as 'female lawyer' and 'woman judge', identifying their existence as counter to societal expectations [29].

Another manifestation of gender bias that is in decline is the use of androcentric terms such as 'he', 'him', 'man' and 'mankind' to refer to both men and women [3, 16]. However, in referring to groups, where there is an expectation that the individuals in question are more likely be of a particular gender, that gender will be used to refer to both men and women in the group [19]. For example in reference to a group of fire-fighters individuals are more likely to be referred to in male terms. In the context of machine learning, while certain linguistic features may be used less in current textual sources, machine learning algorithms that are trained on older corpora may reflect outdated ways of referring to men and women.

Women are described as girls more often than men are described as boys [28]. In an analysis of the use of the terms girl(s) and boy(s) in a corpus of text of British, American and New Zealand English, [29] found that the term 'girl' is 3 times more likely than the term 'boy' to refer to an adult and that women were described as girls in in order to characterize them as immature, innocent, of youthful appearance, subordinate status, emotionally weak or financially dependent. Using 'girl' in conjunction with occupations also reduced the status of the jobs. In [3] it was found that the terms 'boy' and 'girl' occurred with equal frequency in an analysis of examples of British English texts including literature and media content from 2006. However, 52 percent of uses of the term'girl' referenced women while 28 percent of the uses of 'boy' pertained to men. The term 'girl' was also used in more disparaging and sexual contexts . This demonstrates how techniques to analyze not only the frequency of mentions but the broader context of the use of terms for men and women in texts could detect gender bias in training data for machine learning.

Honorific titles such as 'Miss' and 'Mrs' reflect the marital status of women but the male equivalent does not, demonstrating how women are portrayed in terms of their relationships to others [21, 23]. In the 1970s 'Ms' was introduced as an equivalent for 'Mr' to address this asymmetry. However, there is evidence that 'Ms' is being used to replace 'Mrs' but not 'Miss' [16].

3.2 Ordering

Gender bias in language is evident in the ordering of items in lists. In English, it is convention when naming pairs of each gender, to name the male first (eg. son and daughter, husband and wife, Mr and Mrs) [23]. This practice demonstrates a bias which presents a gender-based social order [23–25, 31]. This practice of naming the most powerful of a pair first is evidenced by the following common pairs : 'master/servant' , 'teacher/pupil' and 'doctor/nurse' [24].

A comprehensive study of the ordering of personal binomials in the British National Corpus uncovered examples of word pairs studied included 'man/woman', 'girl/boy', nobility titles such as 'lady/gentleman', 'princess/prince', kingship terms such as 'wife/husband', occupations such as 'actress/actor' and pronouns such as 'he/she' [25]. While there were variances in the order of naming the pairs, gender was the most important influencing factor regarding which of the pair of terms was named first.

3.3 Biased Descriptions

In an analysis of adjectives used to describe men and women in British newspapers, [8] found that men were more frequently described in terms of their behavior while women were described in terms of their appearance and sexuality. In an analysis of the context of the use of the term 'girl', research has shown that girls and boys are represented differently with girls being more objectified [30] and portrayed in more negative contexts [3]. Extraction of adjectives used to describe women in training data could therefore be incorporated as part of gender proofing the textual data that is used to training machine learning algorithms.

How word embedding learns stereotypes has been the focus of recent research on gender bias and artificial intelligence [5]. Evaluating what constitutes a stereotypical association has largely been a result of researcher interpretation. However [27] analyzed the British national Corpus and extracted collocates of men and women and identifying those that were just used for each gender, revealing striking gender stereotypes (Table 1). Other kinds of stereotypes have been identified in relation to sexuality, beauty [13] and levels of agency [23].

3.4 Metaphor

Metaphor is difficult to identify automatically but is a powerful tool in the construction of gender in society [15, 18, 23]. In an endeavor to systematize the identification of metaphors in text [20] outlined five steps that could be applied to linguistic features of a text to identify whether their use was metaphorical or not. Research on the kind of metaphors used to portray men and women has identified a

Gender	Adjectives
Female	Bossy, chattering, gossiping, submissive, bitchy, hysterical, weeping
Male	Gregarious, cautious, affable, amiable, avuncular, funniest, good-natured, jovial, likable, mild-mannered, personable, cruel, dour, insufferable, braver, humane, law-worthy, patient, sincere, tolerant, trustworthy, truthful, upstanding, anxious, insane, astute, scholarly, self-educated, ignorant

Table 1: Gendered Personality Adjectives from the BNC

gender bias whereby those metaphors used to portray women are "more prolific and more derogatory than those used exclusively for men" [15, 18].

3.5 Presence of Women in Text

Straightforward frequency counts of women in text can be a powerful indicator of gender bias. In the British National Corpus, 'Mr' occurs more often than 'Mrs', 'Miss' and 'Ms' combined [28]. Furthermore, mentions of individual men, as distinct from mentions of men as a general category, occurred twice as often as mentions of individual women [27]. In an analysis of business literature, [14] also found that mentions of men occurred 10 times more often than mentions of women and that of the total mentions of terms of terms of address (including Mr, Ms, Mrs, and Miss), 93.5 percent were occurrences of 'Mr'. In a study of 3.5 million articles from British newspapers, automated methods were devised to identify the gender of subjects referenced in newspaper articles [1] . It was found that men were referenced in 49 percent of top stories while women were referenced in 18 percent. Based on this, a simple quota system for the gender balance in in training data for machine learning algorithms may serve to combat much of the latent bias in text based sources of training data.

4 CONCLUSIONS

Identifying gender bias in training data for machine learning algorithms is a complex but not an insurmountable task. The previous section shows how such an abstract concept can be operationalized and captured in computationally identifiable linguistic features of language. While the fact that machine learning algorithms can learn gender bias can be of interest to researchers looking to understand its prevalence in society, it is not an advantage in practical applications making decisions about people's lives. There is an emerging focus on fairness in machine learning generally and it is is essential that women are at the core of who defines the concept of fairness. Advancing women's careers in the area of Artificial Intelligence is not only a right in itself; it is essential to prevent advances in gender equality supported by decades of feminist thought being undone.

REFERENCES

[1] Omar Ali, Ilias Flaounas, Tijl De Bie, Nick Mosdell, Justin Lewis, and Nello Cristianini. 2010. Automating news content analysis: An application to gender bias and readability. In *Proceedings of the First Workshop on Applications of Pattern Analysis*. 36–43.

[2] J Angwin, J Larson, S Mattu, and L Kirchner. 2016. Machine bias risk assessments in criminal sentencing. *ProPublica https://www. propublica. org* (2016).

[3] Paul Baker. 2008. *Sexed texts: language, gender and sexuality.* Equinox.

[4] Sandra L Bem and Daryl J Bem. 1971. *Training the woman to know her place: The social antecedents of women in the world of work.* Department of Psychology, Stanford University.

[5] 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*. 4349–4357.

[6] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency*. 77–91.

[7] Judith Butler. 1990. Gender trouble and the subversion of identity. *New York and London: Routledge* (1990).

[8] Carmen Rosa Caldas-Coulthard and Rosamund Moon. 2010. 'Curvy, hunky, kinky': Using corpora as tools for critical analysis. *Discourse & Society* 21, 2 (2010), 99–133.

[9] Miller Casey, Swift Kate, and Dowrick Stephanie. 1981. *The Handbook of Nonsexist Writing: For Writers, Editors and Speakers.* Women's Press.

[10] Danielle Keats Citron and Frank Pasquale. 2014. The scored society: due process for automated predictions. *Wash. L. Rev.* 89 (2014), 1.

[11] Kate Crawford. 2016. Artificial intelligence's white guy problem. *The New York Times* (2016).

[12] Betty Friedan. 2001. The Feminine Mystique. 1963. *New York* (2001).

[13] Katherine Frith, Ping Shaw, and Hong Cheng. 2005. The construction of beauty: A cross-cultural analysis of women's magazine advertising. *Journal of communication* 55, 1 (2005), 56–70.

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

[15] Caitlin Hines. 1999. Rebaking the pie: the woman as dessert metaphor. *Reinventing identities: The gendered self in discourse* (1999), 145–162.

[16] Janet Holmes. 2002. Gender identity in New Zealand English. *Gender across languages. The linguistic representation of women and men* 1 (2002).

[17] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*. 656–666.

[18] Veronika Koller. 2004. Businesswomen and war metaphors:'possessive, jealous and pugnacious'? *Journal of Sociolinguistics* 8, 1 (2004), 3–22.

[19] Lia Litosseliti and Jane Sunderland. 2002. *Gender identity and discourse analysis*. Vol. 2. John Benjamins Publishing.

[20] Wendy Martyna. 1978. What does 'he' mean? Use of the generic masculine. *Journal of communication* 28, 1 (1978), 131–138.

[21] Casey Miller and Kate Swift. 2001. *The handbook of nonsexist writing.* iUniverse.

[22] Kate Millett. 2016. *Sexual politics.* Columbia University Press.

[23] Sara Mills. 1995. *Feminist stylistics.* Routledge London.

[24] Sandra Mollin. 2012. Revisiting binomial order in English: Ordering constraints and reversibility. *English Language & Linguistics* 16, 1 (2012), 81–103.

[25] Heiko Motschenbacher. 2013. Gentlemen before ladies? A corpus-based study of conjunct order in personal binomials. *Journal of English Linguistics* 41, 3 (2013), 212–242.

[26] Janice Moulton, George M Robinson, and Cherin Elias. 1978. Sex bias in language use:" Neutral" pronouns that aren't. *American Psychologist* 33, 11 (1978), 1032.

[27] Michael Pearce. 2008. Investigating the collocational behaviour of MAN and WOMAN in the BNC using Sketch Engine. *Corpora* 3, 1 (2008), 1–29.

[28] Suzanne Romaine et al. 1998. *Communicating gender.* Psychology Press.

[29] Robert Sigley and Janet Holmes. 2002. Looking at girls in corpora of English. *Journal of English Linguistics* 30, 2 (2002), 138–157.

[30] Charlotte Taylor. 2013. Searching for similarity using corpus-assisted discourse studies. *Corpora* 8, 1 (2013), 81–113.

[31] Gülşen Musayeva Vefali and Fulya Erdentuğ. 2010. The coordinate structures in a corpus of New Age talks:'man and woman'/'woman and man'. *Text & Talk-An Interdisciplinary Journal of Language, Discourse & Communication Studies* 30, 4 (2010), 465–484.