FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY WHEN TREATING GENDER AS A VARIABLE

PRANAV A* & AALIND SINGH[1]

CONTENTS

ABSTRACT

In this paper, the requirements of gender as a variable and its application in machine learning based research work are studied. Treating gender as a variable is an ethical constraint giving fairness, accountability, and transparency as the priority in academic research as well as software engineering are heavily stressed in this paper. Study on how society perceives gender as a binary construct and on how researchers and practitioners can handle it, is done. Highlights are given on why this study is important and what mistakes previous researchers did in the field of machine learning, especially on natural language processing. Guidelines for recognition of gender in machine learning are presented and implementation of these guidelines is presented in this paper.

* *Department of Computer Science and Engineering, Hong Kong University of Science and Technology*
[1] *Department of Information Technology and Engineering, VIT University, Vellore, India*

1 INTRODUCTION

Scientists and engineers in the machine learning or natural language processing fields should study the ethical implications of the careful design of a variable or a category for reporting results or research work. The natural language processing (NLP) field deals with solving linguistic problems computationally. This paper analyzes the role of accountability held by the NLP engineers and researchers and then propose some solutions as an ethical framework. The theory analyzed here lays down some principles for scientists, engineers and peer reviewers which could be applied on other social variables like race. This paper lays down the foundation of fairness, accountability and transparency principles when treating gender as a variable. The objectives of this paper are:

1. To investigate how gender is perceived from different viewpoints.
2. To define how different the role of scientists and engineers are and who should be accountable for their roles.
3. To critically analyze how gender as a variable was treated in the prior works of Natural Language Processing
4. To present guidelines on how to treat gender as a variable.
5. To extrapolate on how these concerns can be extended to race as a variable.
6. To apply the proposed framework solution

In section 2, different viewpoints regarding gender are considered. In section 3, the accountability of the role of scientists and engineers in the natural language processing are described. Section 4 shows a critical analysis of the prior works where analysis on consideration on the gender variables are made. Realizing the problems arose in the prior works, the solutions and guidelines are provided in the section 5. In the section 6, the concerns that were raised for the gender variables are extrapolated on the variables like race. Finally, in the section 7, the work is concluded with defining guidelines.

2 GENDER AS A VIEWPOINT

In this section, three main viewpoints of gender are investigated. The first and most common viewpoint of gender or folk view, on how "gender" variable is treated by people as a "sex" variable with obvious features like outward appearances and behaviour. Sex generally refers to the chromosomal and biological characteristics of a person. Gender on the other hand is the psychological identification with certain characteristics which leads to a certain outward appearance and behavioral conduct. These characteristics are "masculine" and "feminine". This realization of the distinction of male/female from masculine and feminine is new to the western world. In the eastern world gender was always viewed as a psychological and emotional concept. Hence its conflation with sex did not take place. It was realized that these characteristics of masculinity and femininity are present in various proportions in different people. So a male might be more feminine or a female might be more masculine in her behaviour. This growing understanding of gender has lead to the formation of cis- and trans-gender groups. According to New York Times reports there are 1.4 million transgender people in United States of America alone. These transgender people do not refer to their genders as other in day to day conversations. Therefore, the folk view of gender is an important frame for an NLP (Natural Language Processing) researcher, to investigate the responses of the participants when referring to there own gender concepts. [1]

The second viewpoint of the gender is according to the performative viewpoint of the gender. This is based on how gender as a variable *should* be perceived and how the behaviour of the gender variable should perform. According to Defrancisco *et. al* gender is seen as "the behaviors and appearances society dictates a body of a particular sex should perform," organizing "people's understanding of themselves and each other." According to Larson *et al.*, our gender knowledge is comprised of components of our cognitive environment—beliefs about behaviors we expect to have a particular effect or effects on others on the basis of the knowledge about a typical situation in our environment. Language is considered one such behavior.The common observation in all these theories is that the performative viewpoint of gender is not merely performance.Rather it coexists of the behaviors that respond to, or are limited by conventions and that simultaneously reinforce them [2].

The third viewpoint of gender is based on the presumption of the binary social construct of gender and building clusters around it. These clusters represent masculine and feminine qualities such as athleticism, aesthetics and so on. This viewpoint is useful for an NLP practitioner for identification of consumers exhibiting individual characteristics— such as independence and athleticism, so that a particular product could be marketed to those consumers without regard to their gender or sex. Such a methodology may not be available to NLP researchers though, because in their cases participants are required to fill surveys [3].

3 ACCOUNTABILITY OF THE ROLE OF SCIENTISTS AND ENGINEERS

In this section, the roles of NLP researchers and practitioners are defined which are analogous to the roles of scientists and engineers respectively. The differences and contrasts on their roles and how are they accountable for their work would be disscussed in this section. For Engineers the third viewpoint of gender might be quite useful in marketing a particular product to a specific set of consumers. These consumers satisfy the requirement of following certain characteristics which fall under the modes of the binary gender classification. [4]

4 CRITICAL ANALYSIS OF PRIOR WORKS

In this section, the critical remarks regarding the previous works are made where gender as a variable have been disorderly treated.

Rao *et al* have analyzed tweets of the user from the latent variables like gender and age. [5] However, the gender as a latent variable has been vaguely treated by the authors. The intent and conclusions drawn from using gender variable were unclear. Quotes like "For gender, the seed set for the crawl came from initial sources including sororities, fraternities, and male and female hygiene products. This produced around 500 users in each class." show that the classification was performed on only rudimentary heuristics.

Koppel *et al* have treated gender labeling as a text categorization problem. [6] There is a lack of analysis of the training data obtained. No information regarding fairness between the two sexes are considered.Also there is a lack of further investigative or corrective methods for such unfairness. Again, the similar points could be argued there that why binary classification was necessary and why classification system remained ambiguous in the paper.

Herring *et al* have allocated the gender variable to the authors of the blogs with conclusions based on apparent parameters like first names, nicknames, language behaviour and so on. [7] They have neither equalized the distribution between two genders while taking the training data into the consideration nor have they made

any consideration of non binary gender existence . Although, their conclusions indicate that language can not be gender specific.

Burger *et al* have made use of the profile meta-data of the bloggers for the classification of the gender variable. [8] They selected the users randomly and manually examined them for the obvious gender cues. This approach is highly unreliable as there is a little correlation between the tweet and the identity of the tweeter.

Through scraping the usage of the profile account settings of the bloggers, studies have been done in machine learning for the explicit classification of the gender. [9, 10]

These studies do not show how frequently the participants or respondents hold their genders as a significant variable during decision making. The duration for which the participants explore on the possibilities of the options of the gender and how genders other than normal binary classification are treated.

The answers to such questions are vital and form an ethical constraint. Researchers and peer reviewers should be accountable and fair on how to make such studies more valid and transparent for the readers.

5 GUIDELINES ON TREATING GENDER AS A VARIABLE

The guidelines and solutions are presented to the existing problems of treating gender as a variable in this section.

1. The need to define and demonstrate an explicit notion of gender and formulate and build the theories upon it has to be done [11]. This step is crucial to make the research more accountable, valid and transparent. Clarity of the definition and the usage would lead to better reproduction of the research work. For researchers who are doing gender analysis, it is expected to give a brief description on how they have considered gender as a variable. For the engineers, a custom textbox could be provided if they are using a gender variable with non-disclosure options.

2. Barring a few research questions where its absolutely necessary, one could avoid using gender as a variable. Many times while gathering data gender as a field is unconsciously added. This may lead to biased results and unnecessary diversion from context.

3. Gender-neutral language should be used wherever possible, showing transparent assignment of gender variable. Participants and users should know how their gender construct would be treated.

4. How to anticipate the difficulties of the respondent when asked to specify their gender. The respondents who do not identify themselves as one of the binary normative genders also find it disrespectful to be categorized as the "other" gender. They often anchor their identities on this issue as a collective protest against the normative society.

5. The practitioners of NLP who are concerned with designing a product around gender characteristics should do it more responsibly. This is because NLP here is directly intervening and influencing peoples idea of gender.

6. For an NLP researcher keeping in mind the sensitivities of the participants while also maintaining a quality research output is of utmost importance.

7. For a good scientist, doing good science should not be an ethical obligation but a part of job description.

6 EXTRAPOLATING THE CONCERNS TO RACE AS A VARIABLE

Similarly in this section, the concerns that were raised for treating gender as a variable would be extrapolated to the concerns when treating race as a variable. Likewise in the gender variables, the need of race variables should be explicitly defined and anticipate the difficulties of the user when asked.

Often it is found that researchers gather data regarding racial variables in two ways,

1. through the collection of data on the pre-determined values of the racial identities or just through reporting of the variables by the respondents.
2. quantification or a rough estimation of the collection of the elaborated data of their ancestors.

It may be noted that most researchers felt that usage and categorization of racial variables are usually unreliable in a series of experiments done by Giselle *et al* [12].

Researchers mainly view the race using three viewpoints,

1. **Race as a biological construct**: A section of researchers generally favour the idea of viewing race as a biological construct due to genetic variation. Other factors include recovery time of the diseases and variations within the allele frequencies.
2. **Race as a social construct**: Another section of researchers generally favour the idea of viewing race as a social construct due to economic variation. Other factors include recovery time of the diseases and variations within the allele frequencies.
3. **Race as a biological and social construct**: This viewpoint arises due to holding both views of the biological and social constructs. Some of them believed that genetic differences occurred due to social differences.

Most of the researchers believed that race is a futile variable and the data derived could be easily misused. From the collection of the manuscripts, it has been found that 84 percent of the researchers used the word *race* in their manuscripts, but only 10 percent of them only defined the concept of its usage. This shows that if researchers are using race as a variable, then notion of usage of this variable should be well defined.

7 CONCLUSION

This paper represents only a starting point for treating the research variable gender in an ethical fashion. The solutions provided for scientists and researchers are proposed to be clear and basic. A point to note that if someone wants to participate in some research or engineering with high ethical standards or morals, one should not view ethics as a to-do list of the execution process. One should rather anticipate the ethical implications of their work, making their work accountable, transparent and fair whenever possible. More similar concerns related to social variables like ethnicity and religion could be extrapolated to the framework presented here.

REFERENCES

[1] Robin Waterfield et al. *Meno and Other Dialogues: Charmides, Laches, Lysis, Meno*. OUP Oxford, 2005.

[2] Victoria Pruin DeFrancisco, Catherine Helen Palczewski, and Danielle D McGeough. *Gender in communication: A critical introduction*. SAGE Publications, 2013.

[3] Fredda Blanchard-Fields, Lynda Suhrer-Roussel, and Christopher Hertzog. A confirmatory factor analysis of the bem sex role inventory: Old questions, new answers. *Sex Roles*, 30(5):423–457, 1994.

[4] Brian N Larson. Gender as a variable in natural-language processing: Ethical considerations. *EACL 2017*, page 30, 2017.

[5] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM, 2010.

[6] Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412, 2002.

[7] Susan C Herring and John C Paolillo. Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4):439–459, 2006.

[8] John D Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309. Association for Computational Linguistics, 2011.

[9] Xiang Yan and Ling Yan. Gender classification of weblog authors. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 228–230, 2006.

[10] Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12(9), 2007.

[11] Brian N Larson. Gender/genre: The lack of gendered register in texts requiring genre knowledge. *Written Communication*, 33(4):360–384, 2016.

[12] Giselle Corbie-Smith, Gail Henderson, Connie Blumenthal, Jessica Dorrance, and Sue Estroff. Conceptualizing race in research. *Journal of the National Medical Association*, 100(10):1235–1243, 2008.