An Italian dataset for the analysis of gender stereotypes in textual documents*

Silvana Badaloni[1,2,†], Antonio Rodà[1,*,†] and Martino Scagnet[1,†]

[1]Department of Information Engineering, via Gradenigo, 6, 35131 Padova, Italy
[2]Elena Cornaro Center on Gender Studies, University of Padova, Italy

Abstract

The presence of stereotypes associated to historically disadvantaged groups constitutes a strong limitation for the justice and welfare of society. Gender stereotypes are among the most deep-rooted ones and have, over time, given rise to real conventions that permeate various aspects of social life, creating unfairness and sometimes discrimination. This study focuses on the possibility of identifying gender stereotypes in textual documents using Machine Learning and Natural Language Processing tools. To this end, a corpus of Italian language texts was collected and 107 participants were asked to evaluate each sentence by assigning a score that would reveal the presence of gender stereotypes (female or male). The collected data allowed the labelling of the text sections of the corpus, by assigning a "gender score" to each one. The dataset thus developed can be used to foster the development and/or evaluation of automatic tools for detecting gender stereotypes, facilitating the writing of more inclusive texts.

Keywords

gender bias, gendered innovation, fairness, artificial intelligence, machine learning,

1. Introduction

The problem of biased and unfair outcomes of AI-based systems is becoming increasingly clear. One of the main causes is that Machine Learning algorithms, by their intrinsic nature, are trained on the basis of examples, they learn from data, and therefore can subsume and capture the stereotypes related to people sharing a characteristic, for example the gender identity, which run through the data [1]. If used to make automatic decisions, these potentially biased systems could lead to unfair, incorrect decisions that could discriminate for or against some groups over others. There is the risk of being discriminatory for certain categories of users. Moreover, the triangular relationship between algorithm-human-data, which becomes increasingly relevant as collaboration between humans and AI increases, risks continually feeding the spread of biases.

While the concept of bias is very broad, gender-related biases are considered an essential aspect of fairness [2]. In particular, we believe that in the European social-cultural context, the gender biases represents a particularly interesting case study for the Artificial Intelligence community, for several reasons listed below.

2st Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming, BEWARE-23, co-located with AIxIA 2023, Rome 6-9 Novevember, Italy, 2023

*Corresponding author.
†These authors contributed equally.
silvana.badaloni@unipd.it (S. Badaloni); antonio.roda@unipd.it (A. Rodà); martino.scagnet@studenti.unipd.it (M. Scagnet)

First of all, numerous studies have shown that gender biases are deeply rooted in our society. Therefore, the risk that the datasets used for many applications with great social impact (autonomous driving vehicles, recommendation systems, personnel selection systems, etc.) contain biases linked directly or indirectly to gender is very high.

Secondly, gender biases affect more or less half of the population, so their presence has an impact on a large number of people.

Thirdly, given the wide spread of gender bias in our societies, it is relatively easy to find datasets on which to experiment with analysis and debiasing techniques.

Fourthly, in comparison with other types of bias (racial, social, etc.), it is easier to define the categories subject to possible discrimination. Gender studies, while recognising the multiplicity of gender identities, validate the existence of two well-defined polarities, male and female. The existence of two prevailing categories facilitates the definition of experimental protocols for the validation of analysis and debiasing techniques.

Fifthly, following the usual practice of bringing our research experiences back into teaching, promoting studies on gender biases in AI can facilitate the introduction of gender issues into our computer science courses, with a twofold advantage: a) increasing the degree of involvement of our female students, and b) making our male students aware of stereotypes and biases that risk discriminating against their female counterparts, making their university and professional careers more difficult.

In this paper, we will focus on this kind of bias and, in particular, we will deal with gender bias as open issue for applications based on Natural Language Processing [3]. How Word Embeddings learns stereotypes has been the focus of many research on gender bias and artificial intelligence [4]. Since Word Embeddings are used as a knowledge base in many applications, biases in these models can propagate into many NLP applications. In general, gender biases diffused in the textual corpora used for Word-Embedding are subsumed by the model: for example, words related to traditionally male professions are found closer to inherently gendered words, such as he or man, and vice versa. Techniques to reduce these biases have been recently studied [5], but the problem is still open, in particular for those language that are more grammatically gendered, as Italian [6]. Language has a profound impact on how we understand gender roles. A gender-inclusive language is, therefore, a key tool to contribute to the achievement of gender equality. Consequently, having tools to identify gender biases in texts is crucial to mitigating their propagation. However, there is still a shortage of gender bias datasets to automate gender bias detection using machine learning (ML) and natural language processing (NLP) techniques (see [7]). In particular, as far as we know, there is no specific dataset in Italian language.

The present study will focus on the possibility of automatically identifying gender stereotypes in textual documents. To this aim, a corpus of texts in Italian, labelled according to the genre (understood in a conventional way) to which the reading is addressed, has been developed[1]. Texts have been collected from various sources assuming the presence of gender stereotypes in some and gender neutrality in others. Then, voluntary participants were asked to rate the genre to which the text fragment was aimed. In the following sections, we will present the methodology used to collect the corpus and the participants' annotations. Finally, we will provide a statistical analysis and discussion of the results.

[1]The labeled dataset is available at https://doi.org/10.5281/zenodo.10027951

Table 1
Initial textual corpus.

Source	# Articles	# Sections	# Words	Adv. words for each article
www.unipd.it	30	533	25602	853
Female magazine	32	548	25066	783
Male magazine	30	542	28523	950

2. The dataset development

2.1. Materials

The first step for dataset building was the collection of an initial corpus of texts. To ensure the presence of sentences with different degrees of gender bias (both feminine and masculine), a number of articles were selected from magazines explicitly targeting either a female or a male audience. Such a choice stems from the assumption that these magazines tend to have content stereotyped for the gender they target in an attempt to maximize the number of interested readers. Indeed, it is a well-known phenomenon that men and women tend to conform to gender stereotypes in order to align with social expectations. And for this purpose, magazines that dispense advice on fashion, on body care and physical training, on managing family or love relationships, all of which are historically gendered in our society, have proven to be useful.

In addition, to have more gender-neutral content, a number of articles were selected from the website of the University of Padua (www.unipd.it), an institution that has a code of conduct to limit gender stereotypes and to make its communications more inclusive.

A total of 92 articles were collected. Each article was then divided into sections 30 to 70 words long, usually containing 2 or 3 sentences, so as to include some context necessary for understanding the text. Table 1gives details on the composition of the initial text corpus.

2.2. Participants

Each section of the corpus was assessed and labeled using a questionnaire involving 107 participants, who responded to the invitation sent by email. Of these, 31 dropped out before the completion of the questions. An additional 5 participants were excluded because they completed the task in less than 4 minutes, a time considered insufficient to provide reliable answers. Of the remaining 71 participants (mean age 45.95), 57 claimed to be female, 13 male, and 1 preferred not to specify the gender.

This data underlines the need to broaden the pool of participants in an attempt to have a sample that best reflects social reality.

2.3. Procedure

The evaluation activity was carried out using an online questionnaire, developed within the framework PsyToolkit [8, 9]. Each participant is presented with 20 items that include, in a single webpage, a section of text and an assessment scale. 18 of the proposed textual sections are randomly selected from the initial corpus. 2 items are control questions to discard any

Table 2
Scores assigend to the text sections, based on their origin in the initial corpus. M=mean, SD=standard deviation.

Source	# Sections	# Answers per item (M)	Score (M)	Score (SD)	min	max
www.unipd.it	52	6.29	-0.1774	0.4847	-1.5	+0.8
Female magazine	55	6.62	-0.7001	0.6703	-1.857	+1
Male magazine	49	6.02	+0.5109	0.8084	-1.6	2

participants who answer randomly or inattentively. These questions also consist of a section of text, but at some point it is made explicit that the item is a control question and that the participant must give a certain answer regardless of the text. The time required for each participant is about 10 minutes.

For each proposed section of text, the question posed to participants is, "You are asked to assess the gender of the reader you think the text is aimed at." with the aim that each participant, based on their own experience and culture, report the presence of gender stereotypes in the texts.

Gender rating is indicated on the following 5-point Likert scale (in parentheses the numerical value assigned to each response, hidden from the participant): Completely female (-2), More female than male (-1), Neutral (0), More male than female (+1), Completely male (+2).

2.4. Results

To increase the reliability of the dataset labels, items answered by fewer than 5 participants were discarded. The different distribution of responses to different items is due to the random assignment made by the survey system, and the possibility for participants to refuse to answer some items, presumably those with unclear or ambiguous sentences.

In the end, from the initial text corpus of Table 1, only the 156 text sections, which received statistically consistent scores, were included. Of these sections: 55 came from journals addressed to females, 49 from male journals, and 52 from texts extracted from www.unipd.it. For each question, the mean of the responses received was performed. As the Lickert scale was defined, a negative value indicates a text judged to be aimed at female readers, while a positive value indicates a bias toward male readers.

Figure 1 shows the distribution of scores assigned to each item. The average score obtained among all items is close to zero with a slight tendency toward the female end of the range. This is probably due to both the slightly higher number of category texts from female journals and the higher score (in absolute value) obtained from the sections judged as female.

In any case, the fact that the mean value approaches 0 is an indicator of a sufficient balance in the dataset between negative (female), and positive (male) scores.

Another interesting analysis performed is that made with respect to the scores obtained from the texts based on their origin in the initial corpus.

Table 2shows a statistical description of the scores obtained from the three types of sources. The scores confirm the assumption made about the sources: sections from women's journals had a negative average score (-0.70), those from men's journals a positive average score (0.51),

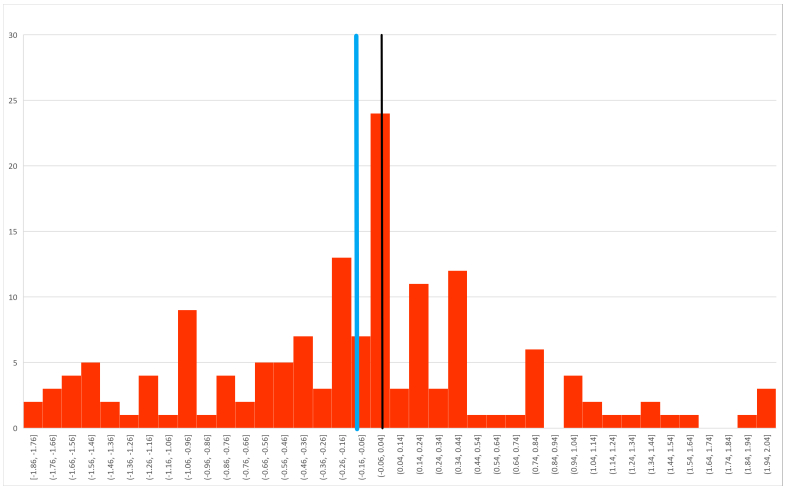

Figure 1: Histogram of the mean scores assigned to the 156 text sections in the dataset. The scale for the evaluation goes from -2 (Completely female) to +2 (Completely male). The black line represents the 0 (Neutral), the blue line the average value of all the scores.

and those from www.unipd.it a slightly negative score, but still very close to 0 (-0.17).

3. Conclusions

As stated in [10], some biases are inevitable in large language models since these models learn from vast amounts of test data and they are exposed to the biases present within human language and culture in different ways of expressions. First, there are inherent biases in language due to the fact that language is the expression of culture. Second, cultural norms and values vary significantly across communities and regions. Third, there are many definitions of fairness as it is a subjective concept. Last, language and culture are constantly evolving, with new expressions, norms and biases emerging over time. Therefore, it is important that developers, researchers and stakeholders continue to work reducing biases by developing strategies for identifying and mitigating them.

The present paper presented a novel labelled dataset to foster the development and/or evaluation of automatic tools for detecting gender stereotypes in Italian texts. The analysis of the results, and in particular the comparison between the participants' scores and the expectations deriving from the sources of the texts, supports the effectiveness of the followed methodology, based on an online questionnaire.

It is worth noting some limitations. The current dataset release includes 156 labeled text sections. A quantity that is certainly insufficient for its use as a training dataset for machine learning models. Its use is therefore more suitable as a test dataset for already trained models or for algorithms aimed at estimating the gender score in texts, such as the one proposed by [11]. Another aspect to pay attention to is that the participants in the dataset annotation are many more females than males. Although it is generally good to have a more balanced gender distribution, in this case we do not believe that this makes the annotation less reliable. Indeed, movements to denounce and raise awareness of discrimination against the females in

our society have made women more alert and aware of stereotypes in texts. In addition, the imbalance in the gender distribution of our participants is due to the fact that many more men dropped out of the annotation task before the end, thus being excluded, confirming the lower awareness and interest of males to whom the invitation to participate had come.

To the best of our knowledge, this is a first dataset of this kind for Italian texts. We plan to continue this work, significantly increasing the size of the dataset, so that it will also be suitable for training tasks and trying to avoid possible imbalances of participants from a gender point of view.

Acknowledgments

This work is partially supported by the project "Creative Recommendations to avoid Unfair Bottlenecks" of the Dept of Information Engineering of the University of Padova.

References

[1] S. Badaloni, A. Rodà, et al., Gender knowledge and artificial intelligence, in: Proceedings of the 1st Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming, BEWARE-22, co-located with AIxIA, 2022.

[2] S. Leavy, U. C. Dublin, Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning, in: Proc. of the ACM/IEEE 1st International Workshop on Gender Equality in Software Engineering, Gothenburg, Sweden, 2018.

[3] J. Doughman, W. Khreich, M. El Gharib, M. Wiss, Z. Berjawi, Gender bias in text: Origin, taxonomy, and implications, in: Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing, 2021, pp. 34–44.

[4] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, Advances in neural information processing systems 29 (2016).

[5] S. L. Blodgett, S. Barocas, H. Daumé III, H. Wallach, Language (technology) is power: A critical survey of "bias" in nlp, arXiv preprint arXiv:2005.14050 (2020).

[6] D. Biasion, A. Fabris, G. Silvello, G. A. Susto, Gender bias in italian word embeddings., in: CLiC-it, 2020.

[7] J. Doughman, W. Khreich, Gender bias in text: Labeled datasets and lexicons, arXiv preprint arXiv:2201.08675 (2022).

[8] G. Stoet, Psytoolkit - a software package for programming psychological experiments using linux, Behavior Research Methods 4 (2010) 1096–1104.

[9] G. Stoet, Psytoolkit: A novel web-based method for running online questionnaires and reaction-time experiments., Teaching of Psychology (2017).

[10] E. Ferrara, Should chatgpt be biased? challenges and risks of bias in large language models, Submitted to Machine Learning with Applications. Preprint on arXiv:2304.03738 (2023).

[11] A. Fabris, A. Purpura, G. Silvello, G. A. Susto, Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms, Information Processing & Management 57 (2020) 102377. doi:10.1016/j.ipm.2020.102377.