STUDENT FORUM

Word embeddings are biased. But whose bias are they reflecting?

Davor Petreski[1] · Ibrahim C. Hashim[2]

Received: 26 July 2021 / Accepted: 24 March 2022 / Published online: 26 May 2022

Abstract

From Curriculum Vitae parsing to web search and recommendation systems, Word2Vec and other word embedding techniques have an increasing presence in everyday interactions in human society. Biases, such as gender bias, have been thoroughly researched and evidenced to be present in word embeddings. Most of the research focuses on discovering and mitigating gender bias within the frames of the vector space itself. Nevertheless, whose bias is reflected in word embeddings has not yet been investigated. Besides discovering and mitigating gender bias, it is also important to examine whether a feminine or a masculine-centric view is represented in the biases of word embeddings. This way, we will not only gain more insight into the origins of the before mentioned biases, but also present a novel approach to investigating biases in Natural Language Processing systems. Based on previous research in the social sciences and gender studies, we hypothesize that masculine-centric, otherwise known as androcentric, biases are dominant in word embeddings. To test this hypothesis we used the largest English word association test data set publicly available. We compare the distance of the responses of male and female participants to cue words in a word embedding vector space. We found that the word embedding is biased towards a masculine-centric viewpoint, predominantly reflecting the worldviews of the male participants in the word association test data set. Therefore, by conducting this research, we aimed to unravel another layer of bias to be considered when examining fairness in algorithms.

Keywords Word embeddings · Androcentrism · Gender bias · Word association test

1 Introduction

Word embeddings are a form of word representations in an n-dimensional space. One of the most commonly used techniques in creating word embeddings is the Google developed Word2Vec. Word2Vec is a neural network that takes a text corpus as input and outputs a vector space where each word of the corpus is assigned a vector in that space. The rationale behind word embeddings is that semantically similar words are placed closer together in the vector space than dissimilar words. This makes word embeddings highly useful by allowing them to represent the semantic relationship between words in mathematical terms, making them an important and widely used component in many Natural Language Processing (NLP) models (Mikolov et al. 2013). However, despite this expansive utility, word embeddings have been found to exhibit strong, ethically questionable human biases. In the past 5 years, there has been extensive research on the topic. Specifically, we have singled out three main areas of previous work on the topic. The first area consists of works primarily dealing with the investigation into morally problematic biases present in word embeddings, such as a bias against a specific race, gender, religions, and sexual preferences among others. This research line asks the question of "What biases are exhibited in word embeddings and how are they present in downstream tasks?" (Caliskan et al. 2017; Garg et al. 2018; May et al. 2019; Zhao et al. 2017, 2018). The second area of work deals with different methods of mitigating the bias in word embeddings, in other words, debiasing (Bolukbasi et al. 2016; Manzini et al. 2019; Zhao et al. 2019). Finally, the third area consists of works focusing on the issues and the improvement of the approaches we have towards detecting and dealing with bias in word embeddings (Gonen and Goldberg 2019; Nissim et al. 2020).

✉ Davor Petreski
davorpetreski2@gmail.com

Ibrahim C. Hashim
i.hashim@maastrichtuniversity.nl

[1] University of Glasgow, Glasgow, UK

[2] Maastricht University, Maastricht, Netherlands

However, the nature, or the "roots" of the biases have seldom been investigated. Examining which social groups' bias is dominantly represented in word embeddings is crucial to further understand the issue at hand. Therefore, in this paper, we aim to investigate whether the world view of one gender is dominantly represented in word embeddings. Concurrently, we propose a novel approach to investigating biases in NLP systems through the use of Word Association Tests. We achieve this by seeking to answer the following research question:

> *To what extent are word embeddings dominated by a single gender worldview?*

In other words, we are seeking to answer whether the semantic relationships in word embeddings are predominantly conforming to a male dominant (androcentric) world view, female dominant world view (gynocentric), or neither.

In line with previous research and work in gender studies and language, we hypothesize that the word embedding will be conforming to, and reproducing an androcentric worldview. To answer our research question we used data from a word association task, which we split based on the gender of the participants. In a word association game, participants are presented with cue words to which they must respond as quickly as possible with words of their own (De Deyne et al. 2019). The similarity between the presented cue words and the responses of participants was calculated using a word embedding trained using the Word2Vec algorithm. We found that responses of male participants were significantly closer to the cue words compared to the responses of female participants. This implies that the word embedding represents an androcentric worldview.

The paper is organized as follows. In section two we attempt to conceptualize and define what bias means for the purpose of this paper, introducing the previous work on gender bias, and the use of masculine or male dominant language (androcentrism in language). In section three we outline some of the notable related research and previous work. We split this section into three subcomponents: (1) work on androcentrism in language; (2) work on bias in word embeddings; (3) work on word association tests. Consequently, in section four we discuss the methodology and results. Namely, we discuss the data and the statistical analysis, and present the results of our study. Finally, in section five we discuss the results of the study, limitations, and present some concluding remarks and directions for further research.

2 Conceptualization of bias

Most humans constantly make biased decisions. We might prefer firm and round foods compared to goey, soft ones; or we might prefer the greener, more open walk home compared to one that is shorter, but grey and narrow. These biases seem to be more of an exercise of our freedom of will, rather than unethical prejudiced decisions. However, not all biases are merely a matter of personal choice and preference, many carry unfair societal implications along. Infact, many biases are unfair or unethical acts, often directed toward a particular individual or group. Unethical, or ethically ambiguous bias is generally underlined by a different form of conceiving or treatment of another person based on their perceived characteristics. Commonly, bias towards a person is closely linked with either physical or societal attributes of that person, such as the person's gender, race, height, weight, age, and sexual orientation. Often, bias is thought to be manifested either as favoritism (positive bias; bias for), or discrimination (negative bias; bias against; Howard & Borenstein, 2018). Similarly, such biases can also be attributed to values, ideas, or words rather than only persons. For example, one might have biased assumptions about "what behaviours are healthy and what behaviors are crazy", or in relation to this research, which words are associated with a given word (e.g. soft; Kaplan 1983, p.788).

When biases, stereotypes, or simply experiences are homogenized over a specific social group, they form normalized world views, and when they are imposed onto other groups they become dominant ways of thinking and doing (Karl Dake 1991). This, in turn, creates another layer of bias, where experiences, thoughts, opinions and biases of one group are over-represented in the social reality. For the purpose of this paper, we are focusing on this layer of bias, specifically, addressing issues of gender.

With gender in mind, in many aspects of our society the world is viewed through a "male lens", meaning that reality is defined primarily from male experiences, perspectives and opinions (Epp et al. 1994, p.452). This form of bias is called androcentrism. Androcentrism is typically prevalent in research, political discourse, digital media, business practices, and educational settings when men's thoughts, opinions, behaviors and experiences are the primary subject of study, the masculine thought is normative and universilized, and the use of masculine language is present, for example, the use of male generic language (Hamilton and Henley 1982). Generally, androcentrism is accompanied by the absence or underrepresentation of female voices, worldviews, experiences, and behaviours. In systems, cultures and organizations with androcentric bias culturally marginalize and 'other' the feminine, and place the masculine at the center of their worldview (Gilman 1970). Opposed to androcentrism, the practice of placing the female worldview at the center is called gynocentrism. Therefore, we want to examine whether an androcentric, gynocentric bias or neither is present in word embeddings.

3 Background and related work

Previous work on this topic can be divided into three categories. Namely, work on androcentrism in language, previous research on bias in word embeddings, and work related to Word Association Tests.

3.1 Androcentrism in Language

Prior to the many current discussions regarding bias in algorithms and NLP models, discussions regarding gender bias and discrimination in language have been held since the inception of critical and feminist theories as early as the 1960's (Leavy 2018). Biases in our language and society are often mirrored in the algorithms we use and produce. Therefore, to better understand and conceptualize the contemporary issues regarding bias in NLP algorithms, it is important to first revisit and familiarize ourselves with previous work on gender bias in language and society.

Although very much related and entangled, work on this problem can be split in two: firstly, works that deal with bias in how women are represented in language mediums such as literature, film, and journalism; and second work on gender biases present in the language itself on a semantic level.

The main issues that were uncovered within works on the former are idealizations, misrepresentations and undermining of femininity and women. For example, multiple scholars, across multiple countries, languages and cultures have argued that in literature characteristics such as hysterical, subordinate, passive, irrational, powerless, and similar others, are more often attributed to women (Bankey 2001; De Valdés 2010; Ramanathan 1996; Millet 2016). The issue stretches beyond literature. In TV commercials in the United States women were more often represented as unemployed, in a domestic setting, and passive. On the other hand, the active role of the narrator was given to men 90% of the time (Bretl and Cantor 1988). Further studies have analyzed the objectification of the feminine and the woman through sexualization and idealization in language. For example, Nanda (2014), has analyzed the portrayal of women in children's stories and fairy tales emphasizing the prominence of the feminie beauty ideal, and the association of beauty with good (beautiful princess), and homeliness with bad (ugly witch) in female characters.

Besides examining gender biases in the use of language in different media, feminist scholars have also questioned and examined the role of language itself in (re)producing gender biases. Seminal works such as Robin Lakoff (1973) and Key (1975) both give a very detailed account of how biases and prejudices against women are linguistically submerged within the English language, and particularly, both paint a picture of the gendered nature of the English language. Moreover, androcentrism in semantics has been addressed in recent research as well. For example, significantly more often men are described with generic, gender-neutral labels such as person or human, whereas descriptions of women are more likely to have gender-specific labels such as lady, miss, or woman. This implies that in semantic terms, men are nested in humanity, whereas women are less associated with humanity and more gender-specific (Bailey et al. 2020). The solution that many linguists and feminist scholars see, is in the production and use of more androgynous language. For instance, in the last chapter of 'Male/Female language', Key addresses androcentrism, and calls for a move towards a more androgynous language (1975).

In this paper, we are concerned with addressing androcentrism in Word Embeddings because the semantic prevalence of androcentrism in everyday language is evident. Consequently, through data, this androcentrism has the potential to translate into gender bias within the Word Embedding.

3.2 Short history of bias in word embeddings

Word embedding is the technique of representing words using vectors for the purpose of rendering the semantic relationships between words in geometric terms. Because bias and stereotypes are present in the large corpora of texts that word embedding models are trained on, word embeddings also exhibit these biases and stereotypes. Further, word embeddings are found to even amplify the pre-existing biases in the training data (Zhao et al. 2017). This is especially worrying due to the widespread use of word embeddings in real-world applications such as search rankings, curriculum vitae parsing, sentiment analysis, product recommendations, conference resolution systems and machine translations (Garg et al. 2018; Zhao et al. 2018).

Caliskan et al. (2017) have found that both morally neutral (e.g. types of flowers) and morally problematic (e.g. race, gender, class) historical biases are present in word representations. Gender bias is specifically and thoroughly addressed in Bolukbasi et al. (2016), where through simple analogies and comparisons regarding professions or gender roles, gender bias in word embeddings is surfaced, e.g. "man" is closer to "doctor" and "woman" is closer to "nurse", or "computer programmer" is to "man" as "homemaker" is to "woman". Word embeddings have further been shown to reflect gender and ethnic stereotypes in the data, making them a compelling tool for research in the social sciences. For example, Garg et al. (2018) uses word embeddings for quantitative social research, where they analyzed and quantified the presence and the changes of gender stereotypes across the past century. Similarly to Bolukbasi et al. (2016), Manzini et al. (2019) showed the existence of not only binary bias (such as binary gender), but multi-class bias as well (such as race and religion). However, the use of analogies for bias detection in word embeddings has concurrently been controversial. Namely, analogies have been shown to be an inaccurate and incompetent diagnostic tool for bias in word embeddings (Gonen & Goldberg, 2019; Nissim et al. 2020). Nissim et al. (2020), discuss several issues such as the methods of implementation, or the subjective choices of researchers when using analogies as a tool to measure and detect bias in word embeddings. These issues might have led to an unrealistic picture of bias in world embeddings, where some non-existing biases are exacerbated, and others, existing ones, are hidden. Nonetheless, Nissim et al. (2020) does not question the existence of gender bias in word embeddings, but the methods used to detect this bias. Moreover, gender bias has been researched and measured in various different types of word representations such as contextualized word embeddings (Zhao et al. 2019), or sentence embeddings (May et al. 2019) without the use of analogies.

Promising direction for research has been work on developing algorithms for debiasing word embeddings (Zhao et al. 2019). For instance, Bolukbasi et al. (2016) have developed a post-processing method to mitigate the bias in word embeddings. Their proposed method removes gendered components from gender-neutral words. Similar method is used by Manzini et al. (2019) for debiasing multi-class biases such as race and religion. However, Gonen and Goldberg (2019), argue that the debiasing methods used in Bolukbasi et al. (2016), are insufficient, and that the original bias can still be recovered from the debiased word embeddings, even though gender components are removed from the biased subspace. In fact, after the removal of biased components, the embeddings with related biases are still clustered together in the vector space. (Gonen and Goldberg 2019). With this in mind, Manzini et al. (2019), extend on the debiasing methods of Bolukbasi et al. (2016) by attempting to mitigate this “cluster bias” as well. Despite the recent focus on debiasing methods and attempts to mitigate bias in word embeddings, it is argued that we should focus on transparency and awareness as well (Caliskan et al. 2017; Gonen & Goldberg, 2019; Nissim et al. 2020).

In this paper, we are not looking to approach bias to uncover and measure gender stereotypes in word embeddings or to mitigate these biases. We use data from word association tasks and word embeddings to pave a new direction in research that aims to uncover where these biases stem from, and whether word embeddings and the data they are trained on is dominated by the worldview of one societal group.

3.3 Word association tasks

The word association task stems from “word association games” which are simple tasks where the subject is presented with a word and they are prompted to quickly respond with the first word that comes to their mind. These games may seem simple, however, when done on a large scale and methodologically in the form of large scale word association tests, they can serve as a research tool to understand “internal representations and processes involved in word meaning and language” (De Deyne et al. 2019). In many accounts, word association tasks are considered as one of the most simple and unbiased approaches to measuring and understanding human semantic knowledge and the meaning of words in the human mind (De Deyne et al. 2019; Jackendoff and Jackendoff 2002; Steyvers and Tenenbaum 2005).

Word association tests have been previously used in gender bias research on word embeddings as well. Du et al. (2019) used word association tests as a representation of real-world bias and then checked whether bias present in word embeddings reflects the real world, “true bias levels”. Similarly to Du et al. (2019) we are using the results of word associations tests to understand word meanings and semantic knowledge in humans and then comparing this with data from word embeddings. However, we are not focusing on gender bias and stereotypes within the data, but we are examining whether the “language” (the meaning of words derived from word association tests) of one gender is more represented in word embeddings than the language of the other.

4 Methodology

In this section, we will present the datasets and methods used for our experiment, as well as the results.

4.1 Data

We evaluated the bias based on data obtained from the preprocessed English dataset of the Small World of Words (SWON-EN2018: Preprocessed) (De Deyne et al. 2019). This dataset contains the responses of 83,864 participants in a word association task to 12,292 cue words. It is currently the biggest dataset of a word association task available in English. Each cue word was judged by exactly 100 participants. Each participant gave three responses. The responses were designated as R1, R2 and R3. Examples are given in Table 1. The data was split according to the gender of the participants (either male or female). Participants where gender was not given (276 participants) were excluded from the analysis. The initial age range of participants was from the ages 16 to 100. To have some control over the age, we

Table 1 Example cue words and responses from the Small World of Words dataset

Cue Word	Participant Gender	R1	R2	R3
Simple	Female	Pilgrim	Shaker	Song
Further	Male	Farther	How	More
Sleep	Female	Time	Lullaby	Tea
Divide	Male	Conquer	Division	Math

This table gives an insight into the data collected as part of the Small World of Words dataset. Each participant gave 3 responses (R1, R2, R3) to each cue word that they were presented with

decided to exclude all participants above the age of the lower (24 years) and below the age of the upper (47 years) interquartile ranges of the original data. This led to the female participants ($M = 33.0$, SD = 6.7) and male participants ($M = 32.3$, SD = 6.6) being similar in age. In the end, the data of 37,645 participants were used in the analysis for this experiment (44.9% of original data). Of these, a total of 23,702 (63.0%) were female.

The word embedding we employed was trained on the GoogleNews corpora (GoogleNews-vectors-negative300 corpora)[1] using the Word2Vec algorithm. For each participant, a normalized mean cosine distance from the cue words to the response words was calculated. Specifically, for every participant the cosine distances between each cue word and their three responses given were calculated and summed up. The lower the cosine distance is between two words, the closer they are in the vector space of the model. As the distance between two words in the vector space can be thought of as their semantic similarity, the lower the cosine distance the closer the words are semantically. The cosine distances will give an indication of the level of representation of the mental lexica of males and females in the model and can allow us to determine if one is more dominantly present. Cues and responses not present in the vocabulary of the trained model were excluded. As different participants had a different number of total responses due to exclusions, summed up cosine distances were normalized to 1 response.

4.2 Statistical analysis

The statistical significance of the bias present in the word embedding trained on the GoogleNews corpora was assessed using an independent *t* test. The continuous independent variable was the mean normalized cosine distance between the cue words and the responses per participant. The dependent variable was gender. Furthermore, a permutation test was carried out for 10,000 permutations where the gender labels were shuffled. The differences in the means between the gender vectors were then evaluated.

4.3 Results

4.3.1 Responses

The mean number of eligible female responses ($M = 42.7$ SD = 8.0) was similar to the mean number of eligible male responses ($M = 41.7$, SD = 8.3). The percentage of eligible male responses compared to the total responses was 97.4% (SD = 2.8%). This was nearly identical for female participants 97.0% (SD = 2.9%).

4.3.2 Mean

The mean distance for the male responses from the cue word in the word embedding ($M = 0.2588$, SD = 0.0557) was found to be significantly lower ($p < 0.001$) than for the female responses ($M = 0.2620$, SD = 0.0528) according to an independent t-test. Furthermore, the 95% confidence interval for males [0.2578, 0.2597] was found to not overlap with that for females [0.2614, 0.2627], further indicating a significant difference between the two (see Fig. 1). Finally, the permutation test also resulted in a significantly higher mean for the male distances compared to the female distances ($p < 0.001$).

5 Discussion and conclusion

The current study is a preliminary investigation into whether results from word association tasks can give an insight into the bias of worldviews present in word embeddings. The main aim was to determine whether the investigated word embedding, trained on the GoogleNews dataset, was prone to an androcentric, gynocentric world view, or neither. The results of the study fit with the expected hypothesis. The word embedding trained on the GoogleNews dataset was found to represent a more androcentric world view. The average response of male participants in the word association task was semantically closer to the cue words, according to the word embedding, compared to the average response of female participants. This indicates that the mental lexicon of the model is closer to that of the average male lexicon than the average female lexicon.

Our findings are not entirely novel, nor unexpected. In fact, much of the previous research on gender bias in word embeddings, and general research on androcentrism in language has already pointed to results suggestive of what we have found. Recent works on gender bias in word embeddings, have discovered a disturbing amount of male–female gender bias, and have further tried to minimize or mitigate

[1] https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit

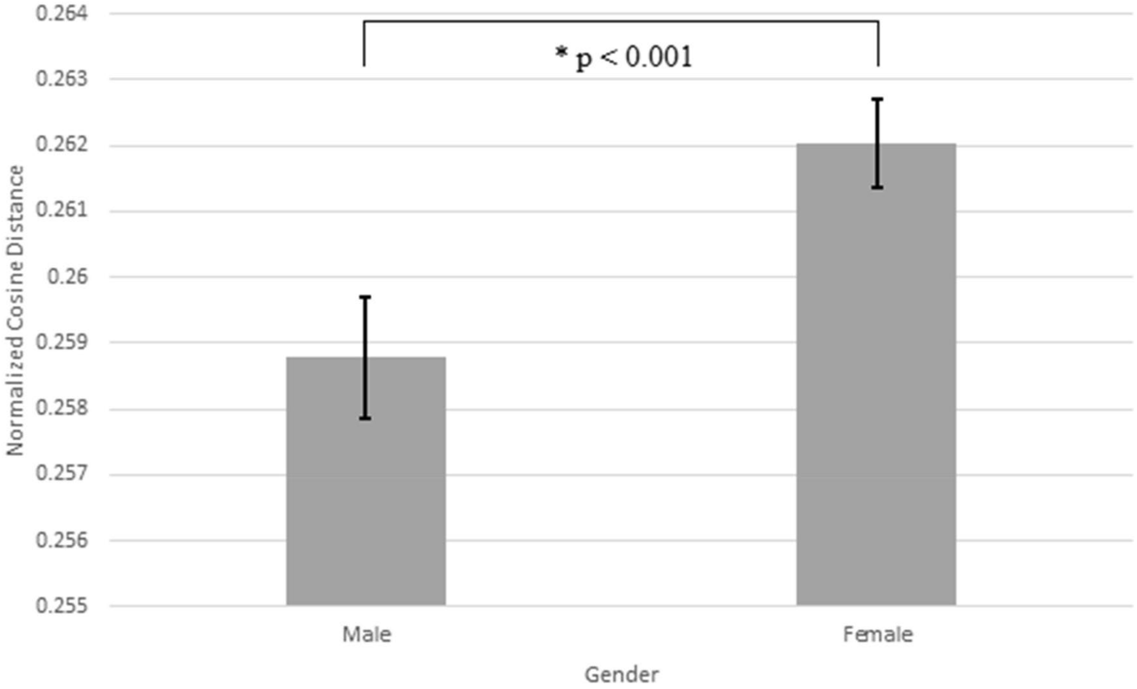

Fig. 1 Comparison of the average normalized cosine distances for male and female responses. *Note.* This figure shows for both male and female participants the mean normalized cosine distance between the responses and the cue words based on the word embedding system. The error bars show the 95% confidence interval. The lack of overlap between the error bars signifies a significant difference between the means. The p-value is the result of the independent t-test

these biases (Bolukbasi et al. 2016; Gonen and Goldberg 2019; Zhao et al. 2019). In a different line of research, critical feminists and linguists have been raising the issue of androcentrism, the dominance of the male voice in everyday language, since the 1960's (Leavy 2018). However, our research is still different and distinct from previous work on the topic.

On the one hand, our point of interest when examining word embeddings is not the particular content and context of the gender bias (such as profession, associations, adjectives, etc.), but the domination of androcentric language on a broader scale within word embeddings. Hence, our findings have distinct meaning and implications when it comes to gender bias in word embeddings. Whereas previous work has shown the presence and different manifestations of gender bias in word embeddings, our work might point to clues about where this bias comes from and suggest the presence of a deeper layer of bias within word embeddings.

On the other hand, unlike much of the research mentioned on androcentrism that is situated in contexts such as in literature, media or public discourse, our research examines androcentrism in a different technological and historical setting. Our findings point to the fact that even after 60 years, androcentrism still prevails even within the latest technologies such as NLP and machine learning. This potentially implies that the progress in technology does not necessarily translate into progress in societal fairness. Nevertheless, as observed in the literature on debiasing of word embeddings, practitioners and scholars in the field are predominantly focused on the technical applications and technical solutions to the societal problems that arise from them. Perhaps, we ought to seek alternatives to this techno-solutionist approach, and look towards approaches based on socio-ethical practices and policies such as promoting diversity, transparency and awareness.

Furthermore, our work can potentially contribute to novel avenues of quantitative linguistic and sociological research. Most of the research on androcentrism is qualitative, and the quantitative research mainly focuses either on examining the representation of the feminine in language as a medium of communication, or the semantic manifestations of androcentric biases in the language (Bailey et al. 2020; Bankey 2001; Bretl & Cantor, 1988; De Valdés 2010; Ramanathan, 1996; Millet 2016). In contrast, this work offers an approach to quantitatively examine how much of the language used and produced by certain technologies or the media, is androcentric.

5.1 Limitations

The main limitation of the current study comes from the dataset used. The dataset was not created for the purpose of this study. The participants comprised a wide age range, had different education levels and had different native tongues. Furthermore, there were unequal numbers of male and female participants. These factors were not controlled for in the dataset as that was not required for the study of De Deyne et al. (2019). All of these can play a role in which responses they would give in a word association task. Therefore, the results of the current study are not as strong as they could be if data was collected specifically for the aim. However, one should view the current study as what it is, not an exhaustive investigation into the link between word embeddings and word association tasks, but a preliminary inquiry.

5.2 Future directions

Several future directions of research are available. A repeat of the current investigation using data specifically collected for the aim of the study would be the best next step. The data can be collected for a word association task similar to De Deyne et al. (2019), however, the participants should be controlled for age, education level and native tongue. Furthermore, equal numbers of male and female participants should be collected. Also, only cue words present in the vocabulary of the word embedding system should be used. Finally, the participants should be limited to responding with only words, not phrases. We are also firmly aware that the gender-binary worldview we have taken in this paper might be discriminatory towards some people. However, this choice was made due to the availability of the data. For the future, we propose extending the study to also include non-binary participants. In the end, we have only used one type of word embedding technique on one dataset. It would be of interest to use our technique to test other datasets and word embedding techniques commonly used in NLP.

The current study looked into the influences of gender worldviews onto word embedding systems, however, other biased worldviews exist. Future studies could investigate the similarity between word embeddings and different morally problematic biased worldviews, such as those dependent on social class, education level, race and age among others. Lastly, future research examining the implications of these findings in downstream tasks and the effect of androcentrism on bias in NLP applications is needed.

Author contributions Both authors contributed equally to the paper.

Funding The author(s) received no financial support for the research, authorship, and/or publication of this article.

Availability of data and material (data transparency) The data that support the findings of this study are openly available: In the research resources of smallworldofwords.org at https://smallworldofwords.org/en/project/research. In the Pre-trained word and phrase vectors section of Google's word2vec archive at https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing.

Code availability (software application or custom code) The custom code used for data cleaning, processing and analysis will be made available once the paper has been conditionally accepted.

Declarations

Conflict of interest The author(s) declare(s) that there is no conflict of interest.

References

Bailey AH, LaFrance M, Dovidio JF (2020) Implicit androcentrism: Men are human, women are gendered. J Exp Soc Psychol 89:103980

Bankey R (2001) La Donna é Mobile: Constructing the irrational woman. Gender Place Cult 8(1):37–54

Bolukbasi T, Chang KW, Zou J, Saligrama V, Kalai A (2016) Man is to computer programmer as woman is to homemaker? debiasing word embeddings. arXiv preprint arXiv:1607.06520

Bretl DJ, Cantor J (1988) The portrayal of men and women in US television commercials: a recent content analysis and trends over 15 years. Sex Roles 18(9–10):595–609

Caliskan A, Bryson JJ, Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. Science 356(6334):183–186

Dake K (1991) Orienting dispositions in the perception of risk: an analysis of contemporary worldviews and cultural biases. J Cross Cult Psychol 22(1):61–82

De Valdés ME (2010) The shattered mirror: representations of women in Mexican literature. University of Texas Press, Austin

De Deyne S, Navarro DJ, Perfors A, Brysbaert M, Storms G (2019) The "small world of words" English word association norms for over 12,000 cue words. Behav Res Methods 51(3):987–1006

Du Y, Wu Y, Lan M (2019) Exploring human gender stereotypes with word association test. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp 6135–6145

Epp JR, Sackney LE, Kustaski JM (1994) Reassessing levels of androcentric bias in educational administration quarterly. Educ Adm Q 30(4):451–471

Garg N, Schiebinger L, Jurafsky D, Zou J (2018) Word embeddings quantify 100 years of gender and ethnic stereotypes. Proc Natl Acad Sci 115(16):E3635–E3644

Gilman CP (1970) The man-made world: or, our androcentric culture. Source Book Press, New York

Gonen H, Goldberg Y (2019) Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. arXiv preprint arXiv:1903.03862

Hamilton MC, Henley NM (1982) Sex bias in language: effects on the reader/hearer's cognitions. In: A Conference of the American Psychological Association, Los Angeles

Howard A, Borenstein J (2018) The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. Sci Eng Ethics 24(5):1521–1536

Jackendoff R, Jackendoff RS (2002) Foundations of language: brain, meaning, grammar, evolution. Oxford University Press, Oxford

Kaplan M (1983) A woman's view of DSM-III. Am Psychol 38(7):786–792

Lakoff R (1973) Language and woman's place. Lang Soc 2(1):45–79

Leavy S (2018) Gender bias in artificial intelligence: the need for diversity and gender theory in machine learning. In: Proceedings of the 1st International Workshop on Gender Equality in Software Engineering, pp 14–16

Manzini T, Lim YC, Tsvetkov Y, Black AW (2019) Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. arXiv preprint arXiv:1904.04047

May C, Wang A, Bordia S, Bowman SR, Rudinger R (2019) On measuring social biases in sentence encoders. arXiv preprint arXiv:1903.10561

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781

Millett K (2016) Sexual politics. Columbia University Press, New York

Nanda S (2014) The portrayal of women in the fairy tales. Int J Soc Sci Humanit Invent 1(4):246–250

Nissim M, van Noord R, van der Goot R (2020) Fair is better than sensational: man is to doctor as woman is to doctor. Comput Linguist 46(2):487–497

Ramanathan G (1996) Sexual politics and the male playwright: the portrayal of women in ten contemporary plays. McFarland and Company, Jefferson

Key, M. R. (1975). Male/female language: with a comprehensive bibliography. Scarecrow Press.

Steyvers M, Tenenbaum JB (2005) The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. Cogn Sci 29(1):41–78

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (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. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. https://doi.org/10.18653/v1/d17-1323

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. In NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference (Vol. 2, pp. 15–20). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/n18-2003

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., & Chang, K. (2019). Gender Bias in Contextualized Word Embeddings. arXiv: 1904.03310