Bias in Word Embeddings

Orestis Papakyriakopoulos
Technical University of Munich
Munich, Germany
orestis.p@tum.de

Simon Hegelich
Technical University of Munich
Munich, Germany
simon.hegelich@hfp.tum.de

Juan Carlos Medina Serrano
Technical University of Munich
Munich, Germany
juan.medina@tum.de

Fabienne Marco
Technical University of Munich
Munich, Germany
fabienne.marco@tum.de

ABSTRACT

Word embeddings are a widely used set of natural language processing techniques that map words to vectors of real numbers. These vectors are used to improve the quality of generative and predictive models. Recent studies demonstrate that word embeddings contain and amplify biases present in data, such as stereotypes and prejudice. In this study, we provide a complete overview of bias in word embeddings. We develop a new technique for bias detection for gendered languages and use it to compare bias in embeddings trained on Wikipedia and on political social media data. We investigate bias diffusion and prove that existing biases are transferred to further machine learning models. We test two techniques for bias mitigation and show that the generally proposed methodology for debiasing models at the embeddings level is insufficient. Finally, we employ biased word embeddings and illustrate that they can be used for the detection of similar biases in new data. Given that word embeddings are widely used by commercial companies, we discuss the challenges and required actions towards fair algorithmic implementations and applications.

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**; • **Computing methodologies → Machine learning**; • **Information systems → Data mining**.

KEYWORDS

word embeddings, bias, detection, diffusion, mitigation, fairness, sexism, racism, homophobia

ACM Reference Format:
Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in Word Embeddings. In *Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain.* ACM, Barcelona, Spain, 12 pages. https://doi.org/10.1145/3351095.3372843

1 INTRODUCTION

The growing ubiquity of algorithms in society poses questions about their social, political, and ethical consequences [77]. One of the issues research focuses on is algorithmic bias, which denotes the deviation of the algorithmic results from specific social expectations, based on epistemic or normative reasons [75].

Prior research has shown that algorithmic bias might result in unfair or discriminative decisions and statements, initiating a multilevel debate on the ethical use of algorithms [62, 102]. Under that framework, researchers, decision makers and institutions try to answer the following questions:

- What definitions of fairness and discrimination are appropriate and under what conditions? [15, 62]
- At which part of an algorithm does bias emerge and in what form? [42, 85, 93]
- What are the actual consequences of biased algorithms and who is accountable for them? [6, 68, 76]
- How can researchers and decision makers mitigate the detected bias? [8, 13]

Problem Statement

This study investigates bias in word embeddings, a set of natural language processing techniques for the mapping of words into numerical vectors. These vectors can then be used for the improvement of the predictions and inferences of other machine learning models [91]. Previous work has proven that word embeddings contain bias [13], and researchers have already developed methodologies for tracing, quantifying, and mitigating it [12, 16]. Recently, researchers have also started to develop methods for comparing biases existing in different datasets [40, 64].

Despite recent scientific findings, computer scientists in the industry widely use word embeddings for the development of highly accurate models that perform text generation, translation, classification and regression, without taking into consideration the impact of their inherent biases. Similarly, researchers have not yet investigated the diffusion and impact of biased word embeddings on further machine learning algorithms. Therefore, we want to provide a complete overview of bias in word embeddings: its detection in the embeddings, its diffusion in algorithms using the embeddings, and its mitigation at the embeddings level and at the level of the algorithm that uses them. We also investigate whether the employment of biased word embeddings contributes to the location of the bias in new data. The study raises additional awareness about a technique, whose implementation can lead to unfair algorithmic

FAT '20, January 27–30, 2020, Barcelona, Spain*
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6936-7/20/02.
https://doi.org/10.1145/3351095.3372843

decisions and inferences. We achieve this, by seeking the answer to the following research questions:

RQ1: How can we evaluate and mitigate the word embeddings' bias diffusion in further machine learning algorithms?

RQ2: Can we employ bias in word embeddings for tracing bias in new data?

Original Contribution

- We train state-of-the-art word embeddings based on the German version of Wikipedia and on unique social media data in the German language. For that, we gather over 22 million tweets and Facebook comments related to German politics. We develop a new method for locating biases in gendered languages, trace niches of sexist, xenophobic and homophobic prejudice and stereotypes on the two sets of vectors, and quantify the overall bias for each dataset.
- We transform and compare the vector spaces without distorting the immanent bias by borrowing techniques from embeddings translation [87]. We then compare the spaces and prove that the social media data containd a higher level of intergroup prejudice, while Wikipedia data contain a stronger bias in terms of stereotypes.
- We create a sentiment classifier based on the two embedding datasets and show how the model replicates bias immanent in the embeddings.
- We compare methodologies to mitigate bias without distorting the accuracy of the classifier. We compare debiasing at the embeddings level and at the level of the classifier. We illustrate that the standard technique for mitigating bias at the embeddings level [13] is insufficient for removing biases completely.
- We develop a new sexism dataset by labeling 100.000 Facebook comments as sexist or neutral and illustrate that embeddings with bias similar to the one in the target data perform better on the classification task.
- Finally, we discuss the issues, possibilities and challenges that accompany the use of biased word embeddings.

Paper Organization

The paper is organized as follows. Section 2 presents the theoretic background and related work. Section 3 describes the data and methodology we followed. Section 4 presents the results. Section 5 discusses the results, demonstrates the implications of the study and concludes the analysis.

2 BACKGROUND AND RELATED WORK

2.1 Algorithms, Bias & Fairness

A prerequisite for understanding bias in word embeddings is to evaluate how the method's results deviate from given social expectations. To do so, we adopt the Friedman et al. [38] proposed framework for analyzing algorithmic bias. They state that algorithms might face three types of bias: 1. preexisting, 2. technical, and 3. emergent.

Preexisting bias is related to the input data. Social or personal attitudes integrated in the input dataset might lead to the deviation of algorithmic inferences from a hypothesized social objective. For example, white hosts on online lodging marketplaces charge more than their non-white counterpart hosts [33]. Algorithms might replicate the asymmetry, valuing an apartment as more expensive only because the owner is white.

Technical bias emerges when there are software, hardware or mathematical constraints. An overfitted algorithm is biased, because its inferences are perfect on the training data but non-generalizable for new cases [39]. Different mathematical models trained on the same data have different prediction accuracies and consequently different amounts of bias, because of the different cost function that they optimize [69]. A computer with RAM limitations will not allow the development of a model on the full dataset, leading to the creation of model predictions that might miss important information.

Emergent bias appears at the evaluation of results and the context of their application. Forming a decision based on algorithmic results might pose an ethical problem, when the decision or inference proposed contradicts existing normative values in the society. Research and policy makers are investigating and trying to define when the emergent bias is transformed to unfair or discriminative decisions. Among others, Goodman [44] refers to algorithms as unfair, when a specific group or individual receive unfavorable treatment as a result of algorithmic decision-making. Cowgill and Tucker [23] argue that algorithmic results should always be compared to a counterfactual ideal case, in relation to which it will be decided when and how an algorithm discriminates. Overall, no unique definition of fairness is available, making each algorithmic application a distinct case to be studied.

Word embeddings might face all three types of bias. It is proven that social attitudes such as sexism and ethnic stereotypes in the initial dataset are transferred to the embeddings [13, 40], denoting the presence of a preexisting bias. Technical bias also appears. Word embeddings trained by different models yield different results on benchmark tests [19, 70, 74, 79]. Word embeddings might also result in emergent biases. Generalizations on social relations based on the distance of words immanent in an embedding space, or by inserting the embeddings in another model for prediction or inference might result in the formation of decisions that deviate from given social imperatives.

Word embeddings are used widely in commercial systems, inter alia for ad generation [46, 47], music and hospitality recommender systems [10, 45], and by tech companies who use them to develop models and offer tools [35, 58]. they constitute decisions that influence multiple social groups and individuals. It is important to understand the appearance of bias in them, the related dangers and possible reactions to them. This will not only contribute towards fairness, but will provide the foundations for creating applications that respect the rights of social groups and individuals [43, 51]. Given that the prominent bias form in word embeddings is related to the input dataset, we investigate preexisting biases and their connection to emergent biases in related applications.

2.2 Text & Social Discrimination

The reason why preexisting biases are imprinted in word embeddings is related to the nature of text. Because text is a medium for communicating and projecting human interactions, it carries features that constitute the social world. In human history, text has not only been used to organize and comprehend sociopolitical events, but also to shape the way these events are perceived and interpreted [60]. Therefore, power relationships, social discrimination, and social asymmetries are always imprinted in text.

In this study, we narrow the investigation to bias related to social discrimination. We investigate how existing forms of social discrimination in text are diffused and influence word embeddings and further models. Social discrimination refers to discrimination emerging from members of one social group towards members of another [81], thus forming a self-other duality. By the time the distinction of people into groups takes place, group members automatically start to assign different properties to in-group members and other properties to members of the 'competing' group [89]. Social theory states that attitudes of dominant social groups are imprinted in the use of language [14]. Consequently, the bases of social discrimination are diffused through statements of prejudice and stereotypes in text, directly and indirectly [88]. Social discrimination can be not only hostile, but also benevolent. Depending social conditions and group relations, stereotypes and prejudice might be both positive or negative in nature. Regardless of their polarity, they are always a result of group antagonism [41, 54, 97].

The understanding of how social discrimination is projected into text is not a trivial task. For a thorough understanding of the process, text must be analyzed and so should the conditions of its production, its context, and use [3, 36]. To achieve that, researchers have developed extensive qualitative frameworks that take into consideration the sociopolitical conditions that lead to the emergence and formulation of lingual symbols [20, 37]. By taking into consideration political and social structures, ethical values, biases, predispositions, and social group perceptions [21, 92], relations and intentions of speakers and receivers in a social situation, researchers have studied language to understand sexism, racism, and other forms of social discrimination [52, 83]. Because of the complexity and types of social discrimination, the detection and quantification of social discrimination is not always possible by the use of formal mathematical techniques. Therefore, for the analysis of bias in word embeddings and further models, we restrict our study to the detection of forms of social discrimination that are detectable by concept comparison tests (e.g. the adjective check list [98], Implicit association test [49], Bem Sex-Role Inventory [56], polarity tests [31, 84]). These methods locate regularities such as stereotypes or prejudices, rather than explaining why they emerged. An explanation would require additional qualitative analysis, which is beyond the scope of this paper.

2.3 Social Discrimination & Word Embeddings

Researchers have proven that word embeddings contain forms of prejudice and stereotypes related to sexism and racism [13, 40]. Based on that, we study how and when biases result in further socially discriminative algorithmic behavior. To do that, we develop methodologies for tracing biases in gendered languages. Existing methods [13, 17] for detecting biases in word embeddings are grounded in qualitative techniques of concept associations, on which we also rely [49, 84, 98]. They analyze qualities of groups and their relation to other concepts, assuming that in an ideal society these concepts would have been either equally assigned to these groups or not at all. For example, ideally an occupation should not be connected more to one sex than the other, nor should one sex be treated more positively or negatively than the other. These assumptions might hide the actual reasons and conditions for the emergence of the specific associations and their straightforward connection to social discrimination, but are the same assumptions used in standard models for measuring social discrimination based on qualitative techniques [17, 49, 84, 98]. Because existing methods are developed primarily for the English language, we develop a new model that can account for gendered versions of words.

Until now, researchers have investigated various dimensions of bias in word embeddings. Garg et al. [40] show how prejudice evolving over time is imprinted in word embeddings. Kozlowski et al. [64] use the positional change of word embeddings to describe semantic transformation. Dev et al. [27] show that names in word embeddings function as proxies of bias against social groups. Arora et al. [4] prove that different meanings of words are 'encoded' in word embeddings and can be retrieved. Zhao et al. [101] propose a methodology to train word embeddings without sexist bias in them. Brunet et al. [16] develop a technique to trace the origin of bias in embeddings back to the original text. Caliscan et al. [17] introduce a general methodology to trace bias in word embeddings, while Drozd [30] et al. automatize the process. Drawing from previous research, we want to provide a complete picture of bias in the embeddings, its diffusion and mitigation.

2.4 Bias Prediction

Another objective of the study is to test whether bias in word embeddings can be used constructively. To that end, we also investigate whether biased word embeddings can contribute toward detecting bias in new text. Research shows that bias detection, especially in cases of social discrimination such as sexism or racism is very complicated. Park et al. [78] have attempted to create classifiers that detect abusive language. Dahou et al. [24] developed models for sentiment analysis. Kathrik et al. [28] tried to automatically trace cyberbulling in Youtube, while Levy [67] tried to detect sexism in newspaper covers. Overall, the performed attempts yield moderate results, especially when using only text as classification inputs because of the complex nature of human language [61]. It is a challenge to test if bias in word embeddings would lead to the improvement of classifiers predicting social discrimination.

3 DATA AND METHODS

3.1 Word Embeddings

To be able to investigate bias in word embeddings trained on different datasets, we collected data from Facebook, Twitter and Wikipedia. For Facebook and Twitter, we used the application programming interfaces (APIs) of each platform. From the social media channels, we collected data for the six main political parties in Germany: CDU, Germany's main conservative party; CSU, the sister party of the CDU in Bavaria; Bundnis90/Die Grünen, the green party in Germany; FDP, a neo-liberal party; SPD, Germany's social-democratic party; Die Linke, the radical left party and Alternative für Deutschland, the extreme right populist party. For Facebook, we retrieved the posts from the political parties and their comments during the period between January 2015 and May 2018. For Twitter, we collected the tweets from political parties' Twitter accounts between January and October 2018. We also included tweets from users that mentioned or retweeted the political parties, as well as tweets that included the names of the political parties for the same period. Overall, we gathered 24 million posts, comments, and tweets, which comprised our social media political dataset (485 mil. tokens). For Wikipedia, we collected the complete German wikipedia as bulk file from the official repository. [1] The Wikipedia dataset consisted of 2.2 million articles (850 mil. tokens). All texts were originally written in German. Therefore, related biases existed in the original text and were not added to it by further textual processing (e.g. translation from other languages to German). We present our results in English for readability purposes.

For training the embeddings on the two datasets, we used GloVe, developed by Pennington et al. [79]. The model creates vectors of words by taking into consideration the word co-occurrence frequencies in the dataset and optimizing

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j + logX_{ij})^2,$$

where V is the vocabulary, i and j two words, w_i the word vector of word i, $\tilde{w}_j$ the context vector of word j, b_i and $\tilde{b}_j$ their biases, and X_{ij} the co-occurrence number of the words for a given window. $f(x) = (x/x_{max})^a$ if x lower than a chosen x_{max}, otherwise $f(x) = 1$, with a a hyper-parameter. Following the authors' recommendations, we tokenized the texts using the nltk tokenizer [71], our window size was 10, $a = 3/4$ and $x_{max} = 100$. Overall, the datasets for Wikipedia and social media contained 390.000 and 200.000 word vectors respectively.

3.2 Vector Space Transformation

Optimizing the GloVe cost function results in the nonlinear map

$$N : C, V \mapsto W$$

where C is the corpus, V the vocabulary and W the word embeddings vector space. Given that the corpora for Wikipedia and social media C_w, C_{sm} vary, as well as the two vocabularies V_w and V_{sm}, the generated vector spaces W_w and W_{sm} are not comparable to each other. A comparison presupposes the projection of the one space on the other, given a Transformation matrix T that preserves the bias in the vector spaces. Both Smith et al. [87] in embeddings translation and Hamilton et al. [53] in measuring semantic change obtain the transformation matrix by solving the Orthogal Procrustes problem

$$L = \underset{\Omega}{\operatorname{argmin}} ||\Omega A - B||_F \text{ subject to } \Omega^T \Omega = I,$$

where A and B, two word embeddings vector spaces and Ω the transformation matrix. The problem is solvable by applying a singular value decomposition algorithm as proposed by Schönman [86]. The specific transformation places all words from vector space A as close as possible to their corresponding words in vector space B. As transformation is linear, the normalized distance between words does not change, thereby preserving bias in the embeddings.

3.3 Bias Detection

For detecting bias in word embeddings we must develop a generally applicable formula. The method proposed by Bolukbasi et al. [13] defines an inter-group direction $\vec{g}$. Then it quantifies the bias of a random word by the cosine distance between the word vector and $\vec{g}$. For example, the vector of the word *nurse* should be independent of the inter-group direction between man and woman. Usually it is not, since society stereotypically sees nursing as a female profession. Nevertheless, this definition does not cover cases that words ought to have an inter-group component. For example, words in German are sex-dependent. There is a male and a female version, denoting that mathematically the vectors of the words and of the sex direction should be dependent. To overcome that, we develop an alternate methodology. First, we define pairs of theory specific words for each type of discrimination. [2] Then we introduce a list of concepts for which we want to measure the bias. If concepts change based on the social groups, e.g. they have male and female versions, they are represented by word-pairs. We calculate the general bias in the embeddings by the equation

$$B_g = \frac{1}{NK} \sum_{i=1}^{N} \sum_{j=1}^{K} |cos(w_{j1}, P_{n1}) - cos(w_{j2}, P_{n2})|,$$

where N is the number of concepts, K the number of theory specific pairs, w_{j1} and w_{j2} the embeddings for the jth pair of theory specific words and P_{n1} and P_{n2} the embeddings for nth concept pair in the list. When $P_{n1} = P_{n2}$, i.e. when we investigate concepts that are not variable with respect to the social groups under investigation, we use the equation

$$B_g = \frac{1}{NK} \sum_{i=1}^{N} \sum_{j=1}^{K} |cos(w_{j1}, P_n) - cos(w_{j2}, P_n)|$$

The general bias equation compares the magnitude of dependence between a concept and the two groups. If the concept vector has a higher cosine distance to one group vector than to the other, then the concept is biased in that direction. We apply the above equation to two tasks. We create a list containing 1600 professions for a profession-related bias task. We also use the sentiment list developed by Remus et al.[80] that contains words of positive and negative polarity for a sentiment bias task.

3.4 Bias Diffusion

In order to measure bias diffusion, we need to capture whether a model that takes biased word embeddings as input will also give biased output. The bias of the model needs to be theoretically comparable to the bias in the embeddings. Therefore, we used the sentiment dictionary of Remus et al. [80], which contains a set of positive-laden and negative-laden words. We trained a linear support vector machine classifier. The classifier took as input the

[1] https://dumps.wikimedia.org/dewiki/

[2] Man-Woman, German-Foreigners, Straight-Gay, for sexist, xenophobic and homophobic prejudice respectively. Both gender and sexuality are spectra. We did not analyze here biases related to the rest social groups for simplicity reasons, as the methodology deals with group dualities.

embedding vector of a word and predicted if it had a positive or negative sentiment. We modified the output of the classifier by transforming the class probabilities to a sentiment score by applying the equation

$$S_w = -[log_{10}((P(w = positive) - log_{10}(P(w = negative))],$$

where S_w is the sentiment score for word w, P(w=positive) and P(w=negative) the model's assigned probability that the word is positive and negative respectively. Then, we designed an experimental setting by which we could measure the level of sexist and xenophobic prejudice that the classifier learned. We used the first names list developed by Winkelmann [100], and acquired male and female stereotypical first names for nine population groups: German, Turkish, Polish, Italian, Greek, French, US American, Russian and Arabic. We then fed the embeddings of the words into the algorithm and measured the sentiment score across different sexes and populations. We claim that ideally a name should have a sentiment score equal to zero, because it should be polarity-independent. We then defined the sentiment bias $B_{s,c}$ of the algorithm being equal to the classifiers' sentiment score and the classifiers' social discrimination bias for a specific social discrimination concept as

$$B_c = \left| \frac{1}{N} \sum_{i=1}^{N} B_{s,ci1} - \frac{1}{K} \sum_{j=1}^{K} B_{s,cj2} \right|,$$

where N and K are the number of names for each of the two investigated social groups and $B_{s,ci1}$ and $B_{s,cj2}$ the sentiment bias of the classifier for a word in each group respectively. This metric quantifies the difference in the assigned sentiment of the classifier for the names of each group. For investigating the statistical significance of our results we apply Mann-Whitney U and Kruskal-Wallis H tests to compare biases among two or more groups.

3.5 Bias Mitigation

A sentiment analysis algorithm has no social discrimination bias when it predicts equal sentiment for names of different sexes or populations. In order to achieve that we try two approaches. In the first case, we adopt and extend the methodology proposed by Bolukbasi et al. [13]. As the classifier assigns a sentiment polarity value for each input word, we define a sentiment direction $\vec{s} \in R^d$, where d is the dimension of a word vector $\vec{w}$. The direction is calculated by forming pairs of theory specific dualities (e.g. good - bad, positive - negative, etc., see Table 1) that are theory specific and taking the difference of their word vectors. Afterwards, we apply PCA, with the resulting first component being the sentiment direction $\vec{s}$. We also define the set $N = \{\vec{w_1}, ..., \vec{w_n}\}$ corresponding to the vectors of theory neutral words. Then we hard neutralize these words by applying

$$\vec{w'_i} = \vec{w_i} - \frac{\vec{w_i} \cdot \vec{s}}{\vec{s} \cdot \vec{s}} \vec{s},$$

where $\vec{w'_i}$ is the debiased non-normalized vector for word w_i. By doing this, we make the vectors of theory neutral words orthogonal to the sentiment vector. We then feed the neutralized embeddings into the classifier and calculate the sentiment for the different groups. This methodology tries to mitigate bias at the word embeddings level. As non-neutral words are not debiased, the accuracy of the classifier does not change.

Table 1: Word pairs used for the calculation of the sentiment direction translated from German.

Positive	Negative
good	bad
positive	negative
happy	sad
peace	war
cheap	expensive
love	hate

In the second case, we try to mitigate the bias at the level of the classifier. The linear SVM classifier learns to split the classes given a linear hyperplane, which is defined by a normal vector $\vec{p}$. This vector actually corresponds to the sentiment direction as learned by the classifier. Therefore, we hard-neutralize the theory neutral vectors given vector $\vec{p}$ by applying the same formula as above.

3.6 Bias Prediction

The last part of our study focuses on understanding whether biased word embeddings can help detecting bias in new text. For this scope, we manually labeled 100,000 user comments from German political parties Facebook pages and created a sexism dataset. We categorized each comment as sexist or neutral based on the following criteria: 1. the existence of a sexist buzzword, 2. the formulation of sex-related compliments, 3. the expression of statements against the equality of sexes, and 4. the assignment of stereotypical roles to persons based on their sex. Each of the four categories denoted a different label in the dataset and its formation was based on previous theoretic work. We traced sexist buzzwords under the notions of traditional sexism [32, 73], while we defined and searched sex related compliments given theories of benevolent sexism [9, 59]. We located statements against sex equality in comments that the users explicitly argued about the topic and we defined stereotypical roles of the sexes based on the works of Eckes [32], Tilegea [90], and Benokraitis et al. [9].

To efficiently code the dataset, we created sound recordings of each comment, because it has been shown that hearing a sentence rather than just reading it improves content understanding [34]. Two coders reviewed the sound corpus, assigning to each comment one or more of the four labels and giving a concrete reason for their decision. In cases of coders' disagreement, the comments were reviewed by one additional coder. For these comments we accepted labels assigned by more than one reviewer. Comments that were not assigned a label at all were then classified as non-sexist, while comments having at least one of the four labels as sexist. Overall, we detected 1,988 sexist comments. We then sampled an equal number of neutral comments, creating a balanced dataset, which we split into a train and a test set. We evaluated the biased word embeddings on the classification test. We created models that included long-short-memory network (LSTM) and attention based architectures, and investigated their accuracy on the test set, with 1. random word embeddings, 2. the embeddings from the Wikipedia data, 3. the embeddings from the social media data and 4. embeddings trained on the sexism dataset. Furthermore, we investigate which properties of the word embeddings are responsible for accuracy improvement. For that, we transformed and compared the word embeddings from the sexism dataset to the other embeddings by calculating their mean weighted cosine similarities, as given by the equation

$$Sim_{s,i} = \frac{\sum_{n=1}^{N} f_n cos(\vec{w_{n,s}}, \vec{w_{n,i}})}{\sum_{n=1}^{N} f_n}, \text{ with } n = 1, ..., N \in s \cap i,$$

where s is the word embeddings trained on the sexism dataset, i is another word embeddings dataset, N is the number of common words in the two datasets and f_n is the frequency of appearance of a common word n in the sexism dataset. We also perform the sentiment task with the sexism dataset embeddings and calculate the level and type of sexist prejudice within them.

4 RESULTS

The results are split into three parts. First, we present our findings on bias within the Wikipedia and social media word embeddings. Second, we analyze how the bias was diffused and how we mitigated it. We also illustrate the efficiency of biased word embeddings when used as sexism detection models. In the last part of the section, we evaluate bias in word embeddings.

4.1 Bias in Word Embeddings

The word embeddings generated on the Wikipedia and social media corpora contained 390,000 and 200,000 vectors respectively. In both cases, the profession and sentiment task revealed intensive stereotypical features assigned to each examined social group. In both Wikipedia and social media spaces, women were mostly associated with professions like nurses and secretaries. On the other hand, men were associated with stereotypical male roles, like policemen and commanders. The aforementioned assigned professions highly correlate with the actual profession distribution in society [1], denoting that the actual social asymmetry is imprinted in the vectors. For Wikipedia, women were strongly associated with concepts related to marriage, while men were linked to concepts related to war and power. This could be because Wikipedia extensively includes biographies of historical figures, in which women are typically associated with marriage and familial relations, while men are associated with concepts such as war and governance [95, 96]. In social media, the female sex was closer to positive feelings such as love and maturity, but also to negative ones like stubbornness and agitation. Men were closer to concepts related to aggression and fighting, with most of them being negative. The stereotypes found in the social media dataset comply with previous research findings [99], which found the existence of power related stereotypes for men and sentiment related stereotypes for women.

In both Wikipedia and social media, Germans were intensively associated with jobs related to governance and journalism, while foreigners either to blue collar jobs or to professionals dealing with foreign populations such as aid officials, politicians or tour guides. Foreigners were generally linked to sentiment concepts related to immigration, law and crime, while Germans to positive feelings such as charm and passion (social media), as well as to cooperation and union (Wikipedia). The association of foreigners to immigration related concepts and professions can be traced back to the refugee crisis taking place in Europe over the last few years, which has a prominent position in the public agenda [65]. Similarly, researchers have proven the existence of biased slants related to immigration issues on wikipedia [48]. Given that both German Wikipedia and the German social media discussions are primarily produced by Germans, we can attribute the inherent positivity and negativity on Germans and foreigners on the intergroup prejudice existing in the society [2, 72]

Table 2: Extreme words for each task and group using the embeddings from Wikipedia data

Wikpedia

Sexist prejudice			
Profession		Sentiment	
Woman	Man	Woman	Man
Nurse	Officer	Wedding	Reinforcement
Secretary	Hunter	Divorce	Attack
Teacher	Commander	Anulment	Combat
Saleswoman	Guard	Engagement	Power
Actress	Cameraman	Marry	Decrease
Population Prejudice			
Profession		Sentiment	
Foreigners	German	Foreigners	German
Aid official	Author	Refugee	Champion
Craftsman	Journalist	Unauthorized	Cooperation
Bank Assistant	Historian	Lawful	Union
Tour guide	Director	Tax	New
Foreman	Painter	Accumulate	Assignment
Sexual Orientation Prejudice			
Profession		Sentiment	
Homosexuality	Heterosexuality	Homosexuality	Heterosexuality
Artist	Singing teacher	Corruption	Unserious
Art dealer	Copywriter	Violence	Nice
Actress	Forest manager	Adultery	Fantastic
Cook	Track driver	Known	Smart
Shoemaker	Carpenter	Prohibited	Fair

The stereotypes were equally intensive for sexual orientation. Homosexuals were related to stereotypical roles such as artists (Wikipedia) and hairdressers (social media), while persons of heterosexual orientation were related to blue collar professions or positions in science. Strikingly, homosexuality was related in both datasets with very negative concepts: from violence, prohibition and adultery (Wikipedia), to death sentencing, abuse and harassment (social media). On the complete opposite side, heterosexuality was closely positioned to inherently positive sentiments such as fantastic and smart (Wikipedia) and to concepts like friendship and deliberation (social media). These findings comply with historic negative social attitudes against homosexuality, where conservative groups state that it is abnormal and that should be prohibited by law [26]. Regarding positive concept relations to homosexuality, researchers have found similar associations in concept association tests [94], illustrating that biases in social media and wikipedia

Table 3: Extreme words for each task and group using the embeddings from social media data

Social Media

Sexist prejudice			
Profession		Sentiment	
Woman	Man	Woman	Man
Nurse Secretary Pharmacist Religion teacher Correspondent	Policeman Musician Priest Coach Paramedic	Agitation Mature Love Increase Stubborness	Robber Attacker Injured Fascist Overwhelmed
Population Prejudice			
Profession		Sentiment	
Foreigners	German	Foreigners	German
Newspaper Skilled worker Politician Consultant Teacher	Government Official Correspondent Notary Butler Reporter	Criminal Exclude Refugee Increase Frustration	Mature Beauty Charm Passion Love
Sexual Orientation Prejudice			
Profession		Sentiment	
Homosexuality	Heterosexuality	Homosexuality	Heterosexuality
Artist Scrap dealer Hairdresser Interviewer Consultant	Streetworker Political scientist Political economist Mediator Biologist	Death sentence Discrimination Abuse Harassment Violence	Friendly Moving Deliberation Increasing Unecessary

correspond to the ones found offline. An overview of the most extreme concept associations for all groups can be found in tables 2 and 3. The results demonstrate strong stereotypical associations for all groups. Overall, the calculated general bias was higher for almost all categories and tasks for the Wikipedia dataset (table 4), denoting that Wikipedia introduces more severe stereotypes for each social group than the examined social media content. The calculated scores are of similar magnitude to those calculated by Bolukbasi et al. [13], who calculated a general bias of 0.08 on the profession task for the two sexes on an English Google news corpus.

Table 4: General bias for each intergroup comparison, bias task and embeddings dataset.

	Wikipedia		Social Media	
	Profession	Sentiment	Profession	Sentiment
Sex	0.080	0.087	0.077	0.037
Population	0.066	0.063	0.054	0.056
Sex orientation	0.064	0.087	0.0619	0.084

The presented associations only reveal partial bias in the embeddings. Indeed, stereotypes are a base of social discrimination, and someone can qualitatively evaluate how specific social groups are presented in the datasets by checking the mostly associated concepts. Nevertheless, this does not per se signify that a specific group is generally favored over another, which would provide evidence of prejudice. To achieve that, we calculated the mean polarity score for the sentiment concepts being closer to each social group, and then extracted the difference for each intergroup comparison. The results are given in Figure 1. For both Wikipedia and social media, Germans were depicted much more positively than foreigners. The same applies for heterosexuals in comparison to homosexuals. Both results are in accordance to the sentiment task results, as Germans and heterosexuals were associated with much more positive feelings and concepts, confirming the existence of biases that favor privileged social groups [26, 50].

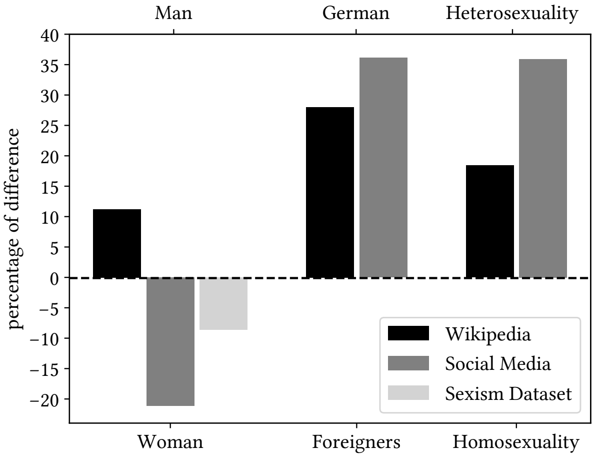

Figure 1: Intergroup positive sentiment difference in the embeddings.

In German Wikipedia, men were generally depicted more positively. On the other hand, in the social media dataset, women were associated with more positive words. One explanation is that in Wikipedia men were described by stereotypical concepts like power, attack and reinforcement, which are labeled as positive in the polarity dictionary. In contrast, the social media data also related men to concepts like fascism and robbery, i.e. words with highly negative sentiment. That could also be rooted in the nature of German language, which uses the male plural when making colloquial general claims. Because negative statements about groups on social media were generated in a male form, this bias could have been replicated by the model. Furthermore, the sentiment difference does not fully replicate bias in text. For example, in social media data, women are often associated with the term ‘mother’, for which the sentiment lexicon assigns a positive score. Nevertheless, the actual combination of words in a political context corresponds to sexist speech, as numerous users refer to female politicians as mothers in order to undermine their political abilities.

The above results illustrate that word embeddings contain a high level of bias in them in terms of group stereotypes and prejudice. The intergroup comparison between sexes, populations, and sexual orientations revealed the existence of strong stereotypes and unbalanced evaluations of groups. Although Wikipedia contained stronger bias in terms of stereotypes, social media contained a higher bias in terms of group prejudice.

4.2 Bias Diffusion, Mitigation & Prediction

Our analysis shows that the above bias was diffused further into the trained sentiment classifiers. We trained one classifier for each embedding dataset, with both having a test set accuracy of around

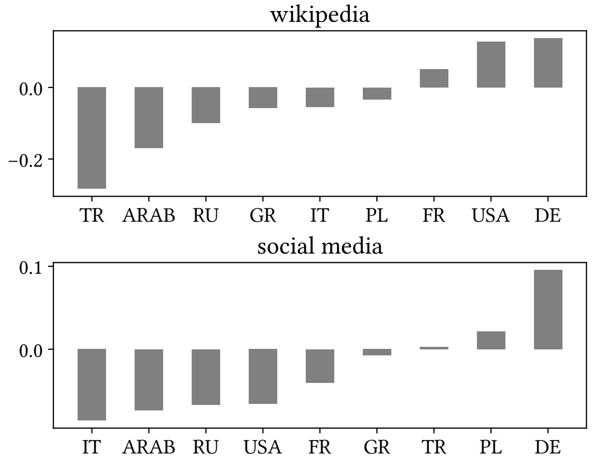

Figure 2: Predicted score of the sentiment classifier for stereotypical names of different populations

85%. The classification task for stereotypical names of different communities illustrated a preference for German names (Figure 2). In both embedding datasets, German names were assigned the highest average sentiment score. In contrast, most of the foreign names were assigned negative sentiment values. Arabic and Russian names were negatively associated in both datasets, which can be grounded both on existing social stereotypes against Russians and Arabs in the society [5, 7], as well as mainstream media representations of the ethnicities [55, 66]. Greek, Polish and Turkish names were seen much more positively by the social media classifier. This comes in contrast to what someone would actually expect, since a large part of contemporary German public opinion holds strong negative stereotypes against Greek, Polish and Turkish populations due to economic and migration issues [5, 11, 63]. French and US-American stereotypical names were classified much more positively by the Wikipedia classifier. The result related to French names was not intuitive, given the historical conflicts between Germany and France that are extensively covered in Wikipedia [57]. In contrary, researchers illustrate that non-English Wikipedia pages on U.S.-American persons generally contain positive cues [18], explaining also the favoritism of the classifier for U.S.-American names. Overall, the classifiers' social discrimination biases for the models trained on the Wikipedia and the social media data were $B_{c,wiki} = 0.23$ and $B_{c,sm} = 0.14$ respectively. The bias of the classifier was similar to the bias in the embeddings, as in both cases German concepts were evaluated much more positively. For both classifiers the Kruskal–Wallis tests were significant (sm classifier: H=101.95, p-value: <0.01; wiki classifer: H=37.36, p-value < 0.01), denoting that the mean bias for each ethnicity varies significantly from the others.

We concluded with similar findings when predicting the sentiment of male and female names. The classifiers exactly replicated the prejudice as measured in the word embeddings (Figure 4). The Wikipedia classifier predicted a higher average sentiment score for male names. In contrast, the social media classifier assigned a much more positive overall score to female names. This complies with the results from the intergroup positive sentiment difference in the embeddings, where women were associated with more positive concepts than men in the social media dataset, while the opposite happened in the Wikipedia embeddings. Hence, we proved that classifiers trained in biased word embeddings replicate the bias existing in the vectors. Overall, the classifiers' social discrimination biases were $B_{c,wiki} = 0.011$ and $B_{c,sm} = 0.068$ respectively. The Mann-Whitney U test was significant for the social media classifier (U = 1027471, p-value < 0.01), but not for the Wikipedia classifier (U = 1069947, p-value = 0.23). This does not mean that there is no bias between sexes in the second case. By breaking down names by ethnicity and comparing them, we get significant results for German (U = 91356, p-value = 0.001), Polish (U = 19, p-value = 0.01), Greek (U = 90, p-value = 0.003) and U.S.-American (U = 63128, p-value = 0.02) names.

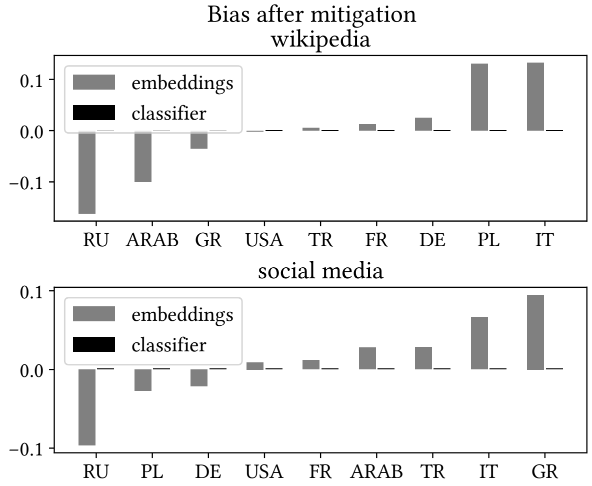

Figure 3: Bias in the sentiment classifier for stereotypical names of various populations after mitigation at (a) the embeddings' level, (b) the level of the classifier.

The study proves that the diffused bias can be mitigated. Both methodologies for bias mitigation reduced bias significantly. Mitigation at the embeddings level resulted in social discrimination biases of the classifiers of $B_{c,wiki} = 0.027$ and $B_{c,sm} = 0.035$ for the population comparison. Similarly, when predicting the sentiment of male and female names, the bias of the classifiers after mitigation was $B_{c,wiki} = 0.009$ and $B_{c,sm} = 0.018$ respectively. Mitigation at the level of the classifier was by far more efficient: In all possible tasks, the overall social discrimination bias vanished. Figure 4 presents an overview of bias before and after mitigation for each case. In order to understand why the second methodology provides better results, we calculated the cosine distance between the sentiment vectors of the embeddings and the classifier, which were used for de-biasing. The value was close to 0.9, denoting that the classifier actually learns a significantly different sentiment direction than the one defined by the methodology proposed by Bolukbasi et al.[13]. Actually, the classifier learns further associations between the vectors, which are not taken into consideration when debiasing at the embeddings level. Debiasing at the embeddings level

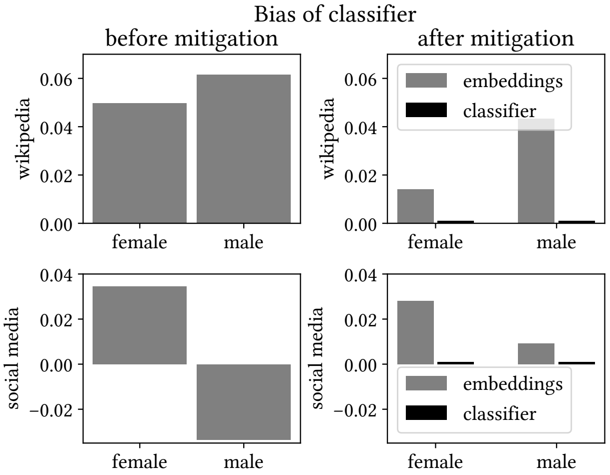

Figure 4: Predicted score of the sentiment classifier for male and female names, before and after mitigation by applying two different methods.

results in the diffusion of a different bias in the classifier. As Figure 3 shows, although bias related to the favored group was highly reduced, remaining patterns in the data resulted in a totally different bias diffusion. This bias was not universally distributed in all cases, but resulted in asymmetries in certain cases. For example, for the classifier trained on the Wikipedia embeddings the mean bias difference between German and Russian (U = 23363, p-value = 0.065), Arabic (U = 90, p-value = 0.045) and Italian (U = 86625, p-value = 0.065) names remained statistically significant. This was not the case for the sex names comparison in either classifier (sm: U=1060104, p-value=0.12; wiki: U=1065656, p-value=0.18) or ethnicity names comparison for the classifier trained on the social media embeddings (Kruscal-Wallis H=4.2, p-value=0.83). Hence, we show that debiasing at the classifier level is a much better and safer methodology to follow. Because of the mathematical definition of the linear support vector classifier, it was straightforward to mitigate the bias in it. For other cases, where non-linearity prevails, more sophisticated methodologies are needed.

Our last finding states that biased word embeddings can be useful for bias prediction tasks. We trained and deployed various models on the sexism prediction task, with and without the trained biased word embeddings. On the first test, we created a simple LSTM model, which had as inputs either a random dataset, Wikipedia, social media, or the sexism dataset word embeddings. We restricted the embeddings from being trainable, in order to evaluate their actual influence on the results. In addition, we only inserted the values of the word embeddings for the words that were common in the datasets. In this way, we could assure that if an embedding dataset had more impact on the results, that it would be because of the type of information encoded into the vectors and not the amount of words existing in the dataset. The models with the trained embeddings provided higher test accuracy and F1 scores. The model with the sexism dataset vectors yielded the best results. The social media embeddings provided better results than the Wikipedia vectors. The calculated weighted mean cosine similarity between the sexism dataset vectors and the social media and the Wikipedia datasets was 0.49 and 0.39 respectively. This denotes that social media vectors are more similar to the vectors of the sexism classifier, which in turn signifies that more similar meanings and, consequently, biases were encoded in them. This is also proven by the sentiment task, for which the sexism dataset vectors had similar prejudice with the social media vectors (Figure 1). Thus, the more similar the bias in the embeddings with the target data, the higher the ability of the classifier to detect the bias.

On the second task, we used additional architectures for the prediction task. We allowed the embeddings to be freely trainable, and used all the available vectors to predict sexism. The best model contained an attention layer and provided an accuracy of 80%. Then, we removed all test observations that contained words that did not appear in the training process, and recalculated the accuracy. We obtained an overall score of 92% on the test data. Given the general difficulty in the detection of sexism and hate-speech by machine learning models [25, 29], the results are more than satisfactory. The model's input was text without any punctuation, nor any other metadata that generally help in detecting social discrimination [82]. Therefore, we showed that biased word embeddings can substantially help in sexism detection, while attention based networks can provide really high accuracy in detecting sexism. An overview of all models can be found in table 5.

Table 5: Classification results for the sexism task

Model	Embeddings	Trainable	Accuracy	F1 - sexist	F1 - neutral
LSTM	Random	False	0.57	0.55	0.62
LSTM	Wiki - common	False	0.68	0.65	0.70
LSTM	SM - common	False	0.70	0.69	0.70
LSTM	Sexism - common	False	**0.75**	**0.75**	**0.75**
Attention	Sexism - all	True	0.80	0.80	0.81
Attention	Sexism - all - filtered	True	**0.92**	**0.92**	**0.91**

4.3 Evaluating biased word embeddings

The analysis provided a thorough description of bias in word embeddings. We proved that the technique replicates biases related to sexism, homophobia, and xenophobia immanent in the original text. We showed that Wikipedia data mediates to the word embeddings stronger stereotypes, while political social media data imprints stronger forms of group favoritism into the vectors.

The study illustrated that the use of biased word embeddings results in the creation of biased machine learning classifiers. Models trained on the embeddings replicate the preexisting bias. Bias diffusion was proven both for sexism and xenophobia, with sentiment classifiers assigning positive sentiments to Germans and negative sentiments to foreigners. In addition, the amount of polarity for men and women in the embeddings was diffused unaltered into the models. We used two methods for bias mitigation, one at the level of the embeddings and one at the level of the classifier. In both cases, we lowered the bias, while mitigation at the level of the classifier was the optimal one.

The analysis also showed that biased word embeddings can be beneficial for bias prediction. Embeddings containing bias similar to the one in the investigated dataset can help in the classification task.

We showed that text-only models for bias prediction can provide more than satisfactory results by using embeddings. Among the various models developed, we found that simple attention-based neural networks yielded the best results. Of course, the developed models are in the position to detect forms of sexism similar to that defined by the inter-subjective coding process and its theoretical assumptions. The models are not generalizable to other forms of sexism that were not taken into consideration at the development of the dataset. Nevertheless, the study provides promising findings for the detection of biases in text by the use of word embeddings and deep neural architectures.

Overall, the study provided a full evaluation of biased word embeddings. It showed how bias can be detected, its diffusion, and how it can be mitigated. It also proved that different forms of bias influence further models differently. In addition, we showed positive aspects of word embeddings. Not only can they be used for bias detection, but most importantly, they can help understand and evaluate sociopolitical relations immanent in text.

5 DISCUSSION

The findings of the study provide a complete picture of the issues, limits, and possibilities of biased word embeddings at the algorithmic level. In the discussion, we go one step further and analyze the societal importance of the aforementioned findings. We illustrate the emerging opportunities for the use of biased word embeddings, while we explain their negative properties. Last but not least, we describe the related challenges that researchers and decision makers need to deal with, in order to assure a just application of algorithmic systems based on word vectors.

On the positive side, the ability of word embeddings to absorb semantic relations of the social world prevails as their main advantage. Being able to quantify bias existing in the society, latent political relations and properties of language and text, has always been a scientific challenge, and until now a privilege of qualitative social science [20]. Word embeddings constitute a way to mathematically grasp and describe sociopolitical relations through the analysis of text, allowing the quantification of phenomena as racism, sexism and social discrimination in general. Based on the vectors, it is possible to evaluate social phenomena, compare and measure their magnitude for different conditions and context. A systematic analysis of word embeddings can result in the creation of new scientific knowledge about the social world, redefining and developing further existing theories. Furthermore, developing models for bias detection by using biased word embeddings can be beneficial. Word embeddings generally improve the accuracy of machine learning models, and we proved that this was also the case in bias prediction, a task which is highly difficult.

On the negative side, the dependence of word embeddings on the nature of the input data is an open methodological issue. There is no such thing as naturally developed neutral text, because the semantic content of words is always bound with the sociopolitical relations of a society [14]. The study illustrates that even text generated in a formal and controlled environment like Wikipedia, results in biased word embeddings. Furthermore, the preexisted bias becomes even more graspable when evaluating the vectors and using them in further algorithms. The algorithms associate stereotypes and concepts to specific social groups, while containing latent prejudice. These associations are usually not directly perceivable in the initial text, nor are they uniformly distributed within it. Nevertheless, the projection of words in a mathematical space by the embeddings consolidates stereotyping and prejudice, assigning static properties to social groups and individuals. Relations are no longer context-dependent and dynamic, and embeddings become deterministic projections of the bias of the social world. This bias is diffused into further algorithms unchanged, resulting in socially discriminative decisions.

Word embeddings are a valuable tool for improving machine learning models and for understanding the social world. Managing their bias prevails as an open challenge for ethical and fair algorithmic applications. Until now, researchers and commercial companies train and integrate word embeddings uncontrollably in their models, without taking into consideration the potential impact and societal implications. The study showed that bias in word embeddings can result in algorithmic social discrimination, yielding negative inferences on specific social groups and individuals. Therefore, it is necessary not only to reflect on the related issues, but also to develop frameworks of action for the just use of word embeddings. To achieve that, it is necessary to develop frameworks that detect bias in concrete algorithmic applications of the embeddings and quantify their impact on individuals and the society [22]. This presupposes commercial companies becoming more transparent regarding the exact algorithms and data they use in their products and decisions. Only through detailed auditing can it be possible to fully understand the issues and start implementing measures that assure algorithmic justice. These measures include the hard mitigation of the bias at the level of the end product, in such a way that no individual is negatively influenced or discriminated against.

It also includes the development of artificial datasets that comply with certain social expectations, on which the embeddings can be trained on. Until now, word embeddings are either trained on text related to a specific algorithmic application or context, or on huge freely accessible corpora. In both cases, bias in the text is always imprinted in the embeddings, and therefore also diffused in further models. It is necessary to search for alternatives, in order to remove preexisting bias in an optimal way.

Our study provides a complete overview on the issue of bias in word embeddings. Not only does it describe the problems and possible solutions, but also initiates an important discussion on the implementation of the vectors in commercial applications. The presented results denote the need for more transparency in the use of word embeddings, in order to ensure their ethical algorithmic implementation. The mathematical tools for model evaluations are already provided; actions from the related stakeholders need to follow.

REFERENCES

[1] [n.d.]. Which jobs do men and women do? Occupational breakdown by gender. https://careersmart.org.uk/occupations/equality/which-jobs-do-men-and-women-do-occupational-breakdown-gender

[2] Richard Alba, Peter Schmidt, and Martina Wasmer. 2004. *Germans or foreigners? Attitudes toward ethnic minorities in post-reunification Germany.* Springer.

[3] Michael W Apple. 1992. The text and cultural politics. *Educational Researcher* 21, 7 (1992), 4–19.

[4] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2018. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association of Computational Linguistics* 6 (2018), 483–495.

[5] Frank Asbrock. 2010. Stereotypes of social groups in Germany in terms of warmth and competence. *Social Psychology* (2010).

[6] Solon Barocas, Sophie Hood, and Malte Ziewitz. 2013. Governing algorithms: A provocation piece. *Available at SSRN 2245322* (2013).

[7] Rupprecht S Baur and Stefan Ossenberg. 2017. Zur Verbindung von Stereotypen und Komik am Beispiel deutsch-russischer Witze. In *(Un) Komische Wirklichkeiten*. Springer, 329–342.

[8] Yahav Bechavod and Katrina Ligett. 2017. Learning fair classifiers: A regularization-inspired approach. *arXiv preprint arXiv:1707.00044* (2017), 1733–1782.

[9] Nijole Vaicaitis Benokraitis and Joe R Feagin. 1995. *Modern sexism: Blatant, subtle, and covert discrimination*. Pearson College Div.

[10] Erik Bernhardsson. 2013. Model benchmarks. https://erikbern.com/2013/11/02/model-benchmarks.html

[11] Hans Bickes, Tina Otten, and Laura Chelsea Weymann. 2014. The financial crisis in the German and English press: Metaphorical structures in the media coverage on Greece, Spain and Italy. *Discourse & Society* 25, 4 (2014), 424–445.

[12] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Quantifying and reducing stereotypes in word embeddings. *arXiv preprint arXiv:1606.06121* (2016).

[13] 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.

[14] Pierre Bourdieu. 1991. *Language and symbolic power*. Harvard University Press.

[15] Danah Boyd, Karen Levy, and Alice Marwick. 2014. The networked nature of algorithmic discrimination. *Data and Discrimination: Collected Essays. Open Technology Institute* (2014).

[16] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2018. Understanding the Origins of Bias in Word Embeddings. *arXiv preprint arXiv:1810.03611* (2018).

[17] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[18] Ewa S Callahan and Susan C Herring. 2011. Cultural bias in Wikipedia content on famous persons. *Journal of the American society for information science and technology* 62, 10 (2011), 1899–1915.

[19] Yanqing Chen, Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2013. The expressive power of word embeddings. *arXiv preprint arXiv:1301.3226* (2013).

[20] Paul Chilton. 2004. *Analysing political discourse: Theory and practice*. Routledge.

[21] Paul Chilton and Christina Schäffner. 2002. *Politics as text and talk: Analytic approaches to political discourse*. Vol. 4. John Benjamins Publishing.

[22] Sasha Costanza-Chock. 2018. Design justice: Towards an intersectional feminist framework for design theory and practice. *Available at SSRN 3189696* (2018).

[23] Bo Cowgill and Catherine Tucker. 2017. *Algorithmic Bias: A Counterfactual Perspective*. Technical Report. Working Paper: NSF Trustworthy Algorithms.

[24] Abdelghani Dahou, Shengwu Xiong, Junwei Zhou, Mohamed Houcine Haddoud, and Pengfei Duan. 2016. Word embeddings and convolutional neural network for arabic sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2418–2427.

[25] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*.

[26] Connie De Boer. 1978. The polls: Attitudes toward homosexuality. *The Public Opinion Quarterly* 42, 2 (1978), 265–276.

[27] Sunipa Dev and Jeff Phillips. 2019. Attenuating Bias in Word Vectors. *arXiv preprint arXiv:1901.07656* (2019).

[28] Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of Textual Cyberbullying. *The Social Mobile Web* 11, 02 (2011), 11–17.

[29] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*. ACM, 29–30.

[30] Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2016. Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 3519–3530.

[31] Alice H Eagly and Antonio Mladinic. 1989. Gender stereotypes and attitudes toward women and men. *Personality and Social Psychology Bulletin* 15, 4 (1989), 543–558.

[32] Thomas Eckes. 2008. Geschlechterstereotype: Von Rollen, Identitäten und Vorurteilen. In *Handbuch Frauen-und Geschlechterforschung*. Springer, 171–182.

[33] Benjamin Edelman, Micahel Luca, et al. 2014. *Digital Discrimination: The Case of Airbnb. com*. Technical Report. Harvard Business School.

[34] K Anders Ericsson and Herbert A Simon. 1984. *Protocol analysis: Verbal reports as data*. the MIT Press.

[35] Facebook. 2018. Research in Brief: Dynamic Meta-Embeddings improve AI language understanding. https://code.fb.com/ai-research/dynamic-meta-embeddings/

[36] Norman Fairclough. 1992. *Discourse and social change*. Vol. 10. Polity press Cambridge.

[37] Michel Foucault. 2013. *Archaeology of knowledge*. Routledge.

[38] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)* 14, 3 (1996), 330–347.

[39] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2001. *The elements of statistical learning*. Vol. 1. Springer series in statistics New York, NY, USA:.

[40] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), E3635–E3644.

[41] Peter Glick and Susan T Fiske. 2018. The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. In *Social Cognition*. Routledge, 116–160.

[42] Bryce Goodman and Seth Flaxman. 2016. European Union regulations on algorithmic decision-making and a" right to explanation". *arXiv preprint arXiv:1606.08813* (2016).

[43] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine* 38, 3 (2017), 50–57.

[44] Bryce W Goodman. 2016. Economic Models of (Algorithmic) Discrimination. In *29th Conference on Neural Information Processing Systems*, Vol. 6.

[45] Mihajlo Grbovic. 2018. Listing Embeddings in Search Ranking. https://medium.com/airbnb-engineering/listing-embeddings-for-similar-listing-recommendations-and-real-time-personalization-in-search-601172f7603e

[46] Mihajlo Grbovic, Nemanja Djuric, Vladan Radosavljevic, Fabrizio Silvestri, Ricardo Baeza-Yates, Andrew Feng, Erik Ordentlich, Lee Yang, and Gavin Owens. 2016. Scalable semantic matching of queries to ads in sponsored search advertising. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 375–384.

[47] Mihajlo Grbovic, Vladan Radosavljevic, Nemanja Djuric, Narayan Bhamidipati, Jaikit Savla, Varun Bhagwan, and Doug Sharp. 2015. E-commerce in your inbox: Product recommendations at scale. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1809–1818.

[48] Shane Greenstein and Feng Zhu. 2012. Is Wikipedia Biased? *American Economic Review* 102, 3 (2012), 343–48.

[49] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology* 74, 6 (1998), 1464.

[50] Louk Hagendoorn. 1995. Intergroup biases in multiple group systems: The perception of ethnic hierarchies. *European review of social psychology* 6, 1 (1995), 199–228.

[51] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2125–2126.

[52] Kira Hall and Mary Bucholtz. 2012. *Gender articulated: Language and the socially constructed self*. Routledge.

[53] William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096* (2016).

[54] Deborah Hellman. 2008. *When is discrimination wrong?* Harvard University Press.

[55] Seth M Holmes and Heide Castañeda. 2016. Representing the "European refugee crisis" in Germany and beyond: Deservingness and difference, life and death. *American Ethnologist* 43, 1 (2016), 12–24.

[56] Cheryl L Holt and Jon B Ellis. 1998. Assessing the current validity of the Bem Sex-Role Inventory. *Sex roles* 39, 11-12 (1998), 929–941.

[57] Michael Howard. 2013. *The Franco-Prussian War: The German Invasion of France 1870–1871*. Routledge.

[58] IBM. 2019. Word Embedding Generator. https://developer.ibm.com/exchanges/models/all/max-word-embedding-generator/

[59] Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the second workshop on NLP and computational social science*. 7–16.

[60] John E Joseph. 2006. *Language and politics*. Edinburgh University Press.

[61] Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya, and Mark Carman. 2016. Are Word Embedding-based Features Useful for Sarcasm Detection? *arXiv preprint arXiv:1610.00883* (2016).

[62] Keith Kirkpatrick. 2016. Battling algorithmic bias: How do we ensure algorithms treat us fairly? *Commun. ACM* 59, 10 (2016), 16–17.

[63] Andreas Klink and Ulrich Wagner. 1999. Discrimination Against Ethnic Minorities in Germany: Going Back to the Field 1. *Journal of Applied Social Psychology* 29, 2 (1999), 402–423.
[64] Austin C Kozlowski, Matt Taddy, and James A Evans. 2018. The Geometry of Culture: Analyzing Meaning through Word Embeddings. *arXiv preprint arXiv:1803.09288* (2018).
[65] Michał Krzyżanowski, Anna Triandafyllidou, and Ruth Wodak. 2018. The mediatization and the politicization of the "refugee crisis" in Europe.
[66] Walter Laqueur. 2018. *Russia and Germany: Century of Conflict.* Routledge.
[67] Susan Leavy. 2014. *Detecting Gender Bias in the Coverage of Politicians in Irish Newspapers Using Automated Text Classification.* Ph.D. Dissertation. Trinity College Dublin.
[68] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2018. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology* 31, 4 (2018), 611–627.
[69] Tjen-Sien Lim, Wei-Yin Loh, and Yu-Shan Shih. 2000. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine learning* 40, 3 (2000), 203–228.
[70] Yang Liu, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2015. Topical Word Embeddings.. In *AAAI.* 2418–2424.
[71] Edward Loper and Steven Bird. 2002. NLTK: the natural language toolkit. *arXiv preprint cs/0205028* (2002).
[72] Bart Maddens, Jaak Billiet, and Roeland Beerten. 2000. National identity and the attitude towards foreigners in multi-national states: the case of Belgium. *Journal of ethnic and migration studies* 26, 1 (2000), 45–60.
[73] Michela Menegatti and Monica Rubini. 2017. Gender bias and sexism in language. In *Oxford Research Encyclopedia of Communication.*
[74] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems.* 3111–3119.
[75] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society* 3, 2 (2016), 2053951716679679.
[76] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How search engines reinforce racism.* NYU Press.
[77] SC Olhede and PJ Wolfe. 2018. The growing ubiquity of algorithms in society: implications, impacts and innovations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2128 (2018), 20170364.
[78] Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206* (2017).
[79] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).* 1532–1543.
[80] Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2010. SentiWS-A Publicly Available German-language Resource for Sentiment Analysis.. In *LREC.*
[81] Katherine J Reynolds, John C Turner, and S Alexander Haslam. 2000. When are we better than them and they worse than us? A closer look at social discrimination in positive and negative domains. *Journal of personality and social psychology* 78, 1 (2000), 64.
[82] Abigail Riemer, Stephenie Chaudoir, and Valerie Earnshaw. 2014. What looks like sexism and why? The effect of comment type and perpetrator type on women's perceptions of sexism. *The Journal of general psychology* 141, 3 (2014), 263–279.
[83] Celia Roberts, Evelyn Davies, and Tom Jupp. 2014. *Language and discrimination.* Routledge.
[84] Paul Rosenkrantz, Susan Vogel, Helen Bee, Inge Broverman, and Donald M Broverman. 1968. Sex-role stereotypes and self-concepts in college students. *Journal of consulting and clinical psychology* 32, 3 (1968), 287.
[85] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* (2014), 1–23.
[86] Peter H Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika* 31, 1 (1966), 1–10.
[87] Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859* (2017).
[88] Dagmar Stahlberg, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2007. Representation of the sexes in language. *Social communication* (2007), 163–187.
[89] Henri Tajfel. 1970. Experiments in intergroup discrimination. *Scientific American* 223, 5 (1970), 96–103.
[90] Cristian Tileaga. 2014. Prejudice as collective defi nition: ideology, discourse and moral exclusion. In *Rhetoric, Ideology and Social Psychology.* Routledge, 85–96.
[91] Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics.* Association for Computational Linguistics, 384–394.
[92] Teun A Van Dijk. 2002. Political discourse and political cognition. *Politics as text and talk: Analytic approaches to political discourse* 203 (2002), 203–237.
[93] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare).* IEEE, 1–7.
[94] Denise C Viss and Shawn M Burn. 1992. Divergent perceptions of lesbians: A comparison of lesbian self-perceptions and heterosexual perceptions. *The Journal of social psychology* 132, 2 (1992), 169–177.
[95] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *Ninth international AAAI conference on web and social media.*
[96] Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. 2016. Women through the glass ceiling: gender asymmetries in Wikipedia. *EPJ Data Science* 5, 1 (2016), 5.
[97] Bernard E Whitley Jr and Mary E Kite. 2016. *Psychology of prejudice and discrimination.* Routledge.
[98] John E Williams and Susan M Bennett. 1975. The definition of sex stereotypes via the adjective check list. *Sex roles* 1, 4 (1975), 327–337.
[99] John E Williams, Robert C Satterwhite, and Deborah L Best. 1999. Pancultural gender stereotypes revisited: The five factor model. *Sex roles* 40, 7-8 (1999), 513–525.
[100] Matthias Winkelmann. 2016. firstname database. https://doi.org/10.5281/zenodo.15991
[101] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496* (2018).
[102] James Zou and Londa Schiebinger. 2018. Design AI so that it's fair. *Nature* 559, 7714 (2018), 324–326.