Using Word Embeddings to Examine Gender Bias in Dutch Newspapers, 1950-1990

Melvin Wevers
DHLab KNAW Humanities Cluster
Oudezijds Achterburgwal 185
1012DK Amsterdam, the Netherlands
melvin.wevers@dh.huc.knaw.nl

Abstract

Contemporary debates on filter bubbles and polarization in public and social media raise the question to what extent news media of the past exhibited biases. This paper specifically examines bias related to gender in six Dutch national newspapers between 1950 and 1990. We measure bias related to gender by comparing local changes in word embedding models trained on newspapers with divergent ideological backgrounds. We demonstrate clear differences in gender bias and changes within and between newspapers over time. In relation to themes such as sexuality and leisure, we see the bias moving toward women, whereas, generally, the bias shifts in the direction of men, despite growing female employment number and feminist movements. Even though Dutch society became less stratified ideologically (depillarization), we found an increasing divergence in gender bias between religious and social-democratic on the one hand and liberal newspapers on the other. Methodologically, this paper illustrates how word embeddings can be used to examine historical language change. Future work will investigate how fine-tuning deep contextualized embedding models, such as ELMO, might be used for similar tasks with greater contextual information.

1 Introduction

In recent years, public and academic debates about the possible impact of filter bubbles and the role of polarization in public and social media have been widespread (Pariser, 2011; Flaxman et al., 2016). In these debates, news media have been described as belonging to particular political ideologies, producing skewed views on topics, such as climate change or immigration. These contemporary debates raise the question to what extent newspapers in the past operated in filter bubbles driven by their own ideological bias.

This paper examines gender bias in historical newspapers. By looking at differences in the strength of association between male and female dimensions of gender on the one hand, and words that represent occupations, psychological states, or social life, on the other, we examine the gender bias in and between several Dutch newspapers over time. Did certain newspapers exhibit a bias toward men or women in relationship to specific aspects of society, behavior, or culture?

Newspapers are an excellent source to study societal debates. They function as a transceiver; both the producer and the messenger of public discourse (Schudson, 1982). Margaret Marshall (1995) claims that researchers can uncover the "values, assumptions, and concerns, and ways of thinking that were a part of the public discourse of that time" by analyzing "the arguments, language, the discourse practices that inhabit the pages of public magazines, newspapers, and early professional journals."

The period 1950-1990 is of particular interest as Dutch society underwent clear industrialization and modernization as well as ideological shifts (Schot et al., 2010). After the Second World War, Dutch society was stratified according to ideological and religious "pillars", a phenomenon known as pillarization. These pillars can be categorized as Catholic, Protestant, socialist, and liberal (Wintle, 2000). Newspapers were often aligned to one of these pillars (Wijfjes, 2004; Rooij, 1974). The newspaper *Trouw*, for example, has a distinct Protestant origin, while *Volkskrant* and *De Telegraaf* can be characterized as, respectively, Catholic and neutral. In recent years, the latter transformed into a newspaper with clear conservative leanings. Newspaper historians have studied the ideological backgrounds of Dutch newspapers using traditional hermeneutic means to which this study adds a computational analysis of language

Figure 1: Female Employment Numbers

use related to gender.

The representation of gender in public discourse is related to ideological struggles over gender equality. Several feminist waves materialized in the Netherlands. The origins of the first feminist wave can be traced back to the mid-nineteenth century and lasted until the interwar period. It took until the 1960s for feminism to flare up again in the Netherlands. In between, confessional parties were vocal in their anti-feminist policies. During the 1960s, the second feminist wave, also known as 'new feminism', focused on gender equality in areas such as work, education, sexuality, marriage, and family (Ribberink, 1987).

The increasing equality between men and women is reflected in growing female employment numbers, which increased from 27.5 percent in 1950 to almost 35 percent in 1990 (Figure 1).[1] Apart from Scandinavia, the Netherlands has the highest levels of equality in Europe. Nonetheless, in terms of education and employment, women are still lagging behind and reports of gender discrimination are not uncommon in the Netherlands (Baali et al., 2018; Ministerie van Onderwijs, 2009).

2 Related Work

Word embedding models can be used for a wide range of lexical-semantic tasks (Baroni et al., 2014; Kulkarni et al., 2015). Hamilton et al. (2016) show how word embeddings can also be used to measure semantic shifts by comparing the contexts in which words are used to denote continuity and changes in language use. More recent work focused on the role of bias in word embeddings, specifically bias related to politics, gender, and ethnicity (Azarbonyad et al., 2017; Bolukbasi et al., 2016; Garg et al., 2018). Gonen et al. (2019) demonstrate that debiasing methods work, but argue that we should not remove them. Azarbonyad et al. (2017) compare semantic spaces related to political views in the UK parliament, effectively comparing biases between embeddings. Garg et al. (2018) turn to biases in embedding to study shifts related to gender and ethnicity.

This study builds upon the work of Garg et al. (2018), and applies it to the context of the Netherlands—represented by Dutch newspapers. We extend their method further by distinguishing between sources, rather than using a comprehensive gold standard data set. We also incorporate external lexicons, such as the emotion lexicon from Cornetto, the *Nederlandse Voornamenbank* (database of Dutch first names), the Dutch translation of LIWC (Linguistic Inquiry and Word Count) and HISCO (Historical International Classification of Occupations) (Vossen et al., 2007; Tausczik and Pennebaker, 2010; Boot et al., 2017; Zijdeman et al., 2013; Bloothooft, 2010).

3 Data

The data set consists of six Dutch national newspapers: *NRC Handelsblad (NRC)*, *Het Vrije Volk (VV)*, *Parool*, *Telegraaf*, *Trouw*, and *Volkskrant (VK)*.[2] These newspapers can be characterized ideologically as liberal, social-democratic, liberal, neutral/conservative, Protestant, and Catholic.

For the analysis, we rely on the articles and not the advertisements in the newspapers. We preprocess the text by removing stopwords, punctuation, numerical characters, and words shorter than three and longer than fifteen characters. The quality of the digitized text varies throughout the corpus due to imperfections in the original material and limitations of the recognition software. Because of the variations in OCR quality, we only retain words that also appeared in a Dutch dictionary.

We use the Gensim implementation of Word2Vec to train four embedding models per newspaper, each representing one decade between 1950 and 1990.[3] The models were trained using C-BOW with hierarchical softmax, with a dimensionality of 300, a minimal word count

[1] https://opendata.cbs.nl/statline/#/CBS/nl/

[2] The digitized newspapers were provided by the National Library of the Netherlands. http://www.delpher.nl

[3] https://radimrehurek.com/gensim/

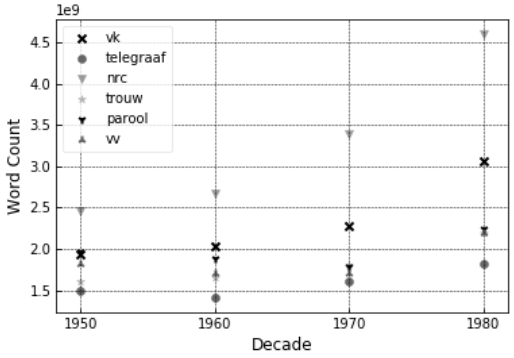

Figure 2: Total number of words per embedding model

and context of 5, and downsampling of 10^{-5}.[4] Figure 2 shows that the size of the vocabulary approximately doubles for some newspapers between 1950 and 1990. The variance of the targets words, however, was small ($\mu \approx 0.003$) and constant ($\sigma[1.3^{-9}, 2.9^{-9}]$), indicating model stability. Since we calculate bias relative to each model, these differences in vocabulary size will have little impact on shifts in bias.

To measure gender bias, we use three sets of targets words. First, we extract a list of approximately 12.5k job titles from the HISCO data set. Second, we select emotion words with a confidence score of 1.0, a positive polarity above 0.5 ($n = 476$) and a negative polarity below -0.5 ($n = 636$) from Cornetto. Third, we rely on the Dutch translation of LIWC2001, which contains lists of words to measure psychological and cognitive states (Pennebaker et al., 2001). We use the following LIWC (sub)categories: Affective and Emotional Processes; Cognitive Processes; Sensory and Perceptual Processes; Social Processes; Occupation; Leisure activity; Money and Financial Issues; Metaphysical Issues; and Physical states.

4 Methodology

For the calculation of gender bias, we construct two vectors representing the gender dimensions (male, female). We do this by creating an average vector that includes words referring to male ('man', 'his', 'father', etc.) or female as well as the most popular first names in the Netherlands

[4]Code can be found here: https://github.com/melvinwevers/historical_concepts and the models here: http://doi.org/10.5281/zenodo.3237380

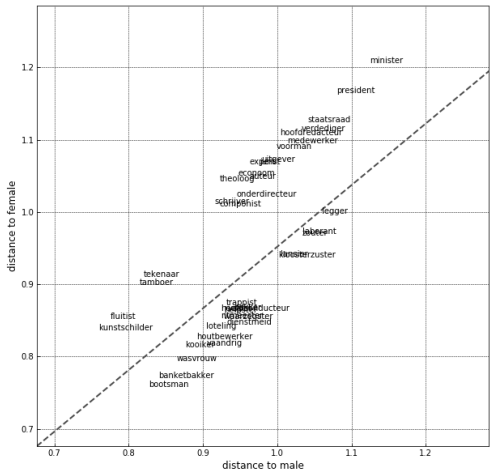

Figure 3: Job titles with strong bias towards men and women in *De Volkskrant*, 1980-1990

for the period 1950-1990.[5] Next, we calculate the distance between each gender vector and every word in a list of target words, for example, words that denote occupations: a greater distance indicates that a word is less closely associated with that dimension of gender. The difference between the distances for both gender vectors represents the gender bias: positive meaning a bias toward women and negative toward men. Figure 3 shows the biases related to forty job titles. Words above the diagonal are biased towards men, and those underneath the diagonal towards women.

Finally, after standardizing and centering the bias values, we apply Bayesian linear regression to determine whether the bias changed over time. The linear model is formulated as:

$$\mu_i = \alpha + \beta * Y_i + \epsilon,$$

with μ_i the bias for each decade (i) and Y_i the coefficient related to each decade (i). The likelihood function is: $X \sim \mathcal{N}(\mu, \sigma)$ with priors defined: $\alpha \sim \mathcal{N}(0, 2)$, $\beta \sim \mathcal{N}(0, 2)$, and $\epsilon \sim \text{HalfCauchy}(\beta = 1)$. For model training, we use a No-U-Turn-Sampler (NUTS) (5k draws, 1.5k tuning steps, Highest Posterior Density (HPD) of .95).[6] For the target words Job Titles, the proposed model (Model B) outperforms a model that only

[5]The word lists for both vectors can be found in Appendix A. The first names were harvested from https://www.meertens.knaw.nl/nvb/

[6]HPD is the Bayesian equivalent of the frequentists confidence interval in Frequentist credible interval. https://docs.pymc.io

	WAIC	pWAIC	dWAIC	weight	SE	dSE
Model B	64624.8	2.9	0	0.99	201.6	0
Model A	64682.1	1.88	57.28	0.01	201.36	15.2

Table 1: Model Comparison

	mean	sd	hpd_2.5	hpd_97.5	n_eff	Rhat
a	-0.164	0.010	-0.185	-0.145	1315.073	1.000
bY	0.046	0.006	0.033	0.055	1261.437	0.999
sigma	1.001	0.005	0.992	1.010	1035.282	1.003

Table 2: Model B Summary

includes the intercept (Model A), indicating that bias changes as a function of time (Table 1 & Table 2).

We compute a linear model that combines all newspapers for the target words Job Titles, Positive Emotions, Negative Emotions, and the selected LIWC columns. Then, for the same categories, we compute individual linear models for each newspaper. The resulting models are reported in Appendix B.

5 Results

The combined linear models, including all newspapers, generally display minimal shifts in bias. While the effects are weak, they fall within a .95 HPD. Partly, the weak trends are related to opposing shifts in the individual newspapers, cancelling each other out. Nonetheless, the bias associated with the categories 'TV', 'Music', 'Metaphysical issues', 'Sexuality' navigate toward women (0.22, 0.12, 0.15, 0.22), with all of them starting from a position that was clearly oriented toward men (-0.36, -0.20, -0.28, -0.39).[7] Conversely, 'Money', 'Grooming', and Negative Emotion words move toward men (-0.24, -0.17, -0.16), which in the 1950s were all more closely related to women (0.33, 0.20, 0.19). For the Job Titles, we see a slight move toward women (0.05), while words from the LIWC category Occupation move marginally in the direction of men (-0.05). This suggests that job titles might be more closely related to women, while the notion of working gravitates toward men.

The linear models for the individual newspapers demonstrate distinct differences between the newspapers. First, *Volkskrant* is the most stable newspapers with 56% of the categories not changing.[8] When bias changes in this newspaper, it

[7]Numbers refer to the slope

[8]Lower confidence interval < 0 and upper > 0

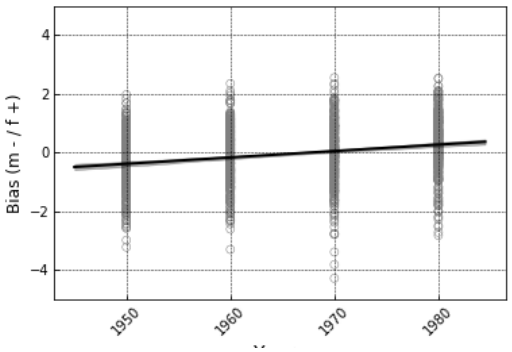

Figure 4: Combined model 'Sexuality'

moves toward women 9 out the 11 categories that change. *Telegraaf*, *NRC*, and *Parool* generally move toward men, respectively (84%, 92%, and 80%). The bias of *Trouw* and *Vrije Volk*, contrarily, move toward women (both 72%).

A noteworthy result is that in all newspapers the bias shifts toward men in the category 'money'. Moreover, they also all exhibit a move toward women for the category 'sexuality', with the clearest shift in *Volkskrant*, *Trouw*, and *Vrije Volk*.

6 Discussion

While the newspaper discourse as a whole is fairly stable, individual newspapers show clear divergences with regard to their bias and changes in this bias. We see that the newspapers with a social-democratic (*Vrije Volk*) and religious background, either Catholic (*Volkskrant*) and Protestant (*Trouw*) demonstrate the clearest shift in bias toward women. The liberal/conservative newspapers *Telegraaf*, *NRC Handelsblad*, and *Parool*, on the contrary, orient themselves more clearly toward men. Despite increasing female employment numbers in the Netherlands, the association with job titles moves only gradually toward women, while words associated with working move toward men. More detailed analysis of the individual trend within each decade is necessary to untangle what exactly is taking place. For example, which words show the biggest shift, and can we identify groups of associated words of which particular words show divergent behavior? Methodologically, this paper shows how word embedding models can be used to trace general shifts in language related to gender. Nevertheless, certain cultural expressions of gender are not captured by distributional semantics represented through word

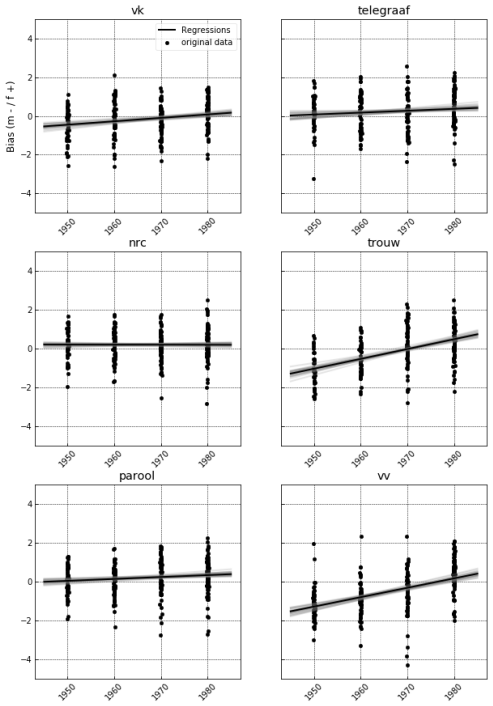

Figure 5: Individual newspaper model 'Sexuality'

embeddings, but rather in syntax, for example, through the use of active of passive sentences. Future work will investigate how fine-tuning state-of-the-art embedding models, such as ELMO and BERT, can be leveraged to gain more contextual knowledge about words and their association with gender (Peters et al., 2018).

Acknowledgments

I would like to thank Folgert Karsdorp for his feedback. This research was part the project "Digital Humanities Approaches to Reference Cultures: The Emergence of the United States in Public Discourse in the Netherlands, 1890-1990", which was funded by the Dutch Research Council (NWO).

References

Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps. 2017. Words are malleable: Computing semantic shifts in political and media discourse. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1509–1518. ACM.

Laila Ait Baali, Roos van Os, and Jantien Kingma. 2018. Overheid moet gendergelijkheid centraal stellen. https://www.volkskrant.nl/gs-b6023212.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 238–247.

Gerrit Bloothooft. 2010. Nederlandse Voornamenbank. https://www.meertens.knaw.nl/nvb/veelgesteldevragen.

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*.

Peter Boot, Hanna Zijlstra, and Rinie Geenen. 2017. The Dutch translation of the Linguistic Inquiry and Word Count (LIWC) 2007 dictionary. *Dutch Journal of Applied Linguistics*, 6(1):65–76.

Seth Flaxman, Sharad Goel, and Justin Rao. 2016. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1):298–320.

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):3635–44.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. *arXiv:1903.03862 [cs]*.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 2116. NIH Public Access.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. International World Wide Web Conferences Steering Committee.

Margaret Marshall. 1995. *Contesting Cultural Rhetorics: Public Discourse and Education, 1890-1900*. University of Michigan Press, Ann Arbor.

Cultuur en Wetenschap Ministerie van Onderwijs. 2009. Vrouwenemancipatie (gendergelijkheid). https://www.rijksoverheid.nl/onderwerpen/vrouwenemancipatie.

Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin, London.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Anneke Ribberink. 1987. *Feminisme*. Stichting Burgerschapskunde, Leiden.

Maarten Rooij. 1974. *Kranten: dagbladpers en maatschappij*. Wetenschappelijke Uitgeverij, Amsterdam.

Johan Schot, Arie Rip, and Harry Lintsen, editors. 2010. *Technology and the Making of the Netherlands: The Age of Contested Modernization, 1890-1970*. MIT Press, Cambridge.

Michael Schudson. 1982. *The Power of News*. Harvard University Press, Cambridge.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

P. Vossen, Katja Hofmann, M. de Rijke, E. Tjong Kim Sang, and Koen Deschacht. 2007. The Cornetto database: Architecture and user-scenarios.

Huub Wijfjes. 2004. *Journalistiek in Nederland, 1850-2000: beroep, cultuur en organisatie*. Boom, Amsterdam.

Michael Wintle. 2000. *An Economic and Social History of the Netherlands, 1800-1920: Demographic, Economic, and Social Transition*. Cambridge University Press, Cambridge.

Richard Zijdeman, Kees Mandemakers, Sanne Muurling, Ineke Maas, Bart Van de Putte, Paul Lambert, Marco Van Leeuwen, Frans Van Poppel, and Andrew Miles. 2013. HSN standardized, HISCO-coded and classified occupational titles.