Contextualized Word Embeddings Expose Ethnic Biases in News

Guusje Thijs
University of Amsterdam
Amsterdam, Netherlands
guusjethijs@hotmail.com

Damian Trilling
University of Amsterdam
Amsterdam, Netherlands
Vrije Universiteit
Amsterdam, Netherlands
d.c.trilling@vu.nl

Anne C. Kroon
University of Amsterdam
Amsterdam, Netherlands
a.c.kroon@uva.nl

ABSTRACT

The web is a major source for news and information. Yet, news can perpetuate and amplify biases and stereotypes. Prior work has shown that training static word embeddings can expose such biases. In this short paper, we apply both a conventional Word2Vec approach as well as a more modern BERT-based approach to a large corpus of Dutch news. We demonstrate that both methods expose ethnic biases in the news corpus. We also show that the biases in the news corpus are considerably stronger than the biases in the transformer model itself.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**;
• **Applied computing** → **Sociology**.

KEYWORDS

static word embeddings, contextualized word embeddings, bias, stereotypes, news, transformer

ACM Reference Format:
Guusje Thijs, Damian Trilling, and Anne C. Kroon. 2024. Contextualized Word Embeddings Expose Ethnic Biases in News. In *ACM Web Science Conference (WEBSCI '24), May 21–24, 2024, Stuttgart, Germany*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3614419.3643994

1 INTRODUCTION

A major issue in today's web is the perpetuation and amplification of stereotypes and biases. By-now famous is the finding that a word embedding model "learns" from a training corpus from Google News that "man is to computer programmer as woman is to homemaker" [5]. Others have made similar observations and found, for example, ethnic biases when training (static) word embedding models on other news corpora [17]. It is especially problematic that these biases are present in the context of *news* – a genre that, based on journalistic norms, should precisely try to minimize biases.

We argue that it is of crucial importance to systematically study and quantify biases in news, for two reasons. First, it allows us to better understand biases that are present in our society, and by extension, where these biases occur and how they develop over time. Second, being able to expose such biases is an important step to address them. As more and more web applications use some form of language modelling, this awareness is needed to work towards a more inclusive web.

In our work, we explore the possibilities to use contextualized word embeddings to expose biases in textual corpora. Contextualized embeddings, in contrast to static embeddings, acknowledge that words can have different meanings depending on their context. Accordingly, they are a more promising approach for research on the content and effects of media [16].

Our approach allows us to do two things: First, we can expose ethnic biases that are encoded in language models (like BERT); second, we can analyze the biases in a corpus of text that we are interested in (such as news). In particular, we report on a case study in which we ask: *How does a contextual word embedding model compare to a static word embedding model in identifying ethnic bias in Dutch news?*

We first compare a Word2Vec model with an additionally trained BERT-model for identifying low-status and high-threat stereotypes of ethnic in- and out-groups in Dutch news articles. Second, we compare two different approaches for measuring stereotypes in the BERT-model: target-neutral templates and target-attribute templates. Third, we investigate in how far the presence of encoded ethnic stereotypes is reduced or enhanced by the additional training of the BERT-model on news articles.

2 RELATED WORK

2.1 Biases and Stereotypes in News

Media play a significant role in shaping attitudes towards social minorities and have the potential to perpetuate stereotypes. Biased beliefs can arise when certain traits, concerns, and opinions of social minorities are selectively presented in news articles, particularly in ethnically segregated societies [2, 18].

Previous research has demonstrated the presence and estimated the strength of stereotypes in news. Early work relied on bag-of-words approaches and focused on the co-occurrence of words like "crime" and "immigration" [13]. Others have used dependency parsing to investigate whether ethnic minorities are related to crime in news reporting [14]. Yet, such approaches by design mainly capture blatant forms of stereotypes where the writer explicitly links, for instance, a minority to crimes. But subtle stereotypes, in which words referring to minorities are used *in the same context* as, for example, a word referring to a criminal, may be more influential in the long run. In particular, while *explicit* stereotypes are increasingly rejected by society, more subtle or *implicit* stereotypes endure, quietly influencing judgments and decisions without provoking resistance, and often developing without conscious awareness [1, 3, 23, 26, 27].

WEBSCI '24, May 21–24, 2024, Stuttgart, Germany

ACM ISBN 979-8-4007-0334-8/24/05
https://doi.org/10.1145/3614419.3643994

To measure subtle or implicit forms of stereotypes, researchers have trained static word embeddings and measured in how far such stereotypes are present in the embeddings, and consequently, also the underlying data [17, 25, 28]. These approaches revealed gender and ethnic stereotypes in 100 years of US news [11], gender bias in Dutch newspapers [28], ethnic stereotypes in Spanish news [25], and ethnic stereotypes in Dutch news [17, 18]. In line with the latter work, we focus specifically on two core dimensions of stereotype content captured by the Stereotype Content Model (SCM) [7]. We investigate to what extent ethnic groups are categorized in terms of *high-threat stereotypes*, which arise from the perception of low warmth (in which ethnic out-groups are framed as a potential threat, such as in relation to criminal behavior) and *low-status stereotypes*, which stem from the perception of a lack of competence (in which they are framed as having a low social of socio-economic status, such as being associated with drug addiction or unemployment). The combination of high-threat and low-status portrayals contributes to a negative characterization and can evoke feelings of fear, hostility, and negative attitudes [6, 7, 9].

One potential drawback of training static word embeddings for detecting bias is that static word embeddings cannot deal with homonyms – after all, words that are spelled the same all share the same embedding. For instance, the Dutch word "oplichten" can mean "to defraud", "to swindle" (obviously relevant in our context, which also deals with crime-related stereotypes), but it can also mean "to lighten up" or even "to lift". Contextualized embeddings, which can deal with such homonyms, also can contain biases [15, 21] – yet, containing less noise, they may give us more accurate estimates of biases in news corpora. We first train – as a baseline – a Word2Vec model on a corpus of Dutch news data, before using the same data to additionally continue training a Dutch-language version of BERT (BERTje, [8]). We ask:

RQ1: Does additional training BERTje lead to different estimates of ethnic stereotypes in news, compared to training Word2Vec?

2.2 Detecting Ethnic Biases Using BERTje

To the best of our knowledge, contextualized word embeddings have not been applied to examine *ethnic* stereotypes in news – but related work has shown that gender biases are present in Transformer models [15, 22]. While using static embeddings to quantify ethnic stereotypes in news is straightforward (train a model on the corpus, look up embedding vectors for words of interest, calculate distances), there are two challenges to address when using contextualized word embeddings to this end.

The first challenge is that conventional techniques used for analysing bias in static embeddings, such as word analogy and association tasks, cannot be applied one-on-one [19]. After all, contextualized embeddings do not provide *one* embedding per token (which we use for such tasks when we have static embeddings), because the position in the embedding space depends on the context. Instead, one can replace tokens in a sentence by "[MASK]" and let the model predict the masked token. While a comprehensive introduction into *masked language modeling* is beyond the scope of this paper, one can intuitively see that one can train a model to learn that the missing (masked) word in a sentence like "the cyclist is ___ her bike" is most likely "riding", and to a (much) lower probability maybe "pushing", "reparing", etc. The predictions the trained model produces provides insights into potential biases in the model as well as in the underlying training data [12, 19].

We can distinguish two possible approaches for creating templates [20], which we will call "target-neutral" and "target-attribute". In our context, an example of a target-neutral template sentence could be "Moroccans are [MASK]"; a target-attribute template sentence could be "[MASK] are [attribute]", where [attribute] is taken from a list of stereotypical attributes. Consequently, we can estimate the likelihood of sentences such as "Moroccans are criminals" versus "Belgians are criminals". We compare both approaches:

RQ2: Does using target-neutral templates lead to different estimates of ethnic stereotypes in news, compared to using target-attribute templates?

The second challenge is to separate the bias in the pre-trained model from the bias in the corpus: In contrast to the static approach, in which *only* our own corpus is seen by the model, BERTje comes pre-trained on 2.4B tokens [8], which already may contain biases. While we are not aware of research specifically aimed at ethnic biases in BERT, gender biases have been found in ELMo [29] and English BERT [4, 19], but not in a German BERT model [4]. A first contribution, therefore, is to explore the presence of ethnic biases in BERTje. Then, to disentangle possible biases within BERTje from biases in the news dataset we study, we compare the stereotypes we find in an out-of-the box BERTje with the stereotypes we find *after* continuing training BERTje with the documents in our corpus – an approach that to the best of our knowledge has not been employed so far.

RQ3: Does additional training of the pretrained BERTje model on a news corpus expose additional ethnic stereotypes in the corpus?

3 DATA AND METHODS

3.1 Resources

We extended the dataset described in [17]. Our corpus differs from that dataset in terms of its temporal and source coverage. Specifically, the present corpus includes news articles spanning from 2000 to 2017, whereas the dataset used by [17] covered 2000–2015. Moreover, the final corpus includes news articles obtained from 28 sources – a notable increase from the five sources in the original dataset – rendering the present dataset larger in size. Originally, the corpus consisted of the five Dutch national newspapers with the highest circulation rate, of which three can be considered high-quality newspapers, and two are often considered as popular tabloid-style newspapers. Alongside these national sources, the corpus was enhanced with regional news sources. Our final corpus is composed of 107,965,966 unique sentences, originating from 7,441,914 Dutch news articles that were published in online and print media.

In order to investigate the existence of ethnic stereotyping, we needed to select specific terms that denote ethnic categories. We chose to duplicate the lexicon as used by [17] for this matter. This list of words represented the eight largest non-Western ethnic groups residing in the Netherlands, namely Turks, Moroccans, Surinamese, Antilleans, Iraqis, Afghans, Syrians, and Somalis. We additionally included Poles, who as fellow EU members may seem closer to the in-group, but are often portrayed as an out-group, especially in the

context of labor migration. Next to the Dutch themselves, Belgians and Germans were considered part of the ethnic in-group due to their geographical, linguistic, and cultural proximity.

To measure low-status and high-threat stereotypes, words that reflect these concepts had to be identified. Because of the similarity in datasets, we used the lexicons representing stereotypes provided by [17], which we slightly revised by correcting spelling mistakes and removing duplicates. The full list is part of the code repository at https://github.com/GuusjeThijs/RM-Thesis.

3.2 Experimental Setup

As outlined above, the static and the dynamic embedding approaches by definition require different analytical approaches.

3.2.1 Word2Vec. Mirroring the approach by [17], we train a Word2Vec model on the entire corpus. In such a model, the target word is predicted based on the surrounding words in the window size. In this study the window size is five, meaning that vocabulary words occurring within five words of each other are considered. Each word is represented as a vector in a 100-dimensional space: the final embedding model returns a distribution of weights over 100 dimensions for each word in the training corpus. We then first conduct a *frequency analysis*: We generate a list of the top 100 most similar words for each ethnicity, in both the singular and plural form. We then, per ethnicity, calculate the proportion of words that appear in the high-threat and low-status stereotype word lists. Second, we do a *cosine similarity comparison*: We calculate the cosine similarity between the vectors for all ethnicities (in both singular and plural forms) and all words in the high-threat and low-status lists. We average these scores per ethnicity.

3.2.2 BERTje. We take an unmodified BERTje model [8] from Hugging Face; additionally, we take a second identical model, which we then continue training using our news corpus. We subject both models to the following analyses. First, we conduct a target-neutral mask experiment. We construct two templates – one for singular and one for plural forms – of neutral sentences containing a mask: "The <ethnicity singular form> is a [MASK]" and "The <ethnicities plural form> are [MASK]." The model assesses the probabilities associated with each potential outcome, where the cumulative probabilities of all outcomes sum up to one. We then retrieve the top 100 masks with the highest probabilities and calculate how many of these words occur in the high-threat and low-status stereotype word lists.

Second, we conduct a target-attribute experiment. We create template sentences that incorporate the target ethnicity and high-threat or low-status attributes. These attributes encompass various word types, such as nouns and adjectives, thus requiring the generation of different types of masked sentences to accommodate all potential attribute words. All templates can be found in the code in the repository at https://github.com/GuusjeThijs/RM-Thesis. In these sentences, all ethnicities are entered at the target position, and for each word in the high-threat and low-status lists, the model assigns a probability indicating the likelihood of the word appearing in the position of [MASK]. This process yields an average probability score for each ethnicity across all sentences with high-threat and low-status words. Using these scores, the mean probability can be calculated for both high-threat and low-status stereotypes within the ethnic in-group and ethnic out-group.

4 RESULTS

4.1 Biases Uncovered Using Word2Vec

Training a Word2Vec model on a corpus of Dutch news reveals striking differences in terms of stereotypes. High-threat and low-status words are almost exclusively found in the vicinity of ethnic out-groups in the embedding space (Fig. 1). The combined ethnic out-group had significantly more occurrences of high-threat stereotypes (171) compared to the combined ethnic in-group (3), resulting in an average occurrence of 19 for the out-group and only 1 for the in-group. Similarly, for low-status stereotypes, the combined out-group had 30 occurrences, while the combined in-group had none, resulting in an average occurrence of 3.33 for the out-group and 0 for the in-group.

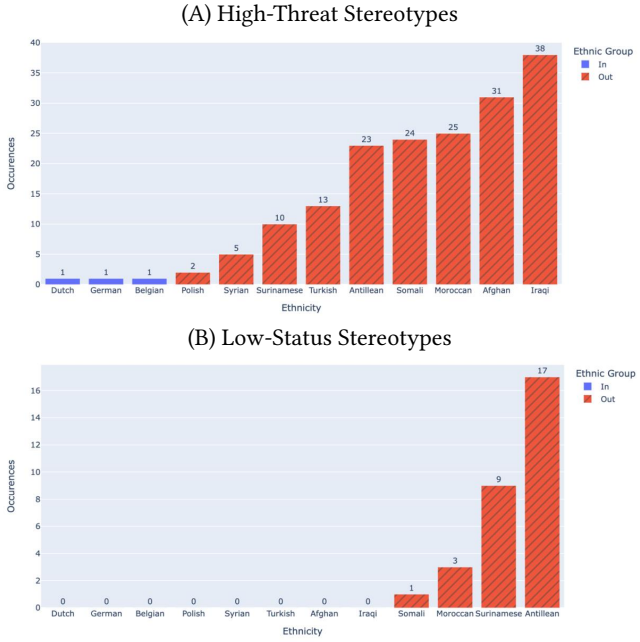

Figure 1: Word2Vec: Frequency of stereotype words in top-100, per ethnicity

Our cosine similarity analysis confirms this picture: the vectors of high-threat and low-status words are clearly more similar to the vectors of ethnic out-groups than ethnic in-groups. If we combine the specific groups into two bins (in-group versus out-group), then the out-group has an average high-threat similarity score of 0.38, while the ethnic in-group scored 0.21. The ethnic out-group has an average low-status similarity score of 0.20, while the ethnic in-group had a score of -0.01.

4.2 Biases Uncovered Using BERTje

Our first observation is that the *target-neutral template* approach turns out not to be effective in uncovering biases in word embedding models. In contrast to the static approach, there was virtually no

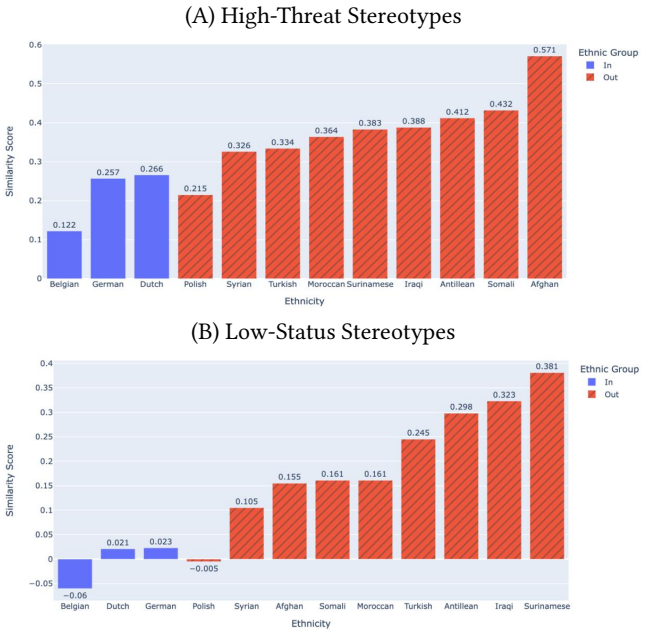

Figure 2: Word2Vec: Average similarity with stereotype words, per ethnicity

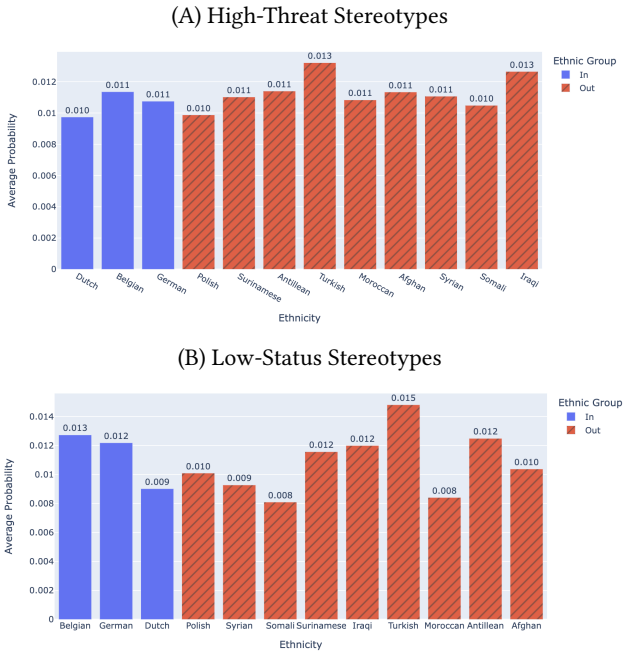

Figure 3: Original BERTje: Average Probabilities per Ethnicity

overlap between predictions and stereotype words. The amount of predicted tokens that match our lists of high-threat and low-status words ranges from 0 to 4 (most values being 0 or 1), depending on ethnicity and whether we use the model with or without additionally training. This does not allow us to make meaningful comparisons.[1]

We therefore focus on the *target-attribute template* approach. Figure 3 shows only limited evidence of ethnic stereotypes in the original, untouched BERTje – the in-group and out-group bars have similar lengths. But once we continue training BERTje with our news corpus, stereotypes clearly emerge, as Figure 4 demonstrates. Numerically, in the experiment using the original BERTje model, the ethnic out-group has an average high-threat probability score of 1.07 times and a low-status probability score of 0.95 times that of the ethnic in-group – hence, the out-group seems to be mentioned even *less* in a low-status context than the in-group. Yet, the size difference seems to be of little practical importance. While we cannot fully rule out the presence of ethnic bias in the model, it also seems that the variation within out-groups (and to a lesser extend also within in-groups) is more substantial than the difference of in- versus out-group (Figure 3).

Once we continue training the model with our news corpus, this picture changes drastically (Figure 4). The ethnic out-group has an average high-threat probability of 2.82 times that of the ethnic in-group. Similarly, the ethnic out-group has a low-status probability of 1.74 times that of the ethnic in-group.

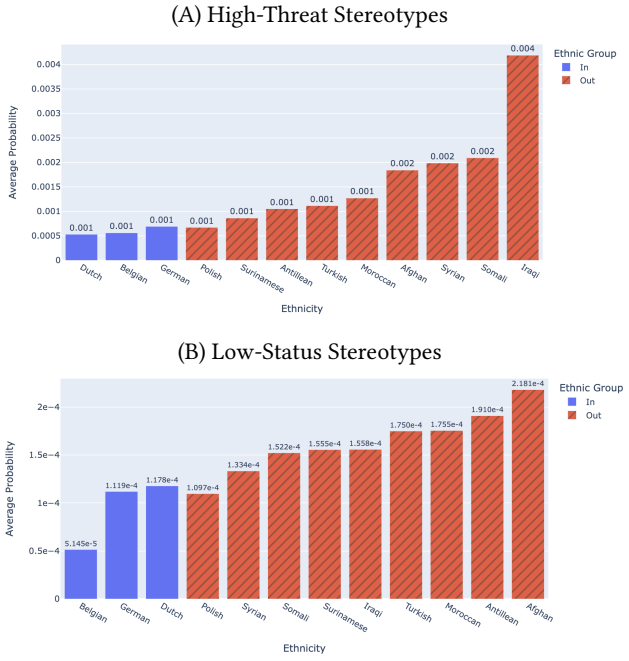

Figure 4: Additionally Trained BERTje: Average Probabilities per Ethnicity

[1]If we compare the numbers anyway, it does seem that additional training *decreases* stereotypes, which seems to contradict our other findings. Yet, these numbers are so small – also compared to the frequencies we find in the other analyses – that it would not be warranted to draw substantive conclusions here.

5 CONCLUSION AND DISCUSSION

5.1 Answers to RQs

We set out to identify ethnic bias in Dutch news, comparing a contextual word embedding model to a static word embedding model. Summarizing our results, we can now give a concise answer to our research questions.

RQ1 asked whether additional training of a contextualized embedding model, specifically BERTje, led to different estimates of ethnic stereotypes in news compared to training a static embedding model, like Word2Vec. Our results show that both approaches can effectively identify low-status and high-threat stereotypes related to in- and out-groups in Dutch news articles. There are some differences regarding the specific ordering of the ethnicities, but the overall picture is remarkably similar: In both approaches, out-groups are shown to be substantively stronger associated with low-status and high-threat stereotypes than in-groups. We view this as a clear indication that training static embeddings or further training a contextualized embedding model can serve as a valuable diagnostic tool for uncovering ethnic stereotypes in large-scale (news) corpora.

RQ2 asked whether using target-neutral templates leads to different estimates of ethnic stereotypes in news, compared to using target-attribute templates. Here, we can be short: At least in our setup, target-neutral templates did not work: There was virtually no overlap between the top-100 predictions and our list of high-threat and low-status words, even though the same lists have proved to work very well in the static embedding approach. We thus recommend using target-attribute sentence templates for measuring ethnic biases.

RQ3 asked whether additional training of the pre-trained BERTje model on a news corpus exposes additional ethnic stereotypes in the corpus. We found that this indeed is the case, to a remarkable degree. While the original BERTje model contained only very limited ethnic biases, additionally training it with our corpus introduced the biases to the model which we also found using the static embedding approach (see answer to RQ1).

5.2 Limitations and Future Work

We demonstrated that Dutch news contains significant ethnic biases, and that these can be exposed both by static as well as by contextualized word embeddings. In this work, we deliberately did not address the question how our findings relate to external real-world data, like for instance crime rates; neither do we discuss the normative question how models should deal with existing real-world biases. Yet, these are important discussions to have where interdisciplinary collaboration is needed.

Whether our results generalize to other languages and media systems requires replication of our work using other datasets. A limitation is that we studied ethnic stereotypes only by considering two dimensions: status and threat. While this is in line with previous research [10, 17, 24], future work may look into other and/or more fine-grained stereotypes and biases.

Also other types of biases still deserve our attention: As we have discussed, there is quite some work on gender bias in news, and we have contributed to the literature on ethnic biases. Yet, one could envision to use comparable methods to also study biases related to, for instance, age, religion, or more.

It needs to be noted that while training the static Word2Vec model took only two hours on a conventional CPU, we needed a server with a dedicated GPU for additionally training BERTje with the same corpus. Training extended over a period of 10 full days. Such considerations may be a good argument for researchers for choosing one method over the other, in particular as we have shown the results to be comparable. These large computation times also mean that training our models multiple times, using resampling techniques like cross-validation or bootstrapping, is not feasible. This would allow us to get a clearer picture of how robust our specific findings are and to calculate some form of error bars for our plots. Unfortunately, the financial costs but also the ecological footprint are often prohibitive here.

Finally, having seen the potential of the target-attribute mask experiments, we plan to explore whether we can leverage more modern models, such as GPT4, to detect biases in news.

ACKNOWLEDGMENTS

This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-5556. DT's contribution is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 947695).

REFERENCES

[1] Florian Arendt. 2013. Dose-dependent media priming effects of stereotypic newspaper articles on implicit and explicit stereotypes. *Journal of Communication* 63, 5 (2013), 830–851.

[2] Florian Arendt. 2015. Effects of Long-Term Exposure to News Stereotypes on Implicit and Explicit Attitudes. *International Journal of Communication* 9 (2015), 2370–2390.

[3] Manuela Barreto, Naomi Ellemers, Sezgin Cihangir, and Katherine Stroebe. 2009. *The self-fulfilling effects of contemporary sexism: How it affects women's well-being and behavior.* American Psychological Association.

[4] Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Barcelona, Spain (Online), 1–16. https://aclanthology.org/2020.gebnlp-1.1

[5] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 4356–4364.

[6] Amy J.C. Cuddy, Susan T. Fiske, and Peter Glick. 2008. Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. In *Advances in Experimental Social Psychology*. Vol. 40. Elsevier, 61–149. https://doi.org/10.1016/S0065-2601(07)00002-0

[7] Amy J. C. Cuddy, Susan T. Fiske, Virginia S. Y. Kwan, Peter Glick, Stéphanie Demoulin, Jacques-Philippe Leyens, Michael Harris Bond, Jean-Claude Croizet, Naomi Ellemers, Ed Sleebos, Tin Tin Htun, Hyun-Jeong Kim, Greg Maio, Judi Perry, Kristina Petkova, Valery Todorov, Rosa Rodríguez-Bailón, Elena Morales, Miguel Moya, Marisol Palacios, Vanessa Smith, Rolando Perez, Jorge Vala, and Rene Ziegler. 2009. Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology* 48, 1 (2009), 1–33. https://doi.org/10.1348/014466608X314935

[8] Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. arXiv:1912.09582 [cs.CL]

[9] Susan T Fiske, Amy JC Cuddy, Peter Glick, and Jun Xu. 2018. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. In *Social Cognition*. Routledge, 162–214.

[10] Yujiro Fujiwara, RCL Velasco, Lee Kenneth Jones, and RL Hite. 2022. Competent and cold: a directed content analysis of warmth and competence dimensions to

identify and categorise stereotypes of scientists portrayed in meme-based GIFs. *International Journal of Science Education* 44, 4 (2022), 694–715.

[11] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word Embeddings as a Lens to Quantify 100 Years of Gender and Ethnic Stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), E3635–E3644. https://doi.org/10.1073/pnas.1720347115 arXiv: 1711.08412 ISBN: 1720347115.

[12] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5491–5501. https://doi.org/10.18653/v1/2020.acl-main.487

[13] Laura Jacobs, Alyt Damstra, Mark Boukes, and Knut De Swert. 2018. Back to Reality: The Complex Relationship Between Patterns in Immigration News Coverage and Real-World Developments in Dutch and Flemish Newspapers (1999–2015). *Mass Communication and Society* 21, 4 (2018), 473–497. https://doi.org/10.1080/15205436.2018.1442479

[14] Azade Esther Kakavand and Damian Trilling. 2022. The Criminal is Always the Foreigner?! A Case Study of Minority Signification in German Crime Reporting. *International Journal of Communication* 16 (2022), 1169–1196.

[15] Styliani Katsarou, Borja Rodríguez-Gálvez, and Jesse Shanahan. 2022. Measuring Gender Bias in Contextualized Embeddings. *Computer Sciences & Mathematics Forum* 3, 1 (2022), 13 pages. https://doi.org/10.3390/cmsf2022003003

[16] Anne Kroon, Kasper Welbers, Damian Trilling, and Wouter van Atteveldt. 2023. Advancing Automated Content Analysis for a New Era of Media Effects Research: The Key Role of Transfer Learning. *Communication Methods and Measures* online first (2023), 1–21. https://doi.org/10.1080/19312458.2023.2261372

[17] Anne C. Kroon, Damian Trilling, and Tamara Raats. 2021. Guilty by Association: Using Word Embeddings to Measure Ethnic Stereotypes in News Coverage. *Journalism & Mass Communication Quarterly* 98, 2 (2021), 451–477. https://doi.org/10.1177/1077699020932304

[18] Anne C. Kroon, Damian Trilling, Toni G. L. A. van der Meer, and Jeroen G. F. Jonkman. 2020. Clouded reality: News representations of culturally close and distant ethnic outgroups. *Communications* 45 (2020), 744–764. Issue s1. https://doi.org/10.1515/commun-2019-2069

[19] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 166–172. https://doi.org/10.18653/v1/W19-3823

[20] Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Pipelines for Social Bias Testing of Large Language Models. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*. Association for Computational Linguistics, virtual+Dublin, 68–74. https://doi.org/10.18653/v1/2022.bigscience-1.6

[21] Jesutofunmi A. Omiye, Jenna C. Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. 2023. Large language models propagate race-based medicine. *npj Digital Medicine* 6, 1 (Oct. 2023), 195. https://doi.org/10.1038/s41746-023-00939-z

[22] Shiva Omrani Sabbaghi, Robert Wolfe, and Aylin Caliskan. 2023. Evaluating biased attitude associations of language models in an intersectional context. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 542–553.

[23] Adam R Pearson, John F Dovidio, and Samuel L Gaertner. 2009. The nature of contemporary prejudice: Insights from aversive racism. *Social and Personality Psychology compass* 3, 3 (2009), 314–338.

[24] Alexander Sink, Dana Mastro, and Marko Dragojevic. 2018. Competent or warm? A stereotype content model approach to understanding perceptions of masculine and effeminate gay television characters. *Journalism & Mass Communication Quarterly* 95, 3 (2018), 588–606.

[25] Danielly Sorato, Diana Zavala-Rojas, and Maria del Carme Colominas Ventura. 2021. Using Word Embeddings to Quantify Ethnic Stereotypes in 12 years of Spanish News. In *Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association*. Australasian Language Technology Association, Online, 34–46. https://aclanthology.org/2021.alta-1.4

[26] Janet K Swim, Kathryn J Aikin, Wayne S Hall, and Barbara A Hunter. 1995. Sexism and racism: Old-fashioned and modern prejudices. *Journal of Personality and Social Psychology* 68, 2 (1995), 199.

[27] Jolien A van Breen, Russell Spears, Toon Kuppens, and Soledad de Lemus. 2018. Subliminal gender stereotypes: Who can resist? *Personality and Social Psychology Bulletin* 44, 12 (2018), 1648–1663.

[28] Melvin Wevers. 2019. Using Word Embeddings to Examine Gender Bias in Dutch Newspapers, 1950-1990. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*. Association for Computational Linguistics, Florence, Italy, 92–97. https://doi.org/10.18653/v1/W19-4712

[29] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender Bias in Contextualized Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 629–634. https://doi.org/10.18653/v1/N19-1064