No Word Embedding Model Is Perfect: Evaluating the Representation Accuracy for Social Bias in the Media

Maximilian Spliethöver [1], Maximilian Keiff [2], and Henning Wachsmuth [1]
[1]Leibniz University Hannover, Institute of Artificial Intelligence
[2]Universität Hamburg, Department of Informatics
{m.spliethoever,h.wachsmuth}@ai.uni-hannover.de,
maximilian.keiff@studium.uni-hamburg.de

Abstract

News articles both shape and reflect public opinion across the political spectrum. Analyzing them for social bias can thus provide valuable insights, such as prevailing stereotypes in society and the media, which are often adopted by NLP models trained on respective data. Recent work has relied on word embedding bias measures, such as WEAT. However, several representation issues of embeddings can harm the measures' accuracy, including low-resource settings and token frequency differences. In this work, we study what kind of embedding algorithm serves best to accurately measure types of social bias known to exist in US online news articles. To cover the whole spectrum of political bias in the US, we collect 500k articles and review psychology literature with respect to expected social bias. We then quantify social bias using WEAT along with embedding algorithms that account for the aforementioned issues. We compare how models trained with the algorithms on news articles represent the expected social bias. Our results suggest that the standard way to quantify bias does not align well with knowledge from psychology. While the proposed algorithms reduce the gap, they still do not fully match the literature.

1 Introduction

Social bias describes prejudices and stereotypical thinking towards certain groups in society, such as genders or ethnicities (Fiske, 1998). *Media bias*, by contrast, refers to the tendency of media entities (e.g., a news outlet) to favor certain facts, views, or framings of events over others (Chen et al., 2021). In this work, we focus on media bias induced by political orientations (henceforth, political bias). While social and political bias differ in appearance, it can be expected that they relate to and mutually influence each other. A particular political bias, for example, may transport ideas of stereotypes, manifesting as social bias, that strengthen specific political ideas in society (Seiter, 1986; Domke et al., 1999). Vice versa, holding particular stereotypical views may make people more susceptible to a political view promoted by a news outlet (Schwarz and Jalbert, 2020). Figure 1 shows excerpts of two news articles, conveying potential gender bias.

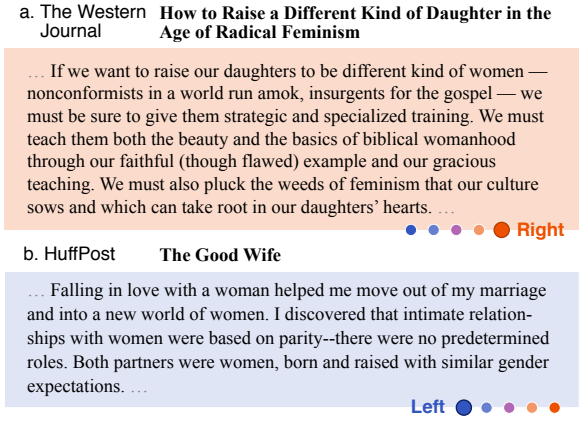

Figure 1: Excerpts of two articles, from a right and a left news outlet according to *allsides.com*. Both show potential gender bias, but of different kind. The articles are included in the corpus presented in Section 4.

The outlined kinds of bias are also relevant to NLP methods that employ news articles to train models (Mikolov et al., 2013) or as a knowledge source (Slonim et al., 2021). For example, bias present in the articles may be learned and amplified by word embeddings if not explicitly accounted for. This impacts generalization performance negatively (Shah et al., 2020) and may have harmful consequences in practical applications (Bender et al., 2021; Joseph and Morgan, 2020). So far, one hurdle to mitigate these problems is the limited reliability of common measures of social bias present in a corpus (Spliethöver and Wachsmuth, 2021), stemming from embedding training algorithms not tailored to low-resource situations (Knoche et al., 2019; Spinde et al., 2021).

In this paper, we investigate how to assess social bias more reliably while empirically studying the interaction of social bias and political bias in

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 2081–2093
December 7-11, 2022

US online news outlets. In particular, we identify *low-resource settings* and *token frequency differences* as two main issues with existing embedding-based bias measures. We consider social bias towards genders, ethnicities and religions, and measure it with the widely used bias measure, WEAT (Caliskan et al., 2017). We restrict our political bias view to the unidimensional spectrum from left to right (Duckitt and Sibley, 2010), ignoring objectivity and fairness aspects (Chen et al., 2020).

In psychology literature, stereotypical views have been shown to coincide with political orientations (Section 2), suggesting that the political views of news outlets coincide with social biases. Under this premise, we aim to find out what word embedding algorithm best serves to reliably measure social bias. We investigate weaknesses of a standard algorithm that stem from the reliance on word lists, infrequent tokens in the data, and the quality of embeddings. We suggest (1) training frequency agnostic embeddings to compensate for lower quality of rare tokens, (2) a fine-tuned language model to account for smaller datasets, and (3) decontextualized embeddings to alleviate the "unnatural input" problem with contextualized models.

For our experiments, we introduce a large-scale media bias corpus in Section 4, covering more than 500,000 news articles from 47 English-language US online news outlets over 12 years (2010–2021). Given the corpus, we evaluate each potential improvement and compare their capability to encode and represent social bias a text corpus (Section 5). To this end, we systematically generate word embedding models from subsets of different political biases. In a second analysis, we explore the development of social bias in outlets over time in a respective manner. We can quantify the considered types of social bias for all models using WEAT.

Our results in Section 5 provide evidence that the general embeddings quality improves notably over standard static embeddings. Additionally, the proposed algorithms better model the expected social bias, though still not fully align with the literature.

This work provides three contributions to computational research on bias in language:

1. Findings on how to combine embedding models and bias measures to adequately quantify social bias in text corpora;
2. a large-scale news resource annotated for political bias; and
3. empirical insights into the interaction of social and media bias in US online news, and its development over time.[1]

2 Related work

We consider social bias that manifests as stereotypes, that is, generalized beliefs about social outgroups based on experiences with single members (Fiske, 1998). Such beliefs may lead to prejudices and discrimination that cause lasting harm. Stereotypes are usually transported through language, uttered either implicitly or explicitly (Wodak, 2008). If entities with high public outreach, such as politicians and media outlets, spread stereotypes, this may therefore profoundly impact their audiences (Seiter, 1986; Domke et al., 1999).

Psychology and political science literature study the relation of stereotypes with political aspects. As part of this, multiple layers of partisan biases have been evaluated (Hayes, 2011; Bauer, 2015; Clifford, 2020). Focusing on social values, Valentino and Sears (2005) find a general shift of public social values related to a shift in voting outcomes. Other works compare social values of conservatives and liberals: While liberals seem more likely to reject "ingroup values", conservatives emphasize tradition and religion (Sylwester and Purver, 2015). Accordingly, Webster et al. (2014) observe a higher level of self-reported prejudices towards social groups that "challenge or violate traditional social values" among conservative probands. Chirumbolo et al. (2016), finally, report that liberals tend to value social equality, whereas conservatives justify social inequality with "the preservation of status quo". We use these connections between political bias and social values as a reference for our analyses.

Media bias can be evaluated from many angles, too. For instance, Chen et al. (2020) explore media bias in political news, automatically detecting incomplete reporting and evaluating its linguistic manifestations. One of their results is that words expressing negative emotions are most correlated with selective biases. Kenix and Jarvandi (2019), in turn, focus on conservative and liberal news articles from the US, Australia, and the UK to understand the construction of media frames. They find that the report framing of specific outlets aligns with their political bias. Rather than unfairness or issue perception, our work targets the interaction of media bias with social bias in news articles.

[1] Code and data at github.com/webis-de/EMNLP-22.

In particular, we quantify social bias in word embeddings with the widely used *Word Embedding Association Test*, WEAT (Caliskan et al., 2017). WEAT's main idea is to calculate the cumulative distance between groups of word vectors that describe a social group and attributes. Similar measures exist, such as ECT (Dev and Phillips, 2019), RNSB (Sweeney and Najafian, 2019), and MAC (Manzini et al., 2019), RIPA (Ethayarajh et al., 2019), WEATVEC (Knoche et al., 2019), the Smoothed First-Order Co-occurrence (Rekabsaz et al., 2021) and SAME (Schröder et al., 2021) but our goal is not to find the best measure. Rather, we seek to learn how measures like WEAT behave for different embedding algorithms. We are not aware of works that have done similar.

Similar to the analysis we carry out, Garg et al. (2018) exploit the properties of word embeddings to evaluate temporal relationships between changes of social bias and empirical demographic changes in the US. They evaluate embedding models trained on texts from different decades, for example finding that gender bias decreased with the women's movement in the 1960's. In a comparable analysis, Rios et al. (2020) find that gender bias reduced in biomedical research over time for some areas, but not in others. In this work, we utilize WEAT to evaluate social bias in news articles. Unlike previous work, however, we compare word embedding algorithms to model social bias in texts and their alignment with the literature reviewed above.

Closest to our work is the research of Knoche et al. (2019) and Spinde et al. (2021). The former use WEAT to compare social biases present word embeddings trained on different ideological online wikis. All wikis are found to have similar biases for gender, race, and religion, but to varying degrees. Spinde et al. (2021) collect US news articles from a liberal and a conservative media outlet. By training one embedding model for each outlet and measuring the differences of all words in the embedding spaces, they determine the most biased words. The underlying hypothesis is that words, for which the context varies more strongly, will also be more biased. We apply a data collection method similar to Spinde et al. (2021), but cover 47 outlets. Additionally, instead of just focusing on two extreme communities, our corpus spans a wider spectrum of political opinions. Our main goal is to deepen the understanding of the social bias in word embeddings for different training algorithms.

3 Method

This paper studies how to best evaluate a text corpus for social bias, harnessing the ability of word embeddings to encode direct contexts. In particular, we quantify the social bias encoded in models trained on a corpus. The models are thus used as a proxy from which we derive the social bias in the original corpus. In the following, we present our evaluation method, discuss potential issues, and describe the employed embedding algorithms.

3.1 Evaluating Social Bias in Embeddings

We seek to analyze to what extent word embedding models encode the social bias of training data. For further insights, we investigate the models quality.

Word Similarity The quality of the semantic space of word embedding models benefits from larger datasets (Pennington et al., 2014). Since most social bias measures rely on this space, better embeddings should also yield more accurate bias evaluations. To gain a better understanding of the quality, we conduct word-similarity evaluations (Spinde et al., 2021) of all models we explore. These evaluations are based on a list of word pairs, human-annotated for similarity. For each pair, the cosine similarity between the vectors generated by a model is computed. The Spearman's ρ between the vector similarities and the annotations represents the score. While this intrinsic evaluation is not able to predict the performance on downstream tasks, it provides insights into the semantic quality of the embeddings (Faruqui et al., 2016). The results also enable us put the social bias evaluation into context. We apply two tests, *MEN* (Bruni et al., 2014) and *WordSim353* (Finkelstein et al., 2001).[2]

Social Bias To quantify social bias, we report results of WEAT (Caliskan et al., 2017). At its core, WEAT relies on four word lists describing a concept. Two lists describe social groups that are evaluated in the context of attributes which represent the other two lists. Common combinations are:

- *Gender.* Male/female and career/family terms
- *Ethnicity.* African-/European-American names and pleasant/unpleasant terms
- *Religion.* Christianity/Islam terms and pleasant/unpleasant terms

[2] https://github.com/EloiZ/embedding_evaluation

Using a given embedding model, all words are transformed into word vectors, in order to measure the cumulative distance between the vectors. Let G and $\tilde{G}$ be the word embeddings for the two social group lists, and A and B those for the attribute lists. Now, let $\Delta(\mathbf{w}, G, \tilde{G})$ be the mean difference between the cosine similarity of a word embedding $\mathbf{w}$ to all word embeddings in G and to the embeddings in $\tilde{G}$. Then, the WEAT score is defined as the effect size of the difference between A and B:

$$\frac{\text{mean}_{\mathbf{a}\in A}\Delta(\mathbf{a}, G, \tilde{G}) \; - \; \text{mean}_{\mathbf{b}\in B}\Delta(\mathbf{b}, G, \tilde{G})}{\text{std\_dev}_{\mathbf{w}\in A\cup B}\Delta(\mathbf{w}, G, \tilde{G})}$$

This results in a value from –2 to 2, where 0 represents the least possible bias. Using WEAT makes our results comparable with related work. We calculate WEAT scores using the implementation of the WEFE framework (Badilla et al., 2020) and use word lists of Spliethöver and Wachsmuth (2021).

Accuracy of Bias Evaluation Evaluating social bias in a word embedding model assumes that its semantic space is meaningful. As different word embedding algorithms achieve this with varying success, they likely also differ in their accuracy in encoding social bias. Assuming that the bias measure at hand (here, WEAT) works as intended, it is possible to evaluate differences between algorithms, given a corpus with known social bias. Below, we thus compare models of different embedding algorithms on training data for which the social bias is known from literature (see Section 2). While we cannot derive exact WEAT values for a corpus, we can infer relative differences for liberal and conservative texts. Together with the results of the word similarity evaluation, we can draw conclusions regarding the reliability of the results.

3.2 Potential Evaluation Issues

As previous research (Spliethöver and Wachsmuth, 2020; Spinde et al., 2021) points out, evaluating text corpora for social bias with static word embeddings (e.g., word2vec) entails three main problems:

1. *Limited Corpus Size.* The training data influences the semantic quality of the embeddings.
2. *Representation Degeneration.* Token frequency differences in the training data entail embeddings of differing quality.
3. *Out-of-Vocabulary Tokens.* Limited vocabularies cause unknown tokens during evaluation.

In the following, we describe these issues in more detail. To alleviate them, we train word embedding models with different algorithms below.

Limited Corpus Size To generate a meaningful semantic space based on context, word embedding models tend to require large datasets. For example, the pre-trained word2vec model (Mikolov et al., 2013) was trained on 100B tokens, the largest GloVe model (Pennington et al., 2014) on 840B. Thus, the quality of the embedding may suffer from small corpora. In turn, the results of the bias evaluation may not be as accurate as with larger corpora.

Representation Degeneration Representation degeneration describes the dependence of meaningful embeddings on the token occurrences, reflecting its available number of contexts (Karampatsis et al., 2020). It implies that infrequent tokens (*rare tokens*) tend to have lower-quality embeddings than more frequent ones (*popular tokens*). While fluctuations are expected due to Zipf's law, they result in less reliable semantic encodings (Gong et al., 2018; Karampatsis et al., 2020; Wolfe and Caliskan, 2021). In the context of social bias measures, this issue is especially relevant, since they implicitly assume a similar quality for all word vectors. The difference between tokens can be high for certain corpora (Spliethöver and Wachsmuth, 2020). Even more problematic, the occurrences also tend to vary within a single test (e.g., more male term occurrences than female ones), potentially influencing the social bias measure results negatively.

While a frequency difference can itself be a form of social bias, it makes the evaluation less straightforward, which is why we ideally seek to abstract from it. A naïve way would be to artificially augment the data by duplicating contexts of rare tokens. As we intend to keep the original signals, though, we explore more direct means of abstraction.

Out-Of-Vocabulary Tokens Static word embedding models have a fixed vocabulary, determined by tokens in their training corpus and are unable to generate embeddings for tokens not included (henceforth, OOV tokens). However, most embedding bias measures rely on pre-defined word lists and assume that an embedding is available for each word. OOV tokens hence need to be ignored in the evaluation, reducing the comparability of multiple models. This can be alleviated by sub-word tokenization, as used for BERT (Devlin et al., 2019).

3.3 Word Embedding Algorithms

We hypothesize that no existing word embedding algorithm is able to account for all issues discussed. Therefore, we train models with multiple algorithms, and we evaluate them against each other. For implementation details on the different algorithms, see Appendix A.

Static As baseline, we train static embedding models with word2vec (Mikolov et al., 2013). An advantage of this method is the fast training process. Also, static word embeddings are by now well researched and interpretable (Bommasani et al., 2020). In turn, the algorithms require large training data to generate high-quality embeddings. Furthermore, due to the representation degeneration problem, the measured bias may be less comparable if the token frequencies vary strongly between two corpora. The static models will be referred to as Static in the following.

Frequency-Agnostic Frequency-agnostic word embeddings (FRAGE) (Gong et al., 2018) aim to approach the representation degeneration problem by accounting for the frequency of tokens. FRAGE does so by training a long short-term memory model (LSTM) on a language modeling task and introducing an adversarial discriminator, classifying tokens as rare or popular. During training, the LSTM tries to minimize the ability of the adversarial to predict the class of each token. While reducing the impact of token frequency, the model is trained from scratch, increasing training time requiring much data to obtain high-quality embeddings. The models will be referred to as FrecAgn.

Fine-Tuned To account for the shortcomings of FRAGE, we additionally fine-tune BERT. On the one hand, it provides a good basis for embeddings, as it is pre-trained on large corpora. This should offer a certain level of base quality for semantic embeddings, potentially reducing the negative effect of size differences in the fine-tuning data. Moreover, it may minimize quality differences between embeddings of rare and popular tokens. Due to sub-word tokenization, OOV tokens are also not an issue. However, BERT contextualizes embeddings dynamically during generation, requiring the context of a token (e.g., the sentence it appears in) as input. Since bias measures usually work with single token embeddings, we need to generate embeddings by querying the model for unnatural inputs (e.g., inputs containing only the token in question without context) (Bommasani et al., 2020). The resulting models will be referred to as Fine-Tuned.

Decontextualized As an alternative to fine-tuned BERT, we employ the averaged pooling strategy presented by Bommasani et al. (2020) to generate decontextualized embeddings. The general idea is to embed all contexts of a specific token in a *context dataset* using a language model. To receive a single embedding per token, the contextualized embeddings are then averaged. Since the final embeddings are contextualized by the context dataset, they can also be expected to encode its social bias. We thus use the corpus we aim to evaluate for social bias as context. Since this method is also based on BERT, we expect the embeddings to have similar advantages over static embeddings, while accounting for the unnatural-input problem. Moreover, since the resulting embeddings are static rather than contextualized, they should retain benefits such as better interpretability. The time needed to generate decontextualized word embeddings is, however, more dependent on the size of the context dataset, since all contexts need to be embedded separately. This results in a potentially long generation time. The models will be referred to as Decontext.

4 Data

We now present the large-scale corpus that we acquired to study the existence of social bias in news articles across the political spectrum in the US.

Source Data Using media bias ratings from news aggregation platform *allsides.com*, we collected articles from liberal (left and lean-left labels) and conservative (right and lean-right labels), as well as neutral (center label) outlets. While this unidimensional view on the political spectrum is limited (Duckitt and Sibley, 2010), it provides us with a clear distinction and makes results easier to interpret. We refer to news articles with liberal, neutral, and conservative labels as *data subsets* in Section 5.

Similar to Spinde et al. (2021), we collected news articles from *Common Crawl*[3]. Since the media bias rating history is not available, we mapped each outlet to its current rating. To extract the pure text from the collected files in WARC format, we used the library *news-please* (Hamborg et al., 2017). For our experiments, we also extracted the articles' date of publication automatically as far as possible.

[3]Common Crawl, https://commoncrawl.org

Orientation	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	No Date	All
Liberal	4559	7953	13969	13474	21685	26238	22900	21302	17641	16542	20613	27114	71408	285398
Left	3955	5847	9679	9133	17186	22018	20767	19301	15962	14892	18494	22193	24043	203470
Lean-left	604	2106	4290	4341	4499	4220	2133	2001	1679	1650	2119	4921	47365	81928
Neutral	4800	3100	5023	3584	6304	7558	7832	6045	7299	8756	11621	10218	7535	89675
Conservative	3878	4746	6624	6666	7259	7922	8438	8715	10867	10224	14797	27326	28263	145725
Lean-right	2392	2521	4083	3203	3617	3659	2715	3814	5259	4434	7583	15382	16298	74960
Right	1486	2225	2541	3463	3642	4263	5723	4901	5608	5790	7214	11944	11965	70765
Total	13237	15799	25616	23724	35248	41718	39170	36062	35807	35522	47031	64658	107206	520798

Table 1: Number of news articles per year for each orientation in our corpus (liberal, neutral, conservative) and their for sub-groups (e.g., left). The total number of articles (*All*) includes those for which *no date* could be extracted.

Preprocessing To filter out non-English articles, we classified the language of each text automatically using the *langdetect* library[4]. In contrast, we intentionally did not filter news categories (e.g., keeping only news articles about politics), in order to avoid selection bias. Furthermore, the different embedding algorithms require varying preprocessing steps. For word2vec, sentence splitting is required. In order to train the FRAGE model, we tokenized the data and replaced ultra-rare tokens with "<unk>", since the model expects the preprocessing of the WikiText-2 corpus (Merity et al., 2016). To do so, we used the *huggingface* tokenizer[5] and ended up with a vocabulary of around 39k tokens.

Statistics In total, we collected 520,798 news articles from 47 different outlets, 19 of which are liberal, 10 neutral, and 18 conservative. Table 1 reports detailed dataset statistics, showing that the number of articles is increasing over time, more or less monotonously. For about 20% of all articles (107,206), no publication date could be extracted.

5 Experiments

We now describe our experiments to evaluate embedding algorithms regarding their capabilities to accurately represent social bias in text corpora. To do so, we assess an algorithm's ability to generate a meaningful embedding space and to avoid the issues detailed in Section 3 arising from sparse data.

In particular, we systematically train models on all news articles with either political bias from our corpus (Section 4), once with each of the four word embedding algorithms from Section 3. To increase the data available for each bias, we aggregate news articles for lean-left and left as *liberal* as well as for lean-right and right outlets as *conservative*.

[4]https://github.com/Mimino666/langdetect

[5]https://github.com/huggingface/tokenizers

	WordSim353			MEN		
Algorithm	Liberal	Neutr.	Cons.	Liberal	Neutr.	Cons.
Static	–0.02	0.05	0.07	0.04	–0.01	–0.03
FrecAgn	0.57	0.56	0.55	0.55	0.46	0.51
Fine-Tuned	0.25	0.48	0.30	0.34	0.46	0.30
Decontext	**0.64**	**0.62**	**0.65**	**0.77**	**0.74**	**0.76**
BERT	0.25			0.21		

Table 2: Spearman's ρ of the word embedding similarity evaluation on the two tests, *WordSim353* and *MEN*. The embedding models were trained using the evaluated algorithms on *liberal*, *neutral* or *conservative* articles. Bold values indicate the best score in each column. For comparison, the values of pre-trained BERT are shown.

5.1 Word Similarity Tests

To better understand the models' quality, we first evaluate their performance on word-similarity tests.

Table 2 indicates that all proposed algorithms produce more meaningful embedding spaces compared to the Static models. The scores of the latter are close to 0.00, suggesting little to no correlation with the actual word similarities. A potential reason for the low scores is the limited training data, as discussed in Section 3.2, which may not be large enough to train high-quality models from scratch. The Decontext models that are pretrained on a larger dataset, on the other hand, achieve the highest scores for all data subsets on both tests (ranging from 0.62 to 0.77), also notably outperforming the underlying BERT model. The fine-tuning process of Fine-Tuned only marginally improves upon the base model. Considering that the liberal and conservative data subsets are notably larger than the neutral subset, it also seems that more data hurts the Fine-Tuned performance. This might be an issue of over-fitting to the fine-tuning data, decreasing the applicability of the resulting embeddings for the general similarity task. Further,

	Gender				Ethnicity				Religion			
Algorithm	**Liberal**	**Neutr.**	**Cons.**	**Δ**	**Liberal**	**Neutr.**	**Cons.**	**Δ**	**Liberal**	**Neutr.**	**Cons.**	**Δ**
Static	–0.151	–0.169	0.230	0.381	0.060	–0.061	0.098	0.038	0.301	–0.002	–0.298	–0.600
FrecAgn	0.632	0.611	0.763	0.131	0.555	0.680	0.658	**0.103**	1.166	0.795	1.181	0.015
Fine-Tuned	0.275	0.036	0.671	**0.396**	0.600	0.659	0.419	–0.182	0.873	1.235	0.442	–0.431
Decontext	0.334	0.409	0.370	0.036	0.419	0.422	0.429	0.010	0.479	0.486	0.519	**0.040**
BERT	0.098				1.234				0.621			

Table 3: WEAT values of the models trained with each evaluated algorithm for the three types of social bias. Δ denotes the difference between the values of the models trained on *conservative* and *liberal* articles respectively; the highest Δ for each bias type is marked bold. For reference, the WEAT values of pre-trained BERT are shown.

the "unnatural" input used to generate Fine-Tuned models, compared to the averaging strategy of the Decontext models, potentially impacts the embedding quality (Bommasani et al., 2020).

These results suggest that, while the size of the training corpus does have an impact on the quality of the word embeddings, it is not the only contributing factor. For example, comparing the results in Table 2 across algorithms for the same dataset, the choice of the algorithm seems to be important as well. That said, some algorithms do seem to benefit from the additional data. While the models trained on the liberal data perform slightly better on MEN tests compared to the other two models trained on smaller data, the benefit seems to be mostly negligible considering the increase in data needed (the liberal dataset contains nearly twice as many articles compared to the conservative dataset) and the additional training time. Furthermore, it is unclear, if this performance difference might partially also due to the selection of tested words in the respective word similarity tests.

Considering consistency, the frequency-agnostic and the decontextualized model appear most stable across all tests and data subsets. As a result, the models are also more comparable in the WEAT evaluation across data subsets, as the quality of the embedding models seems to be less dependent on the corpus size and content.

Overall, the suggested algorithms seem to improve the quality of the embedding space and abstract reasonably from the corpus size. For Decontext and Fine-Tuned, OOV tokens are less of a problem, as they train on sub-word tokens. The impact of fewer OOV tokens seems small in Table 3 than previously assumed. The performance of the FrecAgn model does not vary notably from the two models trained on sub-word tokens. Less OOV tokens should, however, result in more accurate social bias evaluations as more word embeddings exit, from which associations can be measured. As noted before, this can be more important when testing smaller datasets, as done in Section 5.3.

To analyze the representation degeneration, we repeated the evaluation with token pairs for which at least one was among the 100 least used tokens of the respective data subset. In general, results were similar to those in Table 2, indicating that Decontext and FrecAgn also perform well with rare tokens. While the results seem convincing, they must be interpreted with care. The similarity evaluations test word embedding models for general words and meaning rather than for social biases. Furthermore, the relation between these tests and the social bias measures is not fully clear.

5.2 Bias Representation Accuracy

As detailed in Section 3, each model is evaluated for social bias using WEAT. Following Caliskan et al. (2017), positive values indicate potential biases towards women compared to men (*Gender*), African-American compared to European-American names (*Ethnicity*), and Islam compared to Christianity (*Religion*). Based on our literature review presented in Section 2, we expect the liberal models to be biased against men, European-American names and Christianity, which should be reflected in positive WEAT values. Accordingly, we expect the opposite for the conservative models, and the neutral models should receive WEAT values located between the others.

Table 3 shows the results. The Δ columns indicate the difference in WEAT values between the models trained on conservative and on liberal news articles. It is a rough measure of an algorithm's accuracy in encoding social bias. The closer Δ is to the maximum $(2 - (-2) = 4)$, the better the models represent the expected social bias detailed in Section 3. Our discussion relies on this relative measure, as the exact WEAT value of the

data subsets is unknown. We chose not to use absolute values, as a negative Δ highlights cases that contradict our initial expectations, providing additional information on the quality of the word embedding models and applied measures. Since Fine-Tuned and Decontext are based on BERT, we report BERT's WEAT values for reference.

For all three evaluated bias types, at least one of the suggested algorithms receives a better accuracy than Static. While the fine-tuned models achieve the highest Δ in the gender bias evaluation, FrecAgn performs closest to expectation in the ethnicity bias evaluation. For the religion bias evaluation, Decontext shows the highest Δ, even though the absolute differences are comparatively small. It is noteworthy that models trained with Static achieve the second-best accuracy for the gender and ethnicity bias evaluations.

In general, the WEAT values for liberal and conservative models are less divergent than expected. Also, Δ is consistently close to 0 for Decontext and FrecAgn. When comparing the models for a single data subset (e.g., for liberal outlets only), the WEAT value strongly depends on the applied algorithm. The variance for the same data subset across all evaluations is in all cases above 0.7, with an average of 1.005. This is an intriguing finding, indicating that the choice of a particular algorithm is an important parameter when interpreting WEAT results, making exact WEAT values less meaningful and relative comparisons to a reference necessary.

A further interesting result is the fact that the Fine-Tuned and Decontext models have lower WEAT values than the BERT model they are based on. We hypothesize that this is due to our data being less biased, which changed the word associations during the fine-tuning and decontextualization. With the analysis at hand, however, this phenomenon can not be explained conclusively. While there does not seem to be one "best" algorithm for evaluating social bias according to Table 3, the combination of data, algorithm, and bias type seem to matter for the final result. A potential explanation is that the social bias present in the data is not as we hypothesized in Section 2, and the political bias does not correlate with social bias to the expected degree. Neither psychology literature nor our manual inspection of samples of the corpus make this seem likely, though.

Stereotypes, ideas of society, and with that social bias rather may be expressed more implicitly (see the example in Figure 1), potentially drawing word list-based measures to quantify bias ineffective. Similarly, the word lists applied by the WEAT evaluations might not be fully applicable to the evaluated datasets, requiring adaptation to the given linguistic style (Chaloner and Maldonado, 2019). For example, while liberal media may use the term "immigrant" to describe people coming to the US from a different country, conservative media may rather use the term "alien" (Webson et al., 2020). If a word list only includes one of the terms, it cannot properly reflect the associations with the target group and thus the social bias in the data. We suspect that both issues might contribute to the negative Δ values presented in Table 3. In this regard, future work may investigate measures that do not rely on predefined word lists, but adapt to the corpus being evaluated.

5.3 Temporal Evaluation

In our final experiment, we evaluated the change of social bias over time for the three political bias subsets. This also allows for insights into how the presented algorithms work with even fewer data, similar to the analyses of Garg et al. (2018) and Rios et al. (2020). In particular, we trained one word embedding model for articles of each year from each political bias considered (liberal, neutral, and conservative). We excluded all 107,206 articles for which we could not extract dates automatically. Here, we only used the decontextualization algorithm, given that it produced meaningful embeddings across all data subsets above. This is an important property for this evaluation, as the year-based sub-corpora are comparatively small. We evaluate the models for social bias using WEAT.

Figure 2 plots the results for each type of bias. While gender bias doesn't change notably, ethnicity and religious bias increase over the 12 years. We don't attribute this is to the amount of data, as the fluctuations of the neutral model happen mostly during years where the number of articles is similar to the conservative bias (see Table 1). Similar to the evaluation of the full models, we find that the relative social bias levels do represent the expected results to a certain degree. The liberal model generally shows lower WEAT values in the gender and religious evaluation compared to the conservative model. For the ethnicity evaluation, the liberal and conservative models are less distinctive though and show a very similar trend.

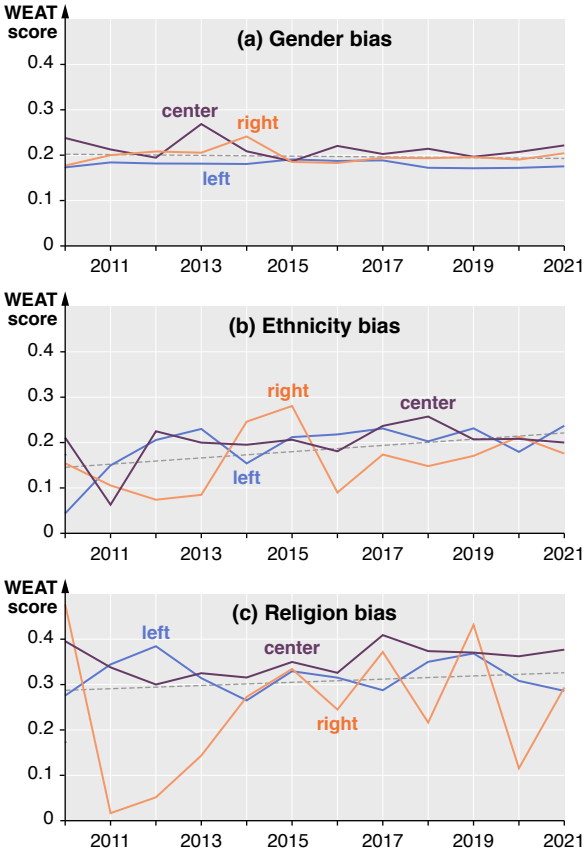

Figure 2: Plots of the development of the WEAT scores of the word embedding models for each bias type over time. Each model was trained on data subsets for each pair of year and political orientation. Gender bias slightly reduces over time, while ethnicity bias and religious increase (dashed regression lines).

Similar to the analysis of the full models, the small differences in WEAT values, compared to the full WEAT scale, might indicate that the absolute WEAT numbers are less meaningful and only work in relative comparisons.

6 Conclusion

In this paper, we have compared word embedding algorithms for the task of evaluating text corpora for social bias. To this end, we have introduced a US online news corpus that covers three political bias directions at five levels. Our literature review has motivated that specific political bias coincides with social bias with respect to gender, ethnicity, and religion. We have taken advantage of this property to train three word embedding algorithms and evaluate them for social bias using WEAT. Lastly, we present an example application, analyzing the development of social bias in news articles over a 12 year period.

We find that the particularly frequency-agnostic and decontextualized embedding spaces are more meaningful and encode the social bias more accurately than word2vec. They fail, however, to do so consistently for all bias types. While the respective algorithms should be more reliable, especially when evaluating sparse datasets, the exact WEAT results should be considered with care. The values do not seem to quantify social bias in the same way for all embedding algorithms. Future research should investigate the relation between WEAT values of an algorithm and the encoded bias.

Our findings give insights into the role of word embedding algorithms within the social bias evaluation of texts, and they demonstrate what type of embedding models work even in sparse data scenarios. Thereby, we contribute to understanding social bias in texts and NLP applications in general.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. Furthermore, we thank the team behind https://allsides.com for providing us with data files of their media bias ratings.

Limitations

One limitation of our evaluation is the distantly supervised approach used to label articles for political bias based on the outlet it was published by. We recognize that not all articles of an outlet are necessarily politically biased in the same way and to the same degree. Similarly, the political bias of an outlet could have changed over the evaluated period. A more refined approach could label articles based on their content, rather than the publishing outlet. Similar can be said for the social bias labels. Ultimately, it is not guaranteed that the social bias present in the analyzed 500k news articles statistically matches knowledge from psychology literature. Under the premise that literature is right, however, we are convinced that our inference from political to social bias to be sound, even if may not apply to the same extent to all articles.

We also acknowledge that we did not account for the completeness of the word lists used in the WEAT evaluations, which might therefore suffer from selection bias, hence not comprehensively representing the target groups. As the WEAT values depend on the contents of the word lists, the presented values might therefore not be fully accurate. A potential improvement to account for representa-

tion issues is to adapt the word lists to the language of each data subset, since outlets might use different vocabularies to describe the same groups.

Lastly, we were only able to evaluate a limited number of word embedding algorithms that account for token frequency issues. Potential alternatives include KAFE (Ashfaq et al., 2022), which relies on a knowledge graph to improve token representations, and AGG (Yu et al., 2022), for which the code was not available at the time of conducting the experiments. Similarly, we chose to fine-tune our BERT model for four epochs in all cases to obtain a comparable setting. Other choices might yield varying results.

Ethical Statement

We generate word embedding models for encoding social bias, as we train explicitly on texts that we expect to be biased. The models might therefore also contain more bias than other pre-trained models. They were, however, solely trained for the purpose of analyzing the training data. Due to the nature of the corpus and the comparatively sparse training data, we believe that the resulting models are not very applicable to other tasks.

We also note that, as already mentioned in the limitations, the word lists that we used in the WEAT evaluation are not complete. They might therefore not represent the social groups to a satisfying degree for real-world applications.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Awais Ashfaq, Markus Lingman, and Slawomir Nowaczyk. 2022. KAFE: Knowledge and Frequency Adapted Embeddings. In *Machine Learning, Optimization, and Data Science*, pages 132–146, Cham. Springer International Publishing.

Pablo Badilla, Felipe Bravo-Marquez, and Jorge Pérez. 2020. WEFE: The Word Embeddings Fairness Evaluation Framework. volume 1, pages 430–436.

Nichole M. Bauer. 2015. Who stereotypes female candidates? Identifying individual differences in feminine stereotype reliance. *Politics, Groups, and Identities*, 3(1):94–110.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.

E. Bruni, N. K. Tran, and M. Baroni. 2014. Multimodal Distributional Semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

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

Kaytlin Chaloner and Alfredo Maldonado. 2019. Measuring Gender Bias in Word Embeddings across Domains and Discovering New Gender Bias Word Categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32, Florence, Italy. Association for Computational Linguistics.

Wei-Fan Chen, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. 2021. Controlled neural sentence-level reframing of news articles. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2683–2693, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wei-Fan Chen, Khalid Al Khatib, Henning Wachsmuth, and Benno Stein. 2020. Analyzing Political Bias and Unfairness in News Articles at Different Levels of Granularity. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 149–154, Online. Association for Computational Linguistics.

Antonio Chirumbolo, Luigi Leone, and Marta Desimoni. 2016. The interpersonal roots of politics: Social value orientation, socio-political attitudes and prejudice. *Personality and Individual Differences*, 91:144–153.

Scott Clifford. 2020. Compassionate Democrats and Tough Republicans: How Ideology Shapes Partisan Stereotypes. *Political Behavior*, 42(4):1269–1293.

Sunipa Dev and Jeff Phillips. 2019. Attenuating Bias in Word vectors. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 879–887.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

Deep Bidirectional Transformers for Language Understanding. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David Domke, Kelley McCoy, and Marcos Torres. 1999. News Media, Racial Perceptions, and Political Cognition. *Communication Research*, 26(5):570–607.

John Duckitt and Chris G. Sibley. 2010. Personality, Ideology, Prejudice, and Politics: A Dual-Process Motivational Model. *Journal of Personality*, 78(6):1861–1894.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding Undesirable Word Embedding Associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems With Evaluation of Word Embeddings Using Word Similarity Tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 30–35, Berlin, Germany. Association for Computational Linguistics.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: the concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, WWW '01, pages 406–414, New York, NY, USA. Association for Computing Machinery.

Susan T. Fiske. 1998. Stereotyping, prejudice, and discrimination. In *The handbook of social psychology*, 4 edition, volume 1-2, pages 357–411. McGraw-Hill, New York, NY, US.

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):E3635–E3644.

Chengyue Gong, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2018. FRAGE: Frequency-Agnostic Word Representation. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. 2017. news-please: A Generic News Crawler and Extractor. In *Everything Changes, Everything Stays the Same? Understanding Information Spaces. Proceedings of the 15th International Symposium of Information Science (ISI 2017)*, pages 218–223. Humboldt-Universität zu Berlin.

Danny Hayes. 2011. When Gender and Party Collide: Stereotyping in Candidate Trait Attribution. *Politics & Gender*, 7(2):133–165.

Kenneth Joseph and Jonathan Morgan. 2020. When do Word Embeddings Accurately Reflect Surveys on our Beliefs About People? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4392–4415, Online. Association for Computational Linguistics.

Rafael-Michael Karampatsis, Hlib Babii, Romain Robbes, Charles Sutton, and Andrea Janes. 2020. Big Code != Big Vocabulary: Open-Vocabulary Models for Source Code. In *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)*, pages 1073–1085.

Linda Jean Kenix and Reza Jarvandi. 2019. The role of ideology in the international mainstream news media framing of refugees: A comparison between conservative and liberal newspapers in United States, United Kingdom and Australia. *Journal of Applied Journalism & Media Studies*, 8(3):349–365.

Markus Knoche, Radomir Popović, Florian Lemmerich, and Markus Strohmaier. 2019. Identifying Biases in Politically Biased Wikis through Word Embeddings. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, HT '19, pages 253–257, New York, NY, USA. Association for Computing Machinery.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in 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)*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer Sentinel Mixture Models. *arXiv:1609.07843 [cs]*. ArXiv: 1609.07843.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations*.

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)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Navid Rekabsaz, Robert West, James Henderson, and Allan Hanbury. 2021. Measuring Societal Biases from Text Corpora with Smoothed First-Order Co-occurrence. *Proceedings of the International AAAI Conference on Web and Social Media*, 15:549–560.

Anthony Rios, Reenam Joshi, and Hejin Shin. 2020. Quantifying 60 Years of Gender Bias in Biomedical Research with Word Embeddings. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 1–13, Online. Association for Computational Linguistics.

Sarah Schröder, Alexander Schulz, Philip Kenneweg, Robert Feldhans, Fabian Hinder, and Barbara Hammer. 2021. Evaluating Metrics for Bias in Word Embeddings. *arXiv:2111.07864 [cs]*.

Norbert Schwarz and Madeline Jalbert. 2020. When (Fake) News Feels True: Intuitions of truth and the acceptance and correction of misinformation. In *The Psychology of Fake News*, 1 edition, pages 73–89. Routledge.

Ellen Seiter. 1986. Stereotypes and the Media: A Reevaluation. *Journal of Communication*, 36(2):14–26.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, et al. 2021. An autonomous debating system. *Nature*, 591(7850):379–384.

Timo Spinde, Lada Rudnitckaia, Felix Hamborg, and Bela Gipp. 2021. Identification of Biased Terms in News Articles by Comparison of Outlet-Specific Word Embeddings. In *Diversity, Divergence, Dialogue*, Lecture Notes in Computer Science, pages 215–224, Cham. Springer International Publishing.

Maximilian Spliethöver and Henning Wachsmuth. 2020. Argument from old man's view: Assessing social bias in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 76–87, Online. Association for Computational Linguistics.

Maximilian Spliethöver and Henning Wachsmuth. 2021. Bias silhouette analysis: Towards assessing the quality of bias metrics for word embedding models. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 552–559. ijcai.org.

Chris Sweeney and Maryam Najafian. 2019. A Transparent Framework for Evaluating Unintended Demographic Bias in Word Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1662–1667, Florence, Italy. Association for Computational Linguistics.

Karolina Sylwester and Matthew Purver. 2015. Twitter Language Use Reflects Psychological Differences between Democrats and Republicans. *PLOS ONE*, 10(9):e0137422.

Nicholas A. Valentino and David O. Sears. 2005. Old Times There Are Not Forgotten: Race and Partisan Realignment in the Contemporary South. *American Journal of Political Science*, 49(3):672–688.

Albert Webson, Zhizhong Chen, Carsten Eickhoff, and Ellie Pavlick. 2020. Are "Undocumented Workers" the Same as "Illegal Aliens"? Disentangling Denotation and Connotation in Vector Spaces. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4090–4105, Online. Association for Computational Linguistics.

Russell J. Webster, Mason D. Burns, Margot Pickering, and Donald A. Saucier. 2014. The Suppression and Justification of Prejudice as A Function of Political Orientation. *European Journal of Personality*, 28(1):44–59.

Ruth Wodak. 2008. The contribution of critical linguistics to the analysis of discriminatory prejudices and stereotypes in the language of politics. In *Handbook of Communication in the Public Sphere*, volume 4, pages 291–316. De Gruyter Mouton.

Robert Wolfe and Aylin Caliskan. 2021. Low Frequency Names Exhibit Bias and Overfitting in Contextualizing Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 518–532, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sangwon Yu, Jongyoon Song, Heeseung Kim, Seongmin Lee, Woo-Jong Ryu, and Sungroh Yoon. 2022. Rare Tokens Degenerate All Tokens: Improving Neural Text Generation via Adaptive Gradient Gating for Rare Token Embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29–45, Dublin, Ireland. Association for Computational Linguistics.

A Model training implementation details

Static We train static embedding models with the gensim implementation of the word2vec algorithm and trained them using the skip-gram method with a window size of five for five epochs.[6] We stick to the commonly used vector size of 300 dimensions.

Frequency-Agnostic To train frequency-agnostic models with the FRAGE algorithm, we used the AWD-LSTM implementation published by Gong et al. (2018). For efficiency reasons, we decrease the number of epochs from 4000 to 500 and increased the batch size from 80 to 600.

[6] https://github.com/RaRe-Technologies/gensim

Fine-Tuned For the fine-tuned language models, we chose uncased BERT as starting point. We fine-tune the model for each political bias for four epochs with a standard masked language modeling objective using the Transformers library[7]. We subsequently extract the embeddings using the flair library (Akbik et al., 2019).

Decontextualized To generate decontextualized embeddings, we again chose uncased BERT as starting point and the flair library for contextualization. For each token of interest, we collect the sentences it occurs in within the context datasets, generate contextualized embeddings for each of the sentences, and average them, as suggested by Bommasani et al. (2020).

[7] https://github.com/huggingface/transformers