Disentangling Document Topic and Author Gender in Multiple Languages: Lessons for Adversarial Debiasing

Erenay Dayanik and **Sebastian Padó**
IMS, University of Stuttgart
Stuttgart, Germany
{erenay.dayanik,sebastian.pado}@ims.uni-stuttgart.de

Abstract

Text classification is a central tool in NLP, including social media analysis. However, when the target classes are strongly correlated with other textual attributes, text classification models can pick up "wrong" features, leading to bad generalization and biases. In social media analysis, this problem surfaces for demographic user classes such as language, topic, or gender, which influence how an author writes a text to a substantial extent. Adversarial training has been claimed to mitigate this problem, but a thorough evaluation is missing.

In this paper, we experiment with text classification of the correlated attributes of *document topic* and *author gender*, using a novel multilingual parallel corpus of TED talk transcripts. Our findings are: (a) individual classifiers for topic and author gender are indeed biased; (b) debiasing with adversarial training works for topic, but breaks down for author gender; (c) gender debiasing results differ across languages. We interpret the result in terms of *feature space overlap*, highlighting the role of *linguistic surface realization* of the target classes.

1 Introduction

Natural language processing, and machine learning more generally, has recently received a significant deal of criticism because of the frequent presence of *bias* in the predictions, where we define bias as a systematic difference in system performance on one set of instances compared to another. Such biases have been identified in NLP tasks such as word representation (Bolukbasi et al., 2016), textual inference (Rudinger et al., 2017), coreference resolution (Zhao et al., 2018), text classification (Dixon et al., 2018) and emotion intensity prediction (Kiritchenko and Mohammad, 2018).

In text classification tasks, a principal source of such biases are demographic attributes of authors, such as gender, age, or race[1]. The reason is that these attributes shape speakers' language use substantially (Hovy, 2015). NLP models are not only able in principle to pick up such cues, as studies on modeling demographic attributes show (Koppel et al., 2004), but they actually have a motivation to do so whenever some demographic attribute is strongly *correlated* with the model's classification target and therefore supports its recognition. As an example, in social psychology, Gross et al. (1997) report that elderly people experience and express their emotions less intensely than younger people. Therefore, in a corpus of emotional expressions across age groups, it is reasonable for a model that predicts emotion intensity to look out for linguistic cues regarding author age, *even if these cues are not really related to emotion intensity per se*, such as typical markers of youth language ("rad", "fam", "FTW", etc.).

This focus is arguably problematic, though, since it can give rise to a form of *age bias* – namely, overestimating emotion intensity for documents exhibiting youth language. More generally, the bias-inducing role of demographic attributes is dangerous for studies that use texts from a multitude of authors – often gathered from social media – to draw inferences about the authors (Sobkowicz et al., 2012; Cheng et al., 2015). In such studies, demographic biases can lead to erroneous causal attributions (as our case will illustrate).

To counteract the presence of biases in NLP, researchers have devised *debiasing* methods. Due to its general applicability and high effectiveness, adversarial debiasing has become one of the most widely used methods for bias mitigation (Elazar and Goldberg, 2018; Zhang et al., 2018; Arduini et al., 2020). Unfortunately, these advances are not accompanied by an analysis of the prerequisites

[1] A subset of these has specific legal protection in many jurisdictions under the name of *sensitive* or *protected* attributes.

Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 50–61

that need to be satisfied for adversarial training to perform successfully. It has been established empirically is that adversarial training works well for many cases in NLP; nevertheless, we demonstrate that there are relatively simple setups where it can fail. We analyze what factors contribute to the failure.

Concretely, we consider the correlated attributes of document topic (scientific / non-scientific) and author gender on a self-collected multilingual corpus of TED talk transcripts in French, German, Spanish, and Turkish. This setup enables us to observe the interplay between linguistic properties and adversarial debiasing.

Our investigation proceeds in three steps. First, we train independent classifiers for each attribute and evaluate them with regard to overall performance and with regard to the bias they exhibit. In the second step, we apply adversarial debiasing to the predicting of each attribute with respect to the other, and re-evaluate the debiased models. Finally, in the third step we discuss the differences observed in the previous step: (a), both document topic and author gender can be classified reasonably well by independent classifiers, but exhibit considerable bias; (b), author gender bias in topic classifiers can be reduced by adversarial training; however, adversarial debiasing in the opposite direction fails completely; (c) this effect is true for all languages except French. Our interpretation is that the failure of adversarial debiasing is due to the fact that feature space for author gender is *subsumed* the topic feature space for all languages except French, where gender is expressed overtly by morphological cues that can be picked up by the model.

2 Related Work

Regarding bias analysis at the representation level, the most important source of bias is arguably formed by corpus-derived embeddings which are used by virtually all current NLP systems. There have been several efforts to investigate the amount of bias within monolingual (Bolukbasi et al., 2016; Caliskan et al., 2017; Garg et al., 2018; Swinger et al., 2019) and multilingual embeddings. (Lauscher and Glavaš, 2019; Zhao et al., 2020). Bias analysis at the system level has investigated a range of applications such as NER (Mehrabi et al., 2020), Machine Translation (Stanovsky et al., 2019), Natural Language Inference (Rudinger et al., 2017), Emotion Intensity Prediction (Kiritchenko and Mohammad, 2018), Coreference Resolution (Rudinger et al., 2018; Zhao et al., 2018) and Text Classification (De-Arteaga et al., 2019).

A number of strategies have been explored for bias mitigation. One common strategy for reducing the bias is to target the representational level again, that is, corpora (Hall Maudslay et al., 2019; Zhao et al., 2018) and word embeddings (Kaneko and Bollegala, 2019; Bolukbasi et al., 2016). Other methods target the model architecture in various ways. For example, Qian et al. (2019) introduce an additional term to be used in loss function of language generation model, seeking to reduce the gender bias exhibited by the model.

A more fundamental idea is to adopt adversarial training (Goodfellow et al., 2014) to other tasks. For instance, Ganin and Lempitsky (2015) adapted adversarial training to the task of domain adaptation by introducing Gradient Reversal Layer (GRL) which acts as an identity function during forward pass and reverses the gradient by multiplying it by a negative scalar during the backward pass. Elazar and Goldberg (2018) apply the idea to the removal of demographic bias; McHardy et al. (2019) remove publication source as a bias variable from a satire detection model. Li et al. (2018) reported that adversarial training with GRL layer can remove unintended bias from the representations of POS tagging and Sentiment analysis models while maintaining task performance. Zhang et al. (2018) show that adversarial training mitigates the bias in word embeddings while maintaining its performance on word analogies task. Finally, Arduini et al. (2020) demonstrate how adversarial learning can be used for debiasing knowledge graph embeddings.

3 Dataset

In order to conduct a study on the relationship between topic and author gender in multiple languages, we require a multilingual comparable corpus for which topic and gender information are available. The corpus should be as parallel as possible so that any differences in outcome across languages are not simply due to differences in the evaluation data. Among the available multilingual parallel data sets, arguably the two most prominent ones are WIT^3 and OPUS. WIT^3 (Cettolo et al., 2012) consists of lecture translations automatically crawled from the TED talks in a variety of languages and was used in the evaluation campaigns

IWSLT 2013 and 2014. OPUS (Tiedemann, 2012) is a collection of data from several sources which provides sentence alignments as well as linguistic markup (for some languages). Unfortunately, neither corpus provides topic or gender labels.

For this reason, we create a new multilingual parallel dataset with these annotations, based on TED talks (http://ted.com/talks). A TED talk is a presentation at the TED conference or one of its international partner events. TED talks are limited to a maximum length of 18 minutes and may be on any topic. TED talks are rehearsed talks and at least semi-formal, while still definitely belonging to the category of spoken language. In this regard, they are comparable to the widely used Europarl corpus (Koehn, 2005). The talks are divided according to the languages, topics and posted dates. All original talks are presented in English, but volunteers provide (and double check) translations into other languages. Authors are identified by name.

Checking for which languages the TED webpage provided substantial numbers of transcripts (as of February 2020) led us to select German (DE), Spanish (ES), French (FR) and Turkish (TR) as target languages. We crawled all 1518 TED talks for which transcripts in all four target languages were available. We conducted some preprocessing: we cleaned transcripts by removing extra line breaks, extra spaces, and punctuation marks. Inspired by the work in open-domain Question Answering (Yang et al., 2019), we then segmented the transcripts into a sequence of segments. Rather than using paragraphs or sentences as segments directly, we split articles into segments with the length of 60 words by sliding window as Wang et al. (2019) demonstrated that splitting articles into non-overlapping fixed-length segments leads to better results in Question Answering.

Finally, we annotated the transcripts with topic and author gender information. For topic, we grouped transcripts into two classes according to the community-provided tags. The instances that have either Technology or Science tag were labeled as *SciTech* while the rest was labeled as *Other*. This grouping strategy led to a balanced dataset (53% Science, 47% Other). For author gender, we assume a binary gender classification (male/female) to be compatible with existing datasets (Verhoeven et al., 2016; Pardo et al., 2016). This should not be understood as a rejection of non-binary gender. We manually determine the author's gender information on the basis of gender indicating pronouns such as *he, she, his, her* that are used to refer to the authors in their biographies published in the authors' TED Talks profile or on other websites, keeping only clear cases. The majority gender is male (69%). Table 1 describes the final dataset.[2] The corpus has very similar properties across languages. The main exception is the lower number of words in Turkish which is due to the agglutinative nature of Turkish morphology. For instance, the English sentence with four words "I am at your house." is translated into a single word Turkish sentence "Evinizdeyim."

# TED Talks	1518			
Author Gender	1042 (Male) / 476 (Female)			
Talk Topic	704 (SciTech) / 814 (Other)			
	DE	ES	FR	TR
# Tokens/doc	2093	2110	2280	1632
# Sentences/doc	115	110	111	114

Table 1: Statistics of TED multilingual corpora.

Document Topic		SciTech	Other
Author Gender	Male	524	518
	Female	180	296

Table 2: Topic–gender correlation: Number of documents in TED corpus for each combination

4 Experimental Design

Table 2 shows a correlation matrix for the two attributes of topic and author gender in our TED corpus. Indeed, the corpus shows a clear correlation between the two: while male authors are represented about equally in TED for scientific-technological topics and other topics, female authors are underrepresented for scientific-technological topics. As motivated in the Introduction, this situation can lead to the model mistakenly picking up linguistic cues from one attribute to predict the other, leading to systematic biases.

We therefore believe that this corpus can serve as a reasonable case study for correlated document attributes. We proceed as follows:

Experiment 1: We learn individual neural models

[2]Code and data are available at http://www.ims.uni-stuttgart.de/data/ted_wassa21. This includes the documents we based our gender determination on, along with the list of gendered pronouns we used.

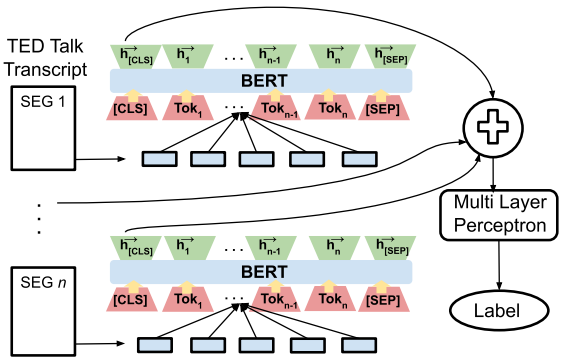


Figure 1: Visualization of classification architecture for topic and author gender

for topic and gender classification. We expect, for each attribute, that predictions are biased regarding the other attribute.

Experiment 2: We debias these models by adversarial training. We expect the models to focus better on features that are predictive of the individual attributes, and to show less bias.

We follow Li et al. (2018) and Zhao et al. (2020) by measuring the *amount of bias* in the models as the average difference in classification performance between documents aggregated by author gender (Male vs Female) or aggregated by topic (SciTech vs Other).

5 Experiment 1: Simple Classification and Bias Analysis

In our first experiment, we set up neural classification models for the two tasks of topic and gender classification individually and evaluate them for the presence of bias in their predictions.

5.1 Method

Figure 1 depicts the model architecture we use for both classification tasks. We use a neural text classifier based on the BERT Transformer (Devlin et al., 2019) with some adjustments. While transformers have shown good performance on many language tasks, most of them can only encode and generate representations for a fixed length token sequence – e.g., BERT implementations are often limited to 512 tokens per sequence. As the average token number per TED talk (cf. Table 1) is much larger. To address this limitation, we encode the input at the paragraph level (cf. Section 3). Specifically, we use the final hidden state corresponding to a special classification token, [CLS], as the representation for the corresponding paragraph. We then obtain the global context vector for the input by summing paragraph representations element-wise. Finally, the global representation of the input is fed through a Multi-Layer Perceptron to a Softmax layer. Our model can be understood as an adaptation of standard transformer classifiers to longer texts.

Language	Overall	By Gender		
		Male	Female	Bias
DE	81.2	80.0	84.0	4.0
ES	80.0	79.0	82.7	3.7
FR	81.5	80.2	83.7	3.5
TR	80.2	78.7	83.0	4.3
Majority BL	37.6	33.6	47.6	14.0

Table 3: F1 scores for topic classification (bottom line: majority baseline, identical for all languages)

5.2 Topic Classification

We first set up the model for topic classification. We approach it as a document-level binary classification task. The input to the model is the full transcript, and the model labels each transcript either as "SciTech" or "Other".

The topic classification results are shown in Table 3, using weighted F_1 score for evaluation. First, we compare the overall performance across languages. A majority baseline performs at 37.6% for all languages, due to the parallel design of the dataset. The neural topic classifiers do substantially better, all showing very similar results around 81% F-Score. Their similar performance may be expected from the parallel nature of the corpus, but it also provides support to our assumption that the texts and transformer models perform comparably across languages. When we break down these results by the other attribute we are interested, namely author gender (Male vs Female), we find that the prediction quality of the topic classifier is an average of 3.6 points lower for male than for female authors. In other words, the topic classifiers show a consistent gender bias across languages, presumably due to the higher-entropy (more equal) topic distribution for male authors (cf. Table 2). While this bias is lower than the bias of a majority baseline (which directly reflects the correlation between the two attributes), it is still substantial and arguably worth mitigating.

Language	Overall	By Topic		
		SciTech	Other	Bias
DE	70.8	69.0	75.0	6.0
ES	72.4	69.2	75.8	6.6
FR	82.4	82.0	83.0	1.0
TR	70.4	66.0	74.8	8.8
Majority BL	57.0	64.4	50.8	13.6

Table 4: F1 scores for gender classification (bottom line: majority baseline, identical for all languages)

DE	Genau hier wurde ich geboren und verbrachte die ersten sieben Jahre meines Lebens.
ES	Esta es la tierra en la que nací y pasé los primeros siete años de mi vida.
FR	Je suis **née** ici même, et j'y ai passé les sept premières années de ma vie.
TR	Doğduğum yer burası ve hayatımın ilk yedi yılını burada geçirdim.

Table 5: Example of inflectional gender marking in different languages (marking only present in French)

5.3 Author Gender Classification

We now address the opposite task, author gender classification, predicting the labels Male and Female, re-using the model architecture from before.

Table 4 summarizes the results. We see a pattern that differs substantially from topic classification, with much larger cross-lingual differences in performance. The results are again substantially above the 57% baseline. We obtain the best result for French (82%), and the worst for Turkish (70%), with a difference of 12% F-Score. This indicates that gender classification builds much more on language-specific information than topic classification. Arguably, for a word piece-based neural model like BERT, a primary source of evidence on author gender are linguistically marked expressions in the text where the author refers to themselves. Thus, prediction of the author gender should be easiest if a language has a frequent and unambiguous mechanism for *gender marking* (Corbett, 1991; Zmigrod et al., 2019). Table 5 shows a multilingual example where French marks gender inflectionally, while the other languages do not. This is indicative of the general case: The languages that we consider in our experiment provide gender marking to different degrees. At one extreme, French marks most adjectives and many nouns consistently for gender. In contrast, Spanish marks gender only for a subset of the lexicon, and morphologically inconsistently (Harris, 1991); German marks only (some) nouns, and marking is sometimes optional. At the other extreme, Turkish does not mark gender at all.

On this basis, we would expect French to perform best, and lower performance for the other three languages – exactly what we find. However, the performances for TR, DE, and ES are surprisingly close to one another, and substantially above the baseline: on the basis of what information in the texts do these classifiers base their predictions?

	DE	ES	FR	TR
SciTech/Female	75.0	75.8	83.0	74.4

Table 6: F1 scores for gender classification on SciTech talks with a female author.

A look at the size of the biases suggests an explanation: The gender classifiers for DE, ES, and TR make substantial use of *topic* cues, which enables them to proceed to some extent due to the correlations between topic and gender, but also lead to biases of 6–9% (highest for Turkish, consistent with the analysis above). In contrast, the French classifier is least biased, indicating that its text contains enough cues for 'proper' gender classification. We illustrate this in Table 6, where we report results on SciTech documents with female authors, that is, the smallest subcategory in our corpus. We find that the gender classifier for FR significantly outperforms the others, which provides additional evidence that the model relies less on the topic cues for gender classification.

Summary. For both tasks, we find that the classification performance shows a bias with respect to the other attribute. The two tasks differ with respect to the cross-lingual component, though: Topic classification works about equally well in all languages. In contrast, author gender classification only works properly in the one language that has consistent linguistic marking of gender, while there is evidence that the other languages fall back on topic features also for this task, which directly leads to biased predictions. These observations motivate experiments into how well these models respond to debiasing.

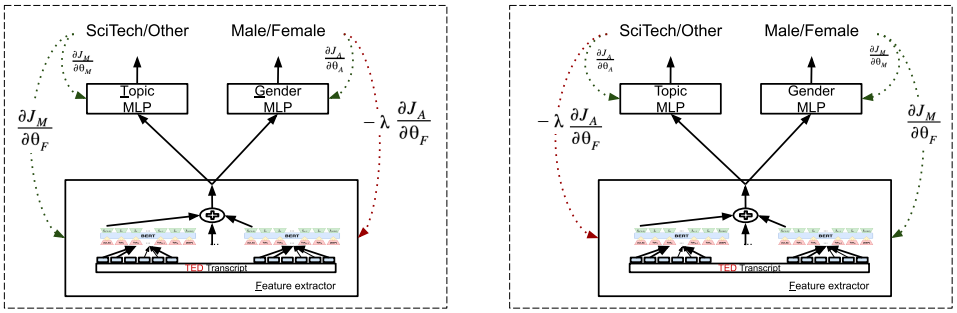


Figure 2: Visualization of debiasing by adversarial training. Left: Adversarial training of topic classifier on author gender, Right: Adversarial training of author gender classifier on topic.

6 Experiment 2: Adversarial Debiasing

Let P be some bias attribute (e.g., gender, race, age etc.) that we want our classifier to ignore while learning to solve another task T. Adversarial debiasing seeks to achieve this by constraining representations in a way so that representations do not rely on P in any substantial way. To this end, the model is trained to simultaneously predict the correct label for task T ("main component") and to prevent a jointly trained adversary ("adversarial component") from predicting P (McHardy et al., 2019). We define the loss functions of the main (J_M) and adversarial (J_A) components as follows:

$$J_A = -\mathbb{E}_{(x,y_A)\sim p_{\text{data}}} \log P_{\theta_A \cup \theta_F}(y_A, x) \quad (1)$$

$$J_M = -\mathbb{E}_{(x,y_M)\sim p_{\text{data}}} \log P_{\theta_M \cup \theta_F}(y_M, x) \quad (2)$$

where θ_A, θ_M are the parameters of adversarial and main components; y_A and y_M are the gold labels for main and adversary tasks. Note that the adversarial and main components share the same feature extractor (i.e., BERT) whose parameters (θ_F) are therefore updated by the gradients coming through the objective functions of both model parts. Let λ be the meta-parameter controlling the intensity of the adversarial training and η the learning rate. Then the following equations describe update rules for each component in the model:

$$\theta_M := \theta_M - \eta \frac{\partial J_M}{\partial \theta_M} \quad (3)$$

$$\theta_A := \theta_A - \eta \frac{\partial J_A}{\partial \theta_A} \quad (4)$$

$$\theta_F := \theta_F - \eta \left(\frac{\partial J_M}{\partial \theta_F} - \lambda \frac{\partial J_A}{\partial \theta_F} \right) \quad (5)$$

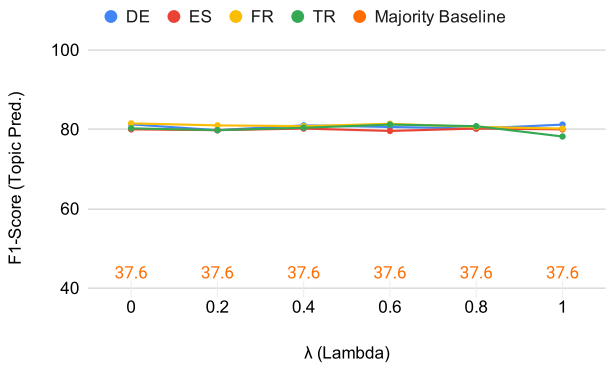
Our application of this training method is shown in Figure 2. We first debias the topic classifier by author gender (left-hand box); then we proceed to debias author gender classifier by topic (right-hand box). For example, to de-bias the topic classification, J_M is the topic loss and J_A the author gender loss; vice versa for author gender de-biasing.

6.1 Topic-Debiased Gender Classification

First, we debias topic classification to reduce the gender bias. The left-hand side of Figure 3 compares overall results across a range of values of λ between 0 (no adversarial training) and 1 (equal weight of main and adversarial loss). We find that, similar to Experiment 1, the results are essentially identical across languages. Furthermore, the choice of λ hardly matters in this interval: adversarial training does not have a major impact on topic classification. We report detailed results for λ=1 in the right-hand side of Figure 3. The small differences between the Overall results of the Original and Debiased models show that topic classification overall does not lose much by debiasing for gender.[3] The breakdown by gender shows that gender bias is substantially reduced overall. However, there are noticeable differences among languages.

For Spanish and German, we see no overall loss of performance in topic classification, and a substantial reduction in gender bias. For French and Turkish, in contrast, we see a decrease of about 1.5 points in topic classification. Gender bias is reduced for French but hardly for Turkish. This is a somewhat surprising result, given the typological differences between the two languages. Our explanation is that in French, as discussed above, many words are morphologically marked for gender. Due to the correlation between the two attributes, these can be re-used by the topic classifier, but when they are penalized through adversarial training, we see a mild decrease in topic classification accuracy.

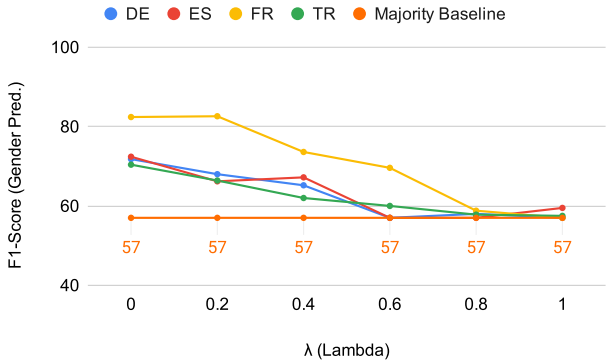
[3] See Appendix for performance on the adversarial task.



Topic classification ($\lambda = 1$)

	Original		Debiased			
	Overall	Bias	Overall	Male	Female	Bias
DE	81.2	4.0	81.2	80.8	81.2	**0.4**
ES	80.0	3.7	80.0	79.2	81.6	**2.4**
FR	81.5	3.5	80.2	79.4	81.4	**2.0**
TR	80.2	4.2	78.4	76.8	80.7	**3.9**

Figure 3: Results for topic classification with adversarial author gender training (F1 scores). Left: Overall results for different λ values. Right: Detail results for λ=1. Original: results from Experiment 1 (cf. Table 3). Lower bias for each language bolded.



Author gender classification ($\lambda = 0.2$)

	Original		Debiased			
	Overall	Bias	Overall	SciTech	Other	Bias
DE	70.8	**6.0**	68.0	63.8	72.2	8.4
ES	72.4	6.6	66.2	62.8	69.2	**6.4**
FR	82.4	1.0	82.6	82.6	82.6	**0.0**
TR	70.4	8.8	66.4	64.0	68.8	**4.4**

Figure 4: Results for author gender classification with adversarial topic training (F1 scores). Left: Overall results for different λ values. Right: Detail results for λ=0.2. Original: results from Experiment 1 (cf. Table 4). Lower bias for each language bolded.

In Turkish, as we have argued in Experiment 1, gender classification depends almost entirely on topic features since there is no linguistic marking of referent gender. Consequently, the adversarial training works against itself to an extent, resulting in a mildly worse topic classification but hardly any decrease in gender bias.

6.2 Gender-Debiased Topic Classification

Now we swap the main and adversarial tasks again, debiasing author gender classification with regard to topic. We use the same setup as in Experiment 1.

The results are shown in Figure 4. The left-hand side shows that varying λ has a substantial effect this time. If we set λ to a value close to 1 – a good choice for gender-debiased topic classification, as we have established in the previous subsection – this leads to a breakdown of the gender classification model. Performance for all languages drops to a F-Score of around 57, the level of the majority baseline (cf. Table 4). Apparently, debiasing author gender classification by adversarial training against topic breaks the author gender classifier for all but small values of λ.

As in the first experiment, we observe differences among languages: French stands out as the language for which the gender classification 'holds out' the longest for high values of λ. Its ultimate failure indicates that even for French, gender marking on its own is not strong enough to support the author gender identification task – or at least our models are not powerful enough to pick up on these cues. The other languages, which, as we have argued in Experiment 1, make substantial use of topic cues for gender classification, fail even earlier.

The right-hand side of Figure 4 reports detailed results for λ=0.2. In line with our analyses above, debiasing works for French but not for the other languages: We find clear decreases in performance (up to 6.2 points, for Spanish), and inconclusive

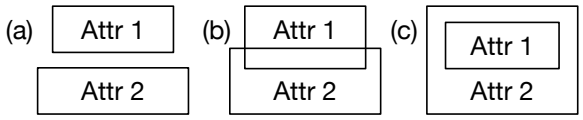


Figure 5: Three cases of latent feature space geometry for two attributes: (a) independent, (b) correlated, (c) subsumed

changes in bias (decrease for Turkish by 4.4 points, increase for German by 1.6 points). While the patterns for these languages are not straightforward to interpret, it seems safe to conclude that topic-debiasing author gender is a failure both with regard to model performance and reduction of bias.

7 Discussion

The results of our two experiments show an intriguing asymmetry between the two tasks of topic and author gender classification when debiased for the respective other attribute. Reducing author gender bias in topic classification with adversarial training proceeds as expected, is relatively robust to the choice of λ in the interval between 0 and 1, and shows a consistent pattern across languages which can be explained by the properties of the languages involved. In contrast, reducing topic bias in author gender classification relies heavily on λ, quickly deteriorating to baseline level for large values of λ, and does not consistently manage to reduce bias in any case. This asymmetry cannot be an artifact of model architecture or data alone, since we use the same model architecture on the same data.

Instead, we believe that these patterns result from an interaction between the representation learning of the model and the information that the model can draw from the data. They can be understood through the *latent feature space* of the final shared layer in our architecture below the two heads (cf. Figure 2), where each class can be characterized by a region of informative features.

Figure 5 shows Venn diagram-style depictions of the three possible cases for a pair of attributes. In the left-hand case, (a), there is no overlap between the latent features of the two attributes. That is, the two attributes are independent of one another, and so is learning. However, this is by definition the case without correlations among attributes that we do not consider. In the center case, (b), there is an intersection between the latent features of the two attributes. The classifiers' use of this overlap potentially creates biases, but adversarial training exactly punishes the use of this region of latent feature space. Thus, debiased classifiers can learn either attribute to the extent that the part of the feature space outside the intersection is still sufficiently informative. The right-hand case, (c), is the limit case when one of the two attributes does not have an independent standing, that is, the informative latent features of attribute 1 are completely contained in the informative feature space of attribute 2. This leads to biases in either classifier just as case (b), but also creates an asymmetry in the effect of adversarial debiasing: Attribute 2 can be debiased by simply 'cutting out' the informative space of attribute 1, but debiasing attribute 1 in the opposite manner results in an empty feature space for attribute 1, and we would expect the classifier to revert to baseline performance.

This set theoretic visualization is a major simplification of the latent feature space in neural models, where the three cases cannot apply categorially — they rather represent different points on a continuum. Nevertheless, the predictions of the subsumption case, (c), match our experimental results well: Assuming that author gender features are included in topic features, we would expect to find successful debiasing of the topic classifier, but breakdown of the debiased author gender classifier. This is exactly the pattern of results that we have observed.

Note that this analysis builds on the behavior of the features of the attributes in the training data, in particular in a representation learning approach like the one we have pursued. In other words, changes of the data – or differences within the data, such as between languages – are expected to influence the outcome. Again, this is what we see: French, due to its consistent morphological marking of gender, is closer to case (b), while the other languages are closer to case (c).

8 Conclusion

This paper was concerned with text classification for correlated attributes, which pose an important but often overlooked challenge to model fairness – in particular, as we have argued, in the case of demographic attributes.

We specifically analyzed the relationship between document topic and author gender. We established that topic classifiers exhibit gender bias and author gender classifiers show topic bias; that adversarial debiasing corrects gender bias in topic classification but breaks down in the opposite di-

rection; and that this effect varies by language.

Beyond the concrete study, our contribution is to draw attention to the general question of prerequisites for successful adversarial debiasing, which, to our knowledge, has not received much attention. Our results indicate that when the target attribute and the bias attribute are too strongly correlated – or, indeed, when the target attribute is subsumed by the bias attribute – adversarial debiasing fails: with a small weight on the bias component, no debiasing takes place; with a large weight, target attribute classification deteriorates to baseline level.

Furthermore, we find that the linguistic expression of the attributes matters greatly: the only language for which we achieved satisfactory results was French, due to the consistent morphological marking of gender which can be captured independently of topic (Zmigrod et al., 2019). This highlights the importance of understanding the differences between languages regarding how they encode content (Dubossarsky et al., 2019), and underscores the importance of cross-lingual methods.

In future work, we plan to develop a diagnostic to recognize potentially problematic constellations of correlated attributes and improve debiasing.

Acknowledgments

Partial funding was provided by Deutsche Forschungsgemeinschaft (DFG) through project MARDY within SPP RATIO. We would like to thank Roman Klinger, Gabriella Lapesa, Vivi Nastase and Michael Roth for valuable comments.

References

Mario Arduini, Lorenzo Noci, Federico Pirovano, Ce Zhang, Yash Raj Shrestha, and Bibek Paudel. 2020. Adversarial learning for debiasing knowledge graph embeddings. In *Proceedings of the 16th International Workshop on Mining and Learning with Graphs (MLG)*, San Diego, California.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29, pages 4349–4357.

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

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *Conference of European Association for Machine Translation*, pages 261–268, Trento, Italy.

Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial behavior in online discussion communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9.

Greville G. Corbett. 1991. *Gender*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

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.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Haim Dubossarsky, Arya D. McCarthy, Edoardo Maria Ponti, Ivan Vulić, Ekaterina Vylomova, Yevgeni Berzak, Ryan Cotterell, Manaal Faruqui, Anna Korhonen, and Roi Reichart, editors. 2019. *Proceedings of TyP-NLP: The First Workshop on Typology for Polyglot NLP*. Association for Computational Linguistics, Florence, Italy.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France. JMLR.org.

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.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, page 2672–2680, Cambridge, MA, USA. MIT Press.

James J Gross, Laura L Carstensen, Monisha Pasupathi, Jeanne Tsai, Carina Götestam Skorpen, and Angie YC Hsu. 1997. Emotion and aging: Experience, expression, and control. *Psychology and aging*, 12(4):590.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5267–5275, Hong Kong, China. Association for Computational Linguistics.

James W. Harris. 1991. The exponence of gender in spanish. *Linguistic Inquiry*, 22(1):27–62.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1641–1650, Florence, Italy. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2004. Automatically categorizing written texts by author gender. *Computing Reviews*, 45(1):43.

Anne Lauscher and Goran Glavaš. 2019. Are we consistently biased? multidimensional analysis of biases in distributional word vectors. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics*, pages 85–91, Minneapolis, Minnesota. Association for Computational Linguistics.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30, Melbourne, Australia. Association for Computational Linguistics.

Robert McHardy, Heike Adel, and Roman Klinger. 2019. Adversarial training for satire detection: Controlling for confounding variables. 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 660–665, Minneapolis, Minnesota. Association for Computational Linguistics.

Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, and Aram Galstyan. 2020. Man is to person as woman is to location: Measuring gender bias in named entity recognition. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, page 231–232, New York, NY, USA. Association for Computing Machinery.

Francisco Manuel Rangel Pardo, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. 2016. Overview of the 4th author profiling task at PAN 2016: Cross-genre evaluations. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016*, volume 1609 of *CEUR Workshop Proceedings*, pages 750–784. CEUR-WS.org.

Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. 2019. Reducing gender bias in word-level language models with a gender-equalizing loss function. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 223–228, Florence, Italy. Association for Computational Linguistics.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 8–14, New Orleans, LA.

Pawel Sobkowicz, Michael Kaschesky, and Guillaume Bouchard. 2012. Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. *Government Information Quarterly*, 29(4):470–479. Social Media in Government - Selections from the 12th Annual International Conference on Digital Government Research (dg.o2011).

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. 2019. What are the biases in my word embedding? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 305–311, New York, NY, USA. Association for Computing Machinery.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ben Verhoeven, Walter Daelemans, and Barbara Plank. 2016. TwiSty: A multilingual twitter stylometry corpus for gender and personality profiling. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1632–1637, Portorož, Slovenia. European Language Resources Association (ELRA).

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Multi-passage BERT: A globally normalized BERT model for open-domain question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5878–5882, Hong Kong, China. Association for Computational Linguistics.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with BERTserini. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 72–77, Minneapolis, Minnesota. Association for Computational Linguistics.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA. Association for Computing Machinery.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

Appendix

A Training of BERT-based document classifiers

In our experiments, for each language we consider we use a cased BERT variant that was trained specifically for the target language.[4] We use the Adam optimizer with learning rates of 5e-5, β_1 = 0.9, β_2 = 0.999, a batch size of 48, a gradient clip threshold of 1.0 and a dropout with p=0.5 on all layers. We train the model for 15 epochs. The Multi Layer Perceptron consists of a single hidden layer with 300 hidden units. We evaluate each classifier using weighted F1-Score which calculate metrics for each label, and find their average weighted by the number of true instances for each label. We repeat every experiment using 5 random train (80%) test (20%) splits and report average of these 5 experiments.

B Adversarial Debiasing: Performance on adversarial tasks

In addition to majority class classifier and non-adversarial model, we use a third baseline model to analyze how adversarial debiasing effects the model's performance on the adversary task. First, we train feature extractor along with the topic classifier head on the topic classification task. Next, we freeze the weights of the feature extractor and

[4] DE: https://deepset.ai/german-bert, ES: https://github.com/dccuchile/beto, FR: https://camembert-model.fr/, TR: https://github.com/dbmdz/berts

	Overall	
	Adv	Baseline
DE	39.8	67.0
ES	22.6	66.2
FR	14.0	67.6
TR	22.6	69.0

Table 7: Gender classification F-scores of gender-debiased topic classifier and baseline model. For main task evaluation, see Figure 3.

	Overall	
	Adv	Baseline
DE	34.0	74.1
ES	37.0	72.0
FR	34.6	78.2
TR	32.8	69.2

Table 8: Topic classification F-scores of topic-debiased gender classifier and baseline model. For main task evaluation, check Figure 4.

train the gender classifier on top of it. Table 7 summarizes the results for gender classification. Significant drop in gender classification performance indicates effectiveness of adversarial training.

As we swapped the main and adversarial tasks, we modify the baseline in the same way too. We start with training the feature extractor and gender classifier head on the topic classification task. Then,we freeze the feature extractor, remove the gender classifier and train the topic classifier on top of freezed feature extractor. Table 8 reports the results on topic classification. Similar to Table 7, adversarial debiasing drops the adversarial task (i.e. topic classification) performance significantly. However, as Figure 4 shows it leads to slight to moderate decrease in gender classification performance and inconclusive changes with regard to topic bias.