Natural Language Engineering (2023), **29**, pp. 449–482
doi:10.1017/S1351324922000122

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Gender bias in legal corpora and debiasing it

Nurullah Sevim[1,2], Furkan Şahinuç[1,3] and Aykut Koç[1,2,*]

[1]Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Turkey, [2]National Magnetic Resonance Research Center (UMRAM), Bilkent University, Ankara, Turkey and [3]ASELSAN Research Center, Ankara, Turkey
*Corresponding author. Email: aykut.koc@bilkent.edu.tr

(Received 9 February 2021; revised 24 February 2022; accepted 28 February 2022; first published online 30 March 2022)

Abstract

Word embeddings have become important building blocks that are used profoundly in natural language processing (NLP). Despite their several advantages, word embeddings can unintentionally accommodate some gender- and ethnicity-based biases that are present within the corpora they are trained on. Therefore, ethical concerns have been raised since word embeddings are extensively used in several high-level algorithms. Studying such biases and debiasing them have recently become an important research endeavor. Various studies have been conducted to measure the extent of bias that word embeddings capture and to eradicate them. Concurrently, as another subfield that has started to gain traction recently, the applications of NLP in the field of law have started to increase and develop rapidly. As law has a direct and utmost effect on people's lives, the issues of bias for NLP applications in legal domain are certainly important. However, to the best of our knowledge, bias issues have not yet been studied in the context of legal corpora. In this article, we approach the gender bias problem from the scope of legal text processing domain. Word embedding models that are trained on corpora composed by legal documents and legislation from different countries have been utilized to measure and eliminate gender bias in legal documents. Several methods have been employed to reveal the degree of gender bias and observe its variations over countries. Moreover, a debiasing method has been used to neutralize unwanted bias. The preservation of semantic coherence of the debiased vector space has also been demonstrated by using high-level tasks. Finally, overall results and their implications have been discussed in the scope of NLP in legal domain.

Keywords: Bias; NLP in law; Legal text processing; Law; Computational law

1. Introduction

1.1 Background

Word embeddings that map words to vectors in multidimensional semantic vector spaces are widely used as underlying word representations in various architectures to tackle natural language processing (NLP) tasks (Mikolov *et al*. 2013b; Peters *et al*. 2018; Devlin *et al*. 2019). The introduction of Word2Vec and GloVe models and later improvements on these models increased the popularity of word embeddings (Mikolov *et al*. 2013b; a; Pennington, Socher, and Manning 2014; Church 2017; Navigli and Martelli 2019). Latterly, the emergence of transformer-based contextualized language models, such as ELMo (Peters *et al*. 2018) and BERT (Devlin *et al*. 2019), has initiated another stage for word representations. Constructing semantic vector spaces is also important on its own for several computational linguistic studies, where word semantics are mathematically represented and utilized for a rich spectrum of purposes including identifying semantic change, gender and ethnic biases, and learning semantic hierarchies (Fu *et al*. 2014; Hamilton, Leskovec, and Jurafsky 2016; Garg *et al*. 2018).

With these advanced renovations, techniques for developing word embeddings have extended to obtain better results in various NLP tasks (Tanaka-Ishii 2007; Üstün and Can 2020; Tezcan, Hoste, and Macken 2020). Word embeddings are now being commonly used for several high-level tasks such as text classification (Lai *et al*. 2015; Joulin *et al*. 2017; Pittaras *et al*. 2020) and sentiment analysis (Tang *et al*. 2014). In conjunction with these, the well-developed word embedding models are also adopted for coreference resolution task (Clark and Manning 2016; Joshi *et al*. 2019).

As better word embeddings and advanced applications of them have been developing and emerging, it has also been shown that they are prone to show human-like biases, such as gender stereotypes and ethnic discrimination (Bolukbasi *et al*. 2016; Caliskan, Bryson, and Narayanan 2017; Garg *et al*. 2018; Manzini *et al*. 2019). As exemplified in the seminal work of Bolukbasi *et al*. (2016), an embedding model may perceive the word `programmer` more likely as a male occupation than a female one. As another example, the word `housekeeper` is linked with Hispanic ethnicity with larger probability than other ethnic origins. The existence of these unacceptable biases in word embeddings, which are at the heart of several NLP applications touching the daily lives of people, has of course evoked ethical concerns. The inherent bias in word embeddings causes several controversies for succeeding applications of them as they inherently contain inequality and unfairness. When, for example, a specific job is explored in search engines, the search engine may have a tendency to show people from a dominant subgroup of society, or to show people from a specific gender, at the top of the search results (De-Arteaga *et al*. 2019). Consequently, people from relatively less dominant subgroups will have difficulties to be recognized for a specific job, since they are underrepresented. This widens the opportunity gap in society and impedes to make accurate decisions (Perez 2019). Being one of the prominent unwanted biases, gender bias in word embeddings can also result in undesired impacts where the gender stereotypes in word embeddings influence high-level computational tasks. Zhao *et al*. (2018a), for example, showed that coreference resolution systems carry gender bias since word embeddings are backbone tools for the state-of-the-art coreference resolution systems. The system can successfully link `he` pronoun with `physician`, whereas the same system fails to make the same connection between `she` and `physician`. In many computational linguistics studies, word embeddings are also a prevalent mechanism that is used for numerous tasks such as investigating interpretability (Murphy, Talukdar, and Mitchell 2012; Faruqui *et al*. 2015; Senel *et al*. 2020). In these studies, possibly existing bias becomes more concerning and eliminating the bias from the word embeddings comes to be substantially important. Therefore, detecting and neutralizing biases have naturally been an important and necessary line of research which is currently in progress (Bolukbasi *et al*. 2016; Caliskan *et al*. 2017; Kiritchenko and Mohammad 2018; Zhang, Lemoine, and Mitchell 2018; Manzini *et al*. 2019; Prost, Thain, and Bolukbasi 2019; Kaneko and Bollegala 2019; Tan and Celis 2019; Zhao *et al*. 2019; Liang *et al*. 2020).

Concurrently, an important line of research has also been developing where NLP and machine learning tools are applied to process legal documents and to develop decision support systems for legal domain (Aleven 2003; Ruger *et al*. 2004; Martin *et al*. 2004; Evans *et al*. 2008; Ashley and Brüninghaus 2009; Katz, Bommarito, and Blackman 2017; Chalkidis and Kampas 2019; Dale 2019). Law, a field that relies mostly on written text, is naturally very open to utilization of NLP applications. Vastly increasing number of case files and the effort of dealing with these cases manually constitute a considerable issue for legal professionals. Consequently, exploiting automated tools has become a requirement to ease human burden, to improve legal services, and to reduce human-induced errors in numerous legal tasks.

The interactions between law and artificial intelligence (AI) fields have indeed a long history, beginning with ideas in 1970s, establishing an active community in 1987, and with further developments (Buchanan and Headrick 1970; Francesconi *et al*. 2010; Bench-Capon *et al*. 2012; Sartor and Rotolo 2013; Casanovas *et al*. 2013). Several NLP methods have been developed for processing legal documents (Ashley 1988; Hafner and Berman 2002; Aleven 2003; Aletras *et al*. 2016; Katz *et al*. 2017; Long *et al*. 2019; Dale 2019; Azarbonyad *et al*. 2021). By using word and n-gram

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

frequencies as features, traditional machine learning techniques such as random forests and support vector machines (SVMs) are exploited for high-level NLP tasks in legal domain (Aletras *et al.* 2016; Katz *et al.* 2017). Advanced machine learning techniques are also being applied in NLP such as long-short term memory networks (LSTMs) (Hochreiter and Schmidhuber 1997) combined with vector representations of words (Chalkidis, Androutsopoulos, and Michos 2018; Chalkidis and Kampas 2019). Combining these powerful machine learning algorithms with NLP techniques, quite high performances on many challenging downstream tasks can be achieved. Predicting decisions of courts (Martin *et al.* 2004; Şulea *et al.* 2017; Virtucio *et al.* 2018; Mumcuoğlu *et al.* 2021), classifying case documents (O'Neill *et al.* 2017), and information retrieval (Sangeetha *et al.* 2017) are some of these tasks that NLP techniques are trying to offer new decision support solutions. *Law2Vec*, which is a specialized word embedding model trained exclusively on legal corpora, has also been introduced (Chalkidis and Kampas 2019). These legal word embeddings offer specialized word representations for several high-level NLP applications in law. They also provide us a framework to computationally study the specialized semantic vector space constructed from legal texts for several other purposes including uncovering the intrinsic properties of these texts (Vo, Privault, and Guillot 2017). Very recently, legal NLP efforts have been amalgamated by Chalkidis *et al.* (2021) in LexGLUE by introducing several legal corpora and standardized benchmarks for the field.

1.2 Research objectives

Among the contemporary application fields, law is probably one of the most influential areas in touching upon lives. It is needless to point out the importance of law and the impacts of consequences of legal processes on people. Therefore, the extent of the aforementioned ethical concerns regarding bias and fairness issues further increases. Investigation of biases in legal text and developing methods to neutralize them are of importance to develop bias-free and fair NLP applications for law. Word embeddings are vital tools to study the biases in legal corpora and uncover possible inequalities in law systems, which may propagate deep into the legal procedures as well as NLP-based decision support applications. Thus, studying possibly existing bias in a legal corpus will give an opportunity to emancipate word embeddings from bias and to increase their fairness.

1.3 Challenges involved

In this article, we study bias in legal texts by systematically compiling legal corpora from various countries and measuring their bias by using legal word embeddings. Although the original Law2Vec embeddings are available in pre-trained format and the corpus it was trained on is specified, the corpus is not readily available in a stand-alone format for further processing. This prevents us to carry a systematic study starting from the training step and also from individually studying subcorpora belonging to different countries. For that reason, we have also compiled an extensive legal corpora and trained our own Law2Vec version on it. We also used original Law2Vec in our analysis for extra comparisons. In addition to adapting and comparing several bias measurement and debiasing methods for the legal domain, we also propose a novel bias measurement method that is specific to legal corpora.

We adopt five methods to evaluate the level of bias legal word embeddings contain. First one is inspired from the method introduced by Bolukbasi *et al.* (2016), which exploits cosine distances and projections of word vectors on predetermined gender vectors. We transform this method to a law specific metric by changing the characteristics of words whose degree of bias is measured. We also introduce a new bias evaluation metric called *CRIme Bias (CriBias)*. To do this, we compiled a word list of criminal acts, called CRIme List (CriList), to measure gender bias present in legal corpora regarding the tendency to commit crime. The objective of the proposed method is to reveal the perspective of legal texts about the relation between gender and criminal acts. The two other

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

baseline methods that we use have been introduced in Gonen and Goldberg (2019), where one of them is based on clustering the words in a target embedding model and the other uses k-nearest neighbors. In the final method, we utilize the Equity Evaluation Corpus (EEC) that is introduced by Kiritchenko and Mohammad (2018). Using EEC, we evaluate the bias of a model designed for emotion intensity regression task. To adopt for legal NLP, we fed the aforementioned model with legal word embeddings and observed bias scores. By using these methods, we achieve to measure the bias in different corpora that contain legal texts from various countries and compare the characteristics of legislation for each country. After obtaining the bias measurements of legal word embeddings, we present how to apply a debiasing procedure. We also measure bias levels after debiasing operations to quantify reduction in bias.

Although having unbiased embeddings is important, one also needs to be sure that the underlying semantic structure is not distorted during the debiasing procedure. After applying the debiasing method to the legal embedding models, we investigate the changes in the semantic space of each embedding model through observing the performances of embeddings in high-level tasks. The reason of this procedure is to be sure that the debiasing techniques do not distort the semantic utility of debiased embeddings. In doing so, our objective is not to increase performance for the high level tasks but only show that there is no degradation after debiasing. We make use of a Part of Speech (POS) tagging task for assessing the semantic utility of the embeddings. In addition, we implement a prediction task where court case decisions are to be predicted by deploying word embeddings.

1.4 Summary of contributions

Our contributions can be summarized as follows. To the best of our knowledge, our study is the first that considers the important issue of bias in the context of recently developing field of NLP in law. We compiled large legal corpora consisting of legislation and regulations from several countries. We developed a specialized bias evaluation method (CriBias) for legal context based on a readily available general bias evaluation approach by introducing a compiled word list of crime related words (CriList) and named this new method as CriBias. We also showed that the debiased legal domain specific word embeddings for high-level NLP applications that perform well can be trained. Finally, as contextualized word embeddings are also very recently being developed for the legal NLP, we provide discussions on extensions to the contextualized word embeddings domain.

The article is organized as follows. We discuss related work in Section 2 under two major titles, bias in NLP and NLP in law. Our materials such as corpora, word lists, legal word embeddings, and test datasets are all presented in Section 3. The methodology of our study and our proposed CriBias are presented in Section 4. In Section 5, our experimental setup and results are demonstrated. Then, in Section 6, we provide discussions on extensions to the domain of contextualized legal word embeddings and possible future research directions. Finally, in Section 7, we conclude with possible impacts of results. We also provide an Appendix for the CriBias.

2. Related work

The previous work related to our paper comes from two major, timely and developing bodies of literature. Below, we review these previous bodies of work under two separate subsections.

2.1 Bias in natural language processing and machine learning

Recent studies have demonstrated that machine learning algorithms and downstream applications that use these algorithms are susceptible to inherit biases such as social, racial, and gender. Zou and Schiebinger (2018) indicated how discriminatory everyday devices with AI algorithms can behave towards some subpopulations in society and came up with several reasons behind that

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

discriminatory behavior. Manzini *et al.* (2019) extended the study of Bolukbasi *et al.* (2016) from binary gender case to multiclass cases such as race and religion. They introduced a multiclass bias measure, namely mean average cosine (MAC) similarity. Caliskan *et al.* (2017) showed that training corpora may encapsulate morally neutral biases towards such as flowers (positive) and insects (negative) while unwanted biases related to race or gender are also encapsulated. May *et al.* (2019) extended the Word Embedding Association Test (WEAT), which is a method to measure bias introduced by Caliskan *et al.* (2017), performed further tests on several sentence encoders and introduced Sentence Encoder Association Test (SEAT). Tan and Celis (2019) also worked on bias in sentence-level encodings and modified SEAT to evaluate bias in contextualized word embeddings such as ELMo, BERT, and GPT (Radford *et al.* 2019). Zhao *et al.* (2017) showed label tagging algorithms also contain gender bias by tagging an image of a cooking person with female tag than male. Stanovsky, Smith, and Zettlemoyer (2019) studied the social bias in machine translation applications by using morphological analysis. They performed tests on four commercial and two academic machine translation models and the findings demonstrated that all of the models are highly prone to show stereotypical behaviors.

There are also several studies to show the existing bias in high level algorithms that use word embeddings. Kiritchenko and Mohammad (2018) introduced the Equity Evaluation Corpus (EEC) to measure the unwanted gender and racial bias in semantic evaluation tasks in SemEval-2018 Task 1: Affect in Tweets (Mohammad *et al.* 2018). They feed the EEC to every model that participated SemEval 2018 and measured the racial and gender bias of models. They concluded that most of the models inherit unwanted bias towards either genders or races. Rudinger *et al.* (2018) demonstrated the existence of gender bias in coreference resolution that uses biased word embeddings. Coreference resolution systems cannot link, for example, `surgeon` occupation with `her` pronoun. De-Arteaga *et al.* (2019) conducted experiments with three semantic representations (bag-of-words, word embeddings, and deep recurrent neural networks) to study gender bias in occupation classification. Garg *et al.* (2018) performed a diachronic study where they used texts coming from a span of 100 years to train embeddings. They showed that biases in embeddings keep track of demographic, social and occupational changes over those years. Brunet *et al.* (2019) attempted to explain the origins of social biases in word embeddings and traced the source of bias encoded in word embeddings back to the documents that cause the most formation of bias within the training corpus.

Researchers also work on elimination of unwanted bias. Bolukbasi *et al.* (2016) suggested *Hard Debiasing* and *Soft Debiasing* algorithms to remove the gender bias. These methods require a set of word pairs to identify the gender subspace of the semantic space. The results show that *Hard Debiasing* performs better than *Soft Debiasing*. Dixon *et al.* (2018) worked on measuring and mitigating the "unintended" bias in text classification algorithms. Zhang *et al.* (2018) utilized adversarial learning to eliminate the bias in classification and analogy completion tasks. Kusner *et al.* (2017) introduced the term "Counterfactual Fairness" which is defined as a decision to be fair towards an individual regardless the demographics. Based on this term, a framework was created and tested on a real-world problem involving fair prediction of success in law schools. Zhao *et al.* (2018b) developed an algorithm to train gender-neutral embeddings by preserving gender attributes in certain dimensions to protect the functionality of the model and eliminating the gender information from other dimensions. Zhao *et al.* (2019) extended the debiasing studies to contextualized word embeddings through analyzing and mitigating gender bias in ELMo. Liang *et al.* (2020) evaluated bias in sentence-level encodings such as BERT and ELMo by using WEAT, SEAT, and MAC, and introduced SENT-DEBIAS to eliminate bias.

2.2 NLP in law

The relation between AI and law has indeed a long history. The first glimpse came from Buchanan and Headrick (1970) in 1970s, but the idea did not get much of a practical implementation until

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

late 1980s. In 1987, the first International Conference on AI and Law was held. The survey of Bench-Capon *et al.* (2012) has superbly presented the summary of the field's earlier days. The initial studies were mainly about utilizing the logical structures of legal debates, and building and exploiting knowledge bases. A system called *case-based reasoning* (CBR), which proposed to work with the help of previous case information, was introduced and later further developed (Ashley 1988; Ashley 1991; Ashley 1992; Hafner and Berman 2002). Through rule-based algorithms, court case prediction and evaluation systems were developed (Aleven 2003; Ashley and Brüninghaus 2009). Galgani, Compton, and Hoffmann (2012) studied automatic summarization and Bach *et al.* (2013) addressed learning logical structures. Francesconi *et al.* (2010) (and the references therein) present a general overview of several NLP applications in law.

Latterly, AI and NLP-based methods that provide automated solutions have accelerated. One of the first studies in legal NLP is on predicting court decisions to avoid potential court congestion issues. It has been shown that automated models can predict outcomes of the US Supreme Court even better than an expert (Martin *et al.* 2004; Ruger *et al.* 2004). Aleven (2003) and Ashley and Brüninghaus (2009) built systems by relying on some hand-crafted programs (like SMILE + IBP in Ashley and Brüninghaus 2009) to predict the outcomes by looking at previous cases. The information of similar cases were extracted by detecting the neighbors of any given case using the features, called "factors" (Ashley and Brüninghaus 2009). In both Aleven (2003) and Ashley and Brüninghaus (2009), where comparisons with simple machine learning techniques are provided, accuracy scores up around 92% were achieved. SVMs turned out to be extremely advantageous for predicting decisions of the European Court of Human Rights (ECHR) (Aletras *et al.* 2016) and of the French Supreme Court (Şulea *et al.* 2017). Katz *et al.* (2017) trained a random forest based algorithm with a corpus from the US Supreme court documents to address legal case classification. Virtucio *et al.* (2018) offered methods to predict the Philippine Supreme Court decisions, where they extracted the features from court cases by gathering n-gram (Mikolov *et al.* 2013a) information to use in SVM and random forest classifiers. Mumcuoğlu *et al.* (2021) developed a framework for predicting Turkish court decisions using decision trees, random forest, and SVM as well as utilizing state-of-the-art models such as LSTM, gated recurrent unit (GRU), and BiLSTM models with attention mechanism embedded to them. Branting *et al.* (2018) trained a neural model to predict administrative adjudications.

O'Neill *et al.* (2017) came up with artificial neural networks (ANN) and distributional semantic model (DSM) representations to address the legal text classification task. Soh, Lim, and Chai (2019) compared the performances of various machine learning algorithms that is fed with topics from the Latent Semantic Analysis and pre-trained language models (BERT). Chalkidis *et al.* (2019) introduced a legal corpus called EURLEX57k that contains 57,000 case documents from EUR-LEX portal. By using EURLEX57k, they implemented several classification methods including the state-of-the-art methods such as Label-wise Attention Network (LWAN), bidirectional gated recurrent unit (BiGRU), Hierarchical Attention Network (HAN), and combinations of them to offer an effective solution to Extreme Multi-Label Text Classification (XMTC). Chalkidis *et al.* (2019) also introduced a method called LW-HAN by combining LWAN and HAN. Chalkidis *et al.* (2018) introduced a modality classifier for legal texts and achieved state-of-the-art performance by using several LSTM-based methods along with law-specific word embeddings provided by Chalkidis and Androutsopoulos (2017). Azarbonyad *et al.* (2021) studied multilabel text classification on documents of the EU that are composed mostly of legal and political content.

Chalkidis and Kampas (2019) worked on semantic feature representations of legal texts and shared famous *Law2Vec* embeddings that are pre-trained over large legal corpora consisting of legislation from UK, EU, Canada, Australia (AU), USA, and Japan. Law2Vec is the first publicly available embedding trained on large legal corpora. Locke and Zuccon (2019) investigated citation treatment task, which is a process of tagging decisions (citations) as applicable, nonapplicable, or no longer current law, for case laws. They classified case law citation treatments using three neural network architectures and SVM classifiers. First neural network-based architecture deployed

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

BERT embedding model following dense layers whereas in the second architecture a skip-gram model was used instead of BERT (Devlin *et al*. 2019). In the third architecture, an LSTM layer is also used. O'Sullivan and Beel (2019) worked on predicting whether there is a violation in ECHR cases. The experiments were performed by using several word embedding models as well as by another model called *Echr2Vec*.

There are also studies in feature extraction and information retrieval (IR). Nguyen *et al*. (2018) recognized parts of sentences that are labeled as *requisite* and *effectuation*. Chalkidis and Androutsopoulos (2017) focused on extracting contract elements based on the dataset provided by Chalkidis, Androutsopoulos, and Michos (2017). IR applications are the following: finding the related law articles for a given query (Kim, Xu, and Goebel 2017; Morimoto *et al*. 2017; Nanda *et al*. 2017; Do *et al*. 2017), finding convenient relations and matching of cases and law provisions (Tang *et al*. 2016), and extracting fact assertions in cases related to a query (Nejadgholi, Bougueng, and Witherspoon 2017). There are also studies on the legal domain-specific NER (Dozier *et al*. 2010; Cardellino *et al*. 2017; Luz de Araujo *et al*. 2018; Sleimi *et al*. 2018; Leitner, Rehm, and Moreno-Schneider 2019; Vardhan, Surana, and Tripathy 2020). Elnaggar, Otto, and Matthes (2018) set a specific example for these studies with their work on the Named Entity Linking (NEL). By utilizing the networks trained for operating NEL method to legal documents, they also experimented on transfer learning. A way of extracting features from legal texts is to recognize *facts*, *obligations*, *prohibitions,* and *principles* as word or sentence-level law-specific features (O'Neill *et al*. 2017; Shulayeva, Siddharthan, and Wyner 2017; Chalkidis *et al*. 2018). More on AI approaches applied to legal domain can also be found in Sartor and Rotolo (2013) and Casanovas *et al*. (2013).

3. Materials: Corpus, embeddings, the CriList, and test dataset[a]

In this study, to conduct various experiments requiring domain-specific data, we compiled a set of relevant material. First, in order to have a proper and field-specific embedding model, the Law2Vec embeddings have been utilized. Then, to create the same embedding model and re-enact the experiments on the replicated model, the corpus that has been used to create Law2Vec model has been collected. We also compiled an even larger corpus and trained our own legal specific embeddings that we called Law2VecNew. A crime-related list of words (CriList) has also been compiled to conduct some law-specific tests on the developed embeddings. We also present the HUDOC Dataset that we utilized to test the debiased word embeddings in high-level court case prediction task. In the following subsections, we elaborate on these aforementioned materials of our study.

3.1 Law2Vec

One of the main tools we use in this paper is the pre-trained word embedding model Law2Vec, which is made publicly available by Chalkidis and Kampas (2019). Although the composition of the corpus used to train Law2Vec is explained in details in Chalkidis and Kampas (2019), the corpus is not directly provided but only the resulting embeddings are available. The content of this corpus is quite rich and varied as described in Table 1.

In total, there are 123,066 documents consisting of 492 million tokens in the Law2Vec corpus.[b] As tabulated in Table 1, the corpus covers legislation and legal case decisions from various countries and organizations. The variety of sources used in training Law2Vec makes it quite representative, even though the weights are not balanced.

[a] All necessary source codes, data and details will be available at: https://github.com/koc-lab/legalbias.
[b] The detailed description can be found in the following web site: https://archive.org/details/Law2Vec.

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

Table 1. Contents of Law2Vec corpus

Country/organization	Number of files
UK legislation	53, 000
European legislation	62, 000
Canadian legislation	5500
Australian legislation	1150
Other EU countries (e.g. Finland, Sweden, France, Germany, etc.)	800
Japanese legislation	780
US supreme court decisions	68
US code	54

Table 2. Categorization of the compiled legal corpus to train Law2VecNew

Corpus content of Law2VecNew		
Country/Organization	Document type	Number of files
UK	Draft statutory instruments	14,431
	Local acts	398
	Ministerial orders	116
	Public general acts	8160
	Statutory instruments	126,636
	Statutory rules and orders	306
EU	Legal acts (1980–2020)	132,354
Canada	Acts	892
	Annual statues	557
	Regulations	4362
Australia	Acts	4290
	Legislative instruments	33,929
Japan	Law	817

3.2 Law2Vec corpus and Law2VecNew

Although the pre-trained Law2Vec is a quite useful tool to start with due to the facts that mentioned in the previous subsection, we also need to reach its originating corpus to make comprehensive and controlled comparisons. Since this legal corpus is not made publicly available as a single collection, we tried to collect all pieces according to the descriptions as much as possible. Since legislation documents are regularly updated by every country, it was impossible to reconstruct exactly the same corpus. However, we made our best to remain loyal to the provided descriptions as well as adding more documents to it to compile a larger corpus. Finally, the description of the corpus that we exactly collected is given in Table 2.

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

Table 3. Parameters used for Word2Vec model from Gensim library

Word2Vec model parameters			
Minimum count	20	Window size	2
Embeddings dimension	300	Down-sampling threshold	6e-5
Alpha	0.03	Minimum alpha	0.0007
Negative sampling	20	Number of cores used	12

The corpus consist of 327,248 legal documents, which is nearly three times larger than the corpus that is used to train Law2Vec. Having obtained a collective corpus similar to the one that gave rise to Law2Vec, we trained a new legal word embeddings model from scratch. Before training the model with our corpus, we implemented a cleaning process in which we deleted tokens that has non-English or nonalphabetic characters. Besides, in cleaning process, tokens that are encountered less than 30 times were deleted from the corpus. The data that we used to train word embeddings has the total of 376 million tokens (unigrams and bigrams) and 228,507 unique tokens after cleaning process. Then we trained our model by using Word2Vec model in Gensim library. The training parameters that we used for the model are reported in Table 3. In this article, we will refer these new embeddings as *Law2VecNew* and refer the original pre-trained embeddings as *Law2Vec*. Since we can now reach the subcorpora for each country separately, we also trained individual word embeddings for each country/organization to comparatively observe their characteristics independently. We also find useful to name each of these legal word embedding models to avoid any confusion. The names are as follows: CAN2Vec, AU2Vec, EU2Vec, UK2Vec, and JAPAN2Vec for embeddings trained on corpora from Canada, Australia, EU, UK, and Japan, respectively. (Note that JAPAN2Vec is trained on Japanese law documents translated to English.)

3.3 CriList

Word lists are common in the literature of bias measurement of word embedding models. Researchers use specific word lists to demonstrate the effects of social bias and validate their findings. In Bolukbasi *et al.* (2016) and Gonen and Goldberg (2019), for example, word lists of professions were used to illustrate the level of bias. Since our purpose is to study the gender bias in legal texts regarding crimes, we also compiled a field-specific word list that consists of 57 words related to criminal acts. Several sources were utilized while collecting the crime related words. The words were collected manually considering the words that are related to crimes in at least one possible context. The full list of these words are given in Appendix and the links to sources used in collecting them are provided online.[c] We will call the proposed list the "CriList".

3.4 HUDOC dataset

As mentioned previously, we assess the semantic utility of the debiased embeddings by observing their performances in a court case decision prediction task. In order to implement this task, we made use of the dataset, which is publicly shared by Medvedeva, Vols, and Wieling (2020), consisting of decisions of the European Court of Human Rights (ECHR). This dataset includes every admissible case available on HUDOC website[d] as of September 11, 2017. The cases which are only available in French are excluded. The content and the size of the dataset provided by Medvedeva

[c] https://github.com/koc-lab/legalbias.
[d] https://hudoc.echr.coe.int/.

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

Table 4. The content of the entire dataset consisting of ECHR decisions

Article	Title	Violation cases	Non-violation cases
Article 2	Right to life	559	161
Article 3	Prohibition of torture	1446	595
Article 4	Prohibition of slavery and forced labour	7	10
Article 5	Right to liberty and security	1511	393
Article 6	Right to a fair trial	4828	736
Article 7	No punishment without law	35	47
Article 8	Right to respect for private and family life	854	358
Article 9	Freedom of thought, conscience and religion	65	31
Article 10	Freedom of expression	394	142
Article 11	Freedom of assembly and association	131	42
Article 12	Right to marry	9	8
Article 13	Right to an effective remedy	1230	170
Article 14	Prohibition of discrimination	195	239
Article 18	Limitation on use of restrictions on rights	7	32

Table 5. The content of the balanced dataset consisting of ECHR decisions

Article	'Violation' cases	'Non-violation' cases	Total	Test set
Article 2	57	57	114	398
Article 3	284	284	568	851
Article 5	150	150	300	1118
Article 6	458	458	916	4092
Article 8	229	229	458	496
Article 10	106	106	212	252
Article 11	32	32	64	89
Article 13	106	106	212	1060
Article 14	144	144	288	44

et al. (2020) are given in Table 4. Since the HUDOC dataset is unbalanced, some of the cases are excluded from the training set and some others are excluded from the whole dataset. Articles 4, 7, 9, 12, and 18 of the ECHR are taken out from the dataset since their sizes are not adequate to train any algorithm. Besides, to have a more balanced training data, the number of cases with 'violation' and 'non-violation' decisions are made equal. To equalize the sizes, random cases from 'violation' cases are excluded to be used as test data ('non-violation' cases were excluded for Article 14, since 'non-violation' cases outnumber 'violation' cases for this article) (Medvedeva *et al.* 2020). The finalized number of cases in HUDOC dataset are tabulated in Table 5.

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

Table 6. Sentence templates used in the EEC

Template	# of sentences
Sentences with emotion words:	
1. <Person> feels <emotional state word>	1200
2. The situation makes <person> feel <emotional state word>	1200
3. I made <person> feel <emotional state word>	1200
4. <Person> made me feel <emotional state word>	1200
5. <Person> found himself/herself in a/an <emotional state word> situation	1200
6. <Person> told us all about recent <emotional state word> events	1200
7. The conversation with <person> was <emotional state word>	1200
Sentences with no emotion words:	
8. I saw <person> in market	60
9. I talked to <person> yesterday	60
10. <Person> goes to the school in our neighborhood	60
11. <Person> has two children	60
Total	8640

3.5 Equity Evaluation Corpus (EEC)

As another approach to the bias measurement, we use the EEC that is introduced by Kiritchenko and Mohammad (2018). The corpus contains 8640 English sentences that are handcrafted for detecting gender and racial bias. It has also 11 sentence templates that use <Person> token and <emotional state word> that can be replaced by suitable words. This allows researchers to generate more sentences that are compatible with their study. The template sentences can be seen in Table 6. The sentences carry four emotions (*anger*, *fear*, *joy*, and *sadness*). Every sentence has also some gender indicating word, either person named-entity or noun phrases. The sentences with person named-entities in them also have the racial labels indicating the origins of the names. However, since our scope is the gender bias, we do not utilize racial information.

4. Methodology

We used four methods to measure bias in Law2Vec, Law2VecNew, and all other embedding models individually trained on country-specific corpora. The first method projects word vectors onto `he-she` gender vector (Bolukbasi *et al*. 2016). The second method calculates a bias score based on the words in the CriList. The third one generates clusters of words to see how well they align with genders. Finally, the last method is to find out the percentage of male/female socially biased words that are taken from the k-nearest neighbors of the specified words (Gonen and Goldberg 2019). For debiasing purpose, we used *Hard Debiasing* method that is introduced by Bolukbasi *et al*. (2016).

4.1 Measuring bias in word embeddings

4.1.1 Projections

In the embedding vector space, a gender vector is defined by the normalized version of the difference of `he` and `she` vectors. Gender vector's direction is chosen to be from `he` to `she`. Next,

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

projections of words onto the gender vector are taken to observe how each word in the corpus is related with the gender attributes. The projection is taken by the inner product of a target word vector and the gender vector:

$$\mathbf{g} \cdot \mathbf{w} = |\mathbf{g}||\mathbf{w}|cos(\theta), \tag{1}$$

where $\mathbf{g}$ is the gender vector, $\mathbf{w}$ is the target word vector, and θ is the angle between the target word vector and the gender vector. Since unit vectors are used in the calculations, the equation turns into:

$$\mathbf{g} \cdot \mathbf{w} = cos(\theta). \tag{2}$$

The projection values can be considered as a continuum where the degree of "maleness" increases towards −1.0 and the degree of "femaleness" increases towards 1.0. For each word, the result of the inner product displays the relation of the word with genders. If the result is closer to −1.0, it is biased towards `he` pronoun and "maleness". Similarly, projections which are close to 1.0 indicate that the target word tends to be associated with "femaleness".

4.1.2 The CriBias

We introduce the CriBias as a new metric to measure bias for our specific purpose. By utilizing the CriList, we propose the CriBias to measure how biased a legal corpus towards genders is when considering criminal acts. As explained previously, the CriList contains crime-related words. Using this attribute of the CriList, one can investigate the perspective of any corpus towards the relation between gender and criminal acts.

The method exploits the projection method described in Section 4.1.1. After calculating the projection of each word in the CriList, the absolute values of projections are summed. Finally, to get a more distinctive measure, the result is amplified with a scalar value as the following:

$$CriBias = \frac{\sum_{\mathbf{w} \in C} |proj(\mathbf{w})|}{|C|} \times 100\,,\ for\ C \subset S, \tag{3}$$

where S is the set of words in the vocabulary and C stands for the set of words in the CriList. This metric helps us to quantify how spread out crime words are from the unbiased point, that is 0 projection.

4.1.3 Clustering

In this method, the most male-biased 500 words and the most female-biased 500 words are determined via previous projection method. By using the famous k-means algorithm, unsupervised clustering is applied to these words and how well these clusters align with genders is investigated. According to the distribution of words to the clusters, socially marked "gendered" words take place close to each other. Although gender bias is not directly visible (i.e. we do not use labels for words anymore), it can still be measured.

4.1.4 K-Nearest Neighbors (KNN)

As explained above, words can carry the biased attribute through their neighboring words. Even if the projections of a particular given word to gender vectors do not explicitly imply a strong bias, its neighbors can indirectly reveal how biased the given word is. For example, `nurse` being close to `receptionist`, `caregiver`, and `teacher` shows the socially implied bias on `nurse` through its neighbors. Moreover, since KNN method uses no trainable parameters, it can only reveal certain attributes of a given dataset. KNN is also deterministic so the results are fixed for a given dataset.

For the implementation, we slightly expanded the counterpart of the KNN method in Gonen and Goldberg (2019). Originally, the method gets 100-nearest neighbors of a target word to classify

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

it within two gender class. Then, it gets the classification result of the k-nearest neighbors method and take the correlation with the classification result of the projection method. The expansion we made for this method is to investigate the classification result of k-nearest neighbors with different k values, instead of considering only 100-nearest neighbors. Through this modification, the method gets more adaptive to different corpora with changing vocabulary sizes.

4.1.5 Bias measurement via EEC

In this method, we train a neural network model for emotion intensity regression task. This task was used as a challenge in the SemEval-2018 Task 1: Affect in Tweets and many participants proposed models to perform the task (Mohammad *et al.* 2018). Among these proposed models, we selected the one with the second rank (Baziotis and Jafari 2018)[e] since the best performing model is not utilizing word embeddings in their architecture. We train this model using all legal word embedding models developed in this paper. After the trainings are done, the models predict emotion intensity scores for every sentence in the EEC. Then, the predicted scores are averaged for male and female-labeled sentences so that every emotion in the EEC gets two average scores, one for male and one for female. The same procedure is repeated with debiased legal word embeddings.

4.2 Debiasing word embeddings

The first step of the hard debiasing method is to identify the gender subspace, in other words, to find the direction that carries the bias. The main goal in hard debiasing is that all pre-determined gender-neutral words have zero projections in the gender subspace. Therefore, the words that are not included to the gender subspace would be arranged such that they are equidistant according to the gender subspace (Bolukbasi *et al.* 2016). Naturally, the gender specific words that have natural gender concept from inherent meanings, such as `mother`, `father`, `man`, `woman`, are of course excluded from this procedure.

More formally, the first step is to determine the gender subspace of the embeddings model. To do this, the difference vector of 10 gender pair words are taken and the principal components for these difference vectors are calculated by using the principal component analysis (PCA). Then, the first principal component is chosen to represent the gender subspace as this vector generally carries most of the gender information present within the vector space. Note that the gender pairs are provided in Bolukbasi *et al.* (2016). After defining the gender subspace, hard debiasing consists of two steps: *Neutralize* and *Equalize*. *Neutralizing* process can be explained as follows: Consider a gender subspace β that is spanned by the set of vectors $\{\mathbf{b_1}, \mathbf{b_2}, \mathbf{b_3}, ..., \mathbf{b_M}\}$. The gender component of a word vector in the embedding model can then be computed as

$$\mathbf{w}_\beta = \sum_{i=1}^{M} \langle \mathbf{w}, \mathbf{b_i} \rangle \mathbf{b_i}, \tag{4}$$

where $\langle ., . \rangle$ is the standard inner product of the vector space.

To neutralize words, the result obtained from Equation (4) is subtracted from the word vector and the resulting vector is then divided by its Euclidean norm to obtain a unit vector:

$$\mathbf{w}' = \frac{\mathbf{w} - \mathbf{w}_\beta}{\|\mathbf{w} - \mathbf{w}_\beta\|}. \tag{5}$$

For the *Equalizing* step, a pre-determined set of equality pairs E are given. The Equality Pairs, which consist of 50 male and 50 female words that are obtained from the Amazon Mechanical Turk, are provided by Bolukbasi *et al.* (2016). First, the mean vector μ of this set is calculated.

[e] https://github.com/cbaziotis/ntua-slp-semeval2018.

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

Table 7. The list of methods and embeddings models used

Methods			Embeddings
Bias measures	Debiasing metdods	Semantic evaluation	Law2Vec
Projection	Hard debiasing	POS tagging	Law2VecNew
CriBias		Court case prediction	UK2Vec
k-Means			EU2Vec
k-Nearest			CAN2Vec
			AU2Vec
			JAPAN2Vec

Then, the gender component of this mean vector is determined by the following:

$$\mu = \frac{1}{|E|} \sum_{\mathbf{w} \in E} \mathbf{w}, \tag{6}$$

$$\mu_\beta = \sum_{i=1}^{M} \langle \mu, \mathbf{b_i} \rangle \mathbf{b_i}. \tag{7}$$

Finally, the word vectors of the set of equality pairs are re-calculated:

$$\mathbf{w}' = (\mu - \mu_\beta) + \sqrt{1 - \|\mu - \mu_\beta\|^2} \frac{\mathbf{w}_\beta - \mu_\beta}{\|\mathbf{w}_\beta - \mu_\beta\|}, \; for \; \mathbf{w} \in E. \tag{8}$$

4.3 Preservation of semantic structure

Without a semantic utility, word embeddings, biased or unbiased, are of no use. Therefore, ensuring that the semantic structure of the vector space of word embeddings is not distorted during the debiasing procedure is necessary. To check the semantic structure preservation, we compare performances of original and debiased legal word embeddings by deploying them in two tasks. The first one is the Part of Speech (POS) tagging. For this task, we utilized the experimental setup provided by Manzini *et al.* (2019).

As the second task, we implemented a high-level court case prediction task, which is an important application of NLP in legal domain. Researchers use various features of court case files such as tf-idf (Medvedeva *et al.* 2020), n-gram frequencies, and word embeddings (O'Sullivan and Beel 2019) for legal case prediction. In this experiment, we deploy the legal word embeddings that are mentioned previously in Sections 3.1 and 3.2. We follow the methodology that is suggested by O'Sullivan and Beel (2019) called *Average Embedding Values* which converts each case file to a single vector by taking the average over the vectors of every word in the corresponding case file document. The vectors are obtained through using word embeddings, so every embedding model provides a distinct set of inputs for the classifier.

Finally, as the summary of the section, Table 7 depicts the methods and embeddings models that are described above and used throughout our study.

5. Experiments[f]

We conduct experiments by using Law2Vec and Law2VecNew, and other models trained on country-specific subcorpora to observe how biased each embedding model is. For every word

[f]All necessary source codes, data, and information to reproduce our experiments will be available at: https://github.com/koc-lab/legalbias.

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

Table 8. Projection values of the CriList Words in Law2Vec and Law2VecNew: Only the most biased 10 words are shown before and after debiasing

Law2Vec				Law2VecNew			
Before debiasing		After debiasing		Before debiasing		After debiasing	
Word	Projection	Word	Projection	Word	Projection	Word	Projection
burglary	−0.2281	kidnapping	$-2.98\ 10^{-8}$	disregard	−0.1507	damage	$-2.608\ 10^{-8}$
cheat	−0.2254	criminal	$-2.328\ 10^{-8}$	felony	−0.124	sexual	$-2.508\ 10^{-8}$
criminal	−0.2176	abuse	$-2.235\ 10^{-8}$	death	−0.1227	escape	$-2.515\ 10^{-8}$
felony	−0.1960	blackmail	$-2.235\ 10^{-8}$	brutality	−0.1012	blackmail	$-2.235\ 10^{-8}$
hijack	−0.1906	brutality	$-2.235\ 10^{-8}$	innocent	−0.0894	attack	$-1.86\ 10^{-8}$
...	...	...	...	...	...	...	...
damage	−0.0360	deliberate	$1.118\ 10^{-8}$	dangerous	0.1116	guilty	$2.049\ 10^{-8}$
alias	−0.0290	eviction	$1.118\ 10^{-8}$	assault	0.119	illegal	$2.98\ 10^{-8}$
combat	−0.0247	escape	$1.816\ 10^{-8}$	danger	0.1451	abuse	$3.35\ 10^{-8}$
jail	−0.0105	assault	$2.235\ 10^{-8}$	kill	0.16	complication	$3.73\ 10^{-8}$
complication	0.0144	arson	$2.608\ 10^{-8}$	escape	0.1631	evil	$3.91\ 10^{-8}$

in the CriList, projections on the gender vector are observed to determine bias in the legal word embeddings. Second, the projections of the words of the CriList are taken to calculate CriBias score. Then, by projecting all vocabulary onto the gender vector, the most biased 1000 words (500 male biased and 500 female biased) are chosen. Those 1000 words are unsupervisedly clustered and the gender direction of every target word is compared with that of its cluster. KNN method with different values of k is applied on the embeddings and the bias measure of this method is compared with the bias measure coming from projections. Finally, we also implemented the bias measurement method using the EEC. Here, we get emotion intensity scores that is averaged over male and female for four emotions.

We also applied "Hard Debiasing" algorithm to all embeddings models except Japan2Vec due to the limited vocabulary size of the corresponding corpus. As we determine the gender subspace for debiasing using the PCA, we also investigate the explained variance for the chosen principal component and observed that more than 60% of the variance is stored in the chosen principal component for every embeddings model. Both before and after executing the "Hard Debiasing" algorithm on word embedding models, all four bias measurement experiments are implemented. The results are presented to make a comparison between bias levels before and after the debiasing procedure. Finally, the preservation of semantic structure is assessed to show that the debiasing procedure does not distort the semantic utility of word embeddings. In the following subsections, we will present results of these experiments in detail.

5.1 Projections

5.1.1 Law2Vec & Law2VecNew

Projection values of the crime related words before and after debiasing are listed in Table 8 for Law2Vec and Law2VecNew. In Table 8, only the words having the most or the least bias values are tabulated while the remaining words reside in between. Upon inspecting Table 8, it can be

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

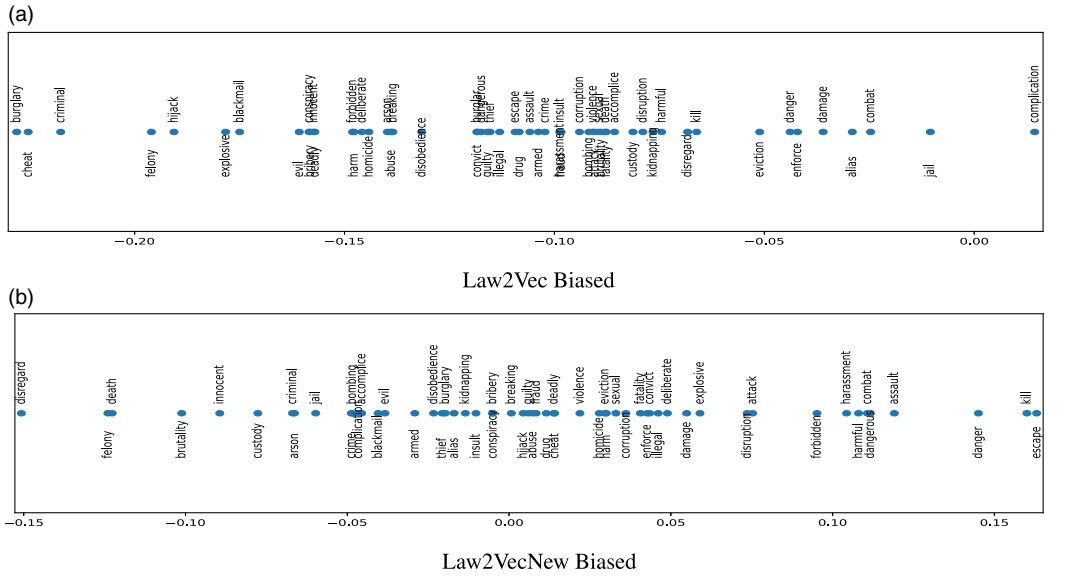

Figure 1. Word projections onto gender vector for Law2Vec and Law2VecNew. Right side represents the female direction, left side represent the male direction, and 0 is the neutral point.

seen that before debiasing of Law2Vec, every word, except one (`complication`), from the crime related word list has a negative projection value, indicating an unwanted bias towards males. A visualization of the projections for several qualitative examples are given as a continuum in Figure 1a, where the right side indicates female and the left side indicates male direction. Note that there is only one point in the right side of 0 that stands as the neutral point. Applying Hard Debiasing to the Law2Vec model has eliminated gender attributes from words as indicated by the almost zero (on the order of 10^{-8}) projections after debiasing. The projections have become negligibly low which means that the goal of debiasing is achieved.

The results of the projection method for Law2VecNew show quite a different behavior. The word projections of the CriList words onto gender vector for Law2VecNew show a balanced behavior towards both male and female directions. However, there is still an unwanted bias distributed towards males or females. The magnitudes of the most male biased and female biased words are close to each other, which implies that although the crime related words are biased in the corpus of Law2VecNew, the direction of the bias cannot be generalized to a single direction. The projections are also given visually in Figure 1b. Again, the right side is female direction and the left side is male direction. In this case, the points are distributed over left and right sides of 0 point more equally than Figure 1a. The effect of Hard Debiasing is obvious that the magnitudes of the projections are again in the order of 10^{-8} or less. This implies that the method successfully debiased embeddings considering the projections of words. As qualitative examples, there are words in the debiased embeddings even with 0 projections onto the gender vector such as `criminal`, `deadly`, and `disregard`.

5.1.2 UK2Vec, EU2Vec, AU2Vec, CAN2Vec, JAPAN2Vec

The results for country-specific legal word embeddings are tabulated in Table 9. As before, only the most male and female biased words are shown. The results of projections for country-specific embeddings are more or less similar to Law2VecNew in terms of the magnitudes of projections and the general direction of the biased words. The bias directions are almost balanced for all embeddings except few of them show small off-balance. It should be noted that, not all words

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

Table 9. Projection values of the CriList words in UK2Vec, EU2Vec, AU2Vec, CAN2Vec, and JAPAN2Vec

UK2Vec				EU2Vec			
Before debiasing		After debiasing		Before debiasing		After debiasing	
Word	Projection	Word	Projection	Word	Projection	Word	Projection
damage	−0.125	bribery	$-2.608\ 10^{-8}$	innocent	−0.126	accomplice	$-2.98\ 10^{-8}$
danger	−0.122	enforce	$-2.28\ 10^{-8}$	criminal	−0.1	burglary	$-2.6\ 10^{-8}$
disobedience	−0.116	eviction	$-1.86\ 10^{-8}$	alias	−0.077	conspiracy	$-2.23\ 10^{-8}$
burglar	−0.114	breaking	$-1.49\ 10^{-8}$	eviction	−0.075	insult	$-1.96\ 10^{-8}$
escape	−0.109	bombing	$-1.118\ 10^{-8}$	custody	−0.068	bombing	$-1.86\ 10^{-8}$
...	...	...	...	...	...	...	...
sexual	0.113	disregard	$1.77\ 10^{-8}$	danger	0.09	fraud	$1.3\ 10^{-8}$
blackmail	0.118	innocent	$1.77\ 10^{-8}$	conspiracy	0.1	harmful	$1.35\ 10^{-8}$
death	0.131	sexual	$2.049\ 10^{-8}$	escape	0.106	arson	$1.513\ 10^{-8}$
disregard	0.152	burglary	$2.61\ 10^{-8}$	disruption	0.115	abuse	$1.86\ 10^{-8}$
harassment	0.181	death	$3.53\ 10^{-8}$	damage	0.147	corruption	$2.608\ 10^{-8}$

CAN2Vec				AU2Vec			
Before debiasing		After debiasing		Before debiasing		After debiasing	
Word	Projection	Word	Projection	Word	Projection	Word	Projection
disregard	−0.166	bribery	$-3.725\ 10^{-8}$	conspiracy	−0.109	homicide	$-2.98\ 10^{-8}$
dangerous	−0.16	fraud	$-2.235\ 10^{-8}$	burglary	−0.097	harm	$-1.86\ 10^{-8}$
damage	−0.156	harmful	$-2.235\ 10^{-8}$	insult	−0.081	violence	$-1.86\ 10^{-8}$
attack	−0.145	homicide	$-1.49\ 10^{-8}$	damage	−0.075	fraud	$-1.68\ 10^{-8}$
alias	−0.122	criminal	$-9.31\ 10^{-9}$	violence	−0.071	disregard	$-1.49\ 10^{-8}$
...	...	...	...	...	...	...	...
deliberate	0.12	violence	$1.58\ 10^{-8}$	alias	0.074	complication	$1.3\ 10^{-8}$
harm	0.12	illegal	$2.42\ 10^{-8}$	harm	0.109	innocent	$1.49\ 10^{-8}$
harmful	0.162	custody	$3.4\ 10^{-8}$	death	0.112	sexual	$1.49\ 10^{-8}$
kill	0.166	corruption	$4.1\ 10^{-8}$	abuse	0.137	explosive	$1.86\ 10^{-8}$
guilty	0.191	arson	$6.7\ 10^{-8}$	escape	0.147	guilty	$2.6\ 10^{-8}$

JAPAN2Vec			
Before Debiasing		After Debiasing	
Word	Projection	Word	Projection
explosive	−0.181	–	–
alias	−0.161	–	–
danger	−0.136	–	–

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

Table 9. Continued.

JAPAN2Vec			
Before Debiasing		After Debiasing	
Word	Projection	Word	Projection
accomplice	−0.103	–	–
disregard	−0.091	–	–
...	...	–	–
disruption	0.092	–	–
custody	0.113	–	–
deliberate	0.129	–	–
enforce	0.147	–	–
abuse	0.192	–	–

in the CriList is included in the vocabulary of country-specific models. The projections for each model are given visually in Figure 2, where the right side represents female direction and the left side represents male direction. The effect of Hard Debiasing is quite clear since the scale of projection magnitudes reduces down to 10^{-8}, and some words have even 0 projection, meaning the bias of these words is totally diminished. However, the Hard Debiasing method failed on JAPAN2Vec since Japanese legislation documents are very limited and the corpus does not include most of the word pairs required to perform Hard Debiasing method. Therefore, results for after debiasing for JAPAN2Vec are not presented in Table 9.

The results we obtained suggest that the legal embeddings contain gender bias about criminal subject. Every criminal word has a bias in male or female direction with a significant magnitude. However, it is also important to observe the number of words that are biased either towards male or female direction. Figure 3 depicts the distribution of male and female directed crime words for each embedding model. The distribution of Law2Vec is quite interesting since it turns out to be that almost every word in the list has male directed bias in contrast to the other embedding models, which have relatively balanced distributions. This significant distinction in the results is probably caused by the documents that are used to build Law2Vec and are not included the corpus of the embeddings that we have formed, especially the US Supreme Court decisions. Since the Supreme Court decisions unintentionally contain the statistical distribution of crime rates among males and females, the embeddings are affected from this statistical distribution. As suggested in Figure 3, legislations of Australia and UK have a bias in associating crimes with females while Canada, EU, and Japan contain an opposite bias.

5.2 The CriBias

Figure 4 depicts the resulting measurements of the CriBias for each embeddings model. According to those results, Law2Vec model inherits the highest rate of gender bias on crime related words with 11.168 CriBias score. CAN2Vec and JAPAN2Vec follow Law2Vec by again with significantly high CriBias scores relative to the rest of the models. Note that some of the models do not have all the CriList words in the corpus they trained on. Therefore, the scores reflect the normalized values of the total projection magnitudes.

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

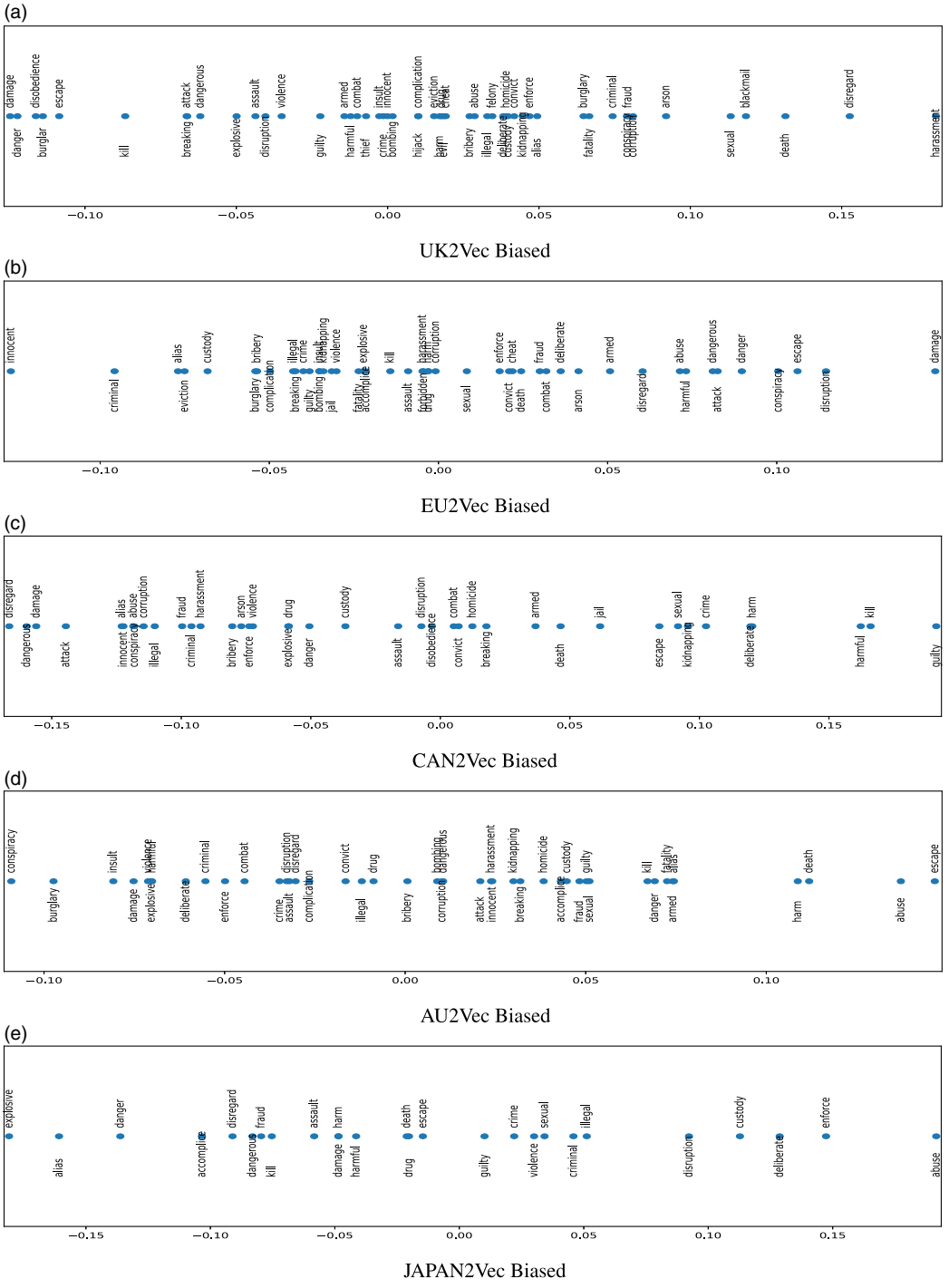

Figure 2. Word projections onto gender vector for UK2Vec,EU2Vec, CAN2Vec, AU2Vec, and JAPAN2Vec. Right side represents the female direction, left side represent the male direction, and 0 is the neutral point.

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

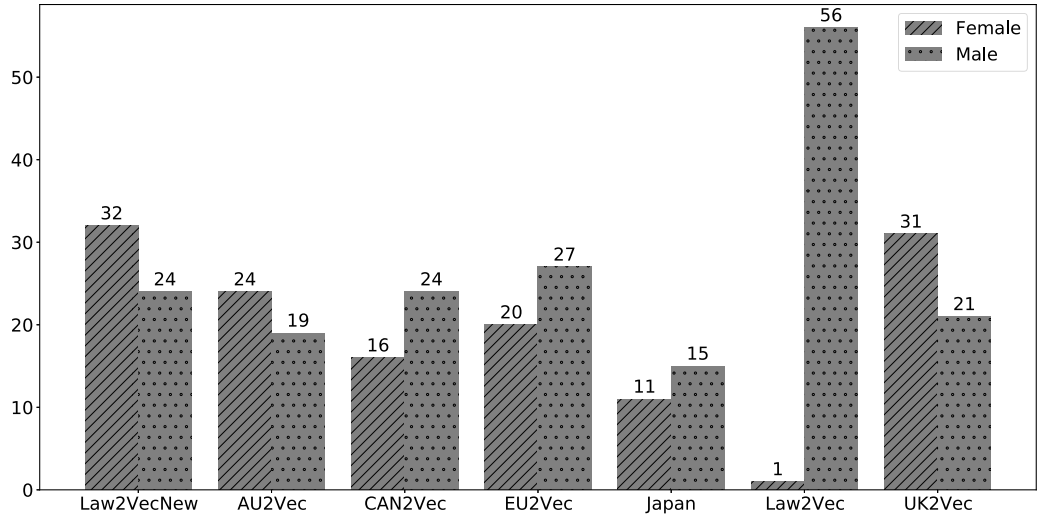

Figure 3. The distribution of bias directions in word embeddings models.

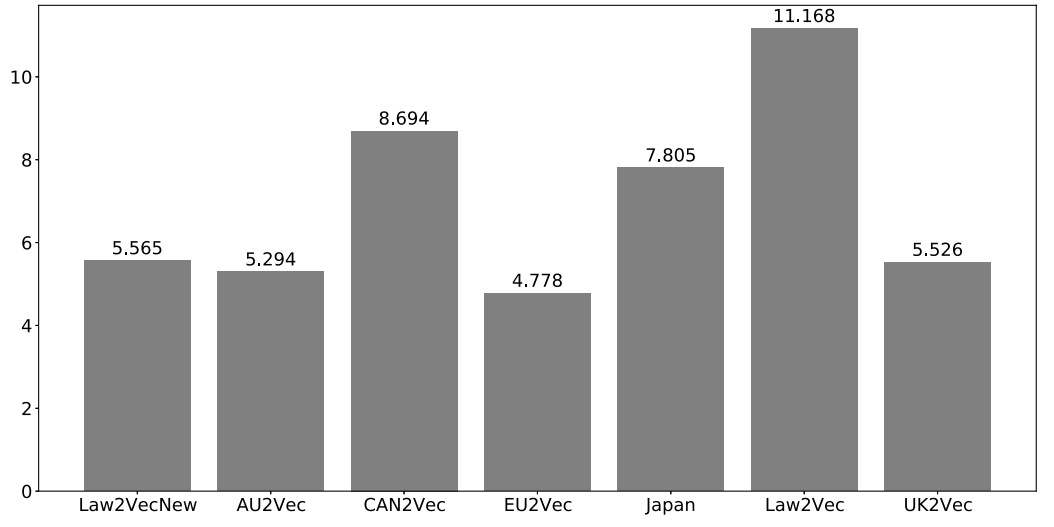

Figure 4. CriBias scores.

5.3 Clustering

We followed the procedure explained in Section 4.1.1 to find the most male-biased 500 words and the most female-biased 500 words from each embeddings model. Note that these words do not include any morally gender-specific words, since such words are excluded in advance. Hence, these 1000 words are expected to be gender-neutral words but still carrying the highest bias projections. Table 10 shows the precision scores of clustering algorithm before and after the debiasing process. High precision means that the 1000 chosen words can be labeled without any prior information except the embedding vectors. Thus, the higher the precision, the more biased the embeddings model is. Considering the precision percentages in Table 10, the embeddings seem to be highly biased and the effect of the debiasing is quite small, almost negligible.

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

Table 10. The precision scores of clustering method

Model	Biased precision (%)	Debiased precision (%)
Law2VecNew	100	99.8
AU2Vec	99.9	97.1
CAN2Vec	100	100
EU2Vec	99	98.6
JAPAN2Vec	99.8	–
Law2Vec	97.9	97.9
UK2Vec	99.9	99.9

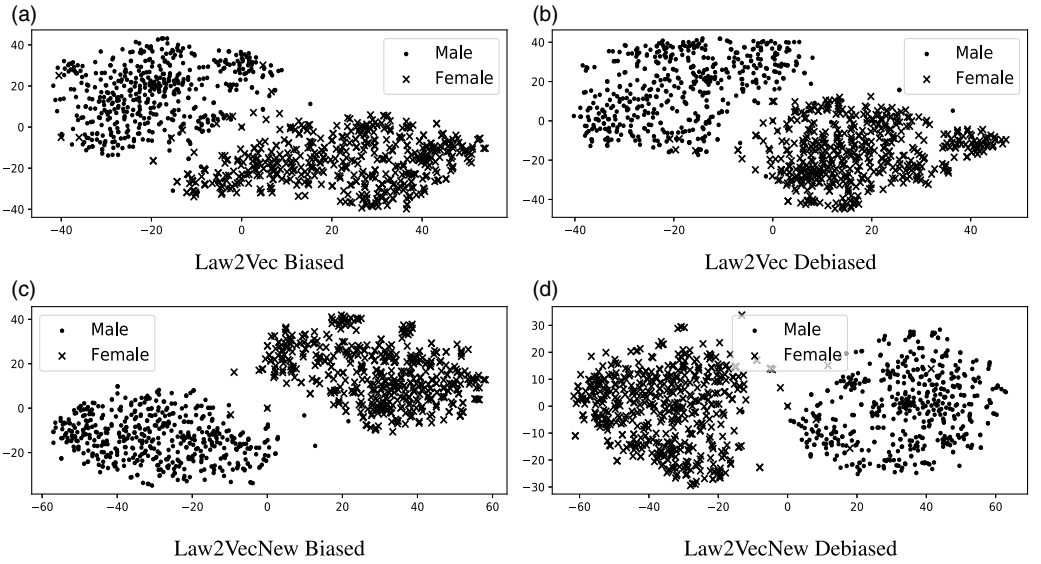

Figure 5. Clustering results of Law2Vec and Law2VecNew.

The results of clustering method are also visualized. After a dimensionality reduction with tSNE method, the embedding vectors of the chosen 1000 words are transformed to a two dimensional space. Corresponding plots for Law2Vec and Law2VecNew are given in Figure 5. As the precision results suggest, the grouping of words in Law2VecNew is more obvious. Plots for the rest of country-specific word embeddings are also shown in Figure 6. For the original embeddings (left sides), the biased words stay close to each other. By observing the placement of words, one can deduce that the models inherit gender bias. By observing the clustering of debiased (right sides) embeddings, it is clear that even though the debiasing has changed the alignment of words, male-oriented, and female-oriented words are still distinguishable. Especially for UK2Vec and CAN2Vec, the gender groups are visually more separable, indicating higher biases.

Thus, in agreement with the precision results, the negligible effect of debiasing can also be seen visually in Figures 5 and 6. Since clustering tries to group out a dataset according to the similar attributes each data point has, it can be concluded that an unsupervised algorithm can also quantify how words in the corpus carry gender attributes and biases. These results imply that

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

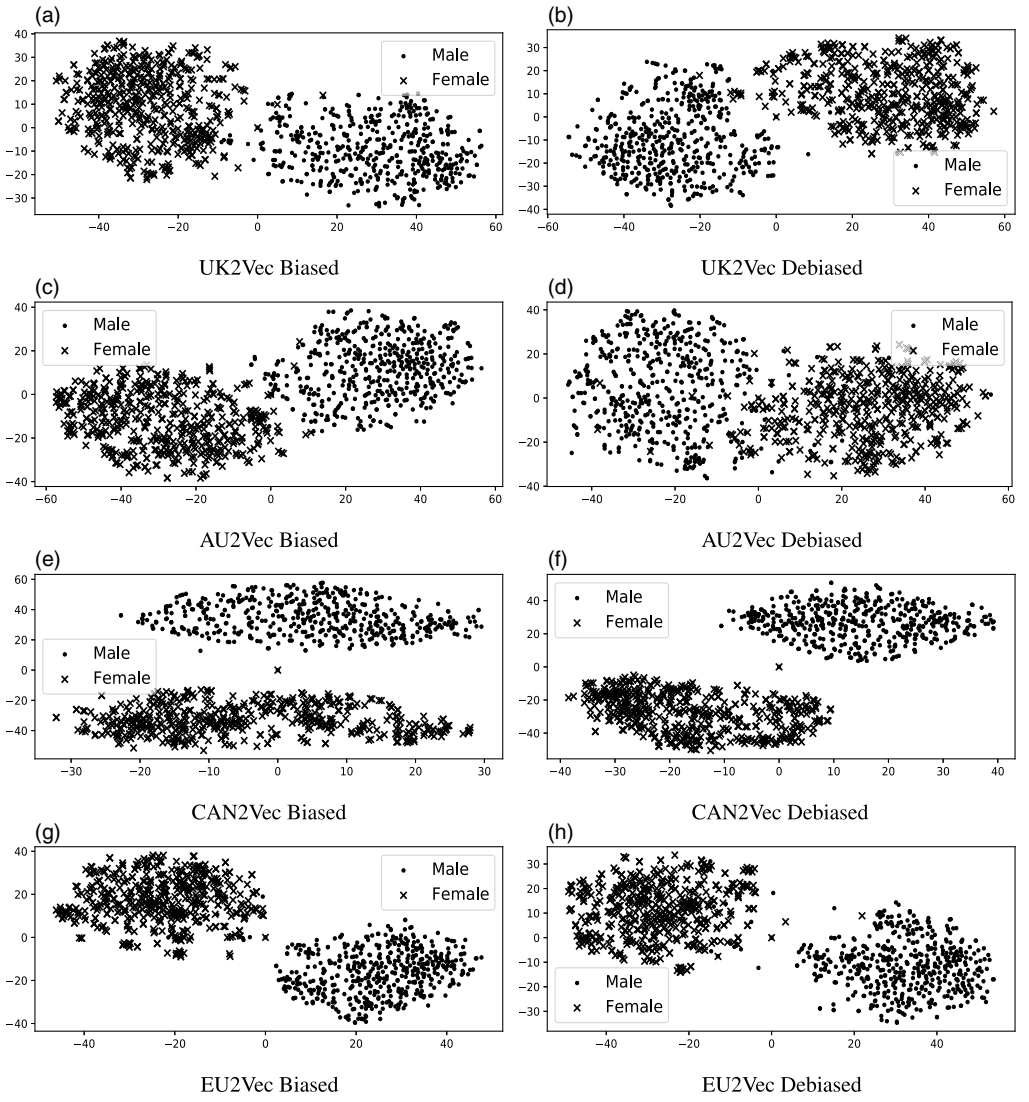

Figure 6. Clustering results for AU2Vec, CAN2Vec, EU2Vec, and UK2Vec.

even if the projections are eliminated, the bias in embeddings can still be found by observing the relations of words relative to each other.

5.4 K-nearest neighbors

We also computed bias by using KNN algorithm to determine the bias rate. The correlation of the bias found with KNN and the bias found with projection method was calculated in order to see the consistency of different bias measuring methods. The same procedure was also applied to the debiased versions of each word embedding model. The parameter for the KNN was taken in a range (from 10 to 100 with stepsize of 10) to observe the effect of number of neighbors that are taken into account. By working with different k values, we also tried to eliminate the effect of varying sizes of our corpora since different k values are more appropriate for different corpus sizes.

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

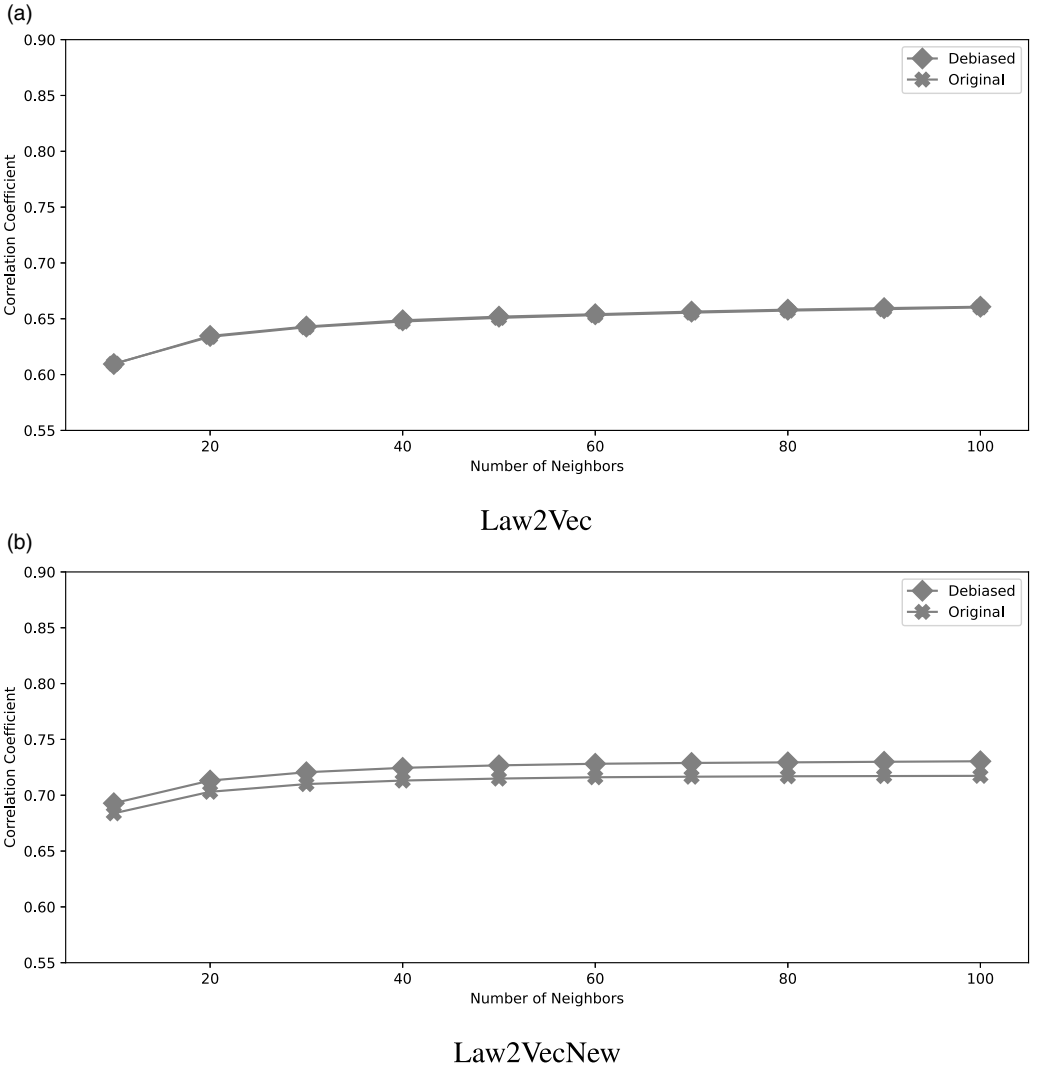

Figure 7. Results based on k-Nearest neighbors for Law2Vec and Law2VecNew.

The plots for Law2Vec and Law2VecNew are depicted in Figure 7, where curves for both the original and debiased embeddings are shown. In these plots, if two curves stay close to each other, it means that the debiasing process is not successful. In the case of Law2Vec, the curves are close to each other indicating a low performing debiasing procedure. On the other hand, the curves are separated on for Law2VecNew. However, the correlations do not differ significantly. Although the debiasing operation works better in Law2VecNew than the original Law2Vec, it is still not sufficient.

The plots for the UK2Vec, AU2Vec, CAN2Vec, and EU2Vec are given in Figure 8. For UK2Vec and CAN2Vec, the curves are almost at the same position, which means debiasing does not effectively reduce the underlying bias that implied by neighbors. On the other hand, the curves of AU2Vec and EU2Vec look slightly separated. It shows that debiasing process reduces underlying gender bias to some degree.

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

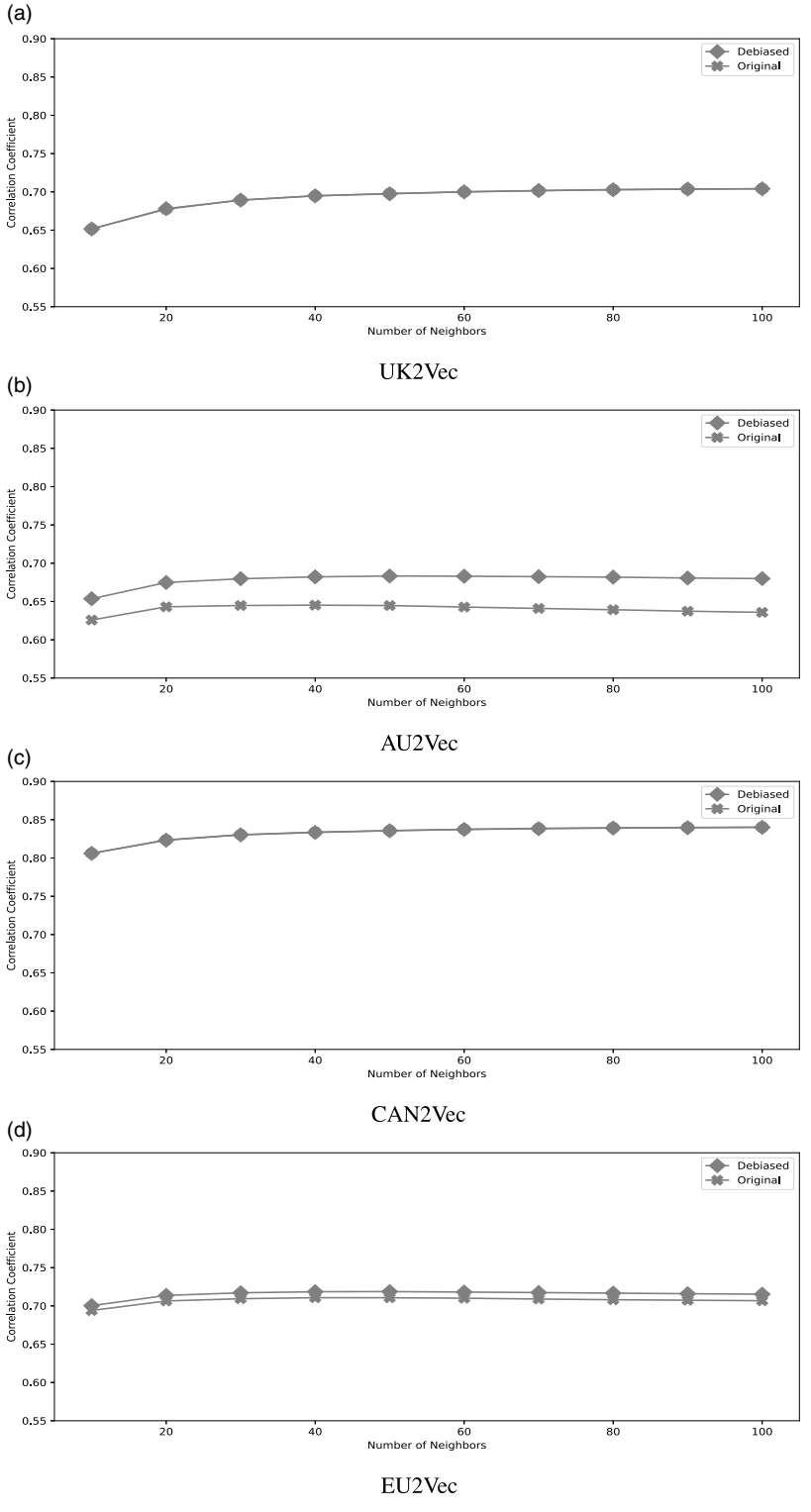

Figure 8. Results based on K-Nearest neighbors for UK2Vec, AU2Vec, CAN2Vec, and EU2Vec.

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

5.5 Bias measurement via EEC

The results for the EEC-based measurement are tabulated in Table 11, where the averages taken along male and female sentences for every emotion are named as *Male* and *Female* scores. *Delta* scores, on the other hand, indicate the difference between average *Male* and *Female* scores. Ideally, we expect *Delta* scores to be 0 since the sentences in the EEC are generated from gender-neutral templates. As the results in Table 11 indicate, *Delta* scores coming from the debiased legal embeddings are closer to 0 than those from biased versions, which shows that the debiasing method work well in general. Lastly, one needs to note that our objective in this experiment is not to increase performance in an emotion intensity regression task by using legal word embeddings, which are not designed for this purpose. Instead, we focus on the statistical inference of the results in observing the effects of gender debiasing.

5.6 Preservation of semantic structure

5.6.1 POS tagging

To ensure that the debiasing procedure does not distort the underlying semantic structure of embeddings, we first tested the biased and debiased word embeddings on the POS tagging task by using the experimental setup provided by Manzini *et al.* (2019). The embeddings were used in CoNLL 2003 shared task (Tjong Kim Sang and De Meulder 2003). The dataset for this task contains 235 training, 58 validation, and 34 test batches with the batch size of 64. The model network was trained for 25 epochs for each task. Initially, the words in the dataset are encoded by using biased and debiased embeddings separately. Then, the model was trained and tested with those encoded data. The F1-scores of the task for each word embeddings are listed in Table 12. The results suggest that the debiasing do not distort the semantic utility of word embeddings, since there is no significant reduction in the F1-scores between biased and debiased embeddings.

5.6.2 Results for predicting court decisions

For the decision prediction experiment, we used the HUDOC dataset (see Section 3.4). Since the cases in HUDOC dataset consists of binary decisions (violation or non-violation), we deploy linear SVMs with hard margin for binary classification. The text files were tokenized by using NLTK[g] library. Having tokenized all cases and mapped them to vectors using word embeddings, the case files transformed to a single vector through implementing *Average Embedding Values* (O'Sullivan and Beel 2019). For each Article of ECHR that is present in HUDOC, a distinct classifier was trained. The accuracy scores on the test set for each Article with biased and debiased word embeddings are given in Table 13. Note that some of the articles were not used due to the inadequate size of data (see Section 3.4 for more details).

The accuracy scores obtained by using embeddings after debiasing are either very close to or same as those when we used biased embeddings. These results from a high-level legal NLP task further show that the utility of legal word embeddings are not distorted after debiasing process. Finally, it should again be noted that our objective in these experiments is not to improve high-level task performance. Rather, we aim to show that the performance of debiased embeddings is on par with that of original embeddings.

6. Discussions on bias in contextualized word embeddings

Recently, transformer-based language models, such as BERT (Devlin *et al.* 2019), have become the state-of-the-art methods in NLP. As we mentioned in Section 2, there are also several studies to investigate and eliminate biases from contextual language models since the available methods

[g] https://www.nltk.org/.

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

Table 11. The results of bias test via the EEC Dataset on Emotion Intensity Regression Task. For each legal embedding model (except JAPAN2Vec) and for each emotion type, the smallest (in absolute means) one of the Delta values from the Biased and Debiased embeddings are emboldened. Ties are underlined. JAPAN2Vec is failed to be debiased due to the lack of sufficient size vocabulary

Model\Emotion			Anger	Fear	Joy	Sadness
Law2Vec	Biased	Male	0.49	0.52	0.47	0.63
		Female	0.47	0.50	0.48	0.61
		Delta	0.02	0.02	−0.01	0.02
	Debiased	Male	0.47	0.51	0.40	0.60
		Female	0.47	0.51	0.40	0.60
		Delta	**0.00**	**0.00**	**0.00**	**0.00**
Law2VecNew	Biased	Male	0.48	0.54	0.45	0.66
		Female	0.49	0.53	0.46	0.64
		Delta	−0.01	**0.01**	−0.01	0.02
	Debiased	Male	0.43	0.52	0.45	0.70
		Female	0.44	0.49	0.45	0.68
		Delta	−0.01	0.03	**0.00**	0.02
AU2Vec	Biased	Male	0.53	0.50	0.42	0.61
		Female	0.51	0.50	0.42	0.61
		Delta	0.02	**0.00**	0.00	0.00
	Debiased	Male	0.53	0.50	0.47	0.56
		Female	0.52	0.52	0.47	0.56
		Delta	**0.01**	−0.02	0.00	0.00
CAN2Vec	Biased	Male	0.54	0.51	0.49	0.59
		Female	0.52	0.50	0.51	0.58
		Delta	0.02	0.01	−0.02	0.01
	Debiased	Male	0.48	0.55	0.51	0.64
		Female	0.48	0.55	0.51	0.64
		Delta	**0.00**	**0.00**	**0.00**	**0.00**
EU2Vec	Biased	Male	0.46	0.47	0.43	0.67
		Female	0.46	0.50	0.42	0.65
		Delta	0.00	−0.03	0.01	0.02
	Debiased	Male	0.46	0.52	0.43	0.63
		Female	0.46	0.53	0.43	0.63
		Delta	0.00	**−0.01**	**0.00**	**0.00**

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

Table 11. Continued

Model\Emotion			Anger	Fear	Joy	Sadness
UK2Vec	Biased	Male	0.48	0.54	0.48	0.55
		Female	0.48	0.55	0.47	0.54
		Delta	0.00	−0.01	0.01	0.01
	Debiased	Male	0.48	0.56	0.46	0.60
		Female	0.48	0.55	0.46	0.60
		Delta	0.00	0.01	**0.00**	**0.00**

Table 12. F1-scores of legal word embedding models on POS tagging task

	Law2Vec	Law2VecNew	EU2Vec	UK2Vec	CAN2Vec	AU2Vec
Biased	0.98	0.95	0.95	0.95	0.95	0.95
Debiased	0.95	0.95	0.95	0.95	0.95	0.95

Table 13. Accuracy (%) results for decision prediction task on ECHR case files with biased and debiased embeddings

		Law2Vec	Law2VecNew	EU2Vec	UK2Vec	CAN2Vec	AU2Vec
Article 2	Biased	84.42	80.90	81.41	83.92	**86.43**	87.44
	Debiased	**85.93**	**83.92**	**82.91**	**85.68**	85.18	87.44
Article 3	Biased	**74.38**	74.74	74.62	73.91	76.26	**71.44**
	Debiased	74.03	74.74	**76.15**	**74.15**	**76.38**	70.27
Article 5	Biased	73.26	**76.65**	78.18	80.14	71.56	73.79
	Debiased	**74.60**	76.48	**78.26**	80.14	**72.81**	**73.97**
Article 6	Biased	80.16	**79.81**	79.33	**76.86**	79.25	**78.35**
	Debiased	**82.65**	79.40	79.33	76.81	**79.33**	78.01
Article 8	Biased	**70.16**	**70.77**	**73.19**	**71.98**	**67.74**	72.78
	Debiased	68.15	70.56	72.78	70.97	66.33	**74.60**
Article 10	Biased	**70.63**	68.25	66.67	**68.65**	**65.48**	69.05
	Debiased	69.44	**69.84**	**67.06**	68.25	65.08	**70.63**
Article 11	Biased	**80.90**	75.28	69.66	74.16	71.91	73.03
	Debiased	79.78	75.28	69.66	74.16	71.91	73.03
Article 13	Biased	**79.53**	**78.30**	79.62	81.98	81.04	76.51
	Debiased	79.43	78.11	**79.72**	**82.17**	**81.42**	**77.17**
Article 14	Biased	72.72	70.45	84.09	84.09	**79.55**	**65.91**
	Debiased	**75.00**	**72.72**	84.09	84.09	72.72	63.64

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

constructed for conventional word embeddings cannot be used to debias contextualized models (Caliskan *et al*. 2017; May *et al*. 2019; Bhardwaj, Majumder, and Poria 2021). Investigating bias in contextualized models requires distinct approaches since every word has a different embedding depending on the sentence. The current works in the literature are mostly focused on analyzing the bias of contextualized language models in general context. As the contextualized language models provide state-of-the-art performances for many NLP tasks, their implementations in legal domain have also started very recently (Chalkidis *et al*. 2020; Chalkidis *et al*. 2021). This opens up several research directions for future work to expand upon our baseline work of investigating unwanted bias in legal language models. We discuss these extensions in detail in the following.

Firstly, to begin with investigating bias in contextualized language models, one needs a framework where a contextualized language models is developed specifically for the legal domain. Chalkidis *et al*. (2020) has very recently introduced a variant of BERT where the model is trained with a legal corpus called LegalBERT. The pre-trained models of LegalBERT are publicly available which eases the setup for further work since the training of a BERT model is extremely time consuming and requires a huge amount of computational power. Having the framework, the next step will be designing a suitable bias measurement method designed specifically for legal context. To this end, one needs to tweak available bias measuring methods with legal domain ingredients to make them operate for NLP in legal domain. A prominent work for this purpose is the *Equity Evaluation Corpus (EEC)* (Kiritchenko and Mohammad 2018). EEC mainly provides a framework to detect gender and racial bias in sentiment analysis task in general context. However, this flexible framework can also be used with various NLP tasks, such as the *Bias Evaluation Corpus with Professions (BEC-Pro)* (Bartl, Nissim, and Gatt 2020). BEC-Pro is inspired by EEC, but it measures the bias by using professions that are statistically gendered instead of emotions (Bartl *et al*. 2020). A similar approach can also be followed to measure biases in downstream tasks dealing with different contexts. Merging the CriList, we introduced with the EEC to construct an evaluation corpus similar to BEC-Pro, one can develop a convenient bias measuring method for contextualized language models in legal corpora.

The mitigation of bias, on the other hand, is rather challenging. The available debiasing methods such as the methods proposed by Bolukbasi *et al*. (2016) and Bhardwaj *et al*. (2021) can be used to investigate their effects. These methods are projection-based ones, which work when the word embeddings in the model are global. However, they might be unsuitable for contextualized settings as each word is represented by a different embedding depending on the context in which it occurs (Kurita *et al*. 2019). Thus, building upon the available methods to develop more specialized debiasing techniques for contextualized language models remain as an open problem for future research. Some of the recent methods utilize the *Masked Language Model (MLM)* feature of BERT (Devlin *et al*. 2019; Bartl *et al*. 2020). Fine tuning a BERT model with the MLM task using a gender-neutral corpus is observed to mitigate the bias of the model on the BEC-Pro (Bartl *et al*. 2020). Either way, after implementing any debiasing method, the semantic functionality of bias-free language models should also be monitored. This requires high-level tasks to check the performances of debiased language models. Chalkidis *et al*. (2021) provide a framework where a number of high-level tasks can be implemented to contextualized language models. Once again, the key here is not compromise the functionality of language models when implementing debiasing methods.

7. Conclusions

In this article, we have investigated the gender bias of legal word embeddings that trained on large legal corpora as well as country-specific sub-corpora. For this purpose, we utilized two legal word embedding models, both a publicly available model, Law2Vec, and our own model Law2VecNew. We intentionally excluded the documents that contain cases of the US Supreme Court, which is present in the original corpus for Law2Vec, from our own corpus. Bias in court case documents and legislation documents are two distinct issues. Court case documents partly contain descriptions of real events and may introduce bias to word embeddings due to the statistical

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

properties of crime-gender relations, which are beyond our scope. However, legislation and other law-related documents that do not directly contain real-world cases are strictly expected to be neutral towards gender attributes. To compare bias issues in legal word embeddings across different countries, cultures and traditional attributes, we also studied legal word embeddings trained on several country-specific corpora.

We used various methods in quantifying gender bias to provide a comprehensive study which considers different methods. We also introduced a new method, the CriBias, to evaluate the gender bias within word embeddings in legal context. Our results, irrespective of the bias measurement methods used or countries considered, consistently suggest that there is certainly an unacceptable gender bias in legal word embeddings.

We also remodeled a popular debiasing algorithm to be applicable in legal domain. With this method, we eliminated or at least significantly reduced gender bias found in legal word embeddings. We noted that the size of the corpus is critical to implement the debiasing method since it failed to operate over JAPAN2Vec, which is trained on a relatively smaller corpus than the other country-specific models. After applying the debiasing algorithm, we used the same bias evaluation methods to observe the effect of debiasing. In this, we reported mixed results. The methods based on the vector projections showed that the debiasing process was quite successful and indeed the algorithm reduced the bias to negligible levels. On the other hand, the methods that indirectly rely on the neighbor words of a target word resulted in contradicting indications. When one relies on indirect bias measurement models like clustering and k-nearest neighbors, the debiasing algorithm did not have any significant effect in reducing the gender bias. The cause of the contradicting results is the diverse natures of the evaluation methods. The methods scrutinize different ways in which the word embeddings possess bias and the debiasing algorithm works with respect to only one of the approaches in evaluation. Having a single perspective of bias, the debiasing method ignores the other ways with which bias can occur in word embeddings. This suggests that further future research is necessary in this area in order to develop better debiasing algorithms, even probably legal domain specific ones. We also performed experiments to show that debiasing did not distort the underlying semantic structure of legal word embeddings, which is critical in their semantic utility.

Although the bias and other ethical issues in word embeddings are being extensively studied in the community, the issue has never been addressed in the legal text processing context. Bias issues in legal NLP domain should be taken doubly serious, considering the role and importance of law in our lives. Gender bias present in legal texts and legislation can imply injustice in legal systems. Our results that showed existing gender bias in legislation have two different implications. First, in developing NLP in law applications, utilizing debiased word embeddings is of technical importance. Second, quantitative methods can be of help in revealing such biases in legislation and can establish an awareness for society and lawmakers when creating legislation. We believe that our results can give rise to an interest and that future studies on the issue will follow in this subfield while NLP in law is also continuously and concurrently developing. Definitely, there is quite a bit room for future research and improvements regarding the study of bias in NLP for legal domain. Possible immediate future research can be pursued on developing more effective debiasing methods specific to legal domain and alternative bias quantification methods as well as investigating other possible biases other than gender.

Acknowledgment. This work was supported by TUBITAK 1001 grant (120E346). We would like to thank the anonymous reviewers for their many insightful comments.

References

Aletras N., Tsarapatsanis D., Preotiuc-Pietro D. and Lampos V. (2016). Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. *PeerJ Computer Science* **2**, e93.

Aleven V. (2003). Using background knowledge in case-based legal reasoning: A computational model and an intelligent learning environment. *Artificial Intelligence* **150**, 183–237.

Ashley K.D. (1988). *Modelling Legal Argument: Reasoning with Cases and Hypotheticals*. PhD thesis, University of Massachusetts, USA. Order No: GAX88-13198.

Ashley K.D. (1991). Reasoning with cases and hypotheticals in HYPO. *International Journal of Man-Machine Studies* **34**(6), 753–796.

Ashley K.D. (1992). Case-based reasoning and its implications for legal expert systems. *Artificial Intelligence and Law* **1**, 113–208.

Ashley K.D. and Brüninghaus S. (2009). Automatically classifying case texts and predicting outcomes. *Artificial Intelligence and Law* **17**(2), 125–165.

Azarbonyad H., Dehghani M., Marx M. and Kamps J. (2021). Learning to rank for multi-label text classification: Combining different sources of information. *Natural Language Engineering* **27**(1), 89–111.

Bach N.X., Minh N.L., Oanh T.T. and Shimazu A. (2013). A two-phase framework for learning logical structures of paragraphs in legal articles. *ACM Transactions on Asian Language Information Processing* **12**(1), 1–32.

Bartl M., Nissim M. and Gatt A. (2020). Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, Spain (Online). Barcelona: Association for Computational Linguistics, pp. 1–16.

Baziotis C. and Jafari B. 2018. ntua-slp-semeval2018. https://github.com/cbaziotis/ntua-slp-semeval2018.

Bench-Capon T., Araszkiewicz A.M., Ashley A.K., Atkinson K., Bex F., Borges F., Bourcier D., Bourgine P., Conrad J.G., Francesconi E., Gordon T.F., Governatori G., Leidner J.L., Lewis D.D., Loui R.P., McCarty L.T., Prakken H., Schilder F., Schweighofer E., Thompson P., Tyrrell A., Verheij B., Walton D.N. and Wyner A.Z. (2012). A history of AI and Law in 50 papers: 25 years of the international conference on AI and Law. *Artificial Intelligence and Law* **20**, 215–319.

Bhardwaj R., Majumder N. and Poria S. (2021). Investigating gender bias in BERT. *Cognitive Computation* **13**, 1008–1018.

Bolukbasi T., Chang K.-W., Zou J., Saligrama V. and Kalai A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS)*, Red Hook, NY, USA. Curran Associates Inc., pp. 4356–4364.

Branting K.L., Yeh A., Weiss B., Merkhofer E. and Brown B. (2018). Inducing predictive models for decision support in administrative adjudication. In **Pagallo U., Palmirani M., Casanovas P., Sartor G. and Villata S.** (eds), *AI Approaches to the Complexity of Legal Systems*. Springer International Publishing, pp. 465–477.

Brunet M.-E., Alkalay-Houlihan C., Anderson A. and Zemel R. (2019). Understanding the origins of bias in word embeddings. In **Chaudhuri K. and Salakhutdinov R.** (eds), *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, vol. 97. PMLR, pp. 803–811.

Buchanan B.G. and Headrick T.E. (1970). Some speculation about artificial intelligence and legal reasoning. *Stanford Law Review* **23**, 40–62.

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

Cardellino C., Teruel M., Alemany L.A. and Villata S. (2017). A low-cost, high-coverage legal named entity recognizer, classifier and linker. In *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law (ICAIL)*, New York, NY, USA. Association for Computing Machinery, pp. 9–18.

Casanovas P., Pagallo U., Palmirani M. and Sartor G. (eds) (2013). *AI Approaches to the Complexity of Legal Systems (AICOL)*, *Lecture Notes in Computer Science*, vol. **8929**. Belo Horizonte, Brazil: Springer International Publishing.

Chalkidis, I. and Androutsopoulos, I. (2017). A deep learning approach to contract element extraction. In Wyner A.Z. and Casini, G. (eds), *Legal Knowledge and Information Systems - (JURIX): The Thirtieth Annual Conference*, Frontiers in Artificial Intelligence and Applications, vol. 302, Luxembourg. IOS Press, pp. 155–164.

Chalkidis I., Androutsopoulos I. and Michos A. (2017). Extracting contract elements. In *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law (ICAIL)*, New York, NY, USA. Association for Computing Machinery, pp. 19–28.

Chalkidis I., Androutsopoulos I. and Michos A. (2018). Obligation and prohibition extraction using hierarchical RNNs. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Melbourne, Australia. Association for Computational Linguistics, pp. 254–259.

Chalkidis I., Fergadiotis E., Malakasiotis P., Aletras N. and Androutsopoulos I. (2019). Extreme multi-label legal text classification: A case study in EU legislation. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, Minneapolis, Minnesota. Association for Computational Linguistics, pp. 78–87.

Chalkidis I., Fergadiotis M., Malakasiotis P., Aletras N. and Androutsopoulos I. (2020). Legal-bert: 'Ppreparing the muppets for court". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pp. 2898–2904.

Chalkidis I., Jana A., Hartung D., Bommarito M.J., Androutsopoulos I., Katz D.M. and Aletras N. (2021). Lexglue: A benchmark dataset for legal language understanding in English. Available at SSRN 3936759.

Chalkidis I. and Kampas D. (2019). Deep learning in law: Early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law* **27**(2), 171–198.

Church K.W. (2017). Word2vec. *Natural Language Engineering* **23**(1), 155–162.

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

Clark K. and Manning C.D. (2016). Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany. Association for Computational Linguistics, pp. 643–653.

Dale R. (2019). Law and word order: NLP in legal tech. *Natural Language Engineering* **25**(1), 211–217.

De-Arteaga M., **Romanov A.**, **Wallach H.**, **Chayes J.**, **Borgs C.**, **Chouldechova A.**, **Geyik S.**, **Kenthapadi K. and Kalai A.T.** (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, New York, NY, USA. Association for Computing Machinery, pp. 120–128.

Devlin J., **Chang M.-W.**, **Lee K. and Toutanova K.** (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics, pp. 4171–4186.

Dixon L., **Li J.**, **Sorensen J.**, **Thain N. and Vasserman L.** (2018). Measuring and mitigating unintended bias in text classification. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, New York, NY, USA. Association for Computing Machinery, pp. 67–73.

Do P.-K., **Nguyen H.-T.**, **Tran C.-X.**, **Nguyen M.-T. and Nguyen M.-L.** (2017). Legal question answering using ranking svm and deep convolutional neural network. arXiv preprint arXiv:1703.05320.

Dozier C., **Kondadadi R.**, **Light M.**, **Vachher A.**, **Veeramachaneni S. and Wudali R.** (2010). Named entity recognition and resolution in legal text. In *Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language*. Berlin, Heidelberg: Springer-Verlag, pp. 27–43.

Elnaggar A., **Otto R. and Matthes F.** (2018). Deep learning for named-entity linking with transfer learning for legal documents. In *Proceedings of the Artificial Intelligence and Cloud Computing Conference (AICCC)*, New York, NY, USA. Association for Computing Machinery, pp. 23–28.

Evans R., **Piwek P.**, **Cahill L. and Tipper N.** (2008). Natural language processing in CLIME, a multilingual legal advisory system. *Natural Language Engineering* **14**(1), 101–132.

Faruqui M., **Tsvetkov Y.**, **Yogatama D.**, **Dyer**, **C. and Smith**, **N.A.** (2015). Sparse overcomplete word vector representations. 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)*, Beijing, China. Association for Computational Linguistics, pp. 1491–1500.

Francesconi E., **Montemagni S.**, **Peters W. and Tiscornia D.** (eds) (2010). *Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language*, *Lecture Notes in Computer Science*, vol. **6036**. New York, NY: Springer.

Fu R., **Guo J.**, **Qin B.**, **Che W.**, **Wang H. and Liu T.** (2014). Learning semantic hierarchies via word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Baltimore, Maryland. Association for Computational Linguistics, pp. 1199–1209.

Galgani F., **Compton P. and Hoffmann A.** (2012). Combining different summarization techniques for legal text. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data (HYBRID)*, USA. Association for Computational Linguistics, pp. 115–123.

Garg N., **Schiebinger L.**, **Jurafsky D. and Zou J.** (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* **115**(16), 3635–3644.

Gonen H. and Goldberg Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *Computing Research Repository*, arXiv:1903.03862. version 2.

Hafner C.D. and Berman D.H. (2002). The role of context in case-based legal reasoning: Teleological, temporal, and procedural. *Artificial Intelligence and Law* **10**(1–3), 19–64.

Hamilton W.L., **Leskovec J. and Jurafsky D.** (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany. Association for Computational Linguistics, pp. 1489–1501.

Hochreiter S. and Schmidhuber J. (1997). Long short-term memory. *Neural Computation* **9**(8), 1735–1780.

Joshi M., **Levy O.**, **Zettlemoyer L. and Weld D.** (2019). BERT for coreference resolution: Baselines and analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China. Association for Computational Linguistics, pp. 5803–5808.

Joulin A., **Grave E.**, **Bojanowski P. and Mikolov T.** (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Valencia, Spain. Association for Computational Linguistics, pp. 427–431.

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

Katz D.M., **Bommarito M.J. and Blackman J.** (2017). A general approach for predicting the behavior of the Supreme Court of the United States. *PLOS ONE* **12**(4), 1–18.

Kim M.-Y., Xu Y. and Goebel R. (2017). Applying a convolutional neural network to legal question answering. In Otake M., Kurahashi S., Ota Y., Satoh K. and Bekki D. (eds), *New Frontiers in Artificial Intelligence*. Springer International Publishing, pp. 282–294.

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

Kurita K., Vyas N., Pareek A., Black, A.W. and Tsvetkov Y. (2019). Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, Florence, Italy. Association for Computational Linguistics, pp. 166–172.

Kusner M.J., Loftus J., Russell C. and Silva R. (2017). Counterfactual fairness. In Guyon I., Luxburg U.V., Bengio S., Wallach, H., Fergus R., Vishwanathan S. and Garnett, R. (eds), *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., pp. 4066–4076.

Lai S., Xu L., Liu K. and Zhao J. (2015). Recurrent convolutional neural networks for text classification. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI Press, pp. 2267–2273.

Leitner E., Rehm G. and Moreno-Schneider J. (2019). Fine-grained named entity recognition in legal documents. In *International Conference on Semantic Systems*. Springer, pp. 272–287.

Liang P.P., Li I.M., Zheng E., Lim Y.C., Salakhutdinov R. and Morency L.-P. (2020). Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics, pp. 5502–5515.

Locke D. and Zuccon G. (2019). Towards automatically classifying case law citation treatment using neural networks. In *Proceedings of the 24th Australasian Document Computing Symposium (ADCS)*, New York, NY, USA. Association for Computing Machinery.

Long S., Tu C., Liu Z. and Sun M. (2019). Automatic judgment prediction via legal reading comprehension. In Sun M., Huang X., Ji H., Liu Z. and Liu Y. (eds), *Chinese Computational Linguistics (CCL)*, Cham. Springer International Publishing, pp. 558–572.

Luz de Araujo P.H., de Campos T.E., de Oliveira R. R.R., Stauffer M., Couto S. and Bermejo P. (2018). LeNER-Br: A dataset for named entity recognition in Brazilian legal text. In *International Conference on the Computational Processing of Portuguese (PROPOR)*, Lecture Notes on Computer Science (LNCS), Canela, RS, Brazil. Springer, pp. 313–323.

Manzini T., Yao Chong L., Black A.W. and Tsvetkov Y. (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics, pp. 615–621.

Martin A.D., Quinn K.M., Ruger T.W. and Kim P.T. (2004). Competing approaches to predicting Supreme Court decision making. *Perspectives on Politics* **2**(4), 761–767.

May C., Wang A., Bordia S., Bowman S.R., and **Rudinger R.** (2019). On measuring social biases in sentence encoders. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics, pp. 622–628.

Medvedeva M., Vols M. and Wieling M. (2020). Using machine learning to predict decisions of the European Court of Human Rights. *Artificial Intelligence and Law* **28**(2), 237–266.

Mikolov T., Chen K., Corrado G. and Dean J. (2013a). Efficient estimation of word representations in vector space. In Bengio Y. and LeCun Y. (eds), *1st International Conference on Learning Representations (ICLR), Workshop Track Proceedings*, Scottsdale, Arizona, USA.

Mikolov T., Sutskever I., Chen K., Corrado G. and Dean J. (2013b). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS) - Volume 2*, Red Hook, NY, USA. Curran Associates Inc., pp. 3111–3119.

Mohammad S., Bravo-Marquez F., Salameh M. and Kiritchenko S. (2018). SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, New Orleans, Louisiana. Association for Computational Linguistics, pp. 1–17.

Morimoto A., Kubo D., Sato M., Shindo H. and Matsumoto Y. (2017). Legal question answering system using neural attention. In Satoh K., Kim M., Kano Y., Goebel R. and Oliveira T. (eds), *4th Competition on Legal Information Extraction and Entailment (COLIEE), held in conjunction with the 16th International Conference on Artificial Intelligence and Law (ICAIL) in King's College London, UK*, EPiC Series in Computing, vol. 47. EasyChair, pp. 79–89.

Mumcuoğlu E., Öztürk C.E., Ozaktas H.M. and Koç A. (2021). Natural language processing in law: Prediction of outcomes in the higher courts of Turkey. *Information Processing & Management* **58**(5), 102684.

Murphy B., Talukdar P. and Mitchell T. (2012). Learning effective and interpretable semantic models using non-negative sparse embedding. In *Proceedings of COLING*, Mumbai, India. The COLING 2012 Organizing Committee, pp. 1933–1950.

Nanda R., John A.K., Caro L.D., Boella G. and Robaldo L. (2017). Legal information retrieval using topic clustering and neural networks. In Satoh K., Kim M.-Y., Kano Y., Goebel R. and Oliveira T. (eds), *4th Competition on Legal Information Extraction and Entailment (COLIEE)*, EPiC Series in Computing, vol. 47. EasyChair, pp. 68–78.

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

Navigli R. and Martelli F. (2019). An overview of word and sense similarity. *Natural Language Engineering* **25**(6), 693–714.

Nejadgholi I., Bougueng R. and Witherspoon S. (2017). A semi-supervised training method for semantic search of legal facts in Canadian immigration cases. In Wyner, A.Z. and Casini G. (eds), *Legal Knowledge and Information Systems - (JURIX): The Thirtieth Annual Conference, Luxembourg, 13–15 December 2017*, Frontiers in Artificial Intelligence and Applications, vol. 302. IOS Press, pp. 125–134.

Nguyen T.-S., Nguyen L.-M., Tojo S., Satoh K. and Shimazu A. (2018). Recurrent neural network-based models for recognizing requisite and effectuation parts in legal texts. *Artificial Intelligence and Law* **26**(2), 169–199.

O'Neill J., Buitelaar P., Robin C. and O'Brien L. (2017). Classifying sentential modality in legal language: A use case in financial regulations, acts and directives. In *Proceedings of the 16th Edition of the International Conference on Artificial Intelligence and Law (ICAIL)*, New York, NY, USA. Association for Computing Machinery, pp. 159–168.

O'Sullivan C. and Beel J. (2019). Predicting the outcome of judicial decisions made by the european court of human rights. In *In Proceedings of the 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science*, Dublin, Ireland.

Pennington J., Socher R. and Manning C.D. (2014). Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543.

Perez C.C. (2019). *Invisible Women: Exposing Data Bias in a World Designed for Men*. Penguin Random House, South Africa.

Peters M., Neumann M., Iyyer M., Gardner M., Clark C., Lee K. and Zettlemoyer L. (2018). Deep contextualized word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana. Association for Computational Linguistics, pp. 2227–2237.

Pittaras N., Giannakopoulos G., Papadakis G. and Karkaletsis V. (2020). Text classification with semantically enriched word embeddings. *Natural Language Engineering* **27**(4), 391–425.

Prost F., Thain N. and Bolukbasi T. (2019). Debiasing embeddings for reduced gender bias in text classification. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, Florence, Italy. Association for Computational Linguistics, pp. 69–75.

Radford A., Wu J., Child R., Luan D., Amodei D. and Sutskever I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog* **1**(8), 9.

Rudinger R., Naradowsky J., Leonard B. and Van Durme B. (2018). Gender bias in coreference resolution. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana. Association for Computational Linguistics.

Ruger T., Kim P., Martin A. and Quinn K. (2004). The Supreme Court forecasting project: Legal and political science approaches to predicting Supreme Court decisionmaking. *Columbia Law Review* **104**, 1150–1210.

Sangeetha D., Kavyashri R., Swetha S. and Vignesh S. (2017). Information retrieval system for laws. In *2016 Eighth International Conference on Advanced Computing (ICoAC)*, pp. 212–217.

Sartor G. and Rotolo A. (2013). *Agreement Technologies*, Chapter AI and Law. New York: Springer, pp. 199–207.

Senel L.K., Utlu I., Şahinuç F., Ozaktas H.M. and Koç A. (2020). Imparting interpretability to word embeddings while preserving semantic structure. *Natural Language Engineering* **27**(6), 721–746.

Shulayeva O., Siddharthan A. and Wyner A. (2017). Recognizing cited facts and principles in legal judgements. *Artificial Intelligence and Law* 25(1), 107–126. Open access via Springer Compact Agreement.

Sleimi A., Sannier N., Sabetzadeh M., Briand L. and Dann J. (2018). Automated extraction of semantic legal metadata using natural language processing. In *IEEE 26th International Requirements Engineering Conference (RE)*. IEEE, pp. 124–135.

Soh J., Lim H.K. and Chai I.E. (2019). Legal area classification: A comparative study of text classifiers on Singapore Supreme Court judgments. In *Proceedings of the Natural Legal Language Processing Workshop*, Minneapolis, Minnesota. Association for Computational Linguistics, pp. 67–77.

Stanovsky G., Smith N.A. and Zettlemoyer L. (2019). Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics, pp. 1679–1684.

Şulea O.-M., Zampieri M., Vela M. and van Genabith J. (2017). Predicting the law area and decisions of French Supreme Court cases. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP*, Varna, Bulgaria. INCOMA Ltd., pp. 716–722.

Tan Y.C. and Celis L.E. (2019). Assessing social and intersectional biases in contextualized word representations. In Wallach H., Larochelle H., Beygelzimer A., d'Alché Buc F., Fox E. and Garnett R. (eds), *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., pp. 13230–13241.

Tanaka-Ishii K. (2007). Word-based predictive text entry using adaptive language models. *Natural Language Engineering* **13**(1), 51–74.

Tang D., Wei F., Yang N., Zhou M., Liu T. and Qin B. (2014). Learning sentiment-specific word embedding for Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Baltimore, Maryland. Association for Computational Linguistics, pp. 1555–1565.

Tang G., Guo H., Guo Z. and Xu S. (2016). Matching law cases and reference law provision with a neural attention model. In *IBM China Research*, Beijing.

Tezcan A., Hoste V. and Macken L. (2020). Estimating word-level quality of statistical machine translation output using monolingual information alone. *Natural Language Engineering* **26**(1), 73–94.

https://doi.org/10.1017/S1351324922000122 Published online by Cambridge University Press

Tjong Kim Sang E.F. and De Meulder F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL*, pp. 142–147.

Üstün A. and Can B. (2020). Incorporating word embeddings in unsupervised morphological segmentation. *Natural Language Engineering* **27**(5), 609–629.

Vardhan H., Surana N. and Tripathy B. (2020). Named-entity recognition for legal documents. In *International Conference on Advanced Machine Learning Technologies and Applications*. Springer, pp. 469–479.

Virtucio M.B.L., Aborot J.A., Abonita J.K.C., Aviñante R.S., Copino, R. J. B., Neverida M.P., Osiana V.O., Peramo E.C., Syjuco J.G. and Tan G.B.A. (2018). Predicting decisions of the Philippine Supreme Court using natural language processing and machine learning. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, vol. 02, pp. 130–135.

Vo N.P.A., Privault C. and Guillot F. (2017). Experimenting word embeddings in assisting legal review. In *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law (ICAIL)*, New York, NY, USA. Association for Computing Machinery, pp. 189–198.

Zhang B.H., Lemoine B. and Mitchell M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES'18, New York, NY, USA. Association for Computing Machinery, pp. 335–340.

Zhao J., Wang T., Yatskar M., Cotterell R., Ordonez V. and Chang K.-W. (2019). Gender bias in contextualized word embeddings. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics, pp. 629–634.

Zhao J., Wang T., Yatskar M., Ordonez V. and Chang K.-W. (2018a). Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana, USA, pp. 15–20.

Zhao J., Wang T., Yatskar M., Ordonez V. and Chang K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark. Association for Computational Linguistics, pp. 2979–2989.

Zhao J., Zhou Y., Li Z., Wang W. and Chang K.-W. (2018b). Learning gender-neutral word embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics, pp. 4847–4853.

Zou J. and Schiebinger L. (2018). AI can be sexist and racist — it's time to make it fair. *Nature* **559**, 324–326.

Appendix A.

CriList – Crime related words				
burglar	thief	kidnapping	insult	violence
abuse	accomplice	assault	alias	armed
attack	arson	blackmail	bombing	breaking
brutality	burglary	cheat	bribery	damage
combat	complication	conspiracy	convict	corruption
crime	criminal	custody	sexual	danger
deliberate	disobedience	dangerous	deadly	death
disregard	disruption	drug	enforce	escape
felony	forbidden	fraud	guilty	harassment
evil	explosive	fatality	eviction	kill
harm	harmful	hijack	homicide	illegal
innocent	jail			

Cite this article: Sevim N, Şahinuç F and Koç A (2023). Gender bias in legal corpora and debiasing it. *Natural Language Engineering* **29**, 449–482. https://doi.org/10.1017/S1351324922000122