ANALYSIS OF GENDER BIAS IN LEGAL TEXTS USING NATURAL LANGUAGE PROCESSING METHODS

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

ELECTRICAL AND ELECRTONICS ENGINEERING

By

Nurullah Sevim

July 2023

Analysis of Gender Bias in Legal Texts Using Natural Language Processing Methods
By Nurullah Sevim
July 2023

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Aykut Koç(Advisor)

Fazlı Can

Gözde Gül Şahin

Approved for the Graduate School of Engineering and Science:

Orhan Arıkan
Director of the Graduate School

ABSTRACT

ANALYSIS OF GENDER BIAS IN LEGAL TEXTS USING NATURAL LANGUAGE PROCESSING METHODS

Nurullah Sevim
M.S. in Electrical and Elecrtonics Engineering
Advisor: Aykut Koç
July 2023

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. Furthermore, transformer-based contextualized language models constitute the state-of-the-art in several natural language processing (NLP) tasks and applications. Despite their utility, contextualized models can contain human-like social biases as their training corpora generally consist of human-generated text. Evaluating and removing social biases in NLP models have been an ongoing and prominent research endeavor. In parallel, the NLP approaches in the legal area, namely legal NLP or computational law, have also been increasing recently. Eliminating unwanted bias in the legal domain is doubly crucial since the law has the utmost importance and effect on people. We approach the gender bias problem from the scope of legal text processing domain. In the first stage of our study, we focus on the gender bias in traditional word embeddings, like Word2Vec and GloVe. Word embedding models which 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. In the second stage, we study the gender bias encoded in BERT-based models. We propose a new template-based bias measurement method with a bias evaluation corpus using crime words from the FBI database. This method quantifies the gender bias present in BERT-based models for legal applications. Furthermore, we propose a fine-tuning-based debiasing method using the

European Court of Human Rights (ECtHR) corpus to debias legal pre-trained models. We test the debiased models on the LexGLUE benchmark to confirm that the underlying semantic vector space is not perturbed during the debiasing process. Finally, overall results and their implications have been discussed in the scope of NLP in legal domain.

Keywords: Natural Language Processing, Word Embeddings, Transformer, Gender Bias.

ÖZET

HUKUKİ METİNLERDEKİ CİNSİYETÇİ ÖNYARGININ DOĞAL DİL İŞLEME METOTLARIYLA ANALİZİ

Nurullah Sevim
Elektrik-Elektronik Mühendisliği, Yüksek Lisans
Tez Danışmanı: Aykut Koç
Temmuz 2023

Kelime temsilleri, doğal dil işlemede (NLP) sıkça kullanılan önemli yapı taşları haline gelmiştir. Çeşitli avantajlarına rağmen, kelime temsilleri, üzerinde eğitildikleri derlemlerde bulunan bazı cinsiyet ve etnik kökene dayalı önyargıları istemeden barındırabilir. Kelime temsilleri birçok üst düzey algoritmada yaygın olarak kullanıldığından etik kaygılar ortaya çıkmıştır. Ayrıca, dönüştürücü tabanlı bağlamsallaştırılmış dil modelleri, çeşitli doğal dil işleme (NLP) görevlerinde ve uygulamalarında en son teknolojiyi oluşturmaktadır. Yararlarına rağmen, bağlamsallaştırılmış modeller, eğitim derlemleri genellikle insan tarafından üretilen metinlerden oluştuğu için insan benzeri sosyal önyargılar içerebilir. NLP modellerinde sosyal önyargıları değerlendirmek ve ortadan kaldırmak, devam eden ve öne çıkan bir araştırma çabası olmuştur. Buna paralel olarak, hukuk alanındaki NLP yaklaşımları, yani yasal NLP veya hesaplama hukuku da son zamanlarda sıkça çalışılan alanlardır. Hukuk alanında istenmeyen önyargıların ortadan kaldırılması iki kat önemlidir, çünkü hukukun insanlar üzerindeki önemi ve etkisi en üst düzeydedir. Cinsiyet yanlılığı sorununa yasal metin işleme alanı kapsamında yaklaşmaktayız. Çalışmamızın ilk aşamasında, Word2Vec ve GloVe gibi geleneksel kelime yerleştirmelerdeki cinsiyet yanlılığına odaklandık. Yasal belgelerdeki cinsiyet yanlılığını ölçmek ve ortadan kaldırmak için farklı ülkelerdeki yasal belgeler ve mevzuattan oluşan derlemler üzerinde eğitilen kelime temsilleri modelleri kullanılmıştır. Cinsiyet önyargısının derecesini ortaya çıkarmak ve ülkeler arasındaki değişimini gözlemlemek için çeşitli yöntemler kullanılmıştır. Ayrıca, istenmeyen önyargıyı etkisiz hale getirmek için bir önyargısızlaştırma yöntemi kullanılmıştır. Saptırılmış vektör uzayının anlamsal tutarlılığının korunduğu, üst düzey görevler kullanılarak da gösterilmiştir. İkinci aşamada, BERT tabanlı modellerde kodlanan cinsiyet yanlılığını inceledik. FBI veri tabanından suç sözcükleri kullanan bir önyargı değerlendirme külliyatı ile şablon tabanlı yeni bir önyargı ölçüm yöntemi önerdik. Bu

yöntem, yasal uygulamalar için BERT tabanlı modellerde bulunan cinsiyet yanlılığını ölçmektedir. Ayrıca, yasal olarak önceden eğitilmiş modellerin önyargısını gidermek için Avrupa İnsan Hakları Mahkemesi (AİHM) külliyatını kullanan ince ayara dayalı bir önyargı azaltma yöntemi önerdik. Önyargısızlaştırma işlemi sırasında altta yatan semantik vektör uzayının bozulmadığını doğrulamak için LexGLUE kıyaslamasında önyargısız modelleri test ettik. Son olarak, genel sonuçlar ve bunların yansımaları, yasal alanda NLP kapsamında tartışılmıştır.

Anahtar sözcükler: Doğal Dil İşleme, Kelime Temsilleri, Dönüştürücüler, Cinsiyetçi Önyargı.

Acknowledgement

I would like to express my deepest gratitude to my advisor, Dr. Aykut Koç, for his invaluable guidance, support, and expertise throughout the entire process of this research endeavor. Their dedication, patience, and insightful feedback have been instrumental in shaping this thesis and enhancing its quality. I am truly fortunate to have had the opportunity to work under his supervision, benefiting from his vast knowledge and experience in the field.

I am also grateful to my thesis committee members Prof. Dr. Fazlı Can and Dr. Gözde Gül Şahin for kindly accepting to review this thesis.

I would like to express my heartfelt gratitude to my friends and family who have been unwavering pillars of support throughout my academic journey. I want to extend a special thank you to Ahmet Akın, Abdullah Başaran, Doğan Can Çiçek, Furkan Burak Mutlu, Emre Uludağ, Enes Duran, Kağan Kaya, Alperen Tavşancı, Akın Aydoğan, Alperen Nazım Demir, Altan Akat, Arda Can Aras, Mustafa Bozdağ, Tuna Alikaşifoğlu, Emirhan Koç, Berk Şimşekler, Bünyamin Kartal, Ceyhun Emre Öztürk, Ege Zeybek, Furkan Şahinuç, Emre Mumcuoğlu, Enes Bilgin, Enes Uğur, Faruk Gülmez, Halil İbrahim Şener, Kerem Neşet Sarıtepe, Mehmet Ali Tosun, Mahmut Sefa Ergündüz, Oğuz Çetinkaya, Ömer Faruk Kayar, Ömer Faruk Ünal, Tarık Yunus Köroğlu, Tolga Talha Yıldız, Tuğçe Çabuk, and Yusuf İskenderoğlu, whose constant encouragement and unwavering belief in my abilities have been instrumental in my success. Their words of wisdom, motivation, and relentless support have helped me overcome the challenges I encountered during this research endeavor. I also want to acknowledge the love and support of my family whose unwavering faith in me has been a constant source of inspiration. Without their love, understanding, and sacrifices, this achievement would not have been possible. Their presence in my life has made this journey all the more meaningful, and I am truly blessed to have such amazing friends and family.

I would like to thank Vodafone for the support that they provided under the *5G and Beyond Graduate Scholarship Programme*.

I would like to thank TUBITAK for financially supporting me during my Master's Degree with *TÜBİTAK 2211 National Graduate Scholarship Programme*.

This work is supported by TUBITAK 1001 grant (120E346).

Contents

List of Figures

List of Tables

List of Publications

This thesis includes content from following publication(s):

1. N. Sevim, F. Şahinuç, and A. Koç, “Gender bias in legal corpora and debiasing it,” *Natural Language Engineering*, vol. 29, no. 2, pp. 449–482, 2023.

Chapter 1

Introduction

1.1 Preliminaries

Word embeddings that map words to vectors in multi-dimensional semantic vector spaces are widely used as underlying word representations in various architectures to tackle natural language processing (NLP) tasks [5, 6, 7]. The introduction of Word2Vec and GloVe models and later improvements on these models increased the popularity of word embeddings [5, 8, 9, 10, 11].

Latterly, the emergence of contextualized models, such as ELMo, GPT2, and BERT, has initiated another stage for NLP where the problem mentioned above is greatly overcome [12, 13, 14, 15]. These models offer word representations such that they assign each word a representation based on context, capturing different uses of words in the process [13, 14, 15, 16]. As they handle polysemy better, contextualized embeddings achieved groundbreaking performances in several NLP tasks and surpassed the static word embeddings [13, 14, 15, 17, 16]. BERT has the most significant number of variants among these models, which are popular pre-trained models themselves [15, 16]. Variations of BERT and BERT-based models have been developed to improve the original model, decrease the computational cost, and achieve better performance in domain-specific tasks [18, 19, 20, 17, 21, 22].

Despite their advantages for many applications, word representation models can contain and reflect human-like biases on race, gender, and religion. As a result, the ethical aspects of their usage in real-life NLP applications became a concern [23, 24, 25, 26, 27, 28, 29, 30]. The existence of human-like biases in static word embeddings are proven in many studies, and the mitigation of these biases is an ongoing research endeavor. Bolukbasi et al. [23] showed that word embeddings can inherit predispositions for genders in professions, with the example of a computer programmer being perceived as a male occupation. At the same time, the homemaker is linked to the female gender in embeddings generated with Glove, and Word2Vec [31, 32]. Similarly, in the work of Manzini et al. [27], the racial and religious biases encoded in Word2Vec [31] are shown with word analogies such as *black-homeless*, *Muslim-uneducated*, and *Asian-laborer*. Following these works, the human-like biases in contextualized language models are analyzed in various other studies [2, 26, 28, 33, 34, 35, 1, 36, 4]. Zhao et al. [26] evaluated the gender bias exhibited in ELMo [13] by showing the gender imbalance in the training data, applying principal component analysis (PCA) to construct a geometric gender subspace, and analyzing the gender information encoded in the word embeddings by projecting them onto the constructed gender subspace. In the study of Kurita et al. [35], the gender bias encoded in BERT is measured by calculating the association between gender-specific nouns-pronouns and gender-neutral professions. Along with the quantification of bias in contextualized language models, different approaches for debiasing these embeddings are also proposed [2, 33, 35, 37, 1, 36, 4]. Similar to debiasing traditional word embeddings, projection-based methods are also applied on contextualized representations [33, 36, 4] along with other approaches utilizing the fine-tuning capabilities of pre-trained embeddings [2, 37, 1].

Concurrently, a significant area of study on using NLP and machine learning to process legal documents and create decision support systems for the legal domain [38, 39, 40, 41, 42] has also gained traction. Law relies on almost only written text, making it a prominent candidate for real-life applications of popular NLP models [43, 44, 45, 46, 47, 48, 49, 50, 51, 52]. A significant problem for legal practitioners is the vastly expanding amount of case files and the time-consuming nature of handling these matters manually. Utilizing automated tools is now necessary to reduce human

error, improve legal services, and ease the burden on legal practitioners in many legal tasks. To this end, the relationship between law and artificial intelligence dates back several decades, with the first ideas originating in the 1970s and the emergence of a thriving community in 1987 [53, 54]. Since then, a number of natural language processing (NLP) methods have been created expressly with the intention of processing and analyzing legal texts [55, 38, 41, 56, 57]. By using word and n-gram frequencies as features, conventional machine learning approaches are being used in the field of law for complex NLP applications. To address the issue of categorizing judicial cases, a study by Katz et al. [41] used a random forest-based algorithm trained on a corpus of US Supreme Court documents. Support Vector Machines (SVMs) have been furthermore demonstrated as a very efficient method for executing NLP tasks in the legal field [56]. NLP also makes use of important machine learning methods like long-short term memory (LSTMs) networks [58], which function best when paired with word vector representations [57]. High performance on several difficult downstream tasks has been achieved as a consequence of combining these potent machine learning algorithms with NLP approaches in legal domain. Predicting decisions of courts [59, 60, 61, 44], classifying case documents [62], and information retrieval [63] 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 [64]. 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 [65]. Furthermore, contextualized language models have also started to contribute to high-level legal tasks lately, and even contextualized models specific to the legal domain have been designed [17, 66]. Very recently, legal NLP efforts have been amalgamated by [3] in LexGLUE by introducing several legal corpora and standardized benchmarks for the field.

One of the most significant areas where NLP is applied, law has a significant influence on people's lives in many different ways. Ethics issues about prejudice and fairness have taken on more significance in light of the importance of law and the potential repercussions of legal procedures. Identifying biases in juridical language and

creating strategies to counteract them are essential steps in creating impartial and bias-free NLP applications for law. The existence of social bias in legal documents, in the form of either ethnicity or gender, has become evident as many recent studies have offered proofs [67, 68, 69]. Ash et al. [67] claimed that the judges have *gender slant* which affect their decisions. They provided statistical and NLP-based proofs for their claim.

1.2 Contributions

In this thesis, we study gender bias in legal NLP, first on static word embeddings then on the popular contextualized language models. In the first stage where we focus on the static legal word embeddings, 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 sub-corpora 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 for static word embeddings 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 [70], 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 baseline methods that we use have been introduced in [71], 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 [72]. 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.

In the second stage, we study gender bias in legal contextualized language models by measuring the encoded bias in the model and eliminating it. Although there is research on bias in contextualized models, to the best of our knowledge, no study focuses on the legal domain for that purpose. In this respect, we propose a bias measuring method that exploits the masked language model (MLM) feature of the BERT architecture and a debiasing method for legal contextualized models named Legal-Context-Debias (LCD), which utilizes the fine-tuning feature of the contextualized models [15].

For the proposed bias measuring algorithm, we introduce a new bias evaluation

corpus compiled to measure gender bias, specifically in legal models, called BEC-Cri. The BEC-Cri dataset contains sample sentences with crime words from the FBI database [73]. Since our methods are domain-specific, BEC-Cri provides more proper evaluation scores of bias measurement. During our proposed debiasing method, on the other hand, we implement a binary classification task as a fine-tuning process of the legal contextualized model. For this task, we utilize the ECtHR corpus, introduced in the study of Chalkidis et al. [74]. We modified the ECtHR corpus to contain an equal number of male and female applicant cases. Since the ECtHR corpus contains legal documents, the proposed method provides a convenient framework for debiasing legal contextualized models.

We also implemented an alternative domain-general bias measuring algorithm to compare the measurement scores with our proposed method and discussed the advantages of having a domain-specific method for bias analysis. The alternative method uses the BEC-Pro dataset proposed by Bartl et al. [1]. Furthermore, we also implemented two alternative baseline debiasing methods to demonstrate the benefit of having a debiasing method specific to the legal domain. The first alternative method named Gender Preserving Debiasing (GPD) proposed by Kaneko and Bollegala [33] uses a similarity-based loss term to debias the model along with a regularizing loss to preserve the gender information encoded in the model. The second alternative baseline relies on fine-tuning the contextualized model with a coreference resolution task. This baseline utilizes the Gendered Ambiguous Pronouns (GAP) corpus proposed by Webster et al. [2]. By comparing our proposed methods with these general-purpose methods, we demonstrate that designing algorithms for specific domains is crucial to study the effects of gender-bias in contextualized models.

In our experiments, we utilized the legal-specific contextualized language model LegalBERT-Small proposed in the work of Chalkidis et al. [17]. The gender bias in the LegalBERT-Small is measured pre- and post-implementations of the debiasing methods, and the bias measurements after implementing these methods are compared to observe the effect of debiasing algorithms. The measurements are also done with our proposed legal-specific bias measurement algorithm and the other general-purpose bias

evaluation method. Then, the debiased models are evaluated on LexGLUE, the common legal language understanding benchmark, to ensure that the semantic information embedded in the model is preserved after implementing the debiasing algorithms [3]. Finally, a novel framework is proposed where the biased and debiased models' LexGLUE performances are recalculated after applying a penalty based on their bias scores. In this newly proposed framework, the goal is to determine the degree of bias in comparison with the model performances.

The contributions of this thesis 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 proposed a bias measuring method, called BEC-Cri, for contextualized language models which are specialized in legal context.
- We developed a novel debiasing algorithm Legal-Context-Debias for BERT-based models.
- We proposed a framework where the models are penalized in their LexGLUE performances according to how much bias they contain.
- To the best of our knowledge, this study is the first that considers the bias issue in legal contextualized language models, such as LegalBERT.

Chapter 2

Literature Review

2.1 Bias in NLP and Machine Learning

Recent research has shown that social, racial, and gender biases may be inherited by machine learning algorithms and the downstream applications that employ them [75, 76, 77]. Zou and Schiebinger [78] identified various reasons behind the discriminatory behavior of AI algorithms towards some sub-populations in society. One of the reasons they pointed out is the biased perspectives of human annotators who label the training data used by the algorithms. Another reason is the lack of diversity in the data, which can lead to the algorithms making incorrect assumptions about certain groups. For example, computer vision algorithms may label a traditional US wedding with tags such as 'bride' and 'woman', but label a North Indian wedding with 'performance art' and 'costume', which can reflect cultural bias. In the study of Zhao et al. [79], the issue of gender bias is analyzed through a label-tagging task. The researchers discovered that there is considerable gender bias present in the datasets utilized for these tasks, with some activities having a higher likelihood of involving females than males. Additionally, models developed using these skewed datasets have a tendency to exacerbate the preexisting bias, producing even more discrepancies in the outcomes. For instance, it is revealed that cooking was almost 33% more likely to include women than men and a model trained on this dataset increased the gender gap to 68% at test

time. Stanovsky et al. [80] investigated the social bias in machine translation (MT) softwares. Based on morphological analysis, the authors created a system for automatically evaluating gender bias in eight target languages having grammatical gender. The study discovered that gender-biased translation mistakes are common in all examined target languages for four well-known industrial MT systems and two contemporary, cutting-edge academic MT models. Church et al. [81] analyzed the recent developments on fair machine learning and outlined the criticism against fairness of AI. They identified and addressed the additional threats to public health, safety, and security brought on by bias in machine learning.

By using Word Embedding Association Test (WEAT), inspired by the Implicit Association Test (IAT) [82], Caliskan et al. [24] demonstrated that the training corpora in NLP algorithms itself carry such biases. They showed that biases comparable to those held by humans may be produced when machine learning is applied to human language. Their findings imply that biases from the past are stored precisely and retrievably in text corpora, regardless of how morally neutral (such as biases toward insects or flowers) or problematic (such as biases toward race or gender) they are or whether they simply reflect the current gender distributions in relation to careers or first names. In WEAT, the cosine similarity between gender-neutral attribute words (e.g., professions, emotions) and gender-specific target words (gendered nouns and pronouns) are calculated and compared. To further improve, May et al. [83] used WEAT on pooled sentence representations, resulting in the Sentence Encoder Association Test (SEAT), which revealed a degree of social bias when applied to popular sentence encoders.

Similar approaches also exist for contextualized word embeddings where a gender subspace is constructed using the vector representations of gender-specific words (e.g., he, she, mom, dad) and projecting the word embeddings onto these subspaces [33, 36, 4]. However, these approaches cause problems for contextualized models such as BERT [15] as contextualized models generate different embeddings for a word depending on the context, which makes cosine similarity and metrics based on the distance between vectors not suitable for them. Thus, to measure bias in BERT, model-specific approaches are exploited through various methods [33, 4, 35, 1]. In the work of Kurita et al. [35], for example, the masked language model (MLM) feature of BERT

is utilized to measure gender bias in BERT-like contextual language models. By comparing their method to the cosine similarity-based approach, they demonstrated how it is more in line with human biases. Additionally, their technique may reveal a wide range of biases in the model when subjected to different signals.

Bias measurement is primarily performed to detect and quantify encapsulated biases in a model and construct a metric for debiasing. Various methods have been proposed to mitigate the measured gender bias in NLP models. Bolukbasi et al. [23] used two different projection-based methods for debiasing type-level word embeddings. This approach is also unsuitable for contextualized word embeddings due to similar problems in traditional bias measurement methods. The projection-based methods only succeed if each word has a global representation. Furthermore, Gonen et al. [34] also claimed that these methods do not remove the encoded bias even in global embeddings, as the information is still reflected in the distances between the debiased words using K-Nearest Neigbors algorithm. They also exploited unsupervised K-Means algorithm to show the biased beahviour is still hidden in the static word embeddings even after using the existing debiasing methods. This problem is crucial for contextualized models since they learn the meaning of a word depending on the sentence in which the word exists; thence, making such methods unfeasible for contextualized models.

A common approach to eliminate bias in BERT is using the fine-tuning capability of the model. This feature of BERT allows the user to fine-tune a pre-trained model on gender-neutral or gender-swapped data to mitigate the predisposition of the model [2, 37]. In the study of Webster et al. [2], the Gendered Ambiguous Pronouns (GAP) dataset is proposed for this purpose. The GAP dataset consists of 8,908 ambiguous pronoun-name pairs, forming a gender-balanced corpus that can be used for fine-tuning to balance the gender information encoded in the model.

2.2 Legal NLP

There is a long-standing relationship between AI and law. The notion was initially conceived by Buchanan and Headrick [53] in the 1970s, but it was not put into practice

until the late 1980s. The first International Conference on AI and Law took place in 1987. The overview of the early years of the discipline was portrayed in the survey of Bench-Capon et al. in detail [54]. The original research largely focused on creating and applying knowledge bases and leveraging the logical frameworks of legal discussions. It was first introduced and later improved upon a method known as case-based reasoning (CBR), which intended to function with the use of prior case information [55, 84, 85, 86]. Later, court case prediction and evaluation systems were created using rule-based algorithms [38, 40]. Recently, series of articles are published surveying the relationship of AI and law in the last thirty years [87, 88, 89, 90, 91].

With the advancements in machine learning and NLP, the focus on the applications of NLP in law has been increasing as well. In the work of Galgani et al. [92], automatic summarization of legal case reports is done using different machine learning based algorithms to ease the analysis of legal repositories. Bach et al. [93] introduced a new task, learning logical structures of paragraphs in legal text, which aims to extract the logical part of sentences in legal documents. Furthermore, studies on feature extraction and information retrieval from legal texts were also conducted. They proposed a two phased structure where in the first phase the logical part is recognized by Conditional Random Field (CRF) approach and in the second phase the logical parts are grouped through a graph-based method. They reported promising results on Japanese National Pension Law corpus where they outperformed the baselines by a significant margin. Nguyen et al. [94] used recurrent neural network-based (RNN) models to extract requisite and effectuation (RE) parts in legal texts. They proposed an improvement on BiLSTM-CRF model, which is introduced previously in [95], and they claimed to have a better model in recognizing RE parts. Besides, they incorporated in-domain word embeddings to further improve the model performance. Feature extraction of legal texts was also performed using the named-entity recognition (NER) [96, 42, 97, 98]. Chalkidis et al. [96] published a labeled dataset with gold contract element annotations for contract element extraction task and employed several methods such as logistic regression and SVMs on this dataset. They compare the effectiveness of these methods using features like word embeddings and part-of-speech tag embeddings. Eventually, they reported that the most successful method is a hybrid approach that combines machine learning with manual post-processing rules. Elnaggar et al.

[97] combined transfer learning with named-entity linking (NEL) to see if the approach was suitable for legal documents, where they reported consistent improvement in European Union law corpus and proved that transfer learning may be a remedy for data scarcity problem in legal domain. Additionally, legal text classification is another area where NLP methods based on artificial neural networks and word embeddings are utilized [99]. Soh et al. [100] compared the classification performances of methods such as SVMs and BERT on court decisions. In their study, they used a new dataset of 6,227 Singapore Supreme Court judgments to compare the effectiveness of state-of-the-art NLP methods with traditional statistical models in analyzing a legal corpus. They reached the conclusion that all tested approaches, including topic model, word embedding, and language model-based classifiers, performed well even with as little as a few hundred judgments.

One of the major drawbacks in legal NLP was the lack of standardized legal corpora and benchmarks to consolidate this fastly developing field. Recently, different datasets containing legal texts such as court cases, contracts, terms of service agreements, and legislation documents became available, paving the way for better models and methods to be developed for the legal domain [56, 41, 101, 102, 103, 104, 105, 106, 107, 108]. In addition, Chalkidis et al. [3] constructed a collection from available datasets to create the LexGLUE benchmark, to evaluate the performance of NLP models on legal natural language understanding (NLU) tasks.

The feature representation of legal texts in the form of word embeddings is a newly studied research area with only a few implementations available in the literature. In one of such examples, O'Sullivan and Beel [109] incorporated documents from the European Court of Human Rights with Word2Vec. Chalkidis et al. [110] also followed a similar process to create Law2Vec, a legal word embedding model based on Mikolov et al.'s seminal work [31]. Alongside the traditional embeddings, contextualized word embeddings have a great potential to offer as they can capture the semantic information along with the context. This feature of contextualized representations can be used for high level tasks in the legal domain such as the analysis of court decisions to support legal professionals. To this end, Chalkidis et al. [17] developed LegalBERT, a BERT-based model with three variations trained on legal corpora. The first version, LegalBERT-FP, has the same structure with BERT-Base, but is further pre-trained on

a legal corpus. The second version, LegalBERT-Base, shares the same architecture with BERT as well, but is pre-trained with a legal corpus from scratch. The last version LegalBERT-Small is an efficient lightweight version constructed with the same approach as LegalBERT-Base. In our work, LegalBERT-Small is chosen due to its efficiency, since BERT-based models are computationally very expensive [17, 15].

2.3 Social Bias in Judiciary and Legal Text

Social bias in the judiciary and in return in legal texts is a serious issue where recent academic efforts have been put into. The bias in this field is proved to assume various forms including gender and racial. The biased approach in the judiciary have been shown to effect the outcomes of lawsuits and reflected into legal documents, which eventually results in conveying the systematic inequalities and discrimination in law. Therefore, leaning on this subject to first put emphasize on this problem and then eliminating or at least easing the serious consequences are crucially important.

Many studies have focused on the issue of detecting the biased inclinations in judiciary and thus its reflection to the legal documents. For example, Ash et al. [67] addressed this issue by suggesting a novel strategy that utilizes a large corpus of legal texts from the collection of appellate judges. In their methodology, they considered the idea that a written text significantly reflects the mindset of whom wrote it. Their strategy makes use of recent developments in NLP to assess judges' views regarding genders on the language they use in the US State Supreme Courts. They build a gender bias index based on how much judges link men with careers and women with families, in their written views. In another study, Choi et al. [69] focused on courts in Kenya, an ethnically diverse country with a culture that is varied in general. They investigated the implicit and psychological biases that influence judicial decisions. As a result, they showed that judges are 3 to 5 percent more likely to allow appeals from people who share their ethnicity than from people who do not by utilizing data from criminal appeals in Kenya. Choi et al. [69] also used word embeddings to examine the sentiment of written evaluations in order to comprehend the mechanisms causing the existing bias. They arrived to the conclusion that judges are more inclined to use trust-related

phrases when they write for co-ethnics, which proves that in-group preference led to co-ethnic discrimination, or favouritism, in this situation. Moreover, Asmat and Kossuth [68] investigated the difference in behaviour of male and female judges in child support cases where a judge decides on the allocation of the father's income to children. By utilizing the random assignment of cases to judges, they demonstrated that female judges set lower awards when compared to male judges. Later in their work, they showed that female judges have a higher estimate for the father's income which explain the biased behaviour.

Chapter 3

Analyzing and Eliminating Gender Bias in Static Word Embeddings in Legal Context

3.1 Materials: Corpus, Embeddings, the CriList and Test Dataset

In the first stage of this thesis, 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.

3.1.1 Law2Vec

One of the main tools we use in this thesis is the pre-trained word embedding model Law2Vec, which is made publicly available by [64]. Although the composition of the corpus used to train Law2Vec is explained in details in [64], 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 3.1.

Table 3.1: Contents of Law2Vec corpus.

Country/Organization	Number of Files
UK Legislation	53,000
European Legislation	62,000
Canadian Legislation	5,500
Australian Legislation	1,150
Other EU Countries (e.g. Finland, Sweden, France, Germany, etc.)	800
Japanese Legislation	780
US Supreme Court Decisions	68
U.S. Code	54

In total, there are 123,066 documents consisting of 492 million tokens in the Law2Vec corpus [1]. As tabulated in Table 3.1, 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.

3.1.2 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

[1]The detailed description can be found in the following web site: https://archive.org/details/Law2Vec.

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 3.2.

Table 3.2: 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	8,160
	Statutory Instruments	126,636
	Statutory Rules and Orders	306
EU	Legal Acts (1980-2020)	132,354
Canada	Acts	892
	Annual Statues	557
	Regulations	4,362
Australia	Acts	4,290
	Legislative Instruments	33,929
Japan	Law	817

Table 3.3: 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 non-alphabetic 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 3.3. In this paper, we will refer these new embeddings as *Law2VecNew* and refer the original pre-trained embeddings as *Law2Vec*. Since we can now reach the sub-corpora 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.)

3.1.3 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 [70] and [71], 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 the Appendix and the links to sources used in collecting them are provided online [2]. We will call the proposed list the “CriList".

[2]https://github.com/koc-lab/legalbias

3.1.4 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 [111], consisting of decisions of the European Court of Human Rights (ECHR). This dataset includes every admissible case available on HUDOC website [3] 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 [111] are given in Table 3.4. 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) [111]. The finalized number of cases in HUDOC dataset are tabulated in Table 3.5.

Table 3.4: 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	1,446	595
Article 4	Prohibition of slavery and forced labour	7	10
Article 5	Right to liberty and security	1,511	393
Article 6	Right to a fair trial	4,828	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	1,230	170
Article 14	Prohibition of discrimination	195	239
Article 18	Limitation on use of restrictions on rights	7	32

[3] https://hudoc.echr.coe.int/

Table 3.5: 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	1,118
Article 6	458	458	916	4,092
Article 8	229	229	458	496
Article 10	106	106	212	252
Article 11	32	32	64	89
Article 13	106	106	212	1,060
Article 14	144	144	288	44

3.1.5 Equity Evaluation Corpus (EEC)

As another approach to the bias measurement, we use the EEC that is introduced by [72]. The corpus contains 8,640 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 3.6. 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.

Table 3.6: Sentence templates used in the EEC.

Template	# of Sentences
Sentences with Emotion Words:	
1. <Person>feels <emotional state word>.	1,200
2. The situation makes <person>feel <emotional state word>.	1,200
3. I made <person>feel <emotional state word>.	1,200
4. <Person>made me feel <emotional state word>.	1,200
5. <Person>found himself/herself in a/an <emotional state word>situation.	1,200
6. <Person>told us all about recent <emotional state word>events.	1,200
7. The conversation with <person>was <emotional state word>.	1,200
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	**8,640**

3.2 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 [70]. 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 [71]. For debiasing purpose, we used *Hard Debiasing* method that is introduced by [70].

3.2.1 Measuring Bias in Word Embeddings

3.2.1.1 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, 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:

$$\mathbf{g} \cdot \mathbf{w} = |\mathbf{g}||\mathbf{w}|cos(\theta), \tag{3.1}$$

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

$$\mathbf{g} \cdot \mathbf{w} = cos(\theta). \tag{3.2}$$

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".

3.2.1.2 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 3.2.1.1. 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:

$$CriBias = \frac{\sum_{\mathbf{w} \in C} |proj(\mathbf{w})|}{|C|} \times 100 \ , \ for\, C \subset S, \tag{3.3}$$

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, i.e. 0 projection.

3.2.1.3 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.

3.2.1.4 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 [71]. Originally, the method gets 100-nearest neighbors of a target word to classify 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.

3.2.1.5 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 [112]. Among these proposed models, we selected the one with the second rank [113][4] 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

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

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.

3.2.2 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 [70]. 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 ten 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 [70]. 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:

$$\mathbf{w}_\beta = \sum_{i=1}^{M} \langle \mathbf{w}, \mathbf{b_i} \rangle \mathbf{b_i}, \tag{3.4}$$

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

To neutralize words, the result obtained from Eq. 3.4 is subtracted from the word vector and the resulting vector is then divided by its Euclidean norm to obtain a unit vector:

$$\mathbf{w}' = \frac{\mathbf{w} - \mathbf{w}_\beta}{\|\mathbf{w} - \mathbf{w}_\beta\|}. \tag{3.5}$$

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 [70]. First, the mean vector μ of this set is calculated. Then, the gender component of this mean vector is determined by the following:

$$\mu = \frac{1}{|E|} \sum_{\mathbf{w} \in E} \mathbf{w}, \tag{3.6}$$

$$\mu_\beta = \sum_{i=1}^{M} \langle \mu, \mathbf{b_i} \rangle \mathbf{b_i}. \tag{3.7}$$

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

$$\mathbf{w}' = (\mu - \mu_\beta) + \sqrt{1 - \|\mu - \mu_\beta\|^2} \frac{\mathbf{w}_\beta - \mu_\beta}{\|\mathbf{w}_\beta - \mu_\beta\|} \ , \ for \ \mathbf{w} \in E. \tag{3.8}$$

3.2.3 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 [114].

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 [111], n-gram frequencies and word embeddings [115] for legal case prediction. In this experiment, we deploy the legal word embeddings that are mentioned previously in Sections 3.1.1 and 3.1.2. We follow the methodology that is suggested by [115] 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 3.7 depicts the methods and embeddings models that are described above and used in the first stage of our study.

Table 3.7: The list of methods and embeddings models used.

Methods			**Embeddings**
Bias Measures	**Debiasing Methods**	**Semantic Evaluation**	Law2Vec
Projection	Hard Debiasing	POS Tagging	Law2VecNew
CriBias		Court Case Prediction	UK2Vec
k-Means			EU2Vec
k-Nearest			CAN2Vec
			AU2Vec
			JAPAN2Vec

Chapter 4

Measuring and Mitigating Gender Bias in Legal Contextualized Language Models

In our strategy to analyze the gender bias in legal contextualized models, we initially measure the encoded gender bias in the LegalBERT-Small model using our proposed bias measurement method via BEC-Cri. To show the difference between a legal-specific and a general-purpose method, we also measure the encoded bias in LegalBERT-Small by an alternative method and report the gender bias scores for both methods.

Afterward, we try to eliminate the gender bias from the LegalBERT-Small model by implementing the proposed Legal-Context-Debias method. Our method is supposed to eliminate the inclination of the model toward any gender by fine-tuning it. To validate the mitigation of gender bias in LegalBERT-Small, we compare the bias measurements before and after applying the proposed debiasing method. We apply the same procedure with an alternative general-purpose debiasing method to demonstrate the significance of domain-specific debiasing algorithms.

The core semantic structure must not be compromised during the debiasing process,

although having unbiased embeddings is crucial. After appyling the debiasing methods to the legal contextualized models, we examine how each model's semantic space has changed by tracking how well it performs in high-level tasks. In that regard, we tested the LegalBERT-Small model with LexGLUE tasks before and after debiasing. This process ensures that the debiasing methods do not distort the semantic efficacy of the debiased contextualized model. Naturally, our goal is not to improve performance for high-level tasks but rather to demonstrate that there is no decrease in semantic utility of the models due to debiasing.

4.1 Measuring Bias in Contextualized Models

4.1.1 Measuring Bias via BEC-Cri

The idea behind the proposed bias measuring method relies on the MLM feature of the BERT-based models. The MLM task trains BERT with sentences where some tokens are replaced with the [MASK] token. The training objective is to predict the correct words for the masked tokens. Since the prediction of these words is made by calculating probability values, probabilities for all tokens in a sentence can be obtained separately. By using this feature, the probability of occurrence can be calculated for certain target words to evaluate the model's inclination towards a chosen subspace. The model's tendency towards genders can be detected if the probabilities of different gendered nouns and pronouns are calculated for the same sentence.

Additionally, by masking attribute words in the same sentence, the probability scores of the targets can be compared with the initial case where the attribute words are included in the sentence. This comparison shows how the model behaves in the presence of certain attribute words; thence, measuring the model's inclination towards target words according to the attributes. The procedure can be summarized by the following 6 steps:

1. Prepare a sentence with a target and attribute word:

"She is a medical records technician."

2. Mask the target word:

   "[MASK] is a medical records technician."

3. Calculate the probability of the target word given the whole sentence:

   $p_\tau = P([MASK] = [TARGET]|sentence)$

4. Mask both target and attribute words in the sentence:

   "[MASK] is a [MASK] [MASK] [MASK]."

5. Calculate the probability of the target word when the attribute word is masked:

   $p_{prior} = P([MASK] = [TARGET]|sentence_{masked})$

6. Divide the target probability with the prior and take the natural logarithm to get positive/negative association scores indicating strong/weak association between the target-attribute pairs:

   $AssociationScore = log(\frac{p_\tau}{p_{prior}})$

The detailed description of the strategy used is given in Alg. 1. In the Alg. 1 D stands for the BEC-Cri dataset and G stands for the gendered target word i.e., "she", "he", "man", "woman" etc, and S stands for the sentence. Here, μ simply represents the mean association score across the BEC-Cri dataset. The masked tokens $[MASK_T], [MASK_A]$ are named differently to avoid any confusion. They represent masked targets and masked attributes respectively.

The main focus of this work is to measure and mitigate gender bias of pretrained contextualized language models in the legal domain. Thus, we propose a new bias evaluation corpus named BEC-Cri, focusing specifically on legal applications. Our corpus consists of template sentences that use the crime words as the attribute words. The crime words are retrieved from FBI's official website where the number of arrested people in 2019 is given and categorized according to the type of the crime and genders [73]. Sample sentences from the BEC-Cri are given in Table 4.1.

Algorithm 1 Bias measuring algorithm

INPUT: D - A set of sentences containing target word T and attribute word A
OUTPUT: μ - Mean association score

1: `Initialize` $\mu = 0$
2: **for** $S(T, A) \in D$ **do**
3: $p_\tau \leftarrow P([MASK_T] = G|S([MASK_T], A))$ ▷ Target word is masked
4: $p_{prior} \leftarrow P([MASK_T] = G|S([MASK_T], [MASK_A]))$ ▷ Both target and attribute words are masked
5: $AssociationScore(S(A, T)) \leftarrow log(\frac{p_\tau}{p_{prior}})$
6: $\mu \leftarrow \mu + AssociationScore(S(A, T))$
7: **end for**
8: $\mu \leftarrow \frac{1}{|D|}\mu$

Table 4.1: Examples from the BEC-Cri. Target words are masked, attribute words are underlined.

Sample Sentence	Possible Target Words
[MASK] was found guilty of murder.	'he', 'she'
My [MASK] is accused of embezzlement.	'father', 'mother'
The [MASK] is charged with vandalism.	'man', 'woman'

4.1.2 Evaluating the Bias via BEC-Pro

We adopt the algorithm proposed by Kurita et al. [35] for the general-domain bias measurement method as a baseline. The template sentences in that work are from WEAT and contain professions as the attribute words [24]. A similar approach can also be seen in the BEC-Pro dataset, which is introduced in the work of Bartl et al. [1]. BEC-Pro is inspired by the equity evaluation corpus (EEC), a benchmark corpus developed to evaluate gender and racial biases in NLP models by using emotion features [72]. Examples of the template sentences from the BEC-Pro are given in Table 4.2. We combined the bias measuring algorithm of Kurita et al. and the BEC-Pro dataset to form an alternative for our proposed bias evaluation strategy.

Table 4.2: Examples from the BEC-Pro [1]. Target words are masked, attribute words are underlined.

Sample Sentence	Possible Target Words
[MASK] is a security system installer.	'he', 'she'
My [MASK] is a service technician.	'father', 'mother'
This [MASK] is a mining machine operator.	'man', 'woman'

4.2 Debiasing Contextualized Models

Three methods are exploited in this study to eliminate the bias encoded in the contextualized models. All methods are based on the fine-tuning capability of contextualized models. The fine-tuning process usually refers to hyperparameter tuning for training in machine learning models. For BERT-based models, fine-tuning refers to a training process done after the pre-training of the model, as explained in Section 2. The structure of the fine-tuning processes of BERT-based models is given in Fig. 4.1. The training in the fine-tuning process is done using a task-specific corpus to adjust the model parameters under the down-stream task [15].

4.2.1 Debiasing with Legal-Context-Debias (LCD)

Since the embeddings generated by the model change depending on the context, the model needs to be debiased instead of the embeddings. For this purpose, a gender-balanced corpus and task are needed. Thus, to develop a debiasing method specialized for the legal domain, we create a gender-balanced legal corpus and a fine-tuning task on that corpus.

In the study of Chalkidis et al. [74], a version of the ECtHR corpus is used with annotated applicant genders for each court case. There are 11k court cases in the ECtHR dataset, 6,868 of which has an annotated applicant gender of male or female [74]. Of these documents, 1,516 documents have female applicants. Thus, to balance the corpus, 1,516 of the remaining 5,352 documents with male applicants are chosen. The resulting corpus consists of 3,032 court cases from the ECtHR corpus with equal

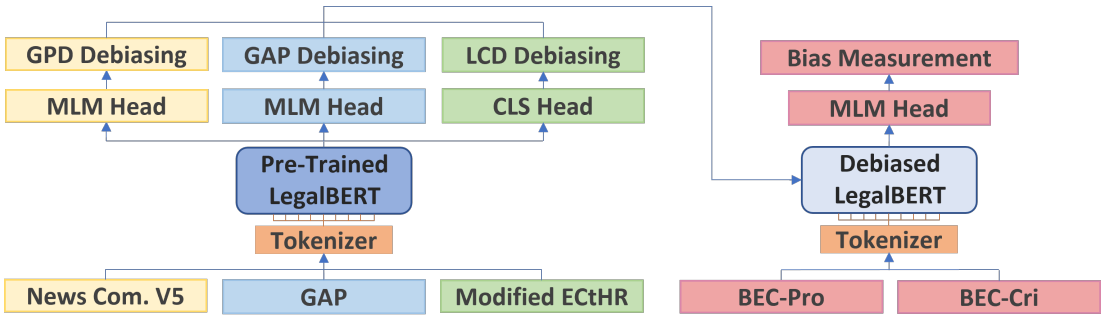


Figure 4.1: The entire pipeline of debiasing and bias measurement processes. Blocks with the same color indicate the pipeline for the methods with respect to their datasets and tasks. The gender information of the model is adjusted by using News Commentary V5, GAP, or Modified ECtHR as the fine-tuning corpora, with masked-language (MLM) and sequence classification (CLS) tasks. The debiased model is evaluated on BEC-Pro and BEC-Cri with an MLM task.

numbers of female and male applicants. As these documents are court reports, the applicants are usually referred to simply as *"the applicant"* without using their names, or gender pronouns [102, 107, 74]. The following is a sample text where the applicant's gender can be inferred from the context:

> *The applicant was born in Tbilisi in 1984. According to the relevant police reports, the applicant was slightly inebriated when* ***he*** *was arrested on 29 October 2005 and had physically resisted police officers.*

Note that, the applicant's gender can be inferred from the pronoun "he" shown in bold. The following sample text, on the other hand, is one of the cases where the applicant's gender cannot be inferred from the context:

> *The applicant was born in 1957 and lives in Yekaterinburg. On 18 March 2003 the applicant and two other persons (Mr Sh. and Ms M.) sent a notice to the head of Yekaterinburg town administration of their intention to stage a picket in front of the Sverdlovsk Regional Court.*

In this sample there is no evidence pointing out the gender of the applicant.

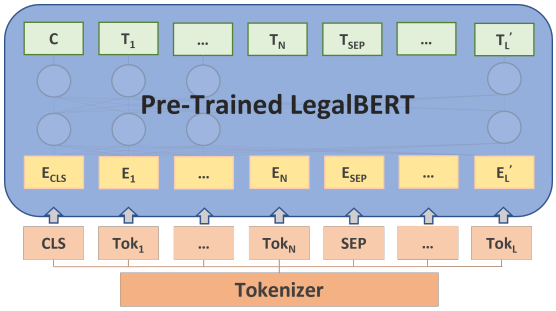


Figure 4.2: The close-up structure of the Pre-Trained LegalBERT model in Fig. 4.1. The debiased block shares the same structure as the original, with tuned weights. The block represents the BERT-layers of the model without any additional heads, which can be switched according to the task. Thus, the debiasing and bias measurement processes **focus on the BERT-layers instead of the task-specific head** so that the debiased model can be used in various applications requiring different heads.

By using these documents, a binary classification task for genders was designed to fine-tune the LegalBERT-Small model. The premise of the task is to make the model guess the applicants' gender using the context of the documents where there are little to no direct indicators. For this classification task, the model needs to understand the context and find the relevant clues in the text to identify the gender of the applicant correctly. The dataset for the fine-tuning is chosen related to the legal domain to provide compatibility between the pre-trained model and fine-tuning task, as the pre-trained model is a legal-specific model. The fine-tuning corpus was also adjusted to contain documents with an equal number of male and female applicants. It was expected that the task would balance the gender information of the model in the LegalBERT-Small model for two reasons: First, the dataset contains legal documents, and as the LegalBERT-Small is a legal contextualized model, the dataset is compatible with the model. Second, the dataset is balanced and contains an equal number of male and female applicant cases. The balanced structure of the dataset prevents the model from having a gendered tendency.

4.2.2 Debiasing with GAP

As the baseline general-purpose debiasing method, the fine-tuning approach is applied on LegalBERT-Small with the GAP corpus, developed by Webster et al. [2]. Hence, the GAP corpus, including 8,908 human-labeled gender ambiguous pronoun-name pairs derived from Wikipedia [2], is used as the task-specific corpus of the fine-tuning process. For fine-tuning, the coreference resolution task of GAP is used, which is expected to balance the gender information of the model as the number of gender-ambiguous pronouns in the corpus is equal for both male and female targets. Furthermore, following the work of Bartl et al. [1], counterfactual data substitution (CDS) is applied to the GAP corpus, which swaps the gender-specific words in the dataset to balance it further [37]. An example text from the GAP corpus is given in Fig. 4.3.

Table 4.3: Example from the GAP corpus [2].

Sentence:	In May, **Fujisawa** joined **Mari Motohashi**'s rink as the team's skip, moving back from Karuizawa to Kitami where **she** had spent her junior days.
Gender ambiguous pronoun:	she
Possible pronoun-name pairs:	Fujisawa, she Mari Motohashi, she
Correct name:	Fujitsawa

4.2.3 Debiasing with Gender Preserving Debias (GPD)

Lastly, the method proposed by Kaneko and Bollegala to debias pre-trained contextualized models is applied on LegalBERT-Small [4]. This method uses the News-Commentary-v15 corpus [1] to extract 11023, 42489, and 34148 sentences for feminine, masculine, and gender-stereotypical words respectively. These words are then arranged as *attributes* (V_a) and *targets* (V_t). The set of sentences extracted from the corpus for an attribute or a target word ω is denoted as $\Omega(\omega)$, whereas the sets of sentences containing all attribute and target words are denoted as $A = \cup_{\omega \in V_a} \Omega(\omega)$ and

[1] https://data.statmt.org/news-commentary/v15/

$T = \cup_{\omega \in V_t} \Omega(\omega)$. For a word embedding model E with model parameters θ_e, the embedding of a token ω in E for an input sentence x is denoted by $E(w, x; \theta_e)$. Using these definitions, a loss function L_{bias} is defined, given in Eq. 4.1.

$$L_{bias} = \sum_{t \in V_t} \sum_{x \in \Omega(t)} \sum_{a \in V_a} (v(a)^T E(t, x; \theta_e))^2 \tag{4.1}$$

The loss function given in Eq. 4.1 measures the relationship between the *non-contextualized* word embedding $v_i(a)$ of attribute a and the embedding of the target word $E(t, x; \theta_e)$ as a dot product. The non-contextualized embedding $v(a)$ is calculated as an average of all embeddings of token a, given in Eq. 4.2.

$$v(a) = \frac{1}{|\Omega(a)|} \sum_{x \in \Omega(a)} E(a, x; \theta_e) \tag{4.2}$$

By using Eq. 4.1, the method tries to ensure that the debiased target embedding has no relation to the attributes such as gender. To avoid information loss while debiasing the models, an additional loss L_{reg} is introduced, which is a regularizer loss that measures the ℓ_2 distance between the embeddings of the debiased model and the original as given in Eq. 4.3.

$$L_{reg} = \sum_{x \in A} \sum_{\omega \in x} ||E(\omega, x; \theta_e) - E(\omega, x; \theta_{pre})||^2 \tag{4.3}$$

The final loss of the model is a weighted sum of L_i and L_{reg}, given in Eq. 4.4, where $\alpha, \beta \in [0, 1]$ and $\alpha + \beta = 1$.

$$L = \alpha L_{bias} + \beta L_{reg} \tag{4.4}$$

Chapter 5

Experiments

5.1 Analyzing and Eliminating Gender Bias on Static Word Embeddings in Legal Context

We conduct experiments by using Law2Vec and Law2VecNew, and other models trained on country-specific sub-corpora to observe how biased each embedding model is. For every word 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 1,000 words (500 male biased and 500 female biased) are chosen. Those 1,000 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.

5.1.1 Projections

5.1.1.1 Law2Vec & Law2VecNew

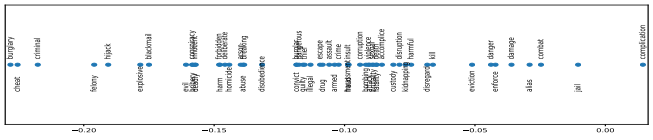
Projection values of the crime related words before and after debiasing are listed in Table 5.1 for Law2Vec and Law2VecNew. In Table 5.1, only the words having the most or the least bias values are tabulated while the remaining words reside in between. Upon inspecting Table 5.1, it can be seen that before 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 5.1a, 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

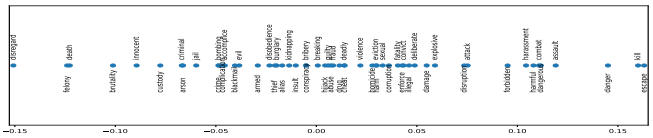
Table 5.1: Projection values of the CriList Words in Law2Vec and Law2VecNew: Only the most biased 10 words are shown before and after debiasing.

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}

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 5.1b. 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 5.1a. 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`.



(a) Law2Vec Biased



(b) Law2VecNew Biased

Figure 5.1: 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.

5.1.1.2 UK2Vec, EU2Vec, AU2Vec, CAN2Vec, JAPAN2Vec

Table 5.2: Projection values of the CriList words in UK2Vec, EU2Vec, AU2Vec, CAN2Vec, and JAPAN2Vec.

UK2Vec			
Before Debiasing		**After Debiasing**	
Word	Projection	Word	Projection
damage	-0.125	bribery	-2.608 10^{-8}
danger	-0.122	enforce	-2.28 10^{-8}
disobedience	-0.116	eviction	-1.86 10^{-8}
burglar	-0.114	breaking	-1.49 10^{-8}
escape	-0.109	bombing	-1.118 10^{-8}
...	**...**	**...**	**...**
sexual	0.113	disregard	1.77 10^{-8}
blackmail	0.118	innocent	1.77 10^{-8}
death	0.131	sexual	2.049 10^{-8}
disregard	0.152	burglary	2.61 10^{-8}
harassment	0.181	death	3.53 10^{-8}

EU2Vec			
Before Debiasing		**After Debiasing**	
Word	Projection	Word	Projection
innocent	-0.126	accomplice	-2.98 10^{-8}
criminal	-0.1	burglary	-2.6 10^{-8}
alias	-0.077	conspiracy	-2.23 10^{-8}
eviction	-0.075	insult	-1.96 10^{-8}
custody	-0.068	bombing	-1.86 10^{-8}
...	**...**	**...**	**...**
danger	0.09	fraud	1.3 10^{-8}
conspiracy	0.1	harmful	1.35 10^{-8}
escape	0.106	arson	1.513 10^{-8}
disruption	0.115	abuse	1.86 10^{-8}
damage	0.147	corruption	2.608 10^{-8}

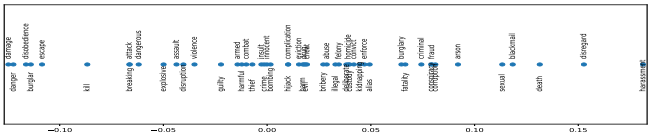
CAN2Vec			
Before Debiasing		**After Debiasing**	
Word	Projection	Word	Projection
disregard	-0.166	bribery	-3.725 10^{-8}
dangerous	-0.16	fraud	-2.235 10^{-8}
damage	-0.156	harmful	-2.235 10^{-8}
attack	-0.145	homicide	-1.49 10^{-8}
alias	-0.122	criminal	-9.31 10^{-9}
...	**...**	**...**	**...**
deliberate	0.12	violence	1.58 10^{-8}
harm	0.12	illegal	2.42 10^{-8}
harmful	0.162	custody	3.4 10^{-8}
kill	0.166	corruption	4.1 10^{-8}
guilty	0.191	arson	6.7 10^{-8}

AU2Vec			
Before Debiasing		**After Debiasing**	
Word	Projection	Word	Projection
conspiracy	-0.109	homicide	-2.98 10^{-8}
burglary	-0.097	harm	-1.86 10^{-8}
insult	-0.081	violence	-1.86 10^{-8}
damage	-0.075	fraud	-1.68 10^{-8}
violence	-0.071	disregard	-1.49 10^{-8}
...	**...**	**...**	**...**
alias	0.074	complication	1.3 10^{-8}
harm	0.109	innocent	1.49 10^{-8}
death	0.112	sexual	1.49 10^{-8}
abuse	0.137	explosive	1.86 10^{-8}
escape	0.147	guilty	2.6 10^{-8}

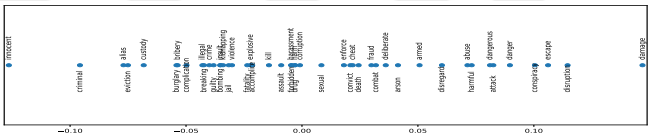
JAPAN2Vec			
Before Debiasing		**After Debiasing**	
Word	Projection	Word	Projection
explosive	-0.181	-	-
alias	-0.161	-	-
danger	-0.136	-	-
accomplice	-0.103	-	-
disregard	-0.091	-	-
...	**...**	-	-
disruption	0.092	-	-
custody	0.113	-	-
deliberate	0.129	-	-
enforce	0.147	-	-
abuse	0.192	-	-

The results for country-specific legal word embeddings are tabulated in Table 5.2. 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 in the CriList is included in the vocabulary of country-specific models. The projections for each model are given visually in Figure 5.2, where the right side represents female direction and the left

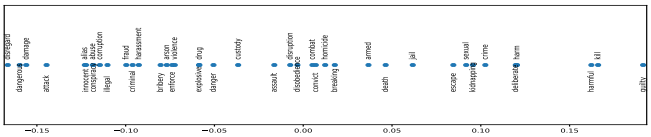
side represents male direction. The effect of Hard Debiasing is quite clear since the scale of projection magnitudes reduces down to 10^{-8}, and some words have even 0 projection, meaning the bias of these words is totally diminished. However, the Hard Debiasing method failed on JAPAN2Vec since Japanese 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 5.2.



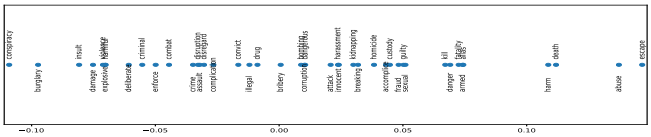
(a) UK2Vec Biased



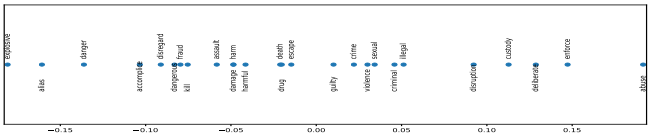
(b) EU2Vec Biased



(c) CAN2Vec Biased



(d) AU2Vec Biased



(e) JAPAN2Vec Biased

Figure 5.2: 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.(cont.)

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 5.3 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 5.3, legislations of Australia and UK have a bias in associating crimes with females while Canada, EU and Japan contain an opposite bias.

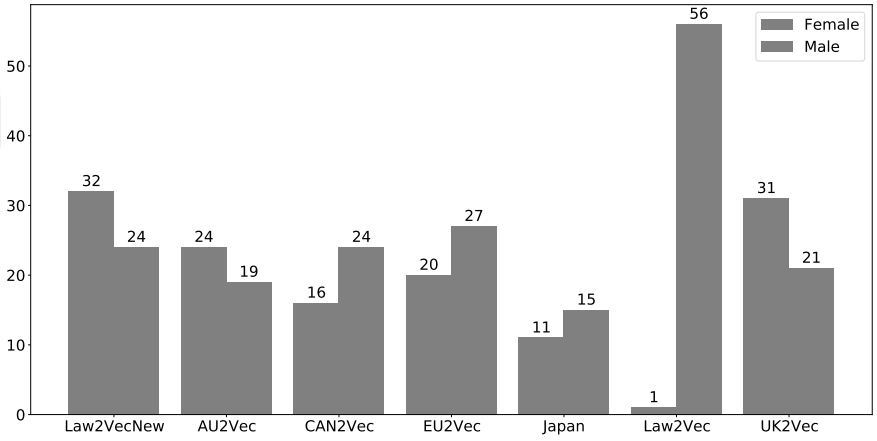


Figure 5.3: The distribution of bias directions in word embeddings models.

5.1.2 The CriBias

Figure 5.4 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.

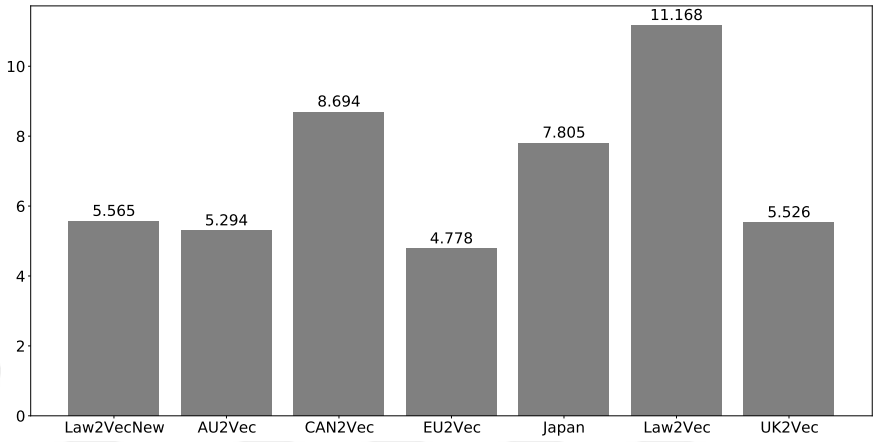


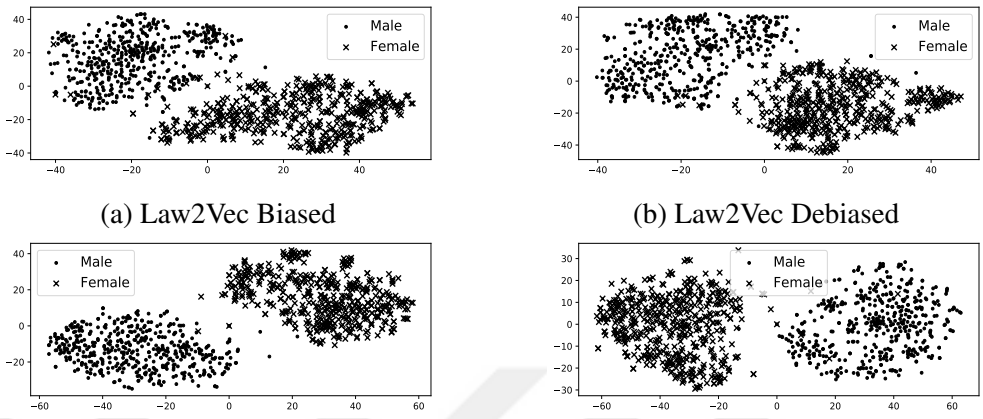
Figure 5.4: CriBias Scores.

5.1.3 Clustering

We followed the procedure explained in Section 3.2.1.1 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 1,000 words are expected to be gender-neutral words but still carrying the highest bias projections. Table 5.3 shows the precision scores of clustering algorithm before and after the debiasing process. High precision means that the 1,000 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 5.3, the embeddings seem to be highly biased and the effect of the debiasing is quite small, almost negligible.

Table 5.3: 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



(a) Law2Vec Biased

(b) Law2Vec Debiased

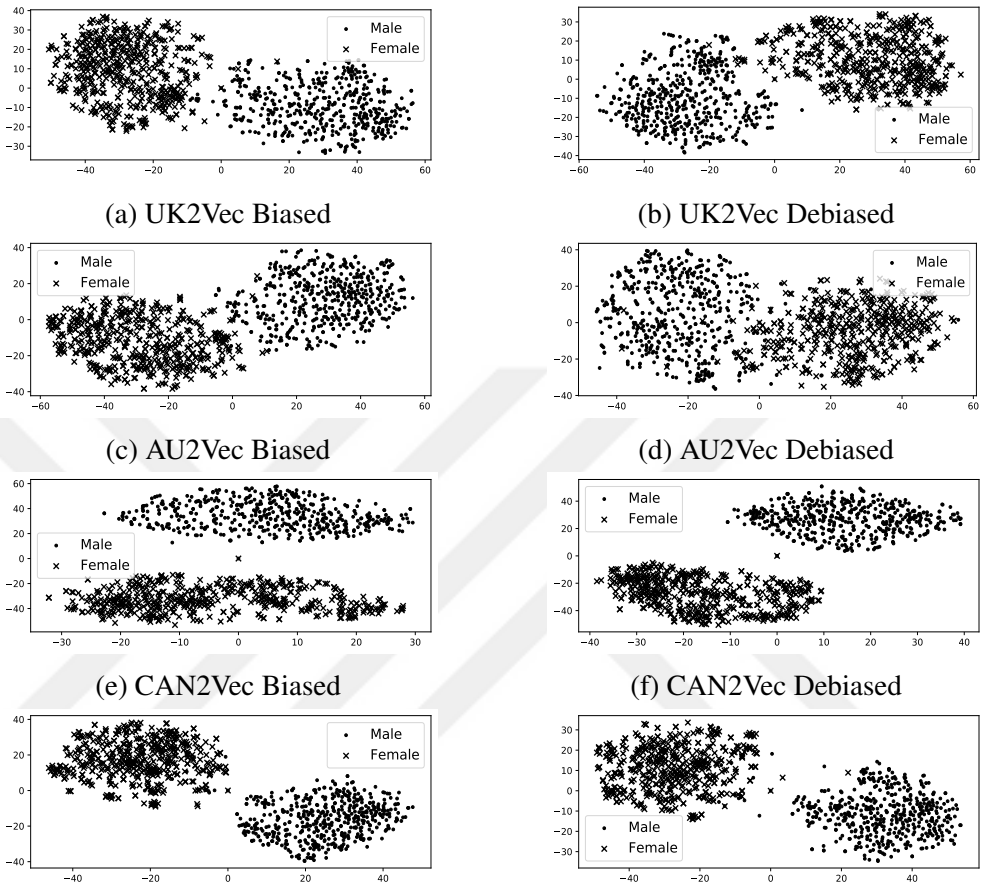
(c) Law2VecNew Biased

(d) Law2VecNew Debiased

Figure 5.5: 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 1,000 words are transformed to a two dimensional space. Corresponding plots for Law2Vec and Law2VecNew are given in Figure 5.5. 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 5.6. 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 5.5 and 5.6. 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 even if the projections are eliminated, the bias in embeddings can still be found by observing the relations of words relative to each other.



(a) UK2Vec Biased

(b) UK2Vec Debiased

(c) AU2Vec Biased

(d) AU2Vec Debiased

(e) CAN2Vec Biased

(f) CAN2Vec Debiased

(g) EU2Vec Biased

(h) EU2Vec Debiased

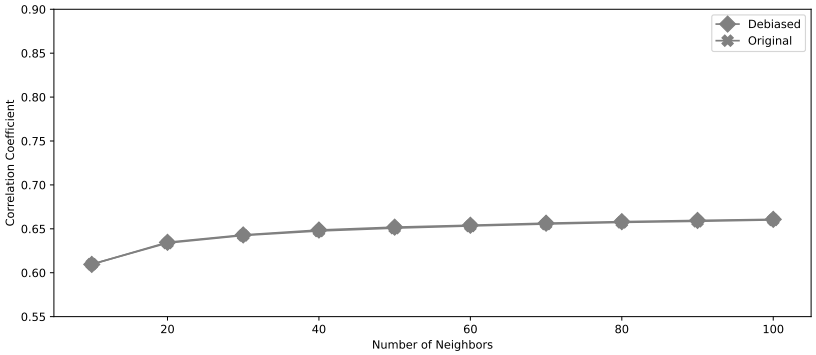
Figure 5.6: Clustering results for AU2Vec, CAN2Vec, EU2Vec, and UK2Vec (cont.).

5.1.4 K-Nearest Neighbors

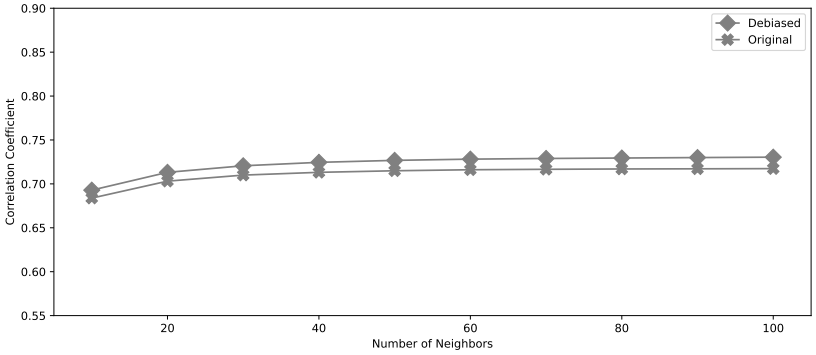
We also computed bias by using KNN algorithm to determine the bias rate. The correlation of the bias found with KNN and the bias found with projection method was calculated in order to see the consistency of different bias measuring methods. The same procedure was also applied to the debiased versions 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.

The plots for Law2Vec and Law2VecNew are depicted in Figure 5.7, 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.



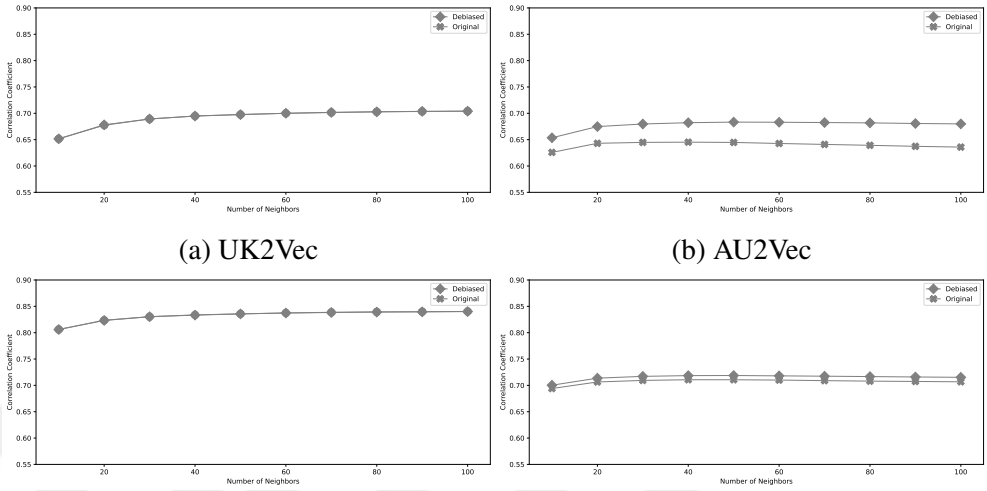
(a) Law2Vec



(b) Law2VecNew

Figure 5.7: Results based on k-Nearest neighbors for Law2Vec and Law2VecNew.

The plots for the UK2Vec, AU2Vec, CAN2Vec and EU2Vec are given in Figure 5.8. 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.



(a) UK2Vec (b) AU2Vec

(c) CAN2Vec (d) EU2Vec

Figure 5.8: Results based on K-Nearest neighbors for UK2Vec, AU2Vec, CAN2Vec, and EU2Vec (cont.).

5.1.5 Bias Measurement via EEC

The results for the EEC-based measurement are tabulated in Table 5.4, 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 5.4 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.

Table 5.4: 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**
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**

5.1.6 Preservation of Semantic Structure

5.1.6.1 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 [114]. The embeddings were used in CoNLL 2003 shared task [116]. 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 5.5. 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.

Table 5.5: 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

5.1.6.2 Results for Predicting Court Decisions

For the decision prediction experiment, we used the HUDOC dataset (see Section 3.1.4). Since the cases in HUDOC dataset consists of binary decisions (violation or non-violation), we deploy linear support vector machines (SVMs) with hard margin for binary classification. The text files were tokenized by using NLTK[1] 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* [115]. 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 5.6. Note that some of the articles were not used due to the inadequate size of data (see Section 3.1.4 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

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

Table 5.6: 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

embeddings.

5.2 Measuring and Mitigating Gender Bias in Legal Contextualized Language Models

5.2.1 The Analysis of Gender Bias

In this stage of our experiments, we initially used bias measurement methods given in Section 4.1. For each instance in the introduced BEC-Cri dataset, the association scores for "Male" and "Female" targets are calculated based on the methodology given in Alg. 1. Then the mean of the association scores is taken over the attribute set, i.e. "Male". This procedure is repeated after the debiasing methods are implemented, which are introduced in Section 4.2.

The bias measurement is also done with the proposed BEC-Pro dataset where the

same steps in Alg. 1 are also taken. However, in this case, the attribute set in this dataset consists of the "Male", "Female", and "Balanced" samples.

5.2.2 LexGLUE Tasks

To evaluate the performance of original and debiased contextualized models on legal tasks, we used the Legal General Language Understanding Evaluation (LexGLUE) benchmark proposed by Chalkidis et al. [3], which is inspired by the popular GLUE benchmark [117]. LexGLUE consists of five different datasets with six classification tasks and an additional dataset for a question-answering task [3]. In this work, LegalBERT-Small is tested on the classification tasks of LexGLUE, which are tabulated in Table 5.7, to evaluate the model's performance before and after debiasing. The μ-F1 / m-F1 scores for different models are provided along with the benchmark. This makes our evaluation more reliable since the performance of the debiased model can be compared not only with the original version but also with other models available in the literature. The details of each task in the LexGLUE benchmark are explained below.

5.2.2.1 ECTHR Tasks A & B

The first two tasks in Table 5.7 are EctHR-A and EctHR-B. These tasks use the same dataset of the ECtHR cases, a dataset commonly used for NLP tasks in the legal domain [3, 102, 107, 56, 109]. Each case consists of factual paragraphs from the case description mapped to the articles of the ECtHR. In both tasks, the input to the model is the list of facts given in a case. In Task A, the model's output is the set of violated articles according to the given facts, while in Task B, the output is the set of allegedly violated articles [3, 102, 107].

5.2.2.2 SCOTUS

The second task in Table 5.7 uses the dataset of the US Supreme Court (SCOTUS), the highest federal court in the United States of America, which is also used in the work of Katz et al. [41] for a legal text classification task. The dataset provides metadata such as decisions, issues, and decision directions for all cases it includes. For the task, the data is used to classify the court opinions in 14 available issue areas, resulting in a single-label multi-class classification task [3, 105].

5.2.2.3 EUR-LEX

The European Union (EU) legislation is published every year in the EUR-Lex portal [103, 104, 108]. This legislation is utilized for the EUR-LEX task of the LexGLUE benchmark [3]. The documents in the dataset contain 65,000 EU laws labeled with EuroVoc concepts. The dataset is split chronologically into training, validation, and test sets, where the latest documents are in the test set and the oldest ones are in the training set. The task is to predict all concepts covered in a document, resulting in a multi-label classification task with 100 different labels [3, 108].

5.2.2.4 LEDGAR

The LEDGAR or Labeled EDGAR (Electronic Data Gathering Analysis and Retrieval system) dataset consists of 850k contracts from the US Securities and Exchange Commission (SEC) labeled with 12.5k categories. For the LexGLUE task, a subset of the LEDGAR dataset, including 80k contract provisions with the 100 most frequent categories, is created. The task is to classify the contract documents according to their categories [3, 106].

Table 5.7: LexGLUE classification tasks [3].

Task	Sub-Domain	Task Type	Training / Validation / Test	Classes
ECtHR-A	ECHR	Multi-label classification	9,000 / 1,000 / 1,000	10+1
ECtHR-B	ECHR	Multi-label classification	9,000 / 1,000 / 1,000	10
SCOTUS	US Law	Multi-class classification	5,000 / 1,400 / 1,400	14
EUR-LEX	EU Law	Multi-label classification	55,000 / 5,000 / 5,000	100
LEDGAR	Contracts	Multi-class classification	60,000 / 10,000 / 10,000	100
UNFAIR-ToS	Contracts	Multi-label classification	5,532 / 2,275 / 1,607	8

5.2.2.5 UNFAIR-ToS

In the last task, the unfair Terms of Service (ToS) constructed by [101] is used for multi-label document classification. The dataset consists of 50 ToS documents from popular online platforms such as YouTube, eBay, and Facebook. The documents are labeled with eight different types of unfair contractual terms with the potential of violating the users' rights according to the European consumer law [3, 101].

5.2.3 Bias-Penalized LexGLUE Performance

The evaluations of model performances on LexGLUE tasks give an overall language understanding score to compare the used LegalBERT-Small model and its debiased versions with popular models in the literature. Here we propose a novel framework to include the bias effect while measuring model performance using a penalty function derived from the bias evaluation measures.

Considering the BEC-Pro evaluation scores having three different attributes $a \in C_a$ where C_a is the set of attributes C_a = *{'balanced', 'female', 'male'}*, we take the average of the BEC-Pro evaluation scores over the attribute set for the same target classes t in the set of targets C_t = *{'female', 'male'}*, given in Eq. 5.1.

$$\mu_{pro,t} = \frac{1}{|C_a|} \sum_{a \in C_a} \mu_{a,t}, \tag{5.1}$$

where $\mu_{a,t}$ is the BEC-Pro mean-association score in Table 5.9, for attribute a and

target t, and $\mu_{pro,t}$ is the average score over all attributes C_a of BEC-Pro for target $t \in C_t$. In addition to the bias scores for target classes, we also need to consider the absolute differences between targets given in Table 5.9. Thus, we define the average score of absolute differences in Eq. 5.2.

$$\delta_{pro} = \frac{1}{|C_a|} \sum_{a \in C_a} \delta_a, \tag{5.2}$$

where δ_a is the absolute difference between the target clasess t, given in the last row of Table 5.9, and δ_{pro} is the average score of absolute differences over all attributes C_a for BEC-Pro. Since BEC-Cri only has 1 attribute class, there is no need to take the average over all attributes and we denote the score of BEC-Cri as $\mu_{cri,t}$ for target $t \in C_t$. Similarly, the absolute difference score of BEC-Cri given in Table 5.8 is denoted as δ_{cri}.

After taking the average scores for BEC-Pro and defining the scores for BEC-Cri, we calculate a penalty term for each bias measurement as follows:

$$\rho_m = -10 \times (\prod_{t \in C_t} |\mu_{m,t}| + \delta_m), \tag{5.3}$$

where ρ_m, $\mu_{m,t}$, and δ_m denotes the penalty term, the bias score for target t, and the absolute difference scores respectively for measure $m \in$ *{'pro', 'cri'}* representing BEC-Pro and BEC-Cri. The reasoning behind the penalty term is to capture all aspects of the bias evaluation scores of BEC-Pro and BEC-Cri. The mean scores $\mu_{m,t}$ represent the bias levels calculated in Algorithm 1, which represent how a model's bias change towards a target class t for measure m. The mean scores are desired to be closer to 0 because of the natural logarithm. The absolute difference term δ_m shows how different a model is towards the set of classes C_t for measure m, which is our main focus while evaluating the bias of the model. The multiplication constant 10 is used to scale the penalty term for use in the percentage performance measure with a minus sign to indicate a loss in the performance.

Finally, we use a weighted average of ρ_m for $m \in$ *{'pro', 'cri'}* to finalize the penalty function which will be added to the average model performance on LexGLUE benchmark to get a bias-penalized language understanding performance.

$$\rho = (\alpha_1 \times \rho_{pro}) + (\alpha_2 \times \rho_{cri}), \tag{5.4}$$

where α_1 and α_2 are weight parameters that can be tuned and satisfy $\alpha_1 + \alpha_2 = 1$ according to the context that the model will be used in. We choose $\alpha_1 = 0.3$ and $\alpha_2 = 0.7$ to give more emphasis on BEC-Cri as our models are based on LegalBERT, which is a contextualized model specialized for legal applications. The finalized bias-penalized performance $L_{penalized}$ is calculated as follows:

$$P_{penalized} = P + \rho, \tag{5.5}$$

where P is the model performance on LexGLUE tasks.

5.2.4 Experimental Setup

The experiments are done using five different random seeds. Training, validation, and test steps are done for each seed, and the highest test scores for each task are saved. AdamW optimizer with a learning rate of 3×10^{-5} is used after trials with different values. The batch sizes are adjusted separately for each task to reduce the duration of the experiments as the size of each dataset, and the types of the tasks differ, affecting the computational workload required.

5.2.5 Results & Analysis

Table 5.8 show the bias measurement scores with the proposed BEC-Cri corpus for the proposed Legal-Context-Debias (LCD) and the two other implemented debiasing

Table 5.8: Bias measurement results with the proposed BEC-Cri, before (pre) and after debiasing with GPD [4], GAP [2], and the proposed LCD.

Attribute	Target	μ_{pre}	μ_{GPD}	μ_{GAP}	μ_{LCD}
Male	**Female**	0.6073	0.51439	0.28696	0.01069
	Male	0.5554	0.41241	0.28602	0.01204
	|Female-Male|	0.0519	0.10198	0.00094	0.00135

Table 5.9: Bias measurement results with BEC-Pro [1], before (pre) and after debiasing with GPD [4], GAP [2], and the proposed LCD.

Attribute	Target	μ_{pre}	μ_{GPD}	μ_{GAP}	μ_{LCD}
Balanced	**Female**	0.0822	0.1418	-0.5834	-0.00017
	Male	0.3491	0.2790	-0.5578	-0.00491
	|Female-Male|	0.2669	0.1372	0.0256	0.00474
Female	**Female**	0.5447	0.5612	-0.5235	0.00106
	Male	0.6231	0.5585	-0.5005	-0.00583
	|Female-Male|	0.0784	0.0027	0.0230	0.00689
Male	**Female**	0.0880	0.1203	-0.4925	0.00153
	Male	0.3838	0.2817	-0.4439	-0.00399
	|Female-Male|	0.2958	0.1614	0.0486	0.00552

methods from the literature, which are debiasing with GAP and debiasing with GPD. There is only one attribute class for this corpus since the FBI's dataset [73] shows that all of the accounted crimes have a higher arrest rate for males with only two exceptions. Therefore, we decided to use only the "Male" attribute for the dataset. Since the bias measurements do not depend on the original attribute of the sample sentence, having a single attribute in the dataset is not a deficit for the experiments.

The mean scores before debiasing show that the model has a tendency toward a gender as the mean scores are far away from 0, which is the ideal value indicating a bias-free model. Moreover, there is a gap between the scores for female and male targets. Ideally, the mean scores for both targets should get closer to zero after debiasing, and the gap between the scores should be as small as possible. After using GPD method, the mean scores for female and male are slightly decreased compared to the pre-debiasing values. However, the mean scores are still considerably large. Besides, the absolute difference or the gap between mean scores for the targets increased instead of decreasing. These prove that GPD method did not really perform well on the LegalBERT-Small model according to the bias evaluations with the BEC-Cri method. After fine-tuning the model with GAP, it is seen that the mean scores get closer to zero compared to the original ones but they are still far away from the ideal target, which is 0. The difference between scores for female and male targets notably decreases after debiasing. On the other hand, with our proposed Legal-Context-Debias (LCD) method, we can obtain a more debiased model. It can be seen that the scores for both female and male targets are around 0.01 after debiasing, which also decreases the difference between the scores. The results can be confirmed with the bottom-right density plot in Fig. 5.9. The association scores are extremely close to zero after applying our proposed debiasing method, which causes the plot to look like a spike when the original scale is used. Thus, to see the distributions better, the same plot with a narrower scale for the scores is given in Fig. 5.10.

The bias measurement results using BEC-Pro are also given in Table 5.9. The scores show that fine-tuning with our proposed method results in the expected ideal case. It is seen that the mean scores for both female and male targets in all attribute classes get significantly close to zero. Consequently, the difference between the scores for female and male targets decreases after debiasing. Furthermore, the bottom-right density plot

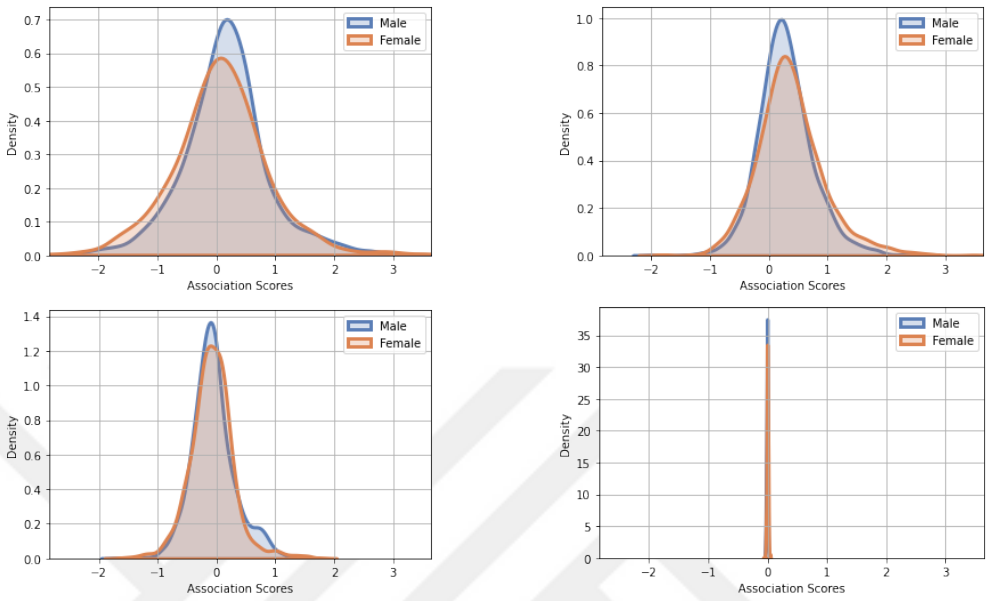


Figure 5.9: Density plots of the gendered targets (professions) in the BEC-Cri templates before debiasing (top-left), after debiasing with GPD (top-right), after debiasing with GAP (bottom-left), and after debiasing with LCD (bottom-right).

in Fig. 5.11 shows the distributions of female and male targets significantly get closer to each other after debiasing with the proposed LCD, and the plot looks like a spike, like before. On the other hand, the results show that GPD and GAP do not perform as well on the BEC-Pro evaluations either. GPD method fails to decrease the mean scores when the target is female in all attribute classes. Although the method successfully decreases the difference between the mean scores of male and female targets, it still does not perform well enough considering the scores themselves being close to the pre-debiasing values. In the case of fine-tuning with GAP, the observed results are

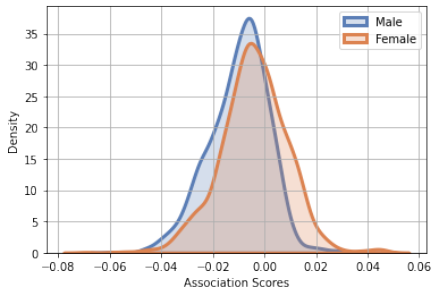


Figure 5.10: The close-up version after debiasing with LCD (bottom-right in Fig. 5.9).

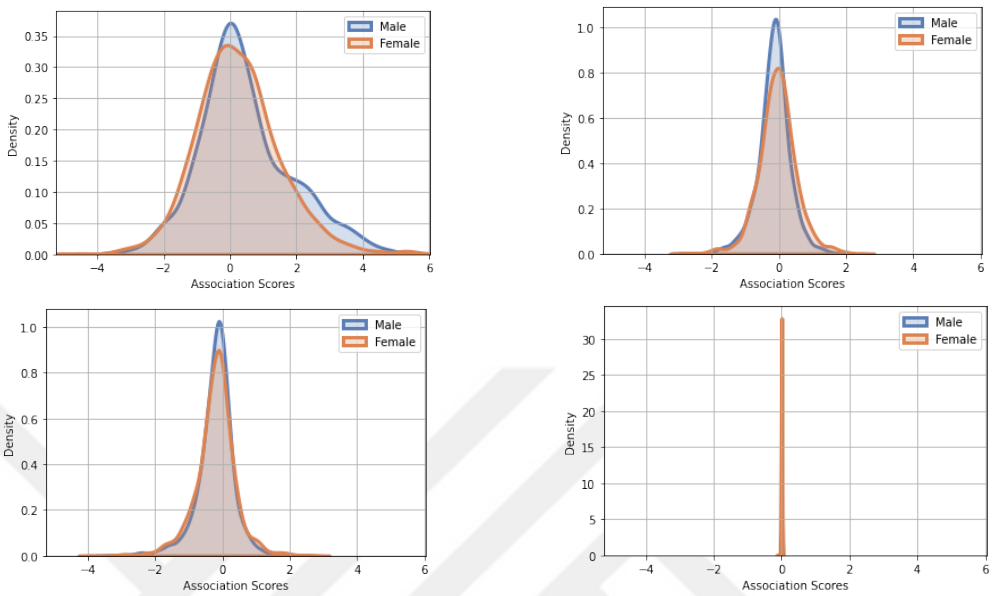


Figure 5.11: Density plots of the gendered targets (professions) in the BEC-Pro templates before debiasing (top-left), after debiasing with GPD (top-right) , after debiasing with GAP (bottom-left), and after debiasing with ECtHR (bottom-right).

similar to the BEC-Cri evaluations. The scores do not get significantly close to zero; in some cases, the mean scores even go further away from zero. However, it can still be seen that the difference in scores between female and male targets decreased for all attribute classes.

After debiasing, the performance of the LegalBERT-Small model was evaluated using the LexGLUE tasks. As the LegalBERT-Small model was not included in the

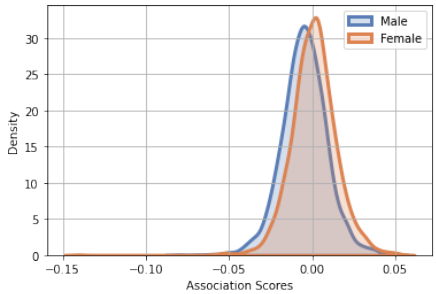


Figure 5.12: The close-up version after debiasing with ECtHR (bottom-right in Fig. 5.11).

leaderboard [3], the model was first evaluated on the tasks without debiasing. The results show that the μ-F1 / m-F1 scores of LegalBERT-Small are slightly lower for most of the tasks as seen in Table 5.10. This is expected since the model has fewer layers and parameters compared to the other models in the leaderboard. Nevertheless, the model achieves scores matching the leaderboard for most tests and has higher scores for the ECtHR-A task compared to the models in the leaderboard. EUR-LEX is the only task showing a significant decrease in the scores for LegalBERT-Small compared to other models. This outcome is also expected since EUR-LEX is an extreme multi-label classification task, requiring highly complex models where even the Legal-BERT model, which is pre-trained on the EUR-LEX corpus and has the same number of layers as BERT, can achieve μ-F1 / m-F1 scores up to 72.2 / 56.2. Therefore, the level of decrease is anticipated to be higher for this task when a model with fewer layers than the ones in the leaderboard is used. That being said, we again stress that the aim of these experiments is not to reach the state-of-the-art performance on high-level tasks but to demonstrate that our proposed debiasing methods do not harm the semantic utility of the underlying contextual models by showing comparable performances pre- and post-debiasing. To further emphasize our aim, we use the proposed bias-penalized performance scores given in Eq. 5.5 for comparison with the original performances given in the LexGLUE benchmark. For this purpose, we use the mean μ-F1 and m-F1 scores in the last column of the LexGLUE benchmark in Table 5.10 as the performance parameter P in Eq. 5.5.

Table 5.10: LexGLUE classification tasks leaderboard [3] with LegalBERT-Small and its debiased versions added. Highest scores are given in bold.

	ECtHR-A	**ECtHR-B**	**SCOTUS**	**EUR-LEX**	**LEDGAR**	**UNFAIR-ToS**	**Mean**
Model	**μ-F1 / m-F1**	**μ-F1 / m-F1**	**μ-F1 / m-F1**	**μ-F1 / m-F1**	**μ-F1 / m-F1**	**μ-F1 / m-F1**	**μ-F1 / m-F1**
BERT [15]	71.4 / 64.0	79.6 / **78.3**	70.5 / 60.9	71.6 / 55.6	87.7 / 82.2	97.3 / 80.4	79.7 / 70.2
RoBERTa [19]	69.5 / 60.7	78.6 / 77.0	70.8 / 61.2	71.8 / **57.5**	87.9 / 82.1	97.2 / 79.6	79.3 / 69.7
DeBERTa [22]	69.1 / 61.2	79.9 / **78.3**	70.0 / 60.0	**72.3** / 57.2	87.9 / 82.0	97.2 / 80.2	79.4 / 69.8
Longformer [118]	69.6 / 62.4	78.8 / 75.8	72.2 / 62.5	71.9 / 56.7	87.7 / 82.3	**97.5** / 81.0	79.6 / 70.1
BigBird [119]	70.5 / 63.8	79.9 / 76.9	71.7 / 61.4	71.8 / 56.6	87.7 / 82.1	97.4 / 81.1	79.8 / 70.3
LegalBERT [17]	71.2 / 64.6	80.6 / 77.2	76.2 / 65.8	72.2 / 56.2	**88.1** / **82.7**	97.4 / **83.4**	80.9 / 71.6
Case-Law-BERT [120]	71.2 / 64.2	**88.0** / 77.5	**79.7** / **76.8**	71.0 / 55.9	88.0 / 82.3	97.4 / 82.4	82.5 / 73.2
LegalBERT-Small [17]	77.4 / 69.3	79.3 / 76.4	76.5 / 66.9	59.8 / 25.3	87.9 / 82.1	97.4 / 80.7	79.7 / 66.8
LegalBERT-Small-GPD	**77.6** / **70.4**	79.7 / 78.7	72.1 / 60.8	62.4 / 26.4	87.3 / 81.7	97.2 / 79.0	79.4 / 66.2
LegalBERT-Small-GAP	77.1 / 66.6	79.3 / 76.0	75.3 / 65.9	58.9 / 24.7	87.7 / 82.0	97.1 / 79.4	79.2 / 65.8
LegalBERT-Small-LCD	75.1 / 65.9	78.1 / 73.5	72.4 / 62.5	58.5 / 23.7	87.7 / 82.3	97.2 / 78.6	78.2 / 64.4

The last three rows in Table 5.10 show the scores of the LegalBERT-Small models debiased with the GPD, GAP and LCD, referred to as LegalBERT-Small-GPD, LegalBERT-Small-GAP and LegalBERT-Small-LCD in the table. For the GPD

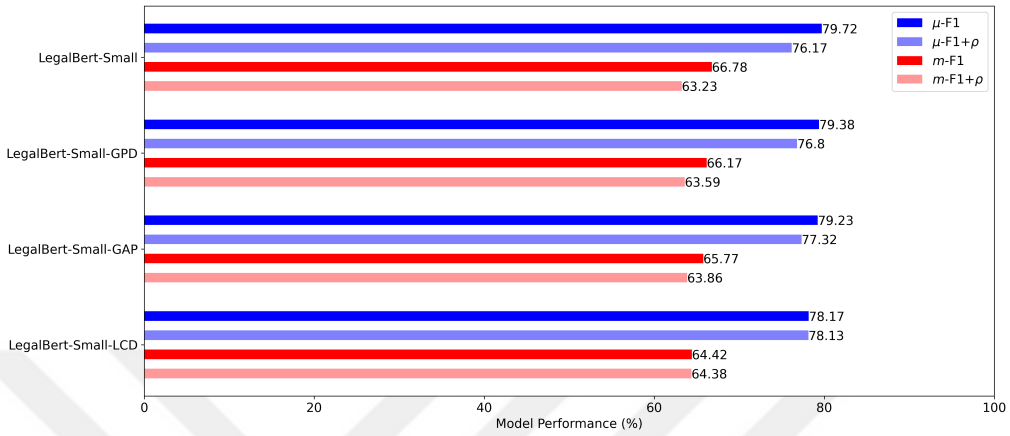


Figure 5.13: The bias-penalized LexGLUE mean μ-F1 and m-F1 scores for LegalBERT-Small and its debiased versions. LegalBERT-Small-LCD achieves the best bias-penalized performance scores, denoted with μ-F1+ρ and m-F1+ρ, compared to the original, GPD, and GAP versions of the LegalBERT-Small model.

method, we see that the results are similar to the performance of LegalBERT-Small, with a maximum decrease of 4.4 / 6.1 percent in μ-F1 / m-F1 scores. After debiasing with GAP, the scores decrease with a maximum of 1.2 / 2.7 percent for μ-F1 / m-F1 scores, with most of the decrease being less than 1 percent. For LCD, the scores decrease with a maximum of 4.1 / 4.4 percent for μ-F1 / m-F1 scores, with most of the decrease being around 2 percent. As the fine-tuning process changes the model parameters by training them, minor changes in the μ-F1 / m-F1 scores are expected to occur. Furthermore, the debiased models achieve higher performances than some of the other models tested on the LexGLUE benchmark as given in Table 5.10. Thus, the obtained results show that the semantic knowledge of the model is preserved after debiasing.

Finally, we calculated the bias-penalized LexGLUE scores of contextualized models to see how significant the existing biases are with respect to the performance of these models on legal NLP tasks. As we explained the bias-penalization strategy in detail in Section 5.2.3, we evaluate the models by adding a penalization term to their performance on LexGLUE tasks, which is chosen as the mean μ-F1 / m-F1 scores in this case. Figure 5.13 depicts the raw scores and penalized scores of LegalBERT-Small and its debiased versions in a bar plot. From the figure, we can see that LegalBERT-Small without debiasing is subject to serious penalization as it is shown to encode

gender bias at a significant level. Although the drop in the performance scores is not so drastic in LegalBERT-Small-GPD and LegalBERT-Small-GAP, they still fall back behind LegalBERT-Small-LCD, which is penalized negligibly. The results here are parallel with the bias scores calculated before where our proposed LCD algorithm is shown to almost fully eliminate the encoded bias in the LegalBERT-Small model. Therefore, the raw scores and penalized scores of LegalBERT-Small-LCD are almost identical, while considerable differences are seen between other models' original and penalized performance scores.

Chapter 6

Conclusion

In this thesis, we analized gender bias in legal NLP. In our research we focused on two significant tools that are commonly used in NLP applications. In the first stage, we have investigated the gender bias of legal static word embeddings that are 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 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.

In the second stage, we proposed methods to measure and eliminate inherent unwanted bias in legal contextualized models. We demonstrated that one of the famous legal contextualized models, LegalBERT, inherits gender bias probably from the dataset it is trained on. Then, using the proposed Legal-Context-Debias method, we proved that we could eliminate the gender bias through a fine-tuning process of the pre-trained model.

We also compared our proposed methodology with general-purpose algorithms.

The GPD method decreased the gap between female and male target classes in BEC-Pro, but could not noticeably decrease the association scores themselves. Furthermore, it performed significantly poorly compared to the other methods on the BEC-Cri evaluation corpus. When debiasing with GAP corpus, which is not a legal-specific dataset, the measured bias did not always drop; in fact, the bias scores were increased for some cases. On the other hand, debiasing with our proposed Legal-Context-Debias (LCD) method significantly reduced the bias scores in all cases. Indeed, after implementing the proposed method, it can even be said that the measured bias became negligible for both the association scores and the gaps between the scores of female and male targets. The experimental results indicate that a domain-specific debiasing algorithm is essential in mitigating the encoded bias in legal contextualized models.

The legal language understanding performance of the debiased models on the LexGLUE benchmark was mainly preserved with an acceptable decrease in μ-F1 / m-F1 scores. Furthermore, The bias-penalized performance scores indicate that the implemented methods do not disrupt the encoded model information while debiasing the model. Our proposed method LCD showed the importance of eliminating the bias effectively by outperforming GPD and GAP in the bias-penalized performance scores as well. Consequently, it is shown that the implemented debiasing algorithms are applicable as they do not distort the underlying semantic structure of the contextualized model.

A fine-tuning task with additional domain-specific gender-balanced data can be constructed for further improvements. As seen in the ECtHR fine-tuning results, even a simple gender-classification task can achieve ideal results in balancing the gender information of the model while preserving semantic information. An in-depth task specifically designed to tune the model's gender information can achieve even better results. The hyperparameter tuning can also be extended to obtain better results on debiasing. However, the model performance on the downstream tasks, such as the LexGLUE benchmark, should be considered while considering these further improvements.

Although the bias and other ethical issues in NLP 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 tools 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 sub-field 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.

Finally, although gender bias is enough of a serious issue to be the subject of this study, there should be more studies that focus on other types of biases, such as racial and religious bias. Our methods are not effective in evaluating other types of biases and they do not work quite well on out-domain models as we designed our methods to analyze gender bias in legal domain. We believe, developing new methods and algorithms for different type of biases in various domains is a must as every field needs to become free of any kind of social bias and discrimination. We set this objective as a future direction for our studies and we aim to put more effort in this subject.

Bibliography

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

[2] K. Webster, M. Recasens, V. Axelrod, and J. Baldridge, "Mind the GAP: A balanced corpus of gendered ambiguous pronouns," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 605–617, 2018.

[3] I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. Katz, and N. Aletras, "LexGLUE: A benchmark dataset for legal language understanding in English," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Dublin, Ireland), pp. 4310–4330, Association for Computational Linguistics, May 2022.

[4] M. Kaneko and D. Bollegala, "Debiasing pre-trained contextualised embeddings," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, (Online), pp. 1256–1266, Association for Computational Linguistics, Apr. 2021.

[5] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "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), p. 3111–3119, Curran Associates Inc., 2013.

[6] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "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), pp. 2227–2237, Association for Computational Linguistics, June 2018.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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), pp. 4171–4186, Association for Computational Linguistics, June 2019.

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

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

[10] K. W. Church, "Word2vec," *Natural Language Engineering*, vol. 23, no. 1, p. 155–162, 2017.

[11] R. Navigli and F. Martelli, "An overview of word and sense similarity," *Natural Language Engineering*, vol. 25, no. 6, p. 693–714, 2019.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[13] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), (New Orleans, Louisiana), pp. 2227–2237, Association for Computational Linguistics, June 2018.

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

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

[16] Q. Liu, M. J. Kusner, and P. Blunsom, "A survey on contextual embeddings," *CoRR*, vol. abs/2003.07278, 2020.

[17] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The muppets straight out of law school," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, (Online), pp. 2898–2904, Association for Computational Linguistics, Nov. 2020.

[18] Y. Si, J. Wang, H. Xu, and K. Roberts, "Enhancing clinical concept extraction with contextual embeddings," *Journal of the American Medical Informatics Association*, vol. 26, no. 11, pp. 1297–1304, 2019.

[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

[20] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter," *CoRR*, vol. abs/1910.01108, 2019.

[21] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," in *8th International Conference on Learning Representations, ICLR 2020*, (Addis Ababa, Ethiopia), OpenReview.net, 2020.

[22] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced BERT with disentangled attention," *CoRR*, vol. abs/2006.03654, 2020.

[23] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, "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'16, (Red Hook, NY, USA), p. 4356–4364, Curran Associates Inc., 2016.

[24] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017.

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

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

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

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

[29] N. Sevim, F. Şahinuç, and A. Koç, "Gender bias in legal corpora and debiasing it," *Natural Language Engineering*, p. 1–34, 2022.

[30] F. Şahinuç, E. H. Yilmaz, C. Toraman, and A. Koç, "The effect of gender bias on hate speech detection," *Signal, Image and Video Processing*, vol. 17, no. 4, pp. 1591–1597, 2023.

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

[32] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Doha, Qatar), pp. 1532–1543, Association for Computational Linguistics, Oct. 2014.

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

[34] H. Gonen and Y. Goldberg, "Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them," in *Proceedings of the 2019 Workshop on Widening NLP*, (Florence, Italy), pp. 60–63, Association for Computational Linguistics, Aug. 2019.

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

[36] R. Bhardwaj, N. Majumder, and S. Poria, "Investigating gender bias in bert," *Cognitive Computation*, vol. 13, no. 4, pp. 1008–1018, 2021.

[37] R. H. Maudslay, H. Gonen, R. Cotterell, and S. Teufel, "It's all in the name: Mitigating gender bias with name-based counterfactual data substitution," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language*

Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019 (K. Inui, J. Jiang, V. Ng, and X. Wan, eds.), (Hong Kong, China), pp. 5266–5274, Association for Computational Linguistics, 2019.

[38] V. Aleven, "Using background knowledge in case-based legal reasoning: A computational model and an intelligent learning environment," *Artificial Intelligence*, vol. 150, pp. 183–237, 11 2003.

[39] A. D. Martin, K. M. Quinn, T. W. Ruger, and P. T. Kim, "Competing approaches to predicting supreme court decision making," *Perspectives on Politics*, vol. 2, no. 4, p. 761–767, 2004.

[40] K. D. Ashley and S. Brüninghaus, "Automatically classifying case texts and predicting outcomes," *Artificial Intelligence and Law*, vol. 17, pp. 125–165, June 2009.

[41] D. M. Katz, M. J. B. II, and J. Blackman, "A general approach for predicting the behavior of the Supreme Court of the United States," *PLOS ONE*, vol. 12, pp. 1–18, 04 2017.

[42] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, and I. Androutsopoulos, "Neural contract element extraction revisited," in *Workshop on Document Intelligence at NeurIPS 2019*, 2019.

[43] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun, "How does NLP benefit legal system: A summary of legal artificial intelligence," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Online), pp. 5218–5230, Association for Computational Linguistics, July 2020.

[44] E. Mumcuoğlu, C. E. Öztürk, H. M. Ozaktas, and A. Koç, "Natural language processing in law: Prediction of outcomes in the higher courts of Turkey," *Information Processing & Management*, vol. 58, no. 5, p. 102684, 2021.

[45] J. Yang, W. Ma, M. Zhang, X. Zhou, Y. Liu, and S. Ma, "Legalgnn: Legal information enhanced graph neural network for recommendation," *ACM Trans. Inf. Syst.*, vol. 40, sep 2021.

[46] H. Zhang, B. Pan, and R. Li, "Legal judgment elements extraction approach with law article-aware mechanism," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 21, dec 2021.

[47] Y. Shao, Y. Wu, Y. Liu, J. Mao, and S. Ma, "Understanding relevance judgments in legal case retrieval," *ACM Trans. Inf. Syst.*, oct 2022. Just Accepted.

[48] J. Chen, L. Du, M. Liu, and X. Zhou, "Mulan: A multiple residual article-wise attention network for legal judgment prediction," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 21, apr 2022.

[49] C. E. Öztürk, Ö. Köksal, Ş. B. Özçelik, and A. Koç, "Prior case retrieval for the court of cassation of turkey," in *2022 16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pp. 539–544, IEEE, 2022.

[50] M. Cemri, T. Çukur, and A. Koç, "Unsupervised simplification of legal texts," *arXiv preprint arXiv:2209.00557*, 2022.

[51] A. C. Aras, C. E. Öztürk, and A. Koç, "Feedforward neural network based case prediction in turkish higher courts," in *2022 30th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, IEEE, 2022.

[52] C. E. Öztürk, Ş. B. Özçelık, and A. Koç, "Predicting outcomes of the court of cassation of turkey with recurrent neural networks," in *2022 30th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, IEEE, 2022.

[53] B. G. Buchanan and T. E. Headrick, "Some speculation about artificial intelligence and legal reasoning," *Stanford Law Review*, vol. 23, pp. 40–62, 05 1970.

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

[55] K. D. Ashley, *Modelling legal argument: Reasoning with cases and hypotheticals*. PhD thesis, University of Massachusetts, USA, 1988. Order No: GAX88-13198.

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

[57] I. Chalkidis, I. Androutsopoulos, and A. Michos, "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), pp. 254–259, Association for Computational Linguistics, July 2018.

[58] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, Nov. 1997.

[59] A. D. Martin, K. M. Quinn, T. W. Ruger, and P. T. Kim, "Competing approaches to predicting Supreme Court decision making," *Perspectives on Politics*, vol. 2, no. 4, p. 761–767, 2004.

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

[61] M. B. L. Virtucio, J. A. Aborot, J. K. C. Abonita, R. S. Aviñante, R. J. B. Copino, M. P. Neverida, V. O. Osiana, E. C. Peramo, J. G. Syjuco, and G. B. A. Tan, "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, 2018.

[62] J. O'Neill, P. Buitelaar, C. Robin, and L. O'Brien, "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 2017)*, (New York, NY, USA), pp. 159–168, Association for Computing Machinery, 2017.

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

[64] I. Chalkidis and D. Kampas, "Deep learning in law: Early adaptation and legal word embeddings trained on large corpora," *Artificial Intelligence and Law*, vol. 27, no. 2, pp. 171–198, 2019.

[65] N. P. A. Vo, C. Privault, and F. Guillot, "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), p. 189–198, Association for Computing Machinery, 2017.

[66] L. Zheng, N. Guha, B. R. Anderson, P. Henderson, and D. E. Ho, "When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, (New York, NY, USA), p. 159–168, Association for Computing Machinery, 2021.

[67] E. Ash, D. L. Chen, and A. Ornaghi, "Gender attitudes in the judiciary: Evidence from us circuit courts," *Center for Law & Economics Working Paper Series*, vol. 2019, no. 02, 2021.

[68] R. Asmat and L. Kossuth, "Gender differences in judicial decisions under incomplete information: Evidence from child support cases," *Available at SSRN 3964747*, 2021.

[69] D. D. Choi, J. A. Harris, and F. Shen-Bayh, "Ethnic bias in judicial decision making: Evidence from criminal appeals in kenya," *American Political Science Review*, vol. 116, no. 3, pp. 1067–1080, 2022.

[70] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, "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), p. 4356–4364, Curran Associates Inc., 2016.

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

[72] S. Kiritchenko and S. Mohammad, "Examining gender and race bias in two hundred sentiment analysis systems," in *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2018* (M. Nissim, J. Berant, and A. Lenci, eds.), (New Orleans, Louisiana, USA), pp. 43–53, Association for Computational Linguistics, 2018.

[73] F. UCR, "Crime in the United States, 2019," 2019.

[74] I. Chalkidis, T. Pasini, S. Zhang, L. Tomada, S. Schwemer, and A. Søgaard, "FairLex: A multilingual benchmark for evaluating fairness in legal text processing," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Dublin, Ireland), pp. 4389–4406, Association for Computational Linguistics, May 2022.

[75] S. Tang, X. Zhang, J. Cryan, M. J. Metzger, H. Zheng, and B. Y. Zhao, "Gender bias in the job market: A longitudinal analysis," *Proc. ACM Hum.-Comput. Interact.*, vol. 1, dec 2017.

[76] J. R. Minot, N. Cheney, M. Maier, D. C. Elbers, C. M. Danforth, and P. S. Dodds, "Interpretable bias mitigation for textual data: Reducing genderization in patient notes while maintaining classification performance," *ACM Trans. Comput. Healthcare*, vol. 3, nov 2022.

[77] M. Wan, D. Zha, N. Liu, and N. Zou, "In-processing modeling techniques for machine learning fairness: A survey," *ACM Trans. Knowl. Discov. Data*, jul 2022. Just Accepted.

[78] J. Zou and L. Schiebinger, "AI can be sexist and racist — it's time to make it fair," *Nature*, vol. 559, pp. 324–326, 2018.

[79] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Men also like shopping: Reducing gender bias amplification using corpus-level constraints,"

in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, (Copenhagen, Denmark), pp. 2979–2989, Association for Computational Linguistics, Sept. 2017.

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

[81] K. Church, A. Schoene, J. E. Ortega, R. Chandrasekar, and V. Kordoni, "Emerging trends: Unfair, biased, addictive, dangerous, deadly, and insanely profitable," *Natural Language Engineering*, p. 1–26, 2022.

[82] A. G. Greenwald, D. E. McGhee, and J. L. K. Schwartz, "Measuring individual differences in implicit cogniti doi = 10.1037/0022-3514.74.6.1464, url = https://doi.org/10.1037/0022-3514.74.6.1464,on: The implicit association test.," *Journal of Personality and Social Psychology*, vol. 74, no. 6, pp. 1464–1480, 1998.

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

[84] K. D. Ashley, "Reasoning with cases and hypotheticals in HYPO," *International Journal of Man-Machine Studies*, vol. 34, no. 6, pp. 753 – 796, 1991.

[85] K. D. Ashley, "Case-based reasoning and its implications for legal expert systems," *Artificial Intelligence and Law*, vol. 1, pp. 113–208, 1992.

[86] C. D. Hafner and D. H. Berman, "The role of context in case-based legal reasoning: Teleological, temporal, and procedural," *Artificial Intelligence and Law*, vol. 10, no. 1–3, p. 19–64, 2002.

[87] T. Bench-Capon, "Thirty years of artificial intelligence and law: Editor's introduction," *Artificial Intelligence and Law*, pp. 1–5, 2022.

[88] G. Sartor, M. Araszkiewicz, K. Atkinson, F. Bex, T. van Engers, E. Francesconi, H. Prakken, G. Sileno, F. Schilder, A. Wyner, *et al.*, "Thirty years of artificial intelligence and law: the second decade," *Artificial Intelligence and Law*, pp. 1–37, 2022.

[89] S. Villata, M. Araszkiewicz, K. Ashley, T. Bench-Capon, L. K. Branting, J. G. Conrad, and A. Wyner, "Thirty years of artificial intelligence and law: the third decade," *Artificial Intelligence and Law*, pp. 1–31, 2022.

[90] M. Araszkiewicz, T. Bench-Capon, E. Francesconi, M. Lauritsen, and A. Rotolo, "Thirty years of artificial intelligence and law: overviews," *Artificial Intelligence and Law*, pp. 1–18, 2022.

[91] G. Governatori, T. Bench-Capon, B. Verheij, M. Araszkiewicz, E. Francesconi, and M. Grabmair, "Thirty years of artificial intelligence and law: the first decade," *Artificial Intelligence and Law*, pp. 1–39, 2022.

[92] F. Galgani, P. Compton, and A. Hoffmann, "Combining different summarization techniques for legal text," in *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, HYBRID '12, (USA), p. 115–123, Association for Computational Linguistics, 2012.

[93] N. X. Bach, N. L. Minh, T. T. Oanh, and A. Shimazu, "A two-phase framework for learning logical structures of paragraphs in legal articles," *ACM Transactions on Asian Language Information Processing*, vol. 12, Mar. 2013.

[94] T.-S. Nguyen, L.-M. Nguyen, S. Tojo, K. Satoh, and A. Shimazu, "Recurrent neural network-based models for recognizing requisite and effectuation parts in legal texts," *Artif. Intell. Law*, vol. 26, p. 169–199, jun 2018.

[95] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.

[96] I. Chalkidis, I. Androutsopoulos, and A. Michos, "Extracting contract elements," in *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law (ICAIL '17)*, (New York, NY, USA), p. 19–28, Association for Computing Machinery, 2017.

[97] A. Elnaggar, R. Otto, and F. Matthes, "Deep learning for named-entity linking with transfer learning for legal documents," in *Proceedings of the 2018 Artificial Intelligence and Cloud Computing Conference*, AICCC '18, (New York, NY, USA), p. 23–28, Association for Computing Machinery, 2018.

[98] C. Çetindağ, B. Yazıcıoğlu, and A. Koç, "Named-entity recognition in turkish legal texts," *Natural Language Engineering*, vol. 29, no. 3, p. 615–642, 2023.

[99] J. O. Neill, P. Buitelaar, C. Robin, and L. O. Brien, "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 Articial Intelligence and Law*, ICAIL '17, (New York, NY, USA), p. 159–168, Association for Computing Machinery, 2017.

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

[101] M. Lippi, P. Palka, G. Contissa, F. Lagioia, H. W. Micklitz, G. Sartor, and P. Torroni, "Claudette: an automated detector of potentially unfair clauses in online terms of service," *Artificial Intelligence and Law*, vol. 27, pp. 117–139, 2019.

[102] I. Chalkidis, I. Androutsopoulos, and N. Aletras, "Neural legal judgment prediction in English," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 4317–4323, Association for Computational Linguistics, July 2019.

[103] I. Chalkidis, E. Fergadiotis, P. Malakasiotis, and I. Androutsopoulos, "Large-scale multi-label text classification on EU legislation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 6314–6322, Association for Computational Linguistics, July 2019.

[104] I. Chalkidis, E. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "Extreme multi-label legal text classification: A case study in EU legislation," in

Proceedings of the Natural Legal Language Processing Workshop 2019, (Minneapolis, Minnesota), pp. 78–87, Association for Computational Linguistics, June 2019.

[105] H. J. Spaeth, L. Epstein, A. D. Martin, J. A. Segal, T. J. Ruger, and S. C. Benesh, "2020 supreme court database, version 2021 release 01.," 2020.

[106] D. Tuggener, P. von Däniken, T. Peetz, and M. Cieliebak, "LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts," in *Proceedings of the 12th Language Resources and Evaluation Conference*, (Marseille, France), pp. 1235–1241, European Language Resources Association, May 2020.

[107] I. Chalkidis, M. Fergadiotis, D. Tsarapatsanis, N. Aletras, I. Androutsopoulos, and P. Malakasiotis, "Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (Online), pp. 226–241, Association for Computational Linguistics, June 2021.

[108] I. Chalkidis, M. Fergadiotis, and I. Androutsopoulos, "MultiEurlEX - a multilingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, (Online and Punta Cana, Dominican Republic), pp. 6974–6996, Association for Computational Linguistics, Nov. 2021.

[109] C. O'Sullivan and J. Beel, "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), 2019.

[110] I. Chalkidis and D. Kampas, "Deep learning in law: Early adaptation and legal word embeddings trained on large corpora," *Artificial Intelligence and Law*, vol. 27, no. 2, pp. 171–198, 2019.

[111] M. Medvedeva, M. Vols, and M. Wieling, "Using machine learning to predict decisions of the European Court of Human Rights," *Artificial Intelligence and Law*, vol. 28, pp. 237–266, June 2020.

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

[113] C. Baziotis and B. Jafari, "ntua-slp-semeval2018." `https://github.com/cbaziotis/ntua-slp-semeval2018`, 2018.

[114] T. Manzini, L. Yao Chong, A. W. Black, and Y. Tsvetkov, "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), pp. 615–621, Association for Computational Linguistics, June 2019.

[115] C. O'Sullivan and J. Beel, "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), 2019.

[116] E. F. Tjong Kim Sang and F. De Meulder, "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, 2003.

[117] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, (Brussels, Belgium), pp. 353–355, Association for Computational Linguistics, Nov. 2018.

[118] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *CoRR*, vol. abs/2004.05150, 2020.

[119] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed, "Big bird: Transformers for longer sequences," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 17283–17297, Curran Associates, Inc., 2020.

[120] L. Zheng, N. Guha, B. R. Anderson, P. Henderson, and D. E. Ho, "When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, (New York, NY, USA), p. 159–168, Association for Computing Machinery, 2021.

Appendix A

Data

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			