Gender and Racial Stereotype Detection in Legal Opinion Word Embeddings

Sean Matthews[1*], **John Hudzina**[1*], **Dawn Sepehr**[2]

[1] Thomson Reuters Labs, Eagan, Minnesota, USA
[2] Thomson Reuters Labs, Toronto, Canada
sean.matthews@thomsonreuters.com, john.hudzina@thomsonreuters.com, dawn.sepehr@thomsonreuters.com

Abstract

Studies have shown that some Natural Language Processing (NLP) systems encode and replicate harmful biases with potential adverse ethical effects in our society. In this article, we propose an approach for identifying gender and racial stereotypes in word embeddings trained on judicial opinions from U.S. case law. Embeddings containing stereotype information may cause harm when used by downstream systems for classification, information extraction, question answering, or other machine learning systems used to build legal research tools. We first explain how previously proposed methods for identifying these biases are not well suited for use with word embeddings trained on legal opinion text. We then propose a domain adapted method for identifying gender and racial biases in the legal domain. Our analyses using these methods suggest that racial and gender biases are encoded into word embeddings trained on legal opinions. These biases are not mitigated by exclusion of historical data, and appear across multiple large topical areas of the law. Implications for downstream systems that use legal opinion word embeddings and suggestions for potential mitigation strategies based on our observations are also discussed.

Introduction

Recent developments in the field of Artificial Intelligence (AI) have transformed the way data is prepared and turned into information for interpretation in different domains spanning from social media to legal documents. These advancements have predominantly paved the way for creating more accurate predictive models, however, multiple research studies have shown that these systems are not without fault and have inadvertently perpetuated some harmful biases and stereotypes present in society by encoding and replicating patterns of bias present in the data upon which they are trained. Examples of such faulty systems include racial bias detected in hate speech predictive models for social media posts (Mozafari, Farahbakhsh, and Crespi 2020), unequal distribution of health care resources across racial groups due to incorrectly identifying patients who need significant healthcare (Obermeyer et al. 2019), displaying fewer Science, Technology, Engineering, and Mathematics (STEM) job advertisements to women compared to men (Lambrecht and Tucker 2019), and racial disparities demonstrated in recidivism risk prediction algorithms (Dieterich, Mendoza, and Brennan 2016).

While these tools are becoming more and more integrated in our societies and extend great benefits when deployed properly, they also pose high risks of imposing unfair life changing decision making upon minorities and more vulnerable communities. This may be particularly important when developing technologies used within the legal system due to the significant impacts the legal system in general has on individuals, businesses, government entities, and many other aspects of society. Deploying predictive technologies based on biased models into contexts where they are used by individuals interacting with the legal system at various levels could potentially result in a broad array of harmful effects in society including decreased quality of legal representation, increased costs associated with litigation, or even increased likelihood or duration of incarceration for individuals belonging to groups affected by these biases. Hence, it is imperative that we take steps towards identifying, mitigating, and ultimately eliminating these undesired effects in legal technologies relying on predictive systems.

We would like to emphasize that historical bias is not the only form of bias that can be found in AI systems: representation, measurement, aggregation, evaluation, and deployment biases have also been identified at different stages of developing an AI system (Suresh and Guttag 2020). In this article, however, we mainly focus on revealing historical and representational bias found in word embeddings trained on judicial opinions from U.S. case law and the distinct challenges that arise when developing predictive models in the legal domain.

Bias in Word Embeddings

Word embedding approaches such as word2vec (Mikolov et al. 2013a,b), GloVe (Pennington, Socher, and Manning 2014), etc., represent words in an n-dimensional space by encoding contextual co-occurrence statistics for words occurring in large text corpora. Since these associations are obtained from compiling large historical corpora, different types of biases that already exist in these texts will inevitably plague the word representations if appropriate considerations are not anticipated. Multiple previous studies have in-

*These authors contributed equally.

vestigated and shown the presence of these biases in the form of either benign or neutral effects such as associating flowers with pleasant words vs. associating weapons with unpleasant words, or detrimental effects by encoding discrimination based on protected categories such as race, gender, social status, etc. (Bolukbasi et al. 2016; Caliskan, Bryson, and Narayanan 2017).

Different methodologies have been proposed to identify, visualize, and mitigate these effects. One prominent approach draws inspiration from a method originally developed in the field of social psychology to measure implicit bias in humans. The Implicit Association Test (IAT) measures the differential response times of human participants while categorizing sets of target words (e.g., flower and insect names) and attribute terms (e.g., pleasant or unpleasant) when they are paired in stereotypical (e.g. flower-pleasant) or counter-stereotypical (e.g. flower-unpleasant) configurations (Greenwald, McGhee, and Schwartz 1998). Both the stimuli used in the IAT and the general strategy of detecting bias through differential association strength have been adapted to develop bias detection strategies to measure bias encoded in word embeddings such as the Word Embedding Association Test (WEAT; Caliskan, Bryson, and Narayanan 2017). The WEAT measures this difference in association strength between two groups by calculating the similarity of the embeddings in a set of target words used as a proxy for group membership (e.g., common female given names) with the embeddings in two sets of attribute words (e.g., pleasant and unpleasant terms) and computing the difference between these similarities, then comparing the difference in these association strengths to the same difference score calculated for a second target group (e.g., common male given names). Using methods based on this test, Rice et. al found evidence for racial biases being encoded in word embeddings trained on legal texts such as appellate court opinions from US state and federal courts (Rice, Rhodes, and Nteta 2019).

Legal Word Embedding Issues

As mentioned in the previous section, the social impact of encoded bias in word embeddings is becoming more significant in the legal domain which has direct implications on many aspects of our society as legal technologies using these types of representations gain greater adoption. In this section, we review some of the challenges that arise when working with legal text corpora and in the subsequent sections we present our solution for addressing these issues.

Names in Legal Text: Although the WEAT racial and gender stereotype tests relied on given names (Caliskan, Bryson, and Narayanan 2017), legal opinions construct more formal sentences than the wikipedia and news articles used to train the publically available GloVe embeddings (Pennington, Socher, and Manning 2014). For example, Figure 1 demonstrates the co-referencing of a natural person in legal opinions. Note that Gerald Bostock's given name is only referenced once. In most cases, the natural person's full name is typically referenced first followed by the surname and/or pronouns thereafter. If a legal system applied the given name tests only, then bias encoded in surnames and gendered pronoun embeddings would be missed.

> Excerpt from Bostock v. Clayton County:
>
> **Gerald Bostock** worked for Clayton County, Georgia, as a child welfare advocate. Under **his** leadership, the county won national awards for its work. After a decade with the county, **Mr. Bostock** began participating in a gay recreational softball league. Not long after that, influential members of the community allegedly made disparaging comments about **Mr. Bostock's** sexual orientation and participation in the league. Soon, **he** was fired for conduct "unbecoming" a county employee.

Figure 1: Legal Opinion Co-referencing Example

Another concern specific to the legal domain is that legal opinions potentially may introduce gender-occupational stereotypes because they typically state a judge's full name and judicial title. Historically, women only account for 12.3% of federal Title III judicial appointments (Federal Judicial Center 2012). Given that Caliskan, Bryson, and Narayanan (2017) found significant gender-occupation bias in non-legal text and the historical imbalance of female judges, embeddings built upon legal opinions potentially perpetuate this specific stereotype.

Positive & Negative Sentiment: Whereas WEAT evaluated sentiment in a generalized modern web corpus, the legal opinions contain historical domain specific terminology. The WEAT study evaluated several tests measuring positive and negative sentiment for various target groups. These tests are from the IAT with very small vocabularies as required due to fatigue effects in human participants (Caliskan, Bryson, and Narayanan 2017). The sentiment-based test must be adjusted for legal opinions because the general vocabularies used to describe positive and negative sentiment do not align with how positive and negative sentiment is expressed in judicial opinions (Rice and Zorn 2021).

Legal Outcomes: While the IAT tests mainly focus on negative or positive attributes, legal outcome extraction provides a greater risk of harm to protected classes than sentiment analysis. Courts document legal outcomes in docket entries, orders, judgements, and/or opinions. Litigation analytics extract legal outcomes for these free text sources because many jurisdictions do not record outcomes in a structured form at a party level (Vacek et al. 2019). If the word embeddings influence outcome extraction based on a party's gender or race, then embedding-based analytics may amplify racial and gender bias by causing parties to settle for something other than their case's merits.

Contributions

As discussed previously, word embeddings are used in many practical NLP systems which operate on legal language. In this article, we propose an approach for identifying racial

and gender biases encoded in word embeddings that are created using the text of legal opinions. This approach addresses multiple issues specific to legal language that have not been addressed in the previous work. These challenges deal with idiomatic phrases as well as specific considerations for adapting the WEAT tests to legal language for detection of bias. We also investigate how these biases have changed over time as well as their strengths in different topical areas of the law.

Proposed Approach

In this section we describe the main approach proposed for identifying bias in word embeddings created based on legal opinions. We first describe the legal corpus under study and the required data preparation. Next, we briefly discuss the Word Embedding Association Test (WEAT). Finally, we state how we addressed the challenges identified in the previous section with domain adapted tests.

Opinion Preparation & Embedding Construction

For our experiments, we examined embeddings created from a large corpus of U.S. legal opinions. The corpus includes over 12 million opinions from 1,949 current and historic jurisdictions dating back to 1650. The corpus size contains 10x more opinions than a previous legal opinion bias study (Rice, Rhodes, and Nteta 2019). The main corpus includes U.S. federal, state, and territorial courts with the notable exception of tribal courts. The tribal court opinions are handled as a supplemental corpus from the source system. To generate the embeddings for the full corpus, topical subcorpora and historical sub-corpora, we follow the process in Figure 2.

Idiomatic Phrase Extraction: Prior to generating the embeddings, we extracted idiomatic phrases. Non-contextual word embeddings assume phrase meanings are composed from representations of individual words. However, this composability assumption does not always hold true for legal jargon and idiomatic phrases. For example, the Latin phrase *pro hac vice* means "for this time only" and does not have the same semantic meaning as the individual words *pro, hac, & vice*.

Although contextual embeddings handle this issue by representing the relative relationships between words, non-contextual embeddings only represent idiomatic phrases as single tokens (Mikolov et al. 2013b). To avoid overly large n-gram dictionaries, the phrase extractor only combines tokens that commonly appear together. Our phrase extractor used a Normalized Point-wise Mutual Information (NPMI) score to select the n-grams to add in the dictionary (Bouma 2009). NPMI scores range from -1 (never co-occurs) to 1 (always co-occurs), with 0 meaning the tokens are completely independent. We ran two passes of the phrase extractor that selected phrases with a minimum NPMI score of 0.5.

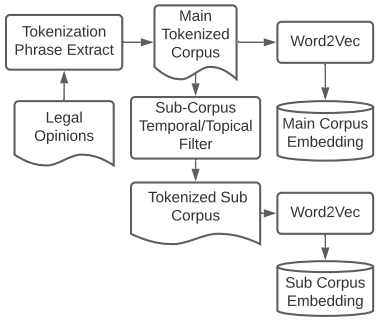

Figure 2: Corpus Prep & Embedding Generation

Embedding Training: Once the phrase extraction was completed, we trained the embeddings against the complete corpus, as well as sub-corpora for temporal cutoff dates, and divided by topic (see section "Experimental Results"). Each embedding followed the same training procedure using a skip-gram word2vec model. For all embeddings, the hyper-parameters included a 300 dimension vector size, a minimum term frequency of 30, a 10^{-4} sampling threshold, a learning rate of 0.05, a window size of 10, and 10 negative samples.

Word Embedding Association Test

Our experiments use Caliskan, Bryson, and Narayanan (2017) original word lists applied to the legal opinion embeddings, as well as tests based on domain specific and expanded word lists. For each test, which includes both the target X and Y word lists, and the attribute A and B word lists, we calculate the effect size (Cohen's d). We also calculate the standard error by sub-sampling the word lists with a simple bootstrapping procedure.

Domain Adaptation

Once the embeddings were trained, we extended the Caliskan, Bryson, and Narayanan tests with new domain specific tests. This section details the methodologies used to generate legal specific target and attributes terms for the new tests. These updates include new attribute lists:

- Positive vs. Negative Legal
- Legal (Motion) Outcome
- Expanded Career vs. Family

The domain updates also include new target lists:

- Surnames by Race
- Male vs. Female Terms
- Judge Given Names

Positive vs. Negative Legal: In order to generate a legal specific sentiment vocabulary, we implemented a minimally supervised approach developed by Rice and Zorn (2021). This work provided a legal specific list of positive, V_p, and negative, V_n, seed terms (Rice and Zorn 2019) and a method for expanding the term sets. We then generated the expanded list based on the legal opinion corpus embeddings. The expanded positive valence terms were found by a cosine similarity search on the vector $\sum \vec{V}_p - \sum \vec{V}_n$. Conversely, The negative valence terms were found on a cosine similarity search on the vector $\sum \vec{V}_n - \sum \vec{V}_p$. The expanded term lists

were then manually reviewed to exclude any terms with obvious race or gender associations (e.g. "gentlemanly")[1].

Legal (Motion) Outcome: Trial motions are discussed and reviewed within legal opinions. This text typically includes the party's surname, motion type, and disposition (Vacek et al. 2019). The manually created "Grant vs. Deny" attribute lists capture the positive and negative outcomes for a given motion.

Expanded Career vs. Family List: The expanded career list employed the same minimally supervised approach as the positive and negative legal attribute list expansion. Instead of finding terms along a positive/negative dimension of interest, we extracted terms along a career versus family axis in the embedding space. The seed lists contained the career and family terms from Caliskan, Bryson, and Narayanan (2017).

Surnames by Race: As noted in the introduction, court documents reference parties by their surnames throughout the document. For the racial stereotype experiments, we used the surnames list from the 2010 U.S. decennial census as a proxy for race similar to medical outcome studies (Kallus, Mao, and Zhou 2020). The census provides an estimated percentage of each race by surname. We sampled names from the list with over a 90% probability for a given race.

Although the U.S. Census provided a list of surnames, the referenced name is not guaranteed to refer to a natural person. Instead, the name may reference a legal person's (i.e., corporation) name, a place name, or a thing. To reduce the potential name overlap with common words, we employed three methods to either reduce and or eliminate multi-sense words from the surname list:

- Title cased the surnames to target proper nouns.
- Idiomatic phrase extraction to exclude non-person names like the *State of Washington* (Mikolov et al. 2013b).
- Centroid-based filtering to remove multi-sense words.

The centroid-based filter removes candidate surnames based on the following procedure developed for WEAT (Caliskan, Bryson, and Narayanan 2017). We computed a centroid vector based on the embedding vectors for all surnames in the U.S. 2010 Census and then computed the cosine similarity for each surname relative to the centroid. Finally, we removed 20% of the least similar names. Once the filter was applied, we created the name lists for each test. While our target sample size was 200 surnames with at least 300 opinions per racial group, those criteria were not achievable for all races. Specifically, Native American and Alaskan names were proportionally underrepresented in the main corpus because fewer tribal court jurisdictions publish to and or are collected by the source system compared to State and Federal jurisdictions. Table 1 shows the sample sizes for each group. We adjusted the sample size for each test pair of surnames based on the smallest sample size in the pair.

[1]See the Supplementary Material for excluded terms: https://arxiv.org/abs/2203.13369

Group	Sample Size	Min. Cases
European	46 - 200	300
African American	164	300
Hispanic	200	300
Asian Pacific Islander	200	300
Native American / Alaskan	46	30

Table 1: Surname Lists by Race

Male vs. Female Terms: A similar problem exists with the gender tests based on given names since individuals are often referred to primarily by their surnames throughout legal opinions. To address this issue, we created a list of gendered pronouns and common gendered nouns (e.g. man/-woman) for use in gender bias WEATs.

Judge Given Name List: In addition to the gendered first name list created by Caliskan, Bryson, and Narayanan (2017), we generated a gendered first name list based on judicial biographical data exported from the free law project's court listener (Free Law Project 2021). The biographical information included both race and gender for both State and Federal Judges. We calculate the percentages of female and male genders for each first name. For each gendered list we select names that occur at least 90% of the time for that gender.

As with the surnames, some first names might overlap with place names, corporations, or other concepts. For example, Virginia might represent a Judge's name or a State. As with the surname we employed the following procedures:

- Title cased the first name to target proper nouns.
- Idiomatic phrase extraction to exclude non-person names, like the *Commonwealth of Virginia*.
- Centroid-base filtering to remove multi-sense words.

Experimental Results

Legal Opinion Corpus

Before discussing the results for the legally adapted tests, we evaluate the opinion-based embedding using the Caliskan, Bryson, and Narayanan (2017) tests. Table 2 shows the baseline results[2]. While the legal opinion embeddings show a smaller effect size for the flower/insect control test than Caliskan, the opinion flower/insect control still exhibits a large effect size. In addition, the instrument/weapons control displays equivalent effect sizes between Caliskan's Common Crawl embeddings and the legal opinion embeddings.

Similar to the baseline tests, the racial and gender stereotype tests show a strong effect size. Note that Table 2 uses the exact same target and attributes as Caliskan, Bryson, and Narayanan (2017). These tests use first names as the targets and non-legal terms as the attributes. Yet, we still see a moderate to strong effect for sentiment. The gender specific test replicated the occupational bias seen in past studies.

[2]The Cohen's d effect size ranges from -2.0 to 2.0 with ± 0.5 representing a medium effect

Test	d_C	d_L
Flowers/Insects		
Pleasant vs Unpleasant	1.50	0.97
Instruments/Weapons		
Pleasant vs Unpleasant	1.53	1.55
Eur. / African American names		
Pleasant3 vs Unpleasant3	1.41	0.88
Male vs Female		
Career vs Family	1.81	1.75

Table 2: Cohen's effect size (d) comparison between Common Crawl GloVe (d_C) and Legal Opinion Word2Vec (d_L)

Test	d	error
Male vs Female Terms		
Pleasant vs Unpleasant	-0.197	0.009
Positive vs. Negative Legal	0.089	0.007
Grant vs Deny	0.457	0.008
Male vs. Female Names (Judges)		
Pleasant vs Unpleasant	-0.495	0.008
Positive vs. Negative Legal	-0.254	0.003
Grant vs Deny	0.603	0.007
Male vs. Female Names (Caliskan)		
Pleasant vs Unpleasant	0.208	0.013
Positive vs. Negative Legal	-0.198	0.003
Grant vs Deny	0.506	0.009

Table 3: Cohen's effect size (d) for gender specific test on the Legal Opinion Corpus

Gender Effects: While the effect sizes were comparable between the Common Crawl corpus and the legal corpus, the legal specific gender tests show some differences. Table 3 includes the original Caliskan and the new Legal attributes. The "Grant vs. Deny" tests all show a medium female negative bias for legal outcome. In comparison, the "Pleasant vs. Unpleasant" shows a positive female bias. In essence, positive sentiment does not necessarily relate to a positive outcome.

Racial Effects: As with the gender tests, we create both legal specific target and attribute word lists. Figure 3 shows the results for the surname-based racial bias experiments. The surname tests demonstrate a large difference in effect size between the "Pleasant vs. Unpleasant" sentiment and the Legal attributes. While the Hispanic surname tests only show a small negative effect for general sentiment, both legal specific tests showed a large negative effect.

The Asian Pacific Islander results provided an even greater disparity in results than the Hispanic surname test. Although the "Pleasant vs. Unpleasant" test showed a large positive bias for Asian Pacific Islanders, the "Positive vs. Negative Legal" test showed a large negative bias for Asian Pacific Islanders. In essence, positive stereotypes do not necessarily translate to group fairness in a legal context.

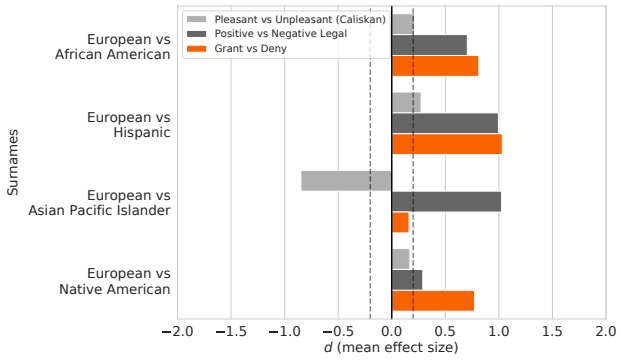

Figure 3: Surname WEAT Cohen's effect sizes

Temporal Effects

Given that the opinion corpus used to train the word embeddings contains opinions dating back to 1650, one possibility is that the observed gender and racial biases are driven by the inclusion of opinions from time periods where these biases were even more explicit and prevalent than in modern society. Our interest in the temporal component of these biases is primarily focused on representational harms that could be caused by NLP systems that may use representations similar to these in legal technology applications rather than in a historical analysis measuring the amount of bias present in any given time period.

To investigate the effect of inclusion of historical opinions on the biases encoded in these representations, we trained word embeddings on temporal subsets of the corpora. These subsets were created by always including modern opinions, but varying the year of cutoffs for inclusion of historical data to incorporate older opinions into the corpus. The year cutoffs we selected were as follows: 2000-2020 (last 20 years), 1980-2020 (last 40 years), 1968-2020 (Post Civil Rights Act), 1954-2020 (Post Brown v. Board of Education), 1930-2020 (Post Great Depression), 1896-2020 (Post Plessy v. Ferguson), and 1865-2020 (Post Civil War). We then applied the legal-adapted WEAT analyses previously described to the embeddings generated for each temporal cutoff. The results of these analyses are shown in Figures 4 and 5 for racial and gender bias WEATs respectively.

In both the gender and racial temporal analyses, it is clear that the biases previously observed in the embeddings trained on the full corpus of judicial opinions were not primarily the result of the inclusion of historical data. For the racial bias tests, we observed WEAT scores with moderate to large effect sizes indicative of negative racial bias in both the positive/negative legal WEAT and the grant/deny WEAT at all time periods for African American and Hispanic surnames as compared to European surnames. The bias effect sizes decreased slightly as less historical data was included for the positive/negative legal attribute WEATs, but remained relatively constant in the grant/deny WEAT. For Asian and Pacific Islander surnames as compared to European surnames, we observed the same pattern of negative biases that decrease slightly over time in the positive/negative

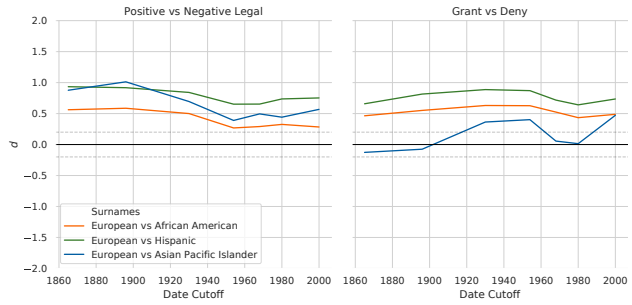

Figure 4: Temporal WEAT scores for race targets (surnames) and legal attributes

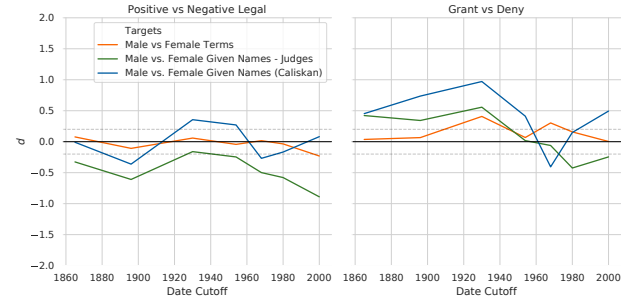

Figure 5: Temporal WEAT scores for gender-related targets and legal attributes

legal WEAT, but only some time periods were observed to have a moderate negative bias on the grant/deny WEAT, with the strongest observed bias being in the embedding trained on opinions from 2000-2020.

Similar to results from the full word embeddings, the bias scores for gender bias legal WEATs were dependent on the set of targets used, with bias scores near neutral for the generic male/female terms and large bias scores (both positive and negative) observed for the given name-based WEATs. While the temporal trends were relatively stable for the generic male/female terms, observed bias scores fluctuated in the given name based measures. This may indicate changes in gender bias over time that swing between positive and negative, but it should be noted that the gender specificity of given names also changes over time, making these results more difficult to interpret.

Since the gender-career bias was the strongest observed in our replication of the original WEAT (see Table 2), we also performed a temporal analysis of gender-career bias using both the original career and family terms from Caliskan et al. and the expanded set previously described. Figure 6 shows that the gender career bias was observed at all time periods and across all gender target types. This effect was extremely strong for the given name based measures and moderately strong for the male/female terms.

Topical Effects

In this section, we further investigate how gender and racial biases change when we only consider cases pertaining to

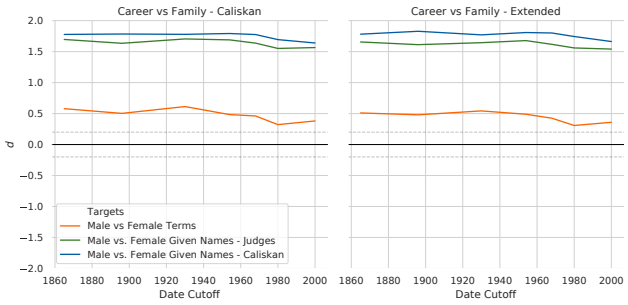

Figure 6: Temporal WEAT scores for gender-related targets and career/family attributes

a specific legal topic (for results related to racial biases, see the Supplementary Material[1]). To categorize the documents in our dataset, we rely on the seven main divisions of law provided by "West's Analysis of American Law" guide (Thomson Reuters Westlaw 2013). We define topical areas for each opinion using the Key Number classification for the headnotes written for the opinion. The seven main categories are contracts, crimes, government, persons, property, remedies, and torts. This guide also provides more granular subdivisions of the main topics, however, for our experiments here we only focus on the main divisions to capture the overall effects observed under each category. To make sure the analyses are not affected by a small sample size, while preparing the dataset for each legal category we removed the words in the target and attribute lists that have a frequency of less than 30 occurrences in the corresponding sub-corpus (see "Embedding Training" in "Proposed Approach"). Figures 7 – 9 illustrate the results of some of these tests (see the Supplementary Material[1] for the results of more experiments).

Figure 7 shows the results of three tests where the target list is "Male vs. Female Terms". We observe that the breakdown of the documents by their legal topic in the case of "Positive vs. Negative Legal" attribute list reveals strong biases in two categories: crimes and property. On the other hand, in the case of "Grant vs. Deny" attribute list we observe a significant bias in all legal topics except for crimes. Finally, as mentioned in the previous results, there exists significant bias in the case of "Expanded Career vs. Family" attribute list (see similar results for the "Career vs. Family (Caliskan)" attribute list in the Supplementary Material[1]) and this bias is consistent in terms of the large magnitude across all the different legal topics.

Figures 8 and 9 illustrate the results of four tests to compare the detected gender bias for the "Male vs. Female (Caliskan)" attribute lists (Figure 8) as a baseline against the legally adapted "Male vs. Female Judge Given Name" lists (Figure 9). These results demonstrate that choosing the legally adapted target lists reveal different type (i.e., sign of the effect size) and magnitude of the bias for each legal topic. Observe, for example, that in the case of "Positive vs. Negative Legal" the magnitude of the effect size of the legally adapted lists is smaller compared to the baseline for topics such as property and remedies, and larger in other

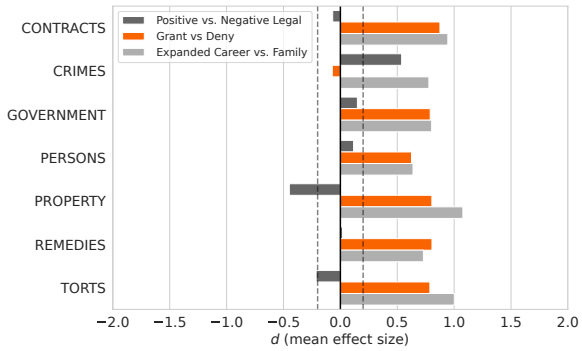

Figure 7: WEAT Cohen's effect sizes for "Male vs. Female Terms" target list. Different attribute lists are shown in different colors.

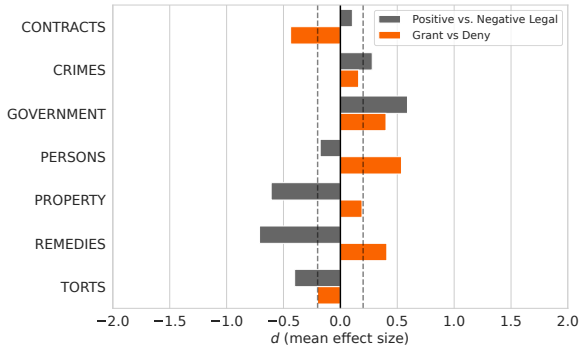

Figure 8: WEAT Cohen's effect sizes for "Male vs. Female (Caliskan)" target list. Different attribute lists are shown in different colors.

topics such as crimes, persons, and torts. We also observe a consistently larger effect size in the case of "Grant vs. Deny" for the legally adapted lists compared to the baseline.

Conclusions and Discussion

In this article we proposed a legally adapted approach for identifying gender and racial biases that are encoded in the word embeddings trained on the text of legal opinions from U.S. case law. This approach considers specific idioms used in legal language and also adapts the general bias detection WEAT method to legal language. The experiments designed in this work demonstrate the importance of domain adaptation for bias detection methods. If general purpose bias identification methods are used to measure gender and racial biases in word embeddings in the legal domain or other domains with specialized vocabularies, the developers of these systems may inadvertently create NLP systems that replicate or even amplify these biases in the world even after trying to screen their word embeddings for potential biases.

Using domain adapted bias detection methods is also important for evaluating the effectiveness of any potential mitigation strategy. We showed that using a date cut off is not an effective strategy for mitigating gender or racial biases present in the legal opinions even though societal opinions regarding these issues have changed over time. Our results also demonstrate that gender-career bias is particularly strong for given names in this domain, suggesting that downstream legal NLP systems that operate on these representations (e.g., coreference resolution) may be particularly likely to make biased predictions. Furthermore, we showed that analyzing the bias across different legal topics not only reveals different types of bias but also signifies the need for evaluating the system for fairness under different topics.

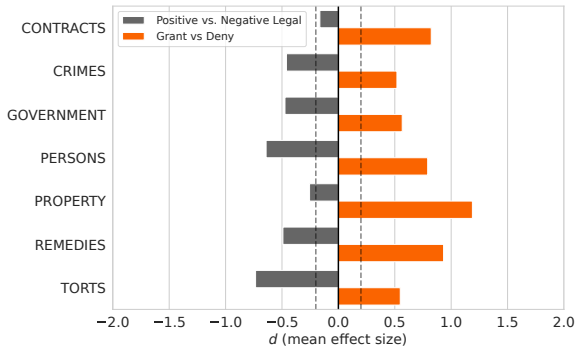

Figure 9: WEAT Cohen's effect sizes for "Male vs. Female Judge Given Name" target list. Different attribute lists are shown in different colors.

Future work in this area should also focus on the downstream effects exhibited by predictive systems that take biased representations as input as well as the effects any mitigation strategies have on these predictions. This work examines only biases in the representations themselves but the way that these biases could potentially cause harm in society is when they are used to make predictions that may be biased and the results of these predictions are displayed to users. The exact nature of the potential harms caused would depend on the specific application, but biased predictions made by these systems could be particularly harmful in contexts where users are not directly viewing the text these models are trained upon but instead are viewing aggregated predictions or summaries of results across many cases.

For example, if a motion outcome prediction system operating on racially biased word representations was deployed within a particularly diverse jurisdiction, it could under count the number of successful motions as compared to model performance in a less diverse jurisdiction. Attorneys representing clients in this jurisdiction might then be less likely to believe that a potential motion in a client's case would succeed based on a summary of historical outcomes and could suggest settlement in scenarios where they would have proposed continuing with the motion if the model had provided a more accurate prediction of outcomes within their jurisdiction. Under-representation of counter-stereotypical scenarios in legal research systems due to biases in predictive models operating on biased representations could ultimately contribute to degradation in the quality of legal representation or increased costs related to additional time required for legal research for individuals in protected classes.

Ethical Statement

This paper leveraged identity characteristics from the U.S. Census and a judicial biographical database to create target lists for the WEAT test. This work examined group fairness for both race and gender in word embeddings built from judicial opinions. While the aim of this work is to measure these potentially harmful representational biases in order to facilitate the creation of mitigation strategies for legal NLP systems that take these types of representations as input, the work could also be used to build intentionally harmful or biased legal NLP tools. Unfortunately, blindness itself leads to unfairness and we need to better understand the impact of stereotypes on legal decisions made by the judiciary (Nielsen 2020).

Acknowledgments

We would like to thank Frank Schilder, Brian Romer, and Nadja Herger for their guidance and support.

References

Bolukbasi, T.; Chang, K.-W.; Zou, J.; Saligrama, V.; and Kalai, A. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, 4356–4364. Red Hook, NY, USA: Curran Associates Inc. ISBN 978-1-5108-3881-9.

Bouma, G. 2009. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the GSCL*, 31–40. Tübingen, Germany: Gunter Narr Verlag.

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

Dieterich, W.; Mendoza, C.; and Brennan, T. 2016. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpoint Inc*, 7(7.4): 1.

Federal Judicial Center. 2012. Biographical Directory of Article III Federal Judges: Export. https://www.fjc.gov/history/judges/biographical-directory-article-iii-federal-judges-export. Accessed: 2021-09-03.

Free Law Project. 2021. Court Listener: Bulk Judicial Database Files. https://www.courtlistener.com/api/bulk-data/people/all.tar.gz. Accessed: 2021-05-03.

Greenwald, A. G.; McGhee, D. E.; and Schwartz, J. L. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6): 1464.

Kallus, N.; Mao, X.; and Zhou, A. 2020. Assessing algorithmic fairness with unobserved protected class using data combination. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 110. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-6936-7.

Lambrecht, A.; and Tucker, C. 2019. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science*, 65(7): 2966–2981.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013a. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013b. Distributed Representations of Words and Phrases and their Compositionality. arXiv:1310.4546.

Mozafari, M.; Farahbakhsh, R.; and Crespi, N. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *PLOS ONE*, 15(8): 1–26.

Nielsen, A. 2020. *Practical Fairness: Achieving Fair and Secure Data Models*. O'Reilly Media, Incorporated. ISBN 978-1-4920-7573-8.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.

Pennington, J.; Socher, R.; and Manning, C. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. Doha, Qatar: Association for Computational Linguistics.

Rice, D.; Rhodes, J. H.; and Nteta, T. 2019. Racial bias in legal language. *Research & Politics*, 6(2): 2053168019848930. Publisher: SAGE Publications Ltd.

Rice, D.; and Zorn, C. 2019. Replication Data for: "Corpus-Based Dictionaries for Sentiment Analysis of Specialized Vocabularies". Harvard Dataverse, V1, https://doi.org/10.7910/DVN/4EKHFM.

Rice, D. R.; and Zorn, C. 2021. Corpus-based dictionaries for sentiment analysis of specialized vocabularies. *Political Science Research and Methods*, 9(1): 20–35.

Suresh, H.; and Guttag, J. V. 2020. A Framework for Understanding Unintended Consequences of Machine Learning. arXiv:1901.10002.

Thomson Reuters Westlaw, ed. 2013. *West's Analysis of American Law*. Westlaw.

Vacek, T.; Song, D.; Molina-Salgado, H.; Teo, R.; Cowling, C.; and Schilder, F. 2019. Litigation Analytics: Extracting and querying motions and orders from US federal courts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 116–121. Minneapolis, Minnesota: Association for Computational Linguistics.