CSC 2611 Project Report: Gender bias word embedding association tests for distributional semantic vector representations of United States Supreme Court opinions corpora

Martin D. Pham

Abstract

We present an approach to legal discourse studies that makes use of semantic embeddings and their operationalization under statistical association tests in order to investigate gender bias in these texts. Firstly, the theoretical motivations for this work are presented. Then, a semantic embedding methodology and relevant data are discussed in order to concretize the motivations with computational techniques. Lastly, a discussion of the statistically insignificant results is presented with avenues of future work.

1 Introduction

Political polarization (PP) (Grossmann and Hopkins, 2016) is a recent topic of interest within political science as it incorporates elements of complex systems theory (Levin et al., 2021) to explain an increasingly dysfunctional form of two-party governance characterized by non-linear positive feedback (Leonard et al., 2021). Although work has been done in the analysis of PP regarding Congressional party strategy (Fishkin and Pozen, 2018), we are interested in a different branch of American federal governance: judicial. At the top of the judicial branch is the Supreme Court of the United States of America (SCOTUS), a set of judges interpreting constitution against laws, resulting in decisions for federal laws. There are two main streams of judicial thought regarding interpretation of the American Constitution (Solum, 2018): originalist (document is to be interpreted as if at the time of its writing) and living (document is to be interpreted as an evolving and updated set of interpretations). Autopoietic theory of law (ATL) is a hybrid of these views, a metaphor derived from research into ecological life (Teubner, 1993; Jacobson, 1989): documents are textual artifacts that refer to previous documents, forming a living body. Put another way: the legal system is a body (of documents describing norms) that reproduces itself by reference to historical cases and generation of legal precedence (i.e. other and more documents). In this way, one may consider the Constitution as initial conditions and decisions/amendments afterwards as updates to the state. One example is that of legal decisions: cases are brought forth to the court, both plaintiff and defendants citing previous relevant and related cases in order to convince the court to rule in their favour. This legal process generates textual artifacts (e.g. opinions, appeals, amicus briefs, etc...) from various agents involved (e.g. opinions are written by justices after a decision while amicus briefs are written by those not party to the legal case but whose information, expertise or insight may be useful to the court). At present, we do not apply the theoretical/philosophical consequences of this position any further than to motivate the linguistic investigation into legal corpora.

We relate ATL to PP through a particular stream of sociolegal thought called legal discourse analysis (DA) (Conley et al., 2019): these textual artifacts of the judicial process capture the linguistic elements of legal events, i.e. discourse in another form. Discourse analysis may be considered and integrated at different scales: e.g. micro, meso and macro (Lempert and Summerson Carr, 2016). Micro-level discourse relates to the text and communications within a conversation or event (such as court proceeding). Meso-level discourse relates to the context and processes of production for such texts. Macro-level discourse relates to the social structure and cultural factors that supervene on the former two. Given the individual effort and time-consumption required to conduct a micro-level discourse analysis for a single text, a computational and quantitative account, such as latent semantic analysis (Landauer et al., 1998), of such documents thus seems comportable with testing conceptual consequences of theories such as PP. This computational semantics approach offers scalability: the analysis may be automated over large corpora, something that would otherwise be

too time-consuming for a domain-expert legal studies researcher. The ability to machine-intelligently summarize features of a legal textual artifacts can provide new insights into legal theory and practice.

To make the connection between PP, ATL and DA more clear, consider the following: wherein rights belong to a subject then that entity also has an identity, however the social understanding of that identity is subject to change through the legal discourse and so linguistic models may be a useful ground upon which to test theories of political polarization (specific to gender identity for the present work). For instance, a person may share identities of both 'woman' and 'mother', however the interfaces of these identities may conflict within the legal space: if a woman has a right to abortion, what does that say about the identity of motherhood? It is here that DA is a useful frame for approaching the question, as these legal decisions regarding the nature of the identities 'woman' and 'mother' may be considered a discourse. These three themes (PP, ATL, DA) thus motivate the presented work: assuming a self-referential ATL generating different textual artifacts from a judicial process (i.e. the legal discourse), is there evidence of PP at the level of individual justices and their opinions and how does this express itself as gender bias in the vector word embeddings of the corpora? One wishes to answer, for example, do the opinions written by different political party-affiliated justices regarding landmark decisions differ in their identity construction and bias of gender? These descriptive models don't seek to displace the previous qualitative work done on micro-level DA in legal texts (which may sometimes include communicative elements beyond semantic content such as conversational form), but explore whether expected theoretical results (from PP, ATL, DA) can be brought to bear on the empirical linguistic evidence generated by judicial process.

The paper is organized as follows. Related work reviews qualitative and quantitative analyses of legal text corpora, vector embeddings methods for text corpora at large, and statistical bias tests for such representations. Computational methodology presents the treatment of embedding legal text corpora and discusses their operationalization under word embedding association tests. Data and other materials discusses the SCOTUS opinions dataset and related metadata. Results presents a comparative set of statistical tests applied for the SCOTUS data, followed by a discussion of the strengths and weaknesses of the method. Finally, we conclude with some possible extensions to address these weaknesses and further take advantage of computational methods to bring legal theory to bear on empirical evidence.

2 Related work

Qualitative analysis of discourse and identity construction (Bamberg et al., 2011; Schwartz et al., 2011) is an established field of research within critical studies that ranges from law to sociolinguistics. (Lipschultz, 2013) discusses social feminism (i.e. first-wave feminism) and legal discourse between the years of 1908-1923. Gendered identities constructed through the legal discourse are further examined by (Smart, 1992) and (Ehrlich, 2007). Of particular interest to this intersection between gender identity and discourse is that of (Berkovitch, 1997) which found strong motherhood-related features to the identity construction of woman in the Israeli legal discourse. Useful for theoretical conceptualization but difficult to conduct case-by-case, these linguistic case studies may be scaled up using automated textual analysis of large legal corpora.

Textual analysis (of opinions, briefs, or any other kind of legal document) is congruent to the professional work of legal scholars (i.e. reading through historical cases to find a common and arguable thread) and so empirical, quantitative methods for systematic content analysis have gained attention (Salehijam, 2018). We are primarily interested in automated, scalable and robust descriptive models of legal documents that may be extended into theoretically grounded predictive models. Current work on textual analysis is mostly based on various metrics differing in their grouping of decisions, i.e. computed summary statistics on documents grouped based either topically (e.g. any decisions related to the First Amendment) or over time (e.g. any decisions during certain decades of interest). (Carlson et al., 2015) provides an analysis of writing styles of entire corpus of SCOTUS decisions from 1972-2008 considering diversity of words (i.e. number of types) and average opinion length by justice; they found that contemporaneous justices wrote more similarly to peers than temporally remote justices (across time). (Coleman and Phung, 2010) evaluated SCOTUS court briefs from 1969-2004, finding a gradual historical trend towards plainer legal writing using Flesch read-

ability scores. (Owens et al., 2013) alternatively used Coleman-Liau readability index to examine readability of 500 randomly selected SCOTUS majority opinions from 1953-2009, suggesting that justices may deploy obscurant language to avoid congressional review. (Okdie and Rempala, 2019) studied unanimous SCOTUS verdicts from 1940-2019 to examine if they contained perceived threat and resistance to change content, and whether this predicted authors' political affiliation. (Dickerson, 1996) conducted textual analysis of six major SCOTUS decisions from 1918-1989 regarding interpretation of First Amendment, identifying six principle metaphors of freedom of expression (i.e. guardianship, foundation, light source, medicine, commodity, space). (Rice, 2019) used structural topic models from 1803-2010 to estimate the effect of time on the prevalence of topics (i.e. amount of attention paid to given issue over time), also looking at correlations between topics.

A new class of machine-learning techniques, word embeddings (Mikolov et al., 2013), represent word symbols as real-valued vectors whose semantic content/meaning is based on the distributional hypothesis of semantics (Harris, 1954). The crux of these techniques rely on the matrix representation densification of word type co-occurences, with Latent Semantic Analysis (LSA) (Landauer et al., 1998) as one of the first methods to explore this by applying singular value decomposition to the pointwise mutual information matrix of a word co-occurence matrix from a text corpus. More modern techniques take a similar dimensional reduction approach but make use of connectionist (i.e. neural networks) models. See (Ethayarajh, 2019) for a comparative interpretation of these more advanced embeddings. Although (Devlin et al., 2018) found that it is possible to pre-train such models and fine-tune their parameters for specific downstream problems, we take the traditional LSA approach without fine-tuning to ground the semantic content of vector representations in only the decision opinions of the SCOTUS (in order to keep with the boundaries drawn by ATL for this analysis). Indeed such pre-trained models could be used for this data but it is important to treat legal artifacts as self-referential (i.e. only include text from documents that appears in the legal discourse).

Using these word embeddings, there has been some work with traces of PP such as (Huang and Lee, 2019) which measured hyperpartisanship in news articles and headlines. A related work is that of (KhudaBukhsh et al., 2020) which makes use of machine translation methods to interpret political polarization as a difference in languages.

Semantic embedding vectors can be operationalized under statistical tests that make use of vector space distance representing semantic similarity (i.e. the closer two vectors are in their vector space, the more they may be considered 'semantic neighbours' that are similar in meaning). Of particular interest to computational sociolinguists is that of social bias in embeddings (Mehrabi et al., 2021; Nissim et al., 2020): does the semantic relations of a particular embedding technique reflect social biases inherent in the processes generating the data? Such biased representations may present themselves along many social dimensions: e.g. moral (Xie et al., 2020), ethnic (Choi et al., 2022), class (Kozlowski et al., 2019), or gender (Bhatia and Bhatia, 2021). Word embedding association test (WEAT) is the word embedding analog to the implicit association test (IAT) introduced by (Caliskan et al., 2017). The associations between two sets of target and attribute words (respectively) are statistically tested, with the null hypothesis that there are no difference in associations between pairs of target-attribute sets. This, in effect, establishes/characterizes bias as a direction in the semantic vector space.

3 Computational methodology

Our methodology is as follows: for a subset corpora (based on political direction or opinion category), an LSA word embedding is constructed and the WEAT is applied to test for gender bias. We make use of the target and attribute topic word sets available in the gender bias tests done by (Charlesworth et al., 2021) (i.e. female, male vs. home, work; female, male vs. good, bad). Dataset preprocessing such as tokenization and cleaning is done while keeping any of the target or attribute words of interest in the corpus. That is, words such as 'he' may usually be considered a stopword with little semantic information that would normally be removed but we keep gendered terms in order to test their associations after embedding. The WEAT test is conducted using (Badilla et al., 2020).

We are interested in investigating PP and so we will compare word embeddings from different opinions across court decision category and political ideological direction. For instance, we may con-

sider testing the word embedding bias of two separate corpora where one may be all majority opinions for a liberal SCOTUS decision and the other may be all dissenting opinions for the same (liberal) decision. This may then be flipped to consider the bias effect present in conservative decisions (and whether those biases are present in majority or dissenting opinions). Alternatively, we may test embedding bias along different metadata such as authoring justice. This would provide an embedding bias tested account of the language deployed by a specific judge of interest.

Here, the three aforementioned themes play out: we make use of legal documents as self-referential text data as per ATL (giving us a legal-text corpora), we seek only those decision opinions related to our interest in DA (gender identity construction), and finally we then split our corpora of interest based on PP (using the metadata available to draw an ideological distinction between the two parties) to conduct statistical tests on them.

Code and Jupyter Lab notebooks producing the results below may be found at `https://github.com/mdpham/CSC2611` under `projects`.

4 Data and other materials

(Fiddler, 2020) is a dataset containing 36000 opinions authored by 98 justices of the SCOTUS spanning the years 1789-2020. It includes useful metadata such as case name, federal citation, year, category of opinion (e.g. majority, dissent, concurring) and authoring justice. One important column of metadata is ideological direction: liberal, conservative, not-applicable. The criteria applied to categorize a decision's direction and assign an ideological direction can be found in (Spaeth and Epstein, 2021). This diachronic dataset may be subsampled to landmark cases related to a topic of interest using case name and federal citation values for each opinion. (American Civil Liberties Union, 2022) is a timeline written by the American Civil Liberties Union describing the chronology of 113 landmark women's rights decisions by the SCOTUS between 1965-2017. The opinions belonging to landmark decisions that match the case name or federal citation from (Fiddler, 2020) are used to construct word embeddings. The resulting ACLU corpora matched 37 of 113 landmark decisions, providing 92 opinions from 20 different authors. Summary plots and tables of the ACLU corpora can be found in Appendix A. Matching the ACLU decisions to opinions available in the SCOTUS dataset is a matter of annotation and curation: matches may be missed due to small differences in case names between the two or unavailability of federal citation values in the latter. For example, the corpora is missing the 1992 landmark case citation '505 U.S. 833' which has full name 'Planned Parenthood of Southeastern Pennsylvania, et al. v. Robert P. Casey, et al.' but is often shortened to 'Planned Parenthood v. Casey'.

We note that the current analysis is applied to decision opinions but similar techniques may be extended to other kinds of documents according to the autopoietic theory of law. Appeals are documents filed in order to seek the reconsideration of a decision from lower courts, and the SCOTUS is responsible with being the final arena of appeals in American law. For instance, one could apply this analysis to (Hurwitz and Kuersten, 2012) which is a dataset of appeals (instead of opinions) from US courts.

5 Results and GitHub

Tables below show the resulting WEAT scores (ranging from 2 to -2), effect size and p-values. We keep the female/male target words and apply the WEAT test to two different groups of attribute dichotomies: home/work and good/bad. For the global ACLU corpora, we compare these results for two different sized LSA embedding vector dimensionalities: 100 and 500.

Attribute	WEAT score	Effect size	p-value
Home/Work	-0.64	-0.59	0.96
Good/Bad	-0.74	-0.35	0.84

Table 1: WEAT test results for female/male target words with different attribute words from (Charlesworth et al., 2021). All opinions matched by ACLU timeline embedded with dimensionality 100.

Attribute	WEAT score	Effect size	p-value
Home/Work	-0.15	-0.37	0.86
Good/Bad	-0.29	-0.22	0.73

Table 2: WEAT test results for female/male target words with different attribute words from (Charlesworth et al., 2021). All opinions matched by ACLU timeline embedded with dimensionality 500.

We then apply this same treatment to subsets of the corpora based on ideological direction and

opinion category for vector dimensionality 500 as seen in table 3 and 4.

Attribute	WEAT score	Effect size	p-value
Home/Work	-0.28	-0.56	0.92
Good/Bad	-0.18	-0.03	0.54

Table 3: WEAT test results for female/male target words with different attribute words from (Charlesworth et al., 2021). Corpora restricted to only those that had liberal direction and were majority opinions (total of 23 opinions).

Attribute	WEAT score	Effect size	p-value
Home/Work	-0.31	-1.05	0.99
Good/Bad	0.06	0.28	0.28

Table 4: WEAT test results for female/male target words with different attribute words from (Charlesworth et al., 2021). Corpora restricted to only those that had conservative direction and were majority opinions (total of 14 opinions).

6 Discussion

Both tables 1 and 2 show that when applied to the entire ACLU opinions corpora, there is a weak association directionality that is statistically insignificant. This poor statistical confidence is only further magnified when subsampling the corpora based on ideological direction and category of opinion as seen in tables 3 and 4. The inaccuracy of the test results may be compared against the frequency of target/attribute words in the text (available in the Jupyter Lab) for some insight into how representative these topical groups from (Charlesworth et al., 2021) are: both target groups (female and male) show up in the order of hundreds but attribute words are only present in the order of tens. The ACLU corpora subset of the SCOTUS dataset may simply be too small to find statistical significance given these particular word groups.

Given this, there are some aspects that may be discussed.

- Recall that only a small subset of matched cases from the SCOTUS are used for embedding (the ones labeled by the ACLU as being landmark decisions for womens' rights): greater annotative work must be done on the dataset in order to include more opinions from the timeline, providing more statistical confidence.
- Note that the p-value of the global corpora embedding tests decreases with an increase in the dimensionality, this suggests there may be an optimal dimensionality to perform the WEAT tests.
- LSA is a basic dense representational transformation on the co-occurence matrix of the corpora: its modern neural-network descendants are much more sophisticated in the mathematical treatment of text embeddings. However, as discussed in the motivation for using LSA, this may become unwieldy when dimension reduction interpretation is required. This may become particularly difficult when discussing possibly useful properties of an embedded semantic space (such as hierarchical meaning, as in (Nickel and Kiela, 2017)).
- The meta-organization of corpora for embedding may be improved by leveraging further domain knowledge: our selection of ideological direction and category of opinion are useful dimensions for investigating PP but we may also look at the individual justices to track which judge contributes to the bias of an embedding.

7 Conclusion and Future Work

By applying WEAT to semantic embeddings of SCOTUS opinions, we found no significant gender bias to dichtonomies such as home/work and good/bad. However, this should not posit that there are no gender biases in these texts. Indeed, given the politically tumultuous history of womens' rights, one should expect an inherent bias that is computationally detectable (e.g. for a case discussing the right to abortion, given the liberal/conservative schismogenesis based on individual freedoms and morality, one would expect some kind of bias from either position).

Returning back to our themes of PP, ATL, and DA: semantic embeddings for identity analysis in the legal discourse making use of different political metadata (such as ideological direction) may be a viable way to computationally reproduce previous qualitative and conceptual results. However, the largest blocker to this may be the availability of large scale legal textual data (relevant to the identity construction of interest). In order to address this, a wider view of the legal system may be taken to include more kinds of documents rather than just decision opinions as in the present paper: ATL includes not just opinions but other kinds of doc-

uments (e.g. appeals) as well. A semantic model of identity construction in the legal discourse must necessarily include these other kinds of artifacts generated within the legal system. Furthermore, in order to test questions of PP, additional metadata must be compiled for the available data (such as president-in-power, or which party nominated the justice, etc...).

Two streams of future work may be identified: (1) the robustness and automation of certain workflow steps, (2) a more sophisticated model that makes use of concepts from the political polarization theory. (1) may be addressed on many fronts: matching more ACLU cases requires an annotation of case names and federal citations (this may be done manually for completeness); target and attribute words are from a more generic dataset of text and so it may be useful to automatically generate relevant sets of attribute words for the identity construction of interest at hand (this may be seen as automatic topic modelling); and more sophisticated embedding methods trained on the legal discourse data (i.e. not pre-trained and fine-tuned). (2) may be addressed by the use of differential bias (Brunet et al., 2019): the marginal contribution of embedding bias may be calculated by selective combination of corpora for embedding depending on the metadata. The axes upon this marginal contribution may be selected based on the metadata previously discussed: one may ask to what degree a specific justice contributed to bias by comparing the embeddings with and without their opinions in the corpora.

Lastly a note on descriptive vs predictive statistical models. Presented above are the beginnings of a more comprehensive descriptive model of historical decisions and the gender bias present in those text artifacts. Nothing has been said of predictive models, indeed this is a task to be done after the descriptive models have been constructed and verified. Such predictive models may have the hypothetical form of: "given a short sample of an opinion and its category, is it possible to predict the ideological direction of the decision?" In order to move towards such predictive models (possibly leveraging the metadata), descriptive models (that are interpretable) must first be generated that leverages both sociolinguistic theory and statistical semantic vector operationalization.

Acknowledgements

This document has been prepared after extensive theoretical and conceptual sociolegal and political linguistic discussions with Joe Hoang and Daisy Liu from the Mind and Body Lab at the University of Toronto.

References

ACLU American Civil Liberties Union. 2022. Timeline of major supreme court decisions on women's rights. www.aclu.org/other/timeline-major-supreme-court-decisions-womens-rights. Accessed: 2022-02-14.

Pablo Badilla, Felipe Bravo-Marquez, and Jorge Pérez. 2020. Wefe: The word embeddings fairness evaluation framework. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 430–436. International Joint Conferences on Artificial Intelligence Organization.

Michael Bamberg, Anna De Fina, and Deborah Schiffrin. 2011. Discourse and identity construction. In *Handbook of identity theory and research*, pages 177–199. Springer.

Nitza Berkovitch. 1997. Motherhood as a national mission: The construction of womanhood in the legal discourse in israel. In *Women's Studies International Forum*, volume 20, pages 605–619. Elsevier.

Nazlı Bhatia and Sudeep Bhatia. 2021. Changes in gender stereotypes over time: A computational analysis. *Psychology of Women Quarterly*, 45(1):106–125.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *International conference on machine learning*, pages 803–811. PMLR.

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

Keith Carlson, Michael A Livermore, and Daniel Rockmore. 2015. A quantitative analysis of writing style on the us supreme court. *Wash. UL Rev.*, 93:1461.

Tessa ES Charlesworth, Victor Yang, Thomas C Mann, Benedek Kurdi, and Mahzarin R Banaji. 2021. Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, 32(2):218–240.

Donghyun Danny Choi, J Andrew Harris, and Fiona Shen-Bayh. 2022. Ethnic bias in judicial decision making: Evidence from criminal appeals in kenya. *American Political Science Review*, pages 1–14.

Brady Coleman and Quy Phung. 2010. The language of supreme court briefs: A large-scale quantitative investigation. *J. App. Prac. & Process*, 11:75.

John M Conley, William M O'Barr, and Robin Conley Riner. 2019. *Just words: Law, language, and power*. University of Chicago Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Donna L Dickerson. 1996. and cultural meaning: An analysis of metaphors in selected supreme court texts. *Communication Law and Policy*, 1(3):367–395.

Susan Ehrlich. 2007. Legal discourse and the cultural intelligiblity of gendered meanings 1. *Journal of Sociolinguistics*, 11(4):452–477.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*.

Garrett Fiddler. 2020. SCOTUS Opinions: full text and metadata of all opinions written by scotus justices through 2020. www.kaggle.com/gqfiddler/scotus-opinions. Accessed: 2022-02-14.

Joseph Fishkin and David E Pozen. 2018. Asymmetric constitutional hardball. *Columbia Law Review*, 118(3):915–982.

Matt Grossmann and David A Hopkins. 2016. *Asymmetric politics: Ideological Republicans and group interest Democrats*. Oxford University Press.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Gerald Ki Wei Huang and Jun Choi Lee. 2019. Hyperpartisan news and articles detection using bert and elmo. In *2019 International Conference on Computer and Drone Applications (IConDA)*, pages 29–32. IEEE.

Mark Hurwitz and Ashlyn Kuersten. 2012. Changes in the circuits: Exploring the courts of appeals databases and the federal appellate courts. *Judicature*, 96:23–34.

Arthur J Jacobson. 1989. *Autopoietic Law: The New Science of Niklas Luhmann*. JSTOR.

Ashiqur R KhudaBukhsh, Rupak Sarkar, Mark S Kamlet, and Tom M Mitchell. 2020. We don't speak the same language: Interpreting polarization through machine translation. *arXiv preprint arXiv:2010.02339*.

Austin C Kozlowski, Matt Taddy, and James A Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949.

Thomas K Landauer, Peter W Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Michael Lempert and E Summerson Carr. 2016. *Scale: Discourse and dimensions of social life*. University of California Press.

Naomi Ehrich Leonard, Keena Lipsitz, Anastasia Bizyaeva, Alessio Franci, and Yphtach Lelkes. 2021. The nonlinear feedback dynamics of asymmetric political polarization. *Proceedings of the National Academy of Sciences*, 118(50).

Simon A Levin, Helen V Milner, and Charles Perrings. 2021. The dynamics of political polarization.

Sybil Lipschultz. 2013. Social feminism and legal discourse, 1908-1923. In *At The Boundaries of Law*, pages 209–225. Routledge.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Maximillian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Malvina Nissim, Rik van Noord, and Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497.

Bradley M Okdie and Daniel M Rempala. 2019. Brief textual indicators of political orientation. *Journal of Language and Social Psychology*, 38(1):106–125.

Ryan J Owens, Justin Wedeking, and Patrick C Wohlfarth. 2013. How the supreme court alters opinion language to evade congressional review. *Journal of Law and Courts*, 1(1):35–59.

Douglas Rice. 2019. Measuring the issue content of supreme court opinions. *Journal of Law and Courts*, 7(1):107–127.

Maryam Salehijam. 2018. The value of systematic content analysis in legal research. *Tilburg Law Review*, 23(1-2):34–42.

Seth J Schwartz, Koen Luyckx, and Vivian L Vignoles. 2011. *Handbook of identity theory and research*. Springer.

Carol Smart. 1992. The woman of legal discourse. *Social & Legal Studies*, 1(1):29–44.

Lawrence B Solum. 2018. Originalism versus living constitutionalism: The conceptual structure of the great debate. *Nw. UL Rev.*, 113:1243.

Harold J Spaeth and Lee Epstein. 2021. 2021 supreme court database, version 2021 release 1. http://supremecourtdatabase.org. Accessed: 2022-04-11.

Gunther Teubner. 1993. *Law as an autopoietic system*. Blackwell Publishers.

Jing Yi Xie, Renato Ferreira Pinto Jr, Graeme Hirst, and Yang Xu. 2020. Text-based inference of moral sentiment change. *arXiv preprint arXiv:2001.07209*.

A ACLU corpora visualizations

Decision direction	Count
Liberal	23
Conservative	14
Unspecifiable	0

Table 5: Ideological direction of decisions for matched ACLU landmark cases.

Opinion Category	Count
Majority	37
Dissenting	24
Concurring	19
Second dissenting	10
Per curiam	2

Table 6: Category of opinion for matched ACLU landmark case decisions.

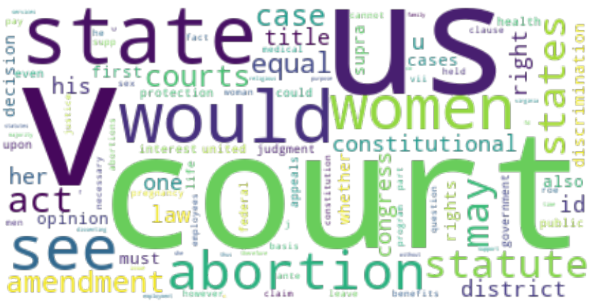

Figure 1: Wordcloud for ACLU corpora

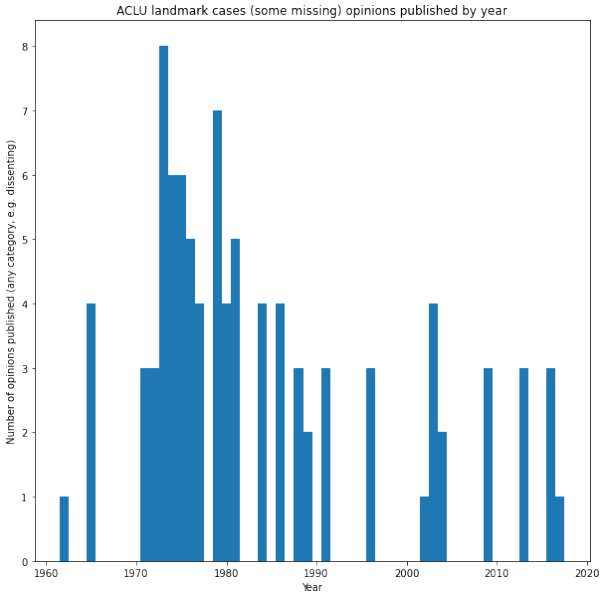

Figure 2: Number of opinions matched from ACLU timeline of decisions. Only 37 of 113 landmark cases were matched between the SCOTUS opinion dataset and ACLU timeline.

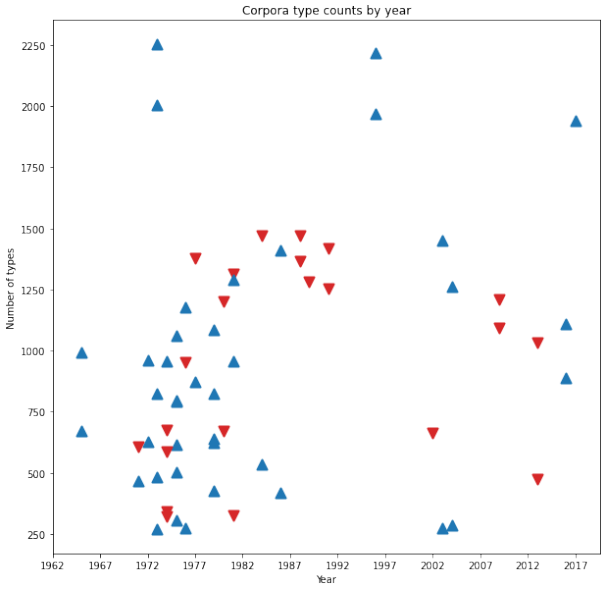

Figure 3: Types scatter plot for majority and dissenting ACLU women's rights landmark decision opinions. Horizontal axis is time, vertical access is number of unique types for each decision. Upward/downward triangles are majority/dissenting opinions, respectively. Blue/red colouring is measured liberal/conservative ideological direction.

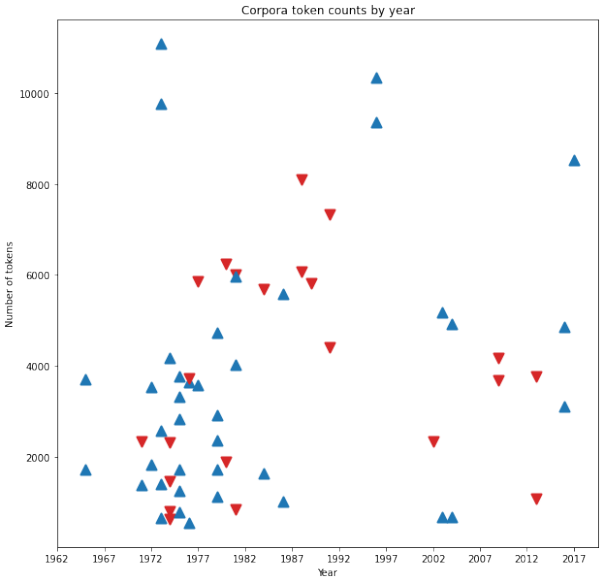

Figure 4: Tokens scatter plot for majority and dissenting ACLU women's rights landmark decision opinions.

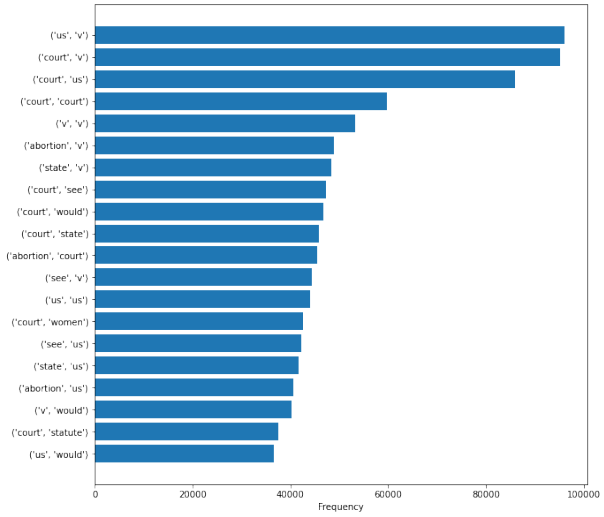

Figure 6: Most common bigram histogram for ACLU corpora.

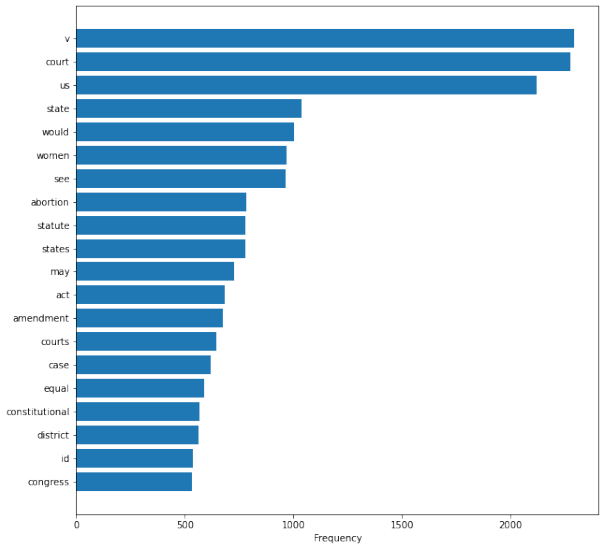

Figure 5: Most common unigram histogram for ACLU corpora.

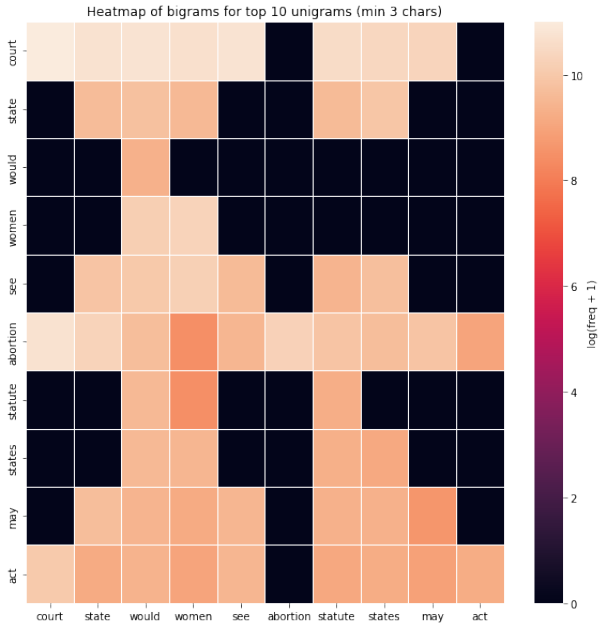

Figure 7: Bigram heatmap for 10 most common unigrams (minimum character length 3) from ACLU corpora.

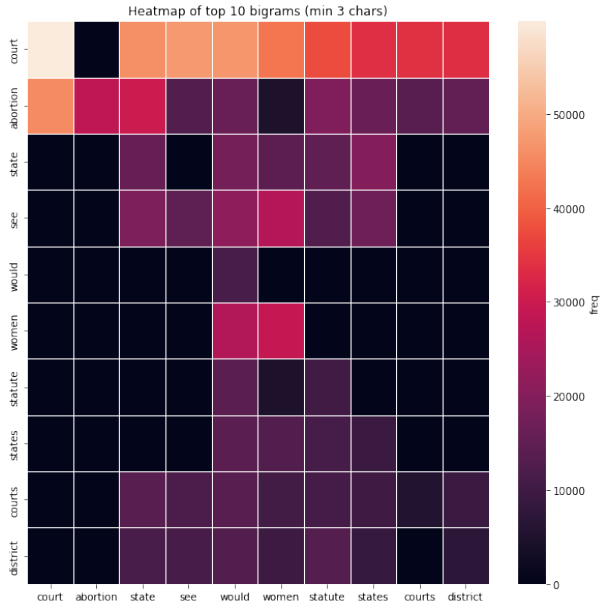

Figure 8: Heatmap of most common bigrams from ACLU corpora.