Analysis of Vocal Implicit Bias in SCOTUS Decisions Through Predictive Modelling

Ramya Vunikili[1], Hitesh Ochani[1], Divisha Jaiswal[1], Richa Deshmukh[1], Daniel L. Chen[2*], Elliott Ash[3*]

[1]Department of Computer Science, Courant Institute of Mathematical Sciences, New York University, NY, USA
[2]University Toulouse Capitole, France
[3]Center for Law & Economics, ETH Zurich, Switzerland
[*]Corresponding authors: daniel.li.chen@gmail.com, elliott.t.ash@gmail.com

Abstract

Several existing pen and paper tests to measure implicit bias have been found to have discrepancies. This could be largely due to the fact that the subjects are aware of the implicit bias tests and they consciously choose to change their answers. Hence, we've leveraged machine learning techniques to detect bias in the judicial context by examining the oral arguments. The adverse implications due to the presence of implicit bias in judiciary decisions could have far-reaching consequences. This study aims to check if the vocal intonations of the Justices and lawyers at the Supreme Court of the United States could act as an indicator for predicting the case outcome.

Key words: Speech analysis, Implicit gender bias, Machine learning, SCOTUS, FAVE

Introduction

Supreme Court of the United States (SCOTUS) is the highest federal court of the country. The cases heard by it are of utmost importance. This gives us the major motivation to employ machine learning techniques to check for the presence of any implicit bias. The SCOTUS comprises of the Chief Justice of United States and eight associate judges. There lies a huge responsibility in their hands to make rational decisions. However, it would be unrealistic to assume that all decisions are rational and unbiased.

This study aims to explore the relationship between implicit gender bias in the oral arguments and the final outcome of the case. In other words, we analyze if the features related to vocal intonations and masculine/feminine style of the speaker is an indicator of their implicit gender bias.

Related Work

According to a study done by Chen *et al.* (2016), it is observed that the perceived masculinity has a negative correlation with the winning of a case. Further, studies done by Klofstad *et al.* (2012) and Tigue *et al.* (2012), claim that individuals with lower-pitched male voices are often associated with higher competence and trustworthiness.

Dataset

The SCOTUS oral arguments have been recorded since October 1995. These recordings along with their transcriptions are available on the Oyez website (See: https://www.oyez.org/). This study uses 1246 cases from the SCOTUS collected during the years 1998, 1999 and 2003-2012. In addition to the recordings and transcriptions of these cases, we also gathered information about the Justices (gender, year of birth, party of appointing President, Segal-cover score etc), lawyers (gender, total number of cases involved, number of cases that involves him\her as a petitioner, number of cases that involves him\her as a respondent etc) along with the case specific information like the issue date, name of the case and the winner etc. In total there are about 2,137 hours of lawyers' recordings and 502 hours of the Justices' recordings.

Based on the type of speakers and their order of speech there are two types of pre-processed datasets – ABA and AxByA. In the AxByA dataset, A and B refer to two different Justices while x and y can represent same or different lawyers. Similarly, in the ABA format, A is always a Justice and B can represent both lawyers or Justices.

Methodology

Data Pre-processing

Using a list of 135 masculine words (such as uncle, man etc) and 135 feminine gendered words (such as sister, waitress etc), we classified all the relevant words spoken by the Justices and lawyers from ABA and AxByA datasets into three classes – masculine, feminine and neutral (neither masculine nor feminine). Among these, 60% were used as the training set, 20% as validation set and the remaining 20% as the test set.

In order to perform hard classification on each dataset, we trained a random forest classifier with hyperparameters optimized based on the validation set. With the number of estimators for the model fixed as 100, we achieved an accuracy of 78.9% on the AxByA test set and an accuracy of 83.3% on the ABA test set.

Further, we added features related to the interruption of a speaker based on the timestamps in the transcriptions i.e., if a Justice has interrupted a lawyer or a Justice has been interrupted by another Justice.

Modelling

In order to predict if the vote of a particular Justice is going to be in favour of or against a lawyer, we've trained two models. They are Extreme Gradient Boosting (XGBoost – Baseline) and Linear Support Vector Machine (SVM – Enhanced Model). For each case, we've extracted features such as the number of masculine and feminine words spoken by the Justice and the lawyer, the number of neutral words spoken by each of them that are classified into masculine and feminine words, the number of times a Justice was interrupted by male/female lawyers and the number of times a Justice interrupts a male/female lawyer, gender of the lawyer and the Justice, the ratio of neutral words that are classified into masculine words for a Justice and the ratio of neutral words that are classified into feminine words for the same Justice. These features were then normalized before training the models.

The best hyperparameters for each model are retrieved by tuning the models on the validation set. These hyperparameters were then used for prediction on the test set. Table 1 and Table 2 give the list of hyperparameter for each model.

Table 1. XGBoost Hyperparameters.

XGBoost Parameter	Value
learning_rate	0.03
max_depth	10
n_estimators	50
objective	binary:logistic

Table 2. SVM Hperparameters.

SVM Parameter	Value
C	0.03
loss	hinge
penalty	L2
tol	0.0001

Results

From Table 3, it can be observed that SVM performs better than XGBoost in predicting the vote of a Justice. While the accuracy of XGBoost is only about 46.85% on the test set, the SVM has a slightly better accuracy of 51.13%.

Table 3. Accuracy of the models on training and test sets.

Model	Train Accuracy	Test accuracy
XGBoost	60.03%	46.85%
SVM	51.57%	51.13%

Though the accuracy of prediction in either case isn't outstanding the most important features that contributed to the vote prediction such as the number of masculine/feminine words spoken by a Judge and their ratio with the neutral words are found to be at the top in both the models. This can be observed in Figure 1.

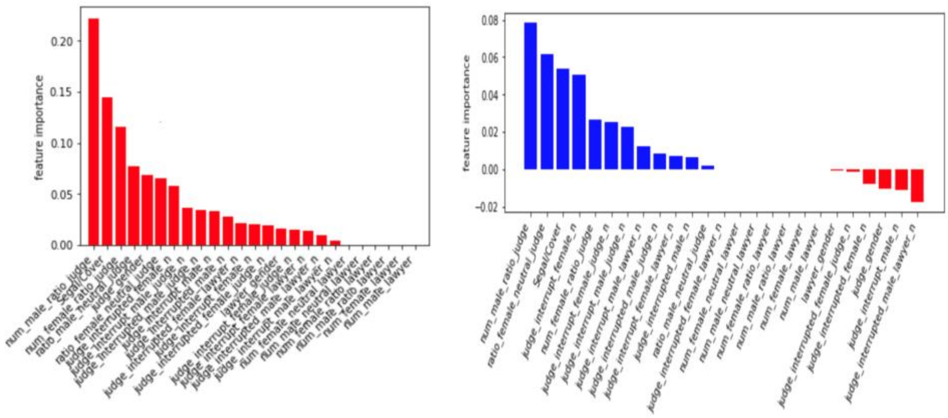

Figure 1. Feature importance of XGBoost (on the left) and SVM (on the right).

Acknowledgements

We've developed this project based on the technical assistance provided by Mufei Li, Kailin Luo, Jiayun Yu, Xuan Wang.

References

Chen, Daniel L., Halberstam, yosh, and Yu, Alan C.L. (a) "*Covering: Mutable Characteristics and Perceptions of Voice in the US Supreme Court",* No. 16-180. Toulouse School of Economics (TSE), 2016.

Klofstad, Casey A., Rindy C. Anderson, and Susan Peters, "*Sounds like a winner: voice pitch influences perception of leadership capacity in both men and women"*, Proceedings of the Royal Society of London B: Biological Sciences, 2012, 279 (1738), 2698–2704. Surname, A., Surname, B. (eds.) 2003. Linguistic Structures. Place, Publisher.

Tigue, Cara C., Diana J. Borak, Jillian J.M. O'Connor, Charles Schandl, and David R. Feinberg, "*Voice pitch influences voting behavior*," Evolution and Human Behavior, 2012, 33, 210–216.