Font-Pomarol et al. *EPJ Data Science* (2024) 13:57
https://doi.org/10.1140/epjds/s13688-024-00494-x

EPJ Data Science
a SpringerOpen Journal

EPJ.org

RESEARCH Open Access

Language and the use of law are predictive of judge gender and seniority

Lluc Font-Pomarol[1], Angelo Piga[1,2], Sergio Nasarre-Aznar[3], Marta Sales-Pardo[1*] and Roger Guimerà[1,4*]

*Correspondence:
marta.sales@urv.cat;
roger.guimera@urv.cat
[1]Department of Chemical Engineering, Universitat Rovira i Virgili, 43007 Tarragona, Spain
[4]ICREA, 08010 Barcelona, Spain
Full list of author information is available at the end of the article

Abstract

There are examples of how unconscious bias can influence actions of people. In the judiciary, however, despite some examples there is no general theory on whether different demographic attributes such as gender, seniority or ethnicity affect case sentencing. We aim to gain insight into this issue by analyzing over 100k decisions of three different areas of law with the goal of understanding whether judge identity or judge attributes such as gender and seniority can be inferred from decision documents. We find that stylistic features of decisions are predictive of judge identities, their gender and their seniority, a finding that is aligned with results from analysis of written texts outside the judiciary. Surprisingly, we find that features based on legislation cited are also predictive of judge identities and attributes. While own content reuse by judges can explain our ability to predict judge identities, no specific reduced set of features can explain the differences we find in the legislation cited of decisions when we group judges by gender or seniority. Our findings open the door for further research on how these differences translate into how judges apply the law and, ultimately, to promote a more transparent and fair judiciary system.

Keywords: Gender differences; Topic model; Judicial decisions

1 Introduction

Social constructs and cultural stereotypes are ubiquitous and lead to unconscious bias in people's actions [1–3], even in scenarios were individuals are explicitly trained and expected to be impartial and objective, such as job interviewing [4, 5], evaluation of college applications [6], peer reviewing [7] or judicial sentencing [8]. In the scope of legal studies, many efforts have been devoted to study the effect of judges' personal attributes on the outcome of cases, showing that gender [9–11], ethnicity [12, 13], age or political affiliation [14, 15] can influence how cases are decided. Despite these efforts, we still lack a general theory; while some attributes such as ideology and partisanship have a clear effect in case sentencing, others such as gender and race present mixed or inconclusive effects [10, 11, 16, 17].

Justices determine the relative position of certain facts, events and actions in relation to the current applicable law. The rationale for the ultimate decision in legal cases is made explicit in the text of the judicial decisions. Thus, going beyond case outcomes by studying

Springer

more subtle differences in the content of such documents can help to understand how individual and group differences of judges intrude in the legal process [18, 19].

Indeed, linguistic and textual differences are predictive of the author of literary texts [20–22], and also predictive demographic group attributes such as gender of social media content authors [23, 24], and age and mental state of anonymous texts [25]. Author profiling is arguably a useful task by itself – for instance in forensics it is a requirement for using written evidence in criminal cases [22, 25]. However, the ability to reveal the demographic attributes of the authors from the written content they generate provides a way to unveil the inherent differences that exist between the corresponding demographic groups. In most of the previously mentioned domains, group differences are more pronounced in those aspects concerning the style of the text [26–28]; however, in a few examples, group differences are more pronounced in content-related aspects such as the main ideas or the topics discussed [29, 30].

When writing decisions, reporting judges tend to display a recognizable style in the form of paraphrasable content which can be expressed in formal or informal language without changing the meaning [31]. Then, given that judges are constrained by the law and that they do not participate in the case assignation process, finding differences that go beyond style might be linked to bias in the judicial process. Here, we explore whether there are measurable differences linked to the attributes of judges that translate into the content of decisions; and we investigate the extent to which these differences are just stylistic or instead substantial to the legal content. To do so, we take three large corpora of almost 100K judicial decisions that correspond to three legal fields in the Spanish judicial system: homicides, condominiums, and housing. We extract features that characterize different aspects of decisions (Fig. 1): (i) stylistic and non-content features, such as function words [32, 33], court id or year of the decision; and (ii) content-related features, such as content words and references to the law. We then consider the attributes of reporting judges (their identity, gender and seniority) and measure the extent to which each of the mentioned

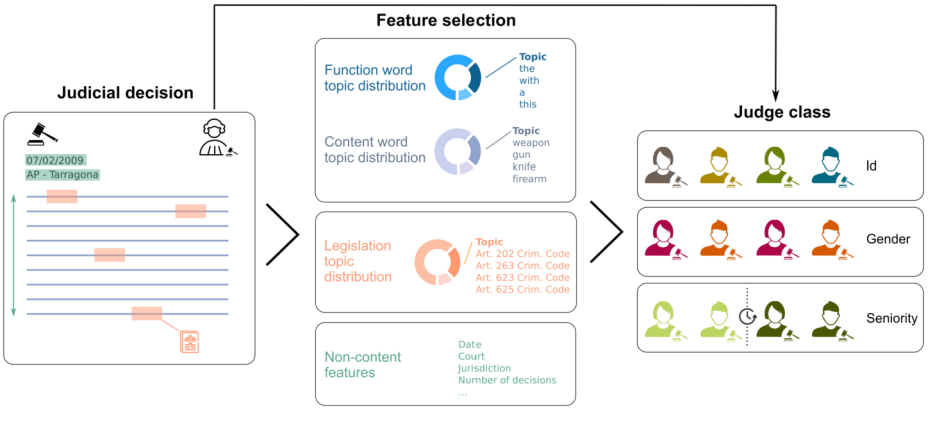

Figure 1 *We use the content of a judicial decision to predict the identity, the gender and the seniority of judges.* From the content and metadata of each decision, we take the words from the text, classifying them into function words (blue) and content words (purple) [32], and the articles in the law cited in the text (orange). We featurize these properties by means of a topic model for each type of content [34], so as to represent each decision as three distributions: one over function word topics, one over content word topics, and one over legislation topics. We also consider non-content features (green) by taking the date of the decision, the court, the jurisdiction and other similar features (see Data and Methods). Then, we use these features to train random forest classifiers and predict the attributes of judges: identity, gender, and seniority

features is predictive of these judge attributes. To do so, we use a random forest classifier that learns the values of the features that best discriminate between attribute groups. Our results show that there are strong individual differences that allow us to clearly predict the identity of the reporting judge. These differences concern stylistic and non-content features as well as content-related features. In the case of gender and seniority, while not so strong, we still find differences that go beyond writing style, allowing us to predict judge attributes more than expected by chance.

2 Results

Our purpose is to evaluate the extent to which the differences in the attributes of judges translate into differences in the content of decisions. To this end, we consider a prediction task of the class of the judge (identity, gender or seniority; see Data and Methods) given different sets of features of the decisions. The features we consider capture aspects of the content of decisions that range from those more linked to the legal practice and legal reasoning to those more linked to writing style. More specifically, we obtain three sets of topics to represent decisions (Fig. 1; Data and Methods): legislation topics (pertaining to the citation of articles of law, thus more linked to the legal practice and legal reasoning); content word topics (pertaining to words that carry the meaning of the legal text); and function word topics (pertaining to the use of words that shape the writing style of the decision).

To control for non-content features that have predictive ability for the same task, we benchmark against non-content features such as the date of the decision, the court, or the length of the decision (Fig. 1; full list in Data and Methods). We perform this class prediction task over three different corpora of judicial decisions corresponding to three broad legal fields: homicides, condominiums and housing.

2.1 Judge identity is highly predictable from language and use of legislation

We start by analyzing the predictability of the identity of a judge from the content of their decisions. Figure 2A-C shows the accuracy in the prediction of the identity of the reporting judge from features that capture the legal aspects of decisions, namely, content word topics and legislation topics (see Data and Methods).

From content word topics, and considering first the homicides corpus, we can predict the exact identity of the judge in 63% of the decisions when using the set of decisions from judges with at least 10 decisions in the corpus (12,618 decisions from 523 judges). When we restrict the analysis to judges with at least 60 decisions (2767 decisions from 24 judges), the accuracy goes up to 93%. The results are similar for the other two corpora: in the condominiums corpus, we obtain 64% accuracy for judges with at least ten decisions (41,907 decisions from 298 judges), and 82% accuracy for judges with at least 60 decisions (13,297 decisions from 52 judges); in the housing corpus, we obtain 50% accuracy (15,331 decisions from 664 judges) and 81% accuracy (1564 decisions from 17 judges), respectively. Legislation topics are also very predictive of judge identity, although less than content word topics. Using a similar selection of decisions (only disregarding a small fraction of decisions with no legislation cited) for each corpus we achieve accuracies in the range from 25% to 44% in homicides; 10% to 22%, in condominiums; and 22% to 65% in housing. In both cases, results are much higher than what would be expected by chance (using a calibrated naive guesser, as described in the Data and Methods section, we obtain

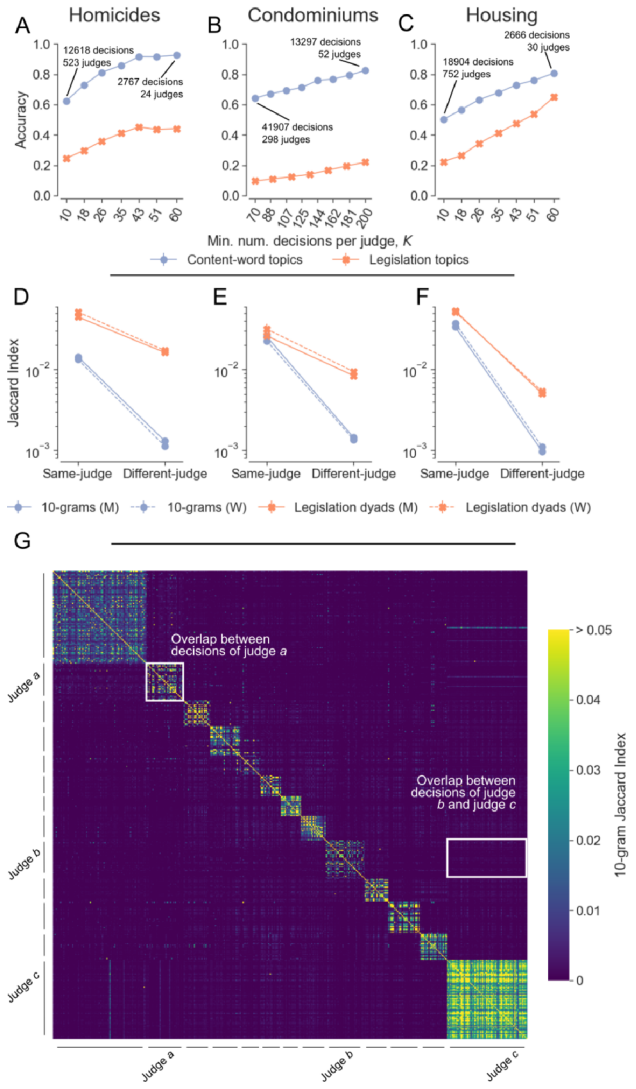

Figure 2 *Differences in language use and law citation are highly predictive of the reporting judge.* (A-C) Accuracy in the prediction of the identity of the reporting judge for the homicides (A), condominiums (B), and housing (C) corpora. We use content-word topics and legislation topics as predictive features (see Data and Methods). For each corpus, we consider different subsets of decisions, each corresponding to decisions from judges with a minimum number of decisions per judge, K. We report the accuracy of the predictions, that is, the fraction of times that the judge identity is predicted correctly. Results are averages over a 10-fold cross-validation. (D-F) Degree of content overlap between pairs of decisions for the homicides (D), condominiums (E), and housing (F) corpora. For each decision, we consider: (i) the list of consecutive 10 words (10-grams); and (ii) the list of cited pairs of legal articles (legislation dyads; see Data and Methods). We compute the average Jaccard index between those lists for pairs of same-judge decisions and different-judge pairs of decisions. We also differentiate between male and female reporting judges (M/W). Standard error bars computed over folds are smaller than symbols in all plots (A-F). (G) 10-grams Jaccard index between decisions from a selection of 13 judges in the homicides corpus

an accuracy of 0.4% to 4% in homicides, 0.4% to 2% in condominiums and 0.2% to 6% in housing). When using alternative machine learning classification algorithms, such as Extreme Gradient Boosting [35], results do not deviate significantly (see Fig. S22 and S23, Additional file 1).

These results suggest that judges must have strong individual signatures that affect both the use of content words and the references to the law, something that makes them very recognizable from a classification point of view. To take a closer look at these individual signatures, we analyze the degree of overlap that exists between pairs of decisions. We hypothesize that this recognizable signatures should arise from a higher overlap between pairs of decisions from the same judge (same-judge pairs) compared to that of pairs of decisions from different judges (different-judge pairs). Again, to compare decisions we use both words and cited legislation, specifically we consider: (i) chains of consecutive 10 words (10-grams); and (ii) combinations of pairs of cited law articles (legislation dyads) (Data and Methods). The results from a selection of judges in the corpus of homicides (see Fig. 2G for 10-grams overlap) show that the degree of overlap between same-judge pairs is much higher than that between different-judge pairs, which results in a distinct diagonal pattern in the matrix of overlaps. Extending the analysis to the other corpora and to legislation dyads, we obtain the same result (Fig. 2D-F). For 10-grams, we observe a 15-fold increase in the degree of overlap between same-judge pairs and different-judge pairs in homicides. For condominiums and housing, the increase in overlap is 20 and 40-fold, respectively. For cited legislation dyads, the increase in overlap is more modest but still sizable: 2-fold in homicides, 3-fold in condominiums, and 6-fold in housing. In terms of the gender of the reporting judge, there is no appreciable difference between men and women.

Among judges, the tendency to reuse more words from their own decisions than from others' (ΔJ_i^W in expression (8); Data and Methods) seems to be positively correlated with the tendency to also reuse cited legislation from own decisions more than from others' (ΔJ_i^L, see Fig. S1, Additional file 1; not significant in the condominiums corpus). This suggests that these two observations are two sides of the same coin of content reuse.

2.2 Judge gender can be predicted from content-related features

Our results clearly show that there are individual traces in each decision that make it possible to guess the identity of the author of each decision. However, this finding does not answer the question of whether there are also more generic group signatures in the content of decisions that allow to identify attributes of judges such as gender or seniority. In what follows, and to prevent the classifier from learning the gender and seniority of judges by learning first their identity, we aggregate all the decisions of each judge, computing the average distributions over word and legislation topics and the average over non-content features (see Data and Methods). In this way, each judge is represented by a single *average decision*. In all forthcoming cross-validation experiments, we split judges (and their average decisions) into training and validation sets. Therefore, the gender and seniority of each judge is predicted from a training set that only includes decisions of other judges, but not their own, so that identity cannot possibly be learned.

Similarly to the case of predicting the identity of a judge, we evaluate how the differences in the gender of the reporting judge translate into differences in the content of decisions. We find that both content word topics and legislation topics can be used to predict the gender of the judge better than expected by chance (Fig. 3A-C). In the case of content word topics, the area under the receiver operating curve (AUROC) ranges from 0.59 in housing to 0.69 in homicides (0.62 in condominiums). In the case of legislation topics, the AUROC ranges from 0.54 in housing to 0.61 in homicides (0.55 in condominiums;

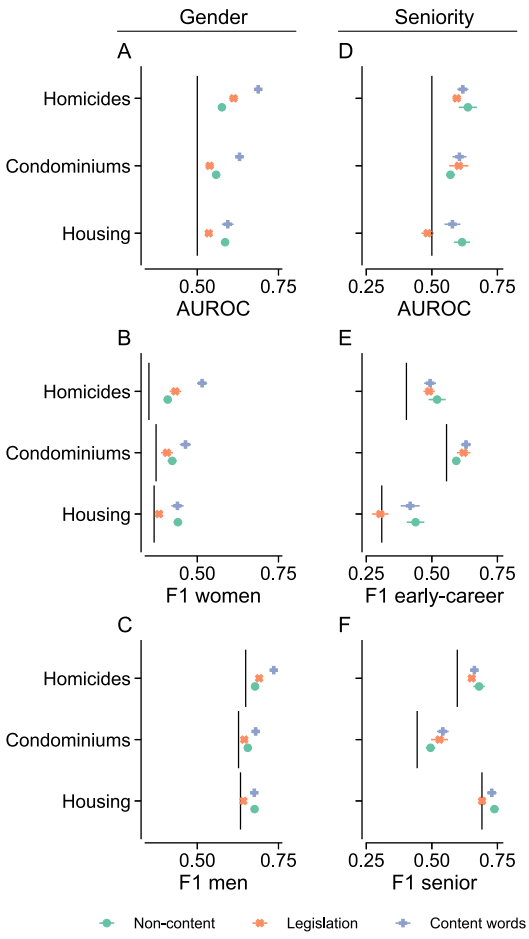

Figure 3 *Judge gender and seniority prediction.* We predict the gender and the seniority of the reporting judge using different features taken from the content of their judicial decisions: topics of content words, topics of legislation, and non-content features (see Data and Methods). We train a random forest classifier and we evaluate the prediction using three different metrics: the area under the receiver operating curve (AUROC - *A, D*), the F1 score for the 'women'/'early career' class (*B, E*) and for the 'men'/'senior' class (*C, F*). We show the score of each metric compared to a calibrated naive guesser (black line, see Data and Methods). Each point corresponds to the average of a *k*-fold cross validation ($k = 20$ for gender, $k = 15$ for seniority) and error bars represent the standard error of the mean. Where not visible, error bars are smaller than symbols

see Fig. 3A). The results for the F1 metric also show both features performing better than chance (Fig. 3B, C). Additionally, we find that content word topics are more predictive than legislation topics for the three corpora: (AUROC is 12% higher in homicides, 17% higher in condominiums and 11% higher in housing; Fig. 3A. Similar results hold for F1 metrics; Fig. 3B, C). These results show that there are inherent differences between male and female judges that permeate into measurable differences in the content of decisions they write, differences that allow us to predict the gender of the judge better than expected by chance.

To benchmark the predictive power of content features (content words and legislation), we compare to the predictive power of non-content features such as the date of the decision, the ruling court and the number of decisions each judge has in the corpus. While there are no significant gender differences in the number of decisions per judge (see Fig. S6, Additional file 1), the differences regarding the ratio between decisions written by men

and women varies considerably, both over years and across courts. For example, in 2001 only 3% of the decisions in the homicides corpus where written by women, while they amounted to 34% of decisions in 2018 see Fig. S2, Additional file 1). Similarly, whereas just 1% of the homicides decisions ruled by the Supreme Court were written by women, the fraction goes up to 50% in Madrid's Court of Appeal (for more details, see Figs. S3-5, Additional file 1). Given these marked differences, it is clear that this information should help considerably to predict the gender of the judge better than expected from chance; we confirm this expectation (Fig. 3A-C). The performance comparison between content and non-content features shows that word topics perform better in the condominiums (13% better) and the homicides corpora (19% better), and similarly in the housing corpus. In the case of legislation topics, the performance is equivalent in homicides and condominiums, and 9% lower in housing (see Fig. 3D-F). Therefore, even though non-content features are intuitively quite predictive because of the large gender disparities in time and geography, content features (especially content words) are often even more predictive or, at least, similarly predictive (with the only the exception of legislation features in the housing corpus).

When using alternative machine learning classification algorithms, such as Extreme Gradient Boosting [35], results do not deviate significantly (see Fig. S20, Additional file 1).

2.3 Judge seniority can be predicted from content-related features

Along the same lines of gender prediction, we explore how the differences in the content of judicial decisions are predictive of the seniority of judges, that is, the length of their careers as measured by the number of years of service. To maintain the structure of the prediction task with respect to gender, we split judges in two groups according to their seniority: early-career judges and senior judges. We establish the group of each judge based on the date of the first decision of the judge in our data set (see Data and Methods). Moreover, we restrict the prediction task to decisions within a narrow window of time (5 years) to avoid the potential confounding effect of the change of the topics over time (see Data and Methods) [36].

Results in Fig. 3D-F show that content word topics can predict the seniority of the judge more than expected by chance (AUROC scores: 0.58 in housing, 0.61 in condominiums, and 0.62 in homicides). In the case of legislation topics, results are similar except for the case of housing, where results are not distinguishable from chance (AUROC scores of 0.60 in condominiums and homicides and 0.48 in housing).

Similarly to the case of gender prediction, we compare the performance of content topics with non-content features. In this case, non-content features include the court and the number of decisions by each judge during the time window we consider. While there are no significant differences in the number of decisions per judge between early-career and senior judges (see Fig. S10, Additional file 1), the ratio of senior to early-career judges varies considerably across courts, going from 31% of early-career judges in the Supreme Court to 85% in Barcelona's Court of Appeal for homicides (for more details, see Figs. S7-9, Additional file 1). The performance for both content-word and legislation topics being statistically indistinguishable to that of non-content features in all three corpora (except for legislation in housing) gives an idea, again, of the extent to which seniority differences affect the legal content of decisions.

Similarly to the case of gender, when using alternative machine learning classification algorithms, such as Extreme Gradient Boosting [35], results do not deviate significantly (see Fig. S21, Additional file 1).

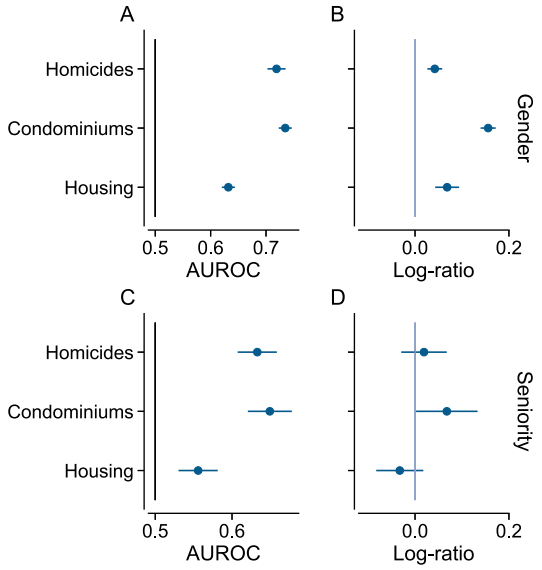

Figure 4 *Function words are as predictive of seniority and more predictive of gender than content words.* (A, B) Using a random forest classifier, we predict the class of the judge (gender or seniority) using function word topics. In particular, we show the area under the receiver operating curve (AUROC) for the prediction of the gender (A) and seniority (C). (B, D) To compare the predictive power of function words to that of content words (as reported in Fig. 3), we show the log-ratio between the AUROC for the prediction with function word topics and that of content word topics. Each point and error bar correspond to the average and standard error of a k-fold cross-validations ($k = 20$ for gender and $k = 15$ for seniority). Where not visible, error bars are smaller than symbols. See Fig. S13-14 for the corresponding results for F1 scores

2.4 Function words are as predictive of seniority and more predictive of gender than content words

Up to this point, we have analyzed how features related to legal reasoning (content words) and legal practice (cited legislation) are predictive of gender and seniority of reporting judges. Next, we analyze the predictive power of features that characterize the stylistic aspects of the text of decisions [27, 37]. Specifically, we consider topics of function words, that is, words with less informational value [32] than what we have defined as content words and have been analyzing so far (see Data and Methods).

Our analysis shows that function words are predictive of the gender of the judge, with AUROC scores of 0.72 in homicides, 0.73 in condominiums and 0.63 in housing (Fig. 4). Similarly, results show that function-word topics are also predictive of judge seniority above what is expected by chance in all three corpora, with AUROC values of 0.63 in homicides, 0.65 in condominiums and 0.56 in housing.

To benchmark these results to those obtained using content words, we calculate the log-ratio between the predictive accuracy of function words and that of content words (Fig. 4B and D). Positive log-ratios indicate that function words are more predictive than content words, and vice versa. We find that function words are as predictive of judge seniority as content words. However, function words are significantly and consistently more predictive of gender than content words, with log-ratio values of 0.04, 0.16 and 0.07 for the homicides, condominiums, and housing corpora, respectively, results are averaged over a k-fold cross-validation ($k = 20$ for gender, $k = 15$ for seniority).

Considering alternative thresholds for the information-content value to separate content words from function words, we find similar results. For gender prediction, using function words is always much more predictive in the condominiums and housing corpora, while the improvement is more modest in the homicides corpus. In the case of seniority, using function words or content words has the same predictive power (see Figs. S11 and S12 and Text S1, Additional file 1).

2.5 Gender and seniority differences are attributed to complex combinations of features

The results we presented in previous sections show that we can predict, better than would be expected by chance, the gender and seniority of judges from the topics that quantify the judicial decisions they write. However, we achieved these predictions by using topic models that involve a considerable number of topics (more than 10^3 for content word topics, more than 10^2 for legislation topics). Therefore, we wonder if there exists a considerably smaller subset of these topics that could achieve a similar prediction performance and thus facilitate the interpretation of the results. For this reason, we took several different-sized subsets of topics and we evaluated their predictive power.

Taking the set of judges to predict $j \in [1, J]$, and the average topic distribution for the decisions of each one of them, we compute the correlation between the weights of each topic and the attributes of each judge, measured as the mutual information between these two functions. Specifically, and for the case of the prediction of judge gender:

$$I(G, T_k^X) = \sum_{i \in J} \sum_{j \in J} P(G, T_k^X | i, j) \log \left(\frac{P(G, T_k^X | i, j)}{P(G|i) P(T_k^X | j)} \right) , \tag{1}$$

where $P(T_k^X|j)$ is the weight of a specific topic k in the average judge topic distribution, $P(G|j)$ is the function that sets the gender of the judge, and $P(G, T_k^X|i,j)$ is the joint probability mass function of both functions. $X \in \{L, W\}$ represents either using content-word topics or legislation topics. Similarly, we can compute the analogous mutual information for the seniority of the judge:

$$I(S, T_k^X) = \sum_{i \in J} \sum_{j \in J} P(S, T_k^X | i, j) \log \left(\frac{P(S, T_k^X | i, j)}{P(S|i) P(T_k^X | j)} \right) , \tag{2}$$

where $P(S|j)$ is the function that sets the seniority of the judge. Once computed the correlation for each topic, we descendantly order them, and then we select subsets, taking the first N.

In Fig. S15, Additional file 1 we show the results regarding the gender prediction performance. When using content-word topics, the performance falls systematically when reducing the number of topics, and when taking 10 topics it falls from an AUROC score of 0.69 to 0.55 in homicides, from 0.64 to 0.55 in condominiums, and from 0.58 to 0.53 in housing. When using legislation topics (see Fig. S16, Additional file 1), the results fluctuate more than those of gender prediction. In homicides, the performance falls from 0.60 to 0.55 when reducing to 10 topics, but the performance is 0.60 for a set of 8 topics and 0.53

when reducing to 4 topics. In condominiums and housing, the performance level fluctuates around the score resulting from considering all topics (0.55 in condominiums, 0.53 in housing) and eventually falls when using 4 topics (0.53 in condominiums, 0.50 in housing).

In Fig. S17, Additional file 1 we show the results regarding the seniority prediction performance. When using content-word topics, the results for different sets of topics fluctuate considerably. In homicides, where the performance using all topics is an AUROC score of 0.58, a selection of 629 topics results in a score of 0.68 whereas a selection of 40 topics gives a score of 0.54. In the case of condominiums and housing we find a similar behavior. In all three cases the performance drops below 10 topics (AUROC score of 0.55 using 7 topics in homicides, 0.51 using 6 topics in condominiums and 0.52 using 7 topics in housing). When using legislation topics (see Fig. S18, Additional file 1), the performance in homicides fluctuates around the score corresponding to using all topics (0.58) and eventually drops to 0.56 using 4 topics. In the case of condominiums, the performance falls from AUROC score of 0.6 using all topics to 0.57 using 4 topics. In the case of housing, the score using all topics falls bellow the expected by chance (0.49) and exploring the performance over the different sets of topics produces results that fluctuate from 0.61 using 30 topics to 0.44 using 16 topics. Additionally, the behavior regarding using function-word topics is similar: in the case of gender, the performance drops systematically for the three corpora when taking less than 10 topics; in the case of seniority, it does not fall systematically, but it fluctuates considerably (see Fig. S19, Additional file 1).

These results show that we are not able to systematically maintain the same prediction performance when reducing the number of topics (either content words or legislation) considerably. In fact, either the performance drops systematically or it fluctuates showing inconclusive results. All in all, these results imply that we are not able to explain the gender and seniority differences in terms of a set of a few topics. Rather, these differences are the result of a complex and intrincated combination of hundreds or thousands of topics. Moreover, while in some cases we are able to retain a significant predictive power by using just 1 or even 2 topics (see the cases of using function-word topics to predict gender and seniority in the homicides corpus in Fig. S19, Additional file 1), the corresponding legal interpretation in these cases is limited. First, because function words tend to lack meaning when there is no context; second, because although the hierarchical level of the model is the lowest and it is the most specific, there are still tens of words in each topic (see Tables S3 and S4 for the case of gender, and Tables S5 and S6 for the case of seniority, in Additional file 1).

3 Discussion

Our results show that there are inherent differences in the way judges write decisions, which make them recognizable, not only at the individual level but also grouping them by gender or seniority. In our analysis, we use a range of features that capture both style and legal reasoning of decisions which allows us to better understand the nature of the differences between judges.

At the individual level, our results show that one of the primary causes for the appearance of these differences is that judges reuse more content from their own past decisions than content of decisions by other judges. Reasonably, this can be expected by the fact that judges have their own way of saying things, using a certain tone, expressions and a given level of technical language [31]. Moreover, when facing a new case, judges might

find it easier to remember similarities with past cases of their own, rather than spending time looking for similarities in the vast archive of available decisions. One can then expect that the reuse of content (some times in the form of copy-paste) translates into individual idiosyncrasies of judges' writing styles, which we are able to reveal by using function words to predict the identity of the judge. However, by only considering the content that underlies the legal reasoning and the framing of the case in the applicable law, that is, considering content words and cited legislation, we still predict very well the identity of the judge, which reveals that content reuse affects the wording of arguments and the choice of supporting legislation as well.

The reuse of content from own documents is common in other domains as well. For instance, in science, scholars tend to reuse text from papers they have authored much more than text from the papers of others [38]. This is to be expected, given that scientists have few constraints in the choice of the subject and the methodology of research, they often draw upon their previous results to move forward. However, in our case of study, the situation is rather different: cases are randomly assigned, so that judges are not free to choose the *subject* of the cases they must decide on.

The individual traits we observe could also come from another source. Normally, courts organize themselves into different sections and chambers to which judges are assigned. According to the Spanish Organic Law of the Judicial Power (LOPJ, art. 152.2), these courts can decide how to assign cases among sections and chambers. This assignation is based on subject criteria; judges are assigned cases depending on their domain of expertise. Thus, it could be possible to find judges whose decisions can be differentiated from those of others by the legal subject. However, we are considering three corpora of cases in very specific fields, which typically would be assigned to experts in their respective courts, and therefore, these thematic differences cannot explain our results.

The results we obtained at the individual level in terms of the reuse of content, constitute a good starting point to further inquire if these practices are the result of a bad practice. In other words, it would be of interest to reveal if the reuse of the same laws, for instance, is legally justified and hence it comes with the reuse of the verdict or the legal reasoning as well.

Beyond the reporting judge, other judges participate in discussions and deliberations of the decision, possibly making these courts mixed both in terms gender and in terms of seniority. A mixed court could blur the footprints of the reporting judge's gender or seniority on the decision's content, potentially hindering the attribute prediction task. The fact that, despite this possible blurring, we are still able to predict these attributes implies that the true differences between individual judges may be even more pronounced than we have reported. Future research efforts could be directed towards gaining a better understanding of the effects of other judge attributes within the court and the interactions among them. Similarly, future research aimed at evaluating the influence of gender or seniority on the final verdict will need to consider the possible mixed composition of the court, as a similar interaction could exist in this context.

Beyond individual differences, our analysis in terms of features of the set of decisions of each judge reveals differences that are predictive of both gender and seniority of judges. We observe these differences for both content and non-content related features. However, despite our reduction of the dimensionality of the description of decisions from tens of thousands of words to hundreds of features, we cannot pinpoint the specific sources for

these differences – indeed, we find that the differences we observe are not attributable to a few features of the decisions but to complex combinations of them. Finding the sources for these differences is thus not trivial, but poses a question that should be investigated in depth. Actually, because these differences cannot be attributed neither to individual differences nor to case assignation criteria, understanding how these observed differences translate into differences in how judges apply the law is a fundamental question that needs to be answered. Further efforts in this direction could enable an intervention in the case allocation policies in the courts, ultimately contributing to the transparency and well-functioning of the judiciary.

4 Data and methods

4.1 Judicial decisions data set

Our data set encompasses three corpora of judicial decisions related to three different areas of law: homicides, condominiums (corresponding to conflicts within multi-unit buildings) and housing (including squatting, abusive clauses in mortgage loan contracts, tenants' evictions and mortgage enforcements). In the three areas, decisions are ruled in the Spanish judicial system, and correspond to courts of appeal (Provincial Courts, *Audiencias Provinciales*, in Spanish, 89%) or higher courts (e.g. Supreme Court, Tribunal Supremo, in Spanish, 8%). We analyze decisions in the period 2001-2018, which results in a total of 22,983 decisions from 2021 judges in housing, 15,648 decisions from 1580 judges in homicides, and 59,516 decisions from 1766 in condominiums.

The data were provided to us by Tirant Online, one of the largest and most comprehensive databases for judicial decisions in Spain. These data provides, for each decision, the full text and a list of metadata. Among other details, these metadata comprises the date of the decision, the ruling court, the identity of the reporting judge and the list of law articles cited in the text.

4.2 Text processing

We process the text of each document by disambiguating specific legal-related terms, (see Text S1 and Table S1, Additional file 1) removing numbers and non-word characters and converting all characters to low case. We also *degenderize* the text by substituting all person names by '_persona_' and by removing the gender declination of certain words that mostly correlate with the gender and identity of the judge, such as magistrado/magistrada (masculine and feminine versions of 'justice', see Table S2 in Additional file 1 for the full list of words *degenderized*). We apply the *degenderization* process when using the data to predict the gender and the identity of the reporting judge but not to predict the seniority.

We also find and substitute the most significant chains of 2 and 3 words (2-grams and 3-grams), which allows us to go beyond the bag-of-words assumption, and consider concepts such as 'código_civil' (civil code) or 'tribunal_supremo' (Supreme Court). For more details see Text S1.2 in Additional file 1.

4.3 Feature selection

In here, we describe the process of converting the content of judicial decisions (processed text and metadata) into features that will be used in the judge class prediction task afterwards. We run this process over each corpus independently.

Content-word topic model We filter words by using an information-theory based method to remove the most entropic words [32]. This method is a universal, corpus-dependent method that removes the so called function words (also stop words in the literature) while keeping the 'content' words, that is, more meaningful words that matter for the substantial content of documents and that improve the quality of the topics inferred afterwards [32]. We also remove words appearing in less than 1% of the documents. See Text S1.4 in Additional file 1 for details on the information-content threshold.

Having each document as a list of terms (filtered words and significant 2,3-grams), we take a topic model approach to reduce the dimensionality of the data. We use the approach by Gerlach et al. [34] to infer the topics present in the corpus and then express each document as a distribution over the topics. Thus, we represent each judicial decision by a vector of weights of each topic. Although the model is hierarchical, we select the lower level for being the one that is more descriptive.

Function-word topic model Because function words have been found in the literature to carry stylistic signatures predictive of the attributes of written text authors, we also consider the stylistic content of decisions by obtaining the corresponding set of topics for this set of words (See Text S1.4, Additional file 1, for details on the information-content threshold). For each decision, we thus use a reduced dimensionality representation, by which we express each judicial decision as a distribution over function-word topics.

Legislation topic model In each judicial decision, reporting judges make references to the current applicable law. Then, taking the list of articles of the law cited in the text, we consider an analogous approach to that of content and function words: we infer legislation topics as groups of articles in the law used similarly over the corpus of decisions. If word topics can be thought as words that can be used in similar contexts, legislation topics are groups of law articles that share some similarity, for instance, they tend to belong to the same law. Thus, we represent each judicial decision by the list of topic weights in the legislation topic distribution.

Non-content features We also consider simple features that are not related to the specific content of judicial decisions to see how just these perform in the prediction task in relation to hundreds of content-related topics. For each judicial decision, we consider: the date, the jurisdiction (civil/criminal) and the court of ruling. Besides, we also consider the number of decisions each judges has in the corpus, as an indicator of how 'prolific' or 'experienced' is a judge in the field.

4.4 Using a random forest classifier to predict the class of judges

Given each set of features, we evaluate the extent to which they are informative to predict the class (identity, gender, seniority) of the judge. To do so, we train a random forest algorithm, a supervised classifier that uses an ensemble of decision trees and learns how to classify the data from the features. The algorithm is well suited for classification problems with high dimensional data, and it has been widely applied in a variety of domains [39, 40]. Then, we validate the trained classifier using a K-fold cross-validation, that is, dividing our data set into K splits, training the classifier on $K-1$ splits while testing in the resting one, and repeating for all K combinations of train an test set. The number of folds chosen in

each case trying to balance performance and computational cost. Decisions are randomly assigned to each split while keeping the proportion of classes equal to the global one. In the same sense, the predictions are calibrated to ensure that the proportion of predicted classes is equal to the proportion of the data.

Reporting judge Since the majority of decisions in our data set originate from courts of appeal and higher courts, they are typically decided through consensus among multiple judges in a jury. However, it is the responsibility of one judge, known as the reporting judge, to draft and present the decision for consideration, after which the other judges may concur or dissent with the decision. In cases where a consensus cannot be reached, another judge assumes the responsibility of drafting the decision, and the dissenting judge may provide an alternative 'dissenting' opinion (*voto particular*, in Spanish). However, such situations are rare, as judges typically arrive at a consensus before finalizing a decision. For this reason, and given the responsibility that comes with the action of writing and proposing the decision, our study focuses on the reporting judge.

Judge identity The identity of each judge is already provided in the metadata. Considering a 10-fold cross-validation, we take a subset of decisions where each judge has at least 10 decisions, o to ensure the presence of each judge in all 10 portions at least once. In the case of condominiums, where we have a much larger corpus, we consider a threshold of 70 decisions, to ensure the computational feasibility of the random forest classifier.

Gender We classify the gender of the reporting judge using the list of masculine and feminine Spanish names. We double check this classification by considering the form of address in the text, which depends on the gender of the judge (*Don*, *Doña*). We did not find any instances of gender-neutral forms of address, which may have indicated non-binary self-attribution of gender by the reporting judge. For each corpus, the fraction of decisions written by women is 0.35 in homicides, 0.37 in condominiums and 0.36 in housing.

When predicting the gender of the judge, we only consider each judge once: we take all the decisions corresponding to the same judge and we compute the average feature value over these decisions. For topic distributions this implies computing the average topic distribution over all decision of a given reporting judge. For non-content features, it implies computing the average of each feature over decisions. This allows us to avoid the effect of predicting the gender by being able to predict the id as well.

Seniority We discretize the seniority of the reporting judge by classifying them between senior or early-career depending on the date of their earliest judicial decision in each corpus. To avoid the effects of the change in the content of decisions over time [36], we restrict this analysis to decisions only published in a 5-year time window, considering decisions ruled in the period 2008-2013. Moreover, to avoid having judges that have their last/first decision in the mentioned time window, we only consider those judges having both decisions ruled within the windows before 2008 and after 2013. This results in a selection of 3145 decisions from 375 judges in omicides (60% of them early-career), 18,133 decisions from 435 judges in condominiums (44% of them early-career) and 3428 decisions from 476 judges in housing (68% of them early-career). When considering legislation topics, we do not consider decisions with no law articles cited, which reduces the sets of decisions in

less than 1%. We tag judges as senior if their first decision was ruled before 2003 and we tag them early-career in the opposite case. As done in the case of predicting the gender of the judge, we also consider each judge once by computing the average features over the decisions ruled by the same judge. For each corpus, the fraction of decisions written by early-career judges is 0.60 in homicides, 0.44 in condominiums and 0.69 in housing.

Naive guesser We compare our results for the prediction tasks of the identity, the gender and the seniority of the judge with a null model characterized by a calibrated naive guesser, which is equivalent to a random assignation of judge attribute labels in the test set, while preserving the ratios of each class, and the subsequent performance evaluation in terms of the accuracy (judge identity) or AUROC (judge gender and seniority).

4.5 Degree of overlap between pairs of judicial decisions

In each corpus of judicial decisions, we measure the degree of overlap between pairs of judicial decisions taking separately the words used in the text and the legislation cited, respectively. To measure the overlap in the use of words, we consider chains of 10 consecutive words (n-grams, n = 10), disregarding chains that include punctuation marks (except before the first word or after the last one) and those only appearing in just one decision. In the case of legislation, we consider the list of references to articles in the law and we take all possible pairs, we call them citation dyads. Thus, we associate each decision with the corresponding set of 10-grams and the corresponding set of citation dyads.

Being W_d and L_d the set of 10-grams and the set of legislation dyads corresponding to decision d, respectively, we measure the normalized intersection between two decisions d and r using the Jaccard index:

$$J_{dr}^{W} = \frac{|W_d \cap W_r|}{|W_d \cup W_r|}, \tag{3}$$

$$J_{dr}^{L} = \frac{|L_d \cap L_r|}{|L_d \cup L_r|}. \tag{4}$$

We then compute the degree of overlap corresponding to each judge. To that end we consider the set of decisions D_i written by judge i. We the compute the degree overlap between decisions in the same set to estimate the reuse of content from own decisions of each judge i

$$J_i^{X_{\text{self}}} = \frac{2}{|D_i|\,(|D_i|-1)} \sum_{(d,r)\in D_i, d\neq r} J_{dr}^{X} \qquad \text{with} \quad X = \{L, W\}. \tag{5}$$

Then, we also consider the overlap between decisions D_i written by judge i and the decisions written by other judges, $D_{\neq i} = \{\bigcup_{j\neq i} D_j\}$,

$$J_i^{X_{\text{other}}} = \frac{1}{|D_i|\,\left|D_{\neq i}\right|} \sum_{d\in D_i; r\in D_{\neq i}} J_{dr}^{X} \quad \text{with} \quad X = \{L, W\}. \tag{6}$$

Globally, we average these quantities over all judges:

$$\langle J^{X_{\text{self}}} \rangle = \frac{1}{N_{\text{judges}}} \sum_i J_i^{X_{\text{self}}} \quad , \qquad \langle J^{X_{\text{other}}} \rangle = \frac{1}{N_{\text{judges}}} \sum_i J_i^{X_{\text{other}}} \quad . \tag{7}$$

Finally, we measure the individual tendency of each judge reuse more their own content than from other as the difference between these quantities as:

$$\Delta J_i^X = J_i^{X_{\text{self}}} - J_i^{X_{\text{other}}} \ . \tag{8}$$

Abbreviations
AUROC, Area under the receiver operating curve; LOPJ, *Ley orgánica del poder judicial*.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-024-00494-x.

Additional file 1. The present article has an accompanying supplementary file. (PDF 749 kB)
Additional file 2. (CSV 3.6 MB)

Acknowledgements
We thank Tirant Online for providing us with the digitized and parsed corpora of legal decisions, and for their help with data processing.

Author contributions
LF-P, MS-P and RG designed the research, analyzed the data and carried out computational experiments. All authors discussed and analyzed the results and wrote and edited the manuscript. All authors read and approved the final manuscript.

Funding
This research was funded by the Social Observatory of the "la Caixa" Foundation as part of the project LCF / PR / SR19 / 52540009, by the Spanish Government Ministerio de Ciencia e Innovación / AEI / 10.13039 / 501100011033 (Projects No. PID2020-112876GB-C31 and PID2022-142600NB-I00) and by the Government of Catalonia (Projects No. 2017SGR-896 and No. 2021SGR-00170).

Data availability
We provide the full list of IDs corresponding to all the judicial decisions studied in this work, see Additional file 2. Using these IDs, the information from each decision can be retrieved downloading the opinions from the following public database https://www.poderjudicial.es/search/indexAN.jsp.

Materials and Code availability
Not applicable.

Declarations

Ethics approval and consent to participate
All the information regarding the judges in the corpora of decisions we analyze is publicly disclosed in the decisions and can be openly accessed. The identity (the name of the judge) is disclosed in the preamble of each decision text. We infer the gender from the disclosed judge name. We derive the seniority of each judge by looking at the publication dates of the decisions authored by each judge.

Consent for publication
Not applicable.

Competing interests
The authors declare no conflict of interest.

Author details
[1]Department of Chemical Engineering, Universitat Rovira i Virgili, 43007 Tarragona, Spain. [2]Departament de Física de la Matèria Condensada, Universitat de Barcelona, 08028 Barcelona, Spain. [3]UNESCO Housing Chair, Universitat Rovira i Virgili, 43003, Tarragona, Spain. [4]ICREA, 08010 Barcelona, Spain.

Received: 6 March 2024 Accepted: 11 August 2024 Published online: 02 September 2024

References

1. Greenwald AG, Mcghee DE, Schwartz JLK (1998) Measuring individual differences in implicit cognition: the implicit association test. J Pers Soc Psychol
2. Devine PG (1989) Stereotypes and prejudice: their automatic and controlled components. J Pers Soc Psychol 56(1):5–18. https://doi.org/10.1037/0022-3514.56.1.5
3. Caliskan A, Bryson JJ, Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. Science 356:183–186

4. Bertrand M, Mullainathan S (2004) Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. In: Inequality in the 21st century: a reader, pp 304–308. https://doi.org/10.4324/9780429499821-53.
5. Bagues M, Sylos-Labini M, Zinovyeva N (2017) Does the gender composition of scientific committees matter? Am Econ Rev 107(4):1207–1238. https://doi.org/10.1257/aer.20151211
6. Moss-Racusin CA, Dovidio JF, Brescoll VL, Graham MJ, Handelsman J (2012) Science faculty's subtle gender biases favor male students. Proc Natl Acad Sci USA 109(41):16474–16479. https://doi.org/10.1073/pnas.1211286109
7. Lee CJ, Sugimoto CR, Zhang G, Cronin B (2013) Bias in peer review. J Am Soc Inf Sci Technol 64:1852–1863. https://doi.org/10.1002/asi
8. Danziger S, Levav J, Avnaim-Pesso L (2011) Extraneous factors in judicial decisions. Proc Natl Acad Sci USA 108(17):6889–6892. https://doi.org/10.1073/pnas.1018033108
9. Asmat R, Kossuth L (2021) Gender Differences in Judicial Decisions under Incomplete Information: Evidence from Child Support Cases. SSRN Electron J, 1–39. https://doi.org/10.2139/ssrn.3964747
10. Collins PM, Manning KL, Carp RA (2010) Gender, critical mass, and judicial decision making. Law Policy 32(2):260–281. https://doi.org/10.1111/j.1467-9930.2010.00317.x
11. Boyd CL, Epstein L, Martin AD (2010) Untangling the causal effects of sex on judging. Am J Polit Sci 54(2):389–411. https://doi.org/10.1111/j.1540-5907.2010.00437.x
12. Crow MS, Goulette N (2022) Judicial diversity and sentencing disparity across U.S. District Courts. J Crim Justice 82:101973. https://doi.org/10.1016/j.jcrimjus.2022.101973
13. Welch S, Combs M, Gruhl J (1988) Do black judges make a difference? Am J Polit Sci 32(1):126–136
14. Kulik CT, Perry EL, Pepper MB (2003) Here comes the judge: the influence of judge personal characteristics on federal sexual harassment case outcomes. Law Hum Behav 27(1):69–86. https://doi.org/10.1023/A:1021678912133
15. Cohen A, Yang CS (2019) Judicial politics and sentencing decisions. Am Econ J: Econ Policy 11(1):160–191. https://doi.org/10.1257/pol.20170329
16. Harris AP, Sen M (2019) Bias and judging. Annu Rev Pol Sci 22:241–259. https://doi.org/10.1146/annurev-polisci-051617-090650
17. Eck K, Crabtree C (2020) Gender differences in the prosecution of police assault: Evidence from a natural experiment in Sweden. PLoS ONE 15 https://doi.org/10.1371/journal.pone.0235894
18. Ash E, Zurich Daniel Chen EL, Ornaghi A, School H, Zurich E, Chen D (2022) Gender attitudes in the judiciary: evidence from U.S. circuit courts. Am Econ J Appl Econ
19. Rice D, Rhodes JH, Nteta T (2019) Racial bias in legal language. Res Polit 6(2). https://doi.org/10.1177/2053168019848930
20. Neidorf L, Krieger MS, Yakubek M, Chaudhuri P, Dexter JP (2019) Large-scale quantitative profiling of the old English verse tradition. Nat Hum Behav 3(6):560–567. https://doi.org/10.1038/s41562-019-0570-1
21. Cafiero F, Camps JB (2019) Why Molière most likely did write his plays. Sci Adv 5(11). https://doi.org/10.1126/sciadv.aax5489
22. Ainsworth J, Juola P (2018) Who wrote this?: modern forensic authorship analysis as a model for valid forensic science. Wash Univ Law Rev 96:1161–1189
23. Hosseini M, Tammimy Z (2016) Recognizing users gender in social media using linguistic features. Comput Hum Behav 56:192–197. https://doi.org/10.1016/j.chb.2015.11.049
24. Bamman D, Eisenstein J, Schnoebelen T (2014) Gender identity and lexical variation in social media. Asian Engl 18(2):135–160. https://doi.org/10.1111/josl.12080. arXiv:1210.4567
25. Argamon S, Koppel M, Pennebaker JW, Schler J (2009) Automatically profiling the author of an anonymous text. Commun ACM 52(2):119–123. https://doi.org/10.1145/1461928.1461959
26. Juola P (2008) Authorship attribution. Found Trends Inf Retr 1(3):233–334. https://doi.org/10.1561/1500000005
27. Kestemont M (2014) Function words in authorship attribution from black magic to theory? In: Proceedings of the 3rd workshop on computational linguistics for literature, pp 59–66
28. Newman ML, Groom CJ, Handelman LD, Pennebaker JW (2008) Gender differences in language use: an analysis of 14,000 text samples. Discourse Process 45(3):211–236. https://doi.org/10.1080/01638530802073712
29. Koning R, Samila S, Ferguson JP (2021) Who do we invent for? Patents by women focus more on women's health, but few women get to invent. Science 372(6548):1345–1348. https://doi.org/10.1126/science.aba6990
30. Jockers ML, Mimno D (2013) Significant themes in 19th-century literature. Poetics 41(6):750–769. https://doi.org/10.1016/j.poetic.2013.08.005
31. Posner RA (1995) Judges' writing styles (and do they matter?). Univ Chicago Law Rev 62(4):1421. https://doi.org/10.2307/1600108
32. Gerlach M, Shi H, Amaral LAN (2019) A universal information theoretic approach to the identification of stopwords. Nat Mach Intell 1(12):606–612. https://doi.org/10.1038/s42256-019-0112-6
33. Manning CD, Raghavan P, Schütze H (2008) An introduction to information retrival. Cambridge Univeristy Press, Cambridge
34. Gerlach M, Peixoto TP, Altmann EG (2018) A network approach to topic models. Sci Adv 4(7). https://advances.sciencemag.org/content/4/7/eaaq1360.full.pdf. https://doi.org/10.1126/sciadv.aaq1360
35. Chen T, Guestrin C (2016) Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. KDD '16. Association for Computing Machinery, New York, pp 785–794. https://doi.org/10.1145/2939672.2939785
36. Font-Pomarol L, Piga A, Garcia-Teruel RM, Nasarre-Aznar S, Sales-Pardo M, Guimerà R (2023) Socially disruptive periods and topics from information-theoretical analysis of judicial decisions. EPJ Data Sci. https://doi.org/10.1140/epjds/s13688-022-00376-0
37. Langford M, Behn D, Lie R (2020) Computational stylometry: predicting the authorship of investment arbitration awards. In: Computational legal studies: the promise and challenge of data-driven research, pp 53–76. https://doi.org/10.4337/9781788977456.00008
38. Citron DT, Ginsparg P (2015) Patterns of text reuse in a scientific corpus. Proc Natl Acad Sci USA 112(1):25–30. https://doi.org/10.1073/pnas.1415135111

39. Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1201/9780429469275-8
40. Boulesteix A-L, Janitza S, Kruppa J, König IR (2012) Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdiscip Rev Data Min Knowl Discov 2(6):323–329. https://doi.org/10.1002/biuz.19920220617

Publisher's Note