

Literary Canonicity and Algorithmic Fairness: The Effect of Author Gender on Classification Models

Ida Marie S. Lassen*, Pascale Feldkamp Moreira, Yuri Bizzoni and Kristoffer Nielbo

Center for Humanities Computing, Aarhus University, Denmark

Abstract

This study examines gender biases in machine learning models that predict literary canonicity. Using algorithmic fairness metrics like equality of opportunity, equalised odds, and calibration within groups, we show that models violate the fairness metrics, especially by misclassifying non-canonical books by men as canonical. Feature importance analysis shows that text-intrinsic differences between books by men and women authors contribute to these biases. Men have historically dominated canonical literature, which may bias models towards associating men-authored writing styles with literary canonicity. Our study highlights how these biased models can lead to skewed interpretations of literary history and canonicity, potentially reinforcing and perpetuating existing gender disparities in our understanding of literature. This underscores the need to integrate algorithmic fairness in computational literary studies and digital humanities more broadly to foster equitable computational practices.

Keywords

bias, algorithmic fairness, gender bias, computational literary studies, canonicity,

1. Introduction


In recent years, computational literary studies have increasingly utilised machine learning (ML) models to analyse and classify literary texts, e.g. to predict reader appreciation [31, 34] or literary success [18, 45, 21, 9] with uptake in applications in the publishing industry.¹ Models often rely on text-intrinsic features, contributing to the study of which text characteristics serve as predictors for a given classification. While other studies have shown that literature assessment can be biased by gender [43, 29] and ethnicity [15], focusing on text-intrinsic characteristics might seem like a way to avoid such biases as it concentrates solely on the text.

However, seemingly objective features can harbour social biases, reflecting disparities in the underlying data. The present work examines gender biases in ML models that predict literary canonicity, demonstrating how the uncritical use of ML models in humanities research can lead to biased knowledge production, potentially skewing our understanding of literary history and the phenomenon of canonicity. This has implications beyond academic research, as these models could influence real-world applications, including the assessment of new manuscripts by publishers based on predicted success or likeness to existing canon. By integrating insights

CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark

✉ idamarie@cas.au.dk (I. M. S. Lassen); pascale.moreira@cc.au.dk (P. F. Moreira); yuri.bizzoni@cc.au.dk (Y. Bizzoni); kln@cas.au.dk (K. Nielbo)

🆔 0000-0001-6905-5665 (I. M. S. Lassen); 0000-0002-2434-4268 (P. F. Moreira); 0000-0002-6981-7903 (Y. Bizzoni); 0000-0002-5116-5070 (K. Nielbo)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹E.g. Edison Online and Marlowe

from algorithmic fairness into our analysis of predictive models, we aim to highlight the potential for hidden biases in seemingly objective computational methods. Our analysis demonstrates how these biases can affect our interpretation of literary history and canon formation, and we emphasise the importance of critical reflection on ML methodologies in DH research.

Our findings underscore that the significance of this work lies not only in the practical application of prediction models but also in exposing the epistemic consequences of using biased ML models to study literary phenomena. This approach invites researchers to consider how computational methods may inadvertently reproduce or amplify existing biases in literary history.

2. Related works

2.1. Predicting canonicity

This study builds on prior research demonstrating the potential of ML classifiers to predict various literary attributes, such as whether a book belongs to the literary canon, is written by a Nobel laureate, is a bestseller, is longlisted for given awards, or receives a high rating on GoodReads [8]. To narrow the scope, we will focus on the attempt to predict canonicity. While various studies focus on classifying canonical works and gauging their textual profile [6, 12, 33], the limited resources in the literary field are rarely openly available. We thus focus on one newly published dataset [10], which served as the foundation for Bizzoni, Feldkamp, Jacobsen, Thomsen, and Nielbo [8] and provides a rich and diverse collection of features of literary works.

In [8], the focus extended beyond classification accuracy to provide insights into the textual features important for the classification models, seeking to understand the characteristics that differentiate canonical from non-canonical books. The study found that “canonical texts have the most distinctive profile across all dimensions and are therefore the easiest to classify in the binary classification task” due to their denser nominal style, lower readability, less predictable sentiment arcs, and higher perplexity.

However, it is well-known that canonical literature – like the literary field more broadly – has historically been dominated by men [37, 30, 36]. Still, studies that seek to predict some form of canonicity or perceived literary quality rarely include reflections on how biases in their data inform their results, and the cultural, temporal, or gendered dimensions of texts are rarely mentioned. While Algee-Hewitt and McGurl [1] show how the “canon” significantly changes depending on the approach taken; our study highlights the critical oversight of gender imbalances inherent in literary datasets, which can inadvertently bias model outcomes.

2.2. Gender differences in literary texts

Previous research on gender differences in literary texts highlights key issues to avoid. One concern is treating these differences as fixed and universal markers of men’s and women’s writing. For example, Burrows [13] shows that gendered patterns in writing styles changed over time, with distinct differences found before 1860 but not after, indicating that gendered styles are historically contingent.

A second concern is the assumption of a binary gender model, where men’s writing is seen as the default. Land [26] critiques such approaches for framing women’s writing as deviant, as seen in studies like [3, 25], which rely on essentialist assumptions and risk reinforcing biased interpretations of literary styles.²

With that being said, studies have found linguistic and stylistic differences between texts by men and women that are independent of topic and genre [41]. In the literary domain, Argamon, Koppel, Fine, and Shimoni [3] shows that a high frequency of pronouns is a “strong female marker”, which is supported by Newman, Groom, Handelman, and Pennebaker [35] who also found that women’s language more frequently includes pronouns, social words, various psychological process references, and verbs, as well as negations and home-related terms. Men, on the other hand, used longer words, more numbers, articles, and prepositions than women (p. 223).

Hiatt [20] examined contemporary (1978) American prose and found that women use twice as many emotional adverbs compared to men, while men use nearly twice as many pace adverbs. She concludes that while there is a distinct feminine writing style, there is “far less basis for labelling the feminine styles as hyperemotional than for labelling the masculine style *hypo-emotional*” [20, p. 226].

Hayward [19] tests whether readers can identify an author’s gender and concludes that gender differences are subtler than genre differences. Koolen [24] goes deeper into the question of genre and examines the interaction of gender and genre, especially with regard to “false labelling”, i.e., that works by women are more often labelled as “women’s books” regardless of genre [40]. The findings suggest that while some romantic novels have distinct styles, novels by women are heterogeneous and not distinguishable from those by men. Considering the prevalence of biased mechanisms in the literary field (e.g., false labelling), it is possible that readers focus on similarities among women authors and differences among men authors rather than the reverse.

The literature reviewed here highlights the complexity of considering gender differences in literary texts and not reducing these differences to essentialist notions about “how women write.” In the following, we will use methods from algorithmic fairness to examine biases in models used to predict canonicity. We do not claim to establish definitive conclusions about the general differences between men’s and women’s writing; rather, we emphasise how *modelling a literary phenomenon inevitably mirrors the underlying data and that results could differ if other datasets were used.*

Considering that questions about bias and fairness are increasingly discussed in ML development and that ML is increasingly applied in DH, algorithmic fairness insights are rarely integrated into computational literary studies. Although Bagga and Piper [4] explored the impact of bias on predictive accuracy and positive prediction balance in literary data, our study presents a more comprehensive bias analysis informed by the methodologies of algorithmic fairness. We aim to answer the following research questions:

- **RQ1:** To what extent do ML models trained on (imbalanced) literary corpora exhibit

²We use the terms “women and men authors” instead of the more commonly used “female and male authors” to distinguish cultural gender (which is examined in this paper) from biological sex.

Table 1

Women/men authors (bottom) represent the number of works written by women or men authors in the canon/noncanon, and in subcategories of the canon.

	Chicago corpus	Canon	OpenSyllabus	Norton	Penguin classics
Texts	9,089	618	476	401	77
♀ / ♂	3,289 / 5,800	166 / 452	132 / 344	93 / 307	7 / 70

biases on author gender in classification tasks, particularly in predicting canonicity?

- **RQ2:** Which features in the dataset significantly differ between books by women and men authors, and how do these features impact the bias in classification models?

These questions are, of course, contingent on the data analysed. Therefore, we zoom out and include a question that addresses a broader concern:

- **RQ3:** How does the use of biased ML models affect the knowledge produced in computational literary studies?

This study focuses on binary gender categories, including only men and women authors. We acknowledge that this does not capture the full spectrum of gender identities and that gender is performative and shaped by discursive practices [14]. However, this approach aligns with historical perspectives and addresses existing biases between men and women in literary canonicity.

3. Methods

3.1. Data

The dataset used in this work is the *Chicago Corpus*, which consists of 9,089 novels from diverse genres published in the US between 1880 and 2000. The data is compiled on the number of libraries holding each novel, with a preference for more circulated works. The dataset was made available with a recent paper [10].³ The canon category is compiled from books by authors in the Norton Anthology, the Penguin Classics series, and the top 1000 authors mentioned in English syllabi (collected by the OpenSyllabus project), as shown in Table 1.

A diverse set of stylistic, syntactic and narrative features were used in [8], which found that “[t]he highest F1 score was achieved when all proposed features were included”. In addition to these features, we have included normalised frequencies of part-of-speech (PoS) features, as they have been highlighted as different in the writings of men and women (see Section 2.2).⁴

³The textual features, including reception categories like ‘canon’, are described on Github

⁴Including features that are potentially strong markers of gender is important because other features can act as ‘proxies’ for these. Ignoring them might not reduce bias, as the model could still pick up on these proxies. Including them allows for a more comprehensive analysis of potential biases [5].

3.2. Modelling

To replicate the experiments in [8], we employed Random Forest (RF) models for the classification task. RF models are known for their robustness to overfitting and ability to handle nonlinear relationships. For fairness analysis, we utilised the Dalex library⁵, which provides tools to explain, explore, and mitigate biases in ML models.

3.3. Algorithmic Fairness

In this work, bias is defined as systematic deviations in predictions that favour or disadvantage one group – here, authors – based on sensitive features (such as gender, ethnicity, religion, etc.). To address this, we incorporate fairness analyses to identify and examine such biases.

Group fairness is particularly relevant in our context as it seeks equitable treatment across different groups of authors. This approach balances the distribution of treatments and resources between groups to ensure that predictions do not disproportionately favour or disadvantage one or multiple social groups [16]. *Equality of opportunity*, *equalised odds*, and *calibration within groups* are metrics used to estimate group fairness in predictive models. Integrating these fairness considerations into DH research is crucial, as biased tools can lead to the misrepresentation of corpora and minority groups, as highlighted in [27].

Equality of opportunity ensures that the opportunity to be classified as a true positive instance is equal for all groups, and for all social groups considered their positive rate (TPR) should be equal:

$$\frac{TP_a}{TP_a + FN_a} = \frac{TP_b}{TP_b + FN_b} \text{ for all groups } a, b \quad (1)$$

In relation to the binary classifier for canonicity, equality of opportunity ensures that the likelihood of correctly recognising a canon book is equal regardless of whether the book is written by a man or a woman.

Equalised odds extends beyond equality of opportunity by ensuring equality of the true negative rate (TNR) and the false positives rate (FPR) for all the specified groups:

$$\frac{TN_a}{TN_a + FP_a} = \frac{TN_b}{TN_b + FP_b} \text{ for all groups } a, b \quad (2)$$

For the canonicity classifier, equalised odds ensure that the likelihood of incorrectly classifying a non-canon book as canon is equal regardless of whether the book is written by a man or a woman.

Calibration Within Groups ensures that the precision of the classifier is balanced, meaning the proportion of correct positive predictions (true positives) out of all positive predictions is the same for all groups:

$$\frac{TP_a}{TP_a + FP_a} = \frac{TP_b}{TP_b + FP_b} \text{ for all groups } a, b \quad (3)$$

⁵<https://dalex.drwhy.ai/>

For the canonicity classifier, this means that the books classified as ‘canon’ are actually canon and that the accuracy of these predictions is consistent across books written by both men and women.

Dalex reports various classification outcomes and calculates the fairness metrics outlined above. The criteria are evaluated using the following:

$$\epsilon \leq \frac{\text{metric for non-privileged group}}{\text{metric for privileged group}} \leq \frac{1}{\epsilon} \quad (4)$$

with $\epsilon = 0.8$, following the four-fifths rule [5]. This threshold is widely used to detect significant disparities in treatment between groups. The benefit of this approach is that it offers a clear and standardised benchmark for assessing fairness, while its limitation is that it may not detect subtle biases and could oversimplify complex fairness issues [39]. The groups considered in our experiments are women and men authors, with men authors being the privileged group.

The outlined criteria have been shown to be impossible to satisfy simultaneously, except for trivial cases [32, 23]. This is a challenging finding because it is difficult to justify sacrificing any of these criteria in a fair classifier. It emphasises the importance of conducting fairness analysis and interpretation within the specific context of use, considering the underlying data foundation. We prioritise equalised odds in the canonicity classifier to ensure fair treatment of men and women authors by balancing FPR and TNR across genders. Without this, one group could disproportionately influence what is deemed canonical. See Section 5 for further discussion.

Dalex was also used to estimate feature importance for the canon classifiers, employing a permutation-based approach to compute feature importance. This assesses the contribution of each feature to classification outcomes by systematically permuting them and calculating their impact on model performance.

4. Results

In the first round of the experiments, we used the same sampling methods as in [8] to ensure balance between the positive and negative class: All 618 canon books are used with a random sub-sample of 618 non-canon books. This process was repeated 20 times, and the average accuracy was 0.72 – somewhat reproducing the accuracy of 0.75 reported in [8]. However, as the gender distribution is not equal for either the positive or negative class, we cannot rule out the effect of class imbalance when examining the fairness results, as models trained on imbalanced datasets often develop a bias favouring the majority class [16].

Considering this, we bootstrapped a 50-50 gender distribution to conduct a more meaningful bias analysis. Since the canon group contains few books by women authors (166 vs. 452 by men authors), we randomly sampled 166 books by men authors from the canon group to achieve gender balance, alongside 166 books by men and women authors from the non-canon group, resulting in a total of $n = 664$. Each sampling selects a random subset of canon books by men authors and non-canon books, and the entire process, including model training and fairness analysis, is repeated 20 times. Sampling is conducted with replacement between rounds to

Fairness Check for Binary Classifier for Canon Literature

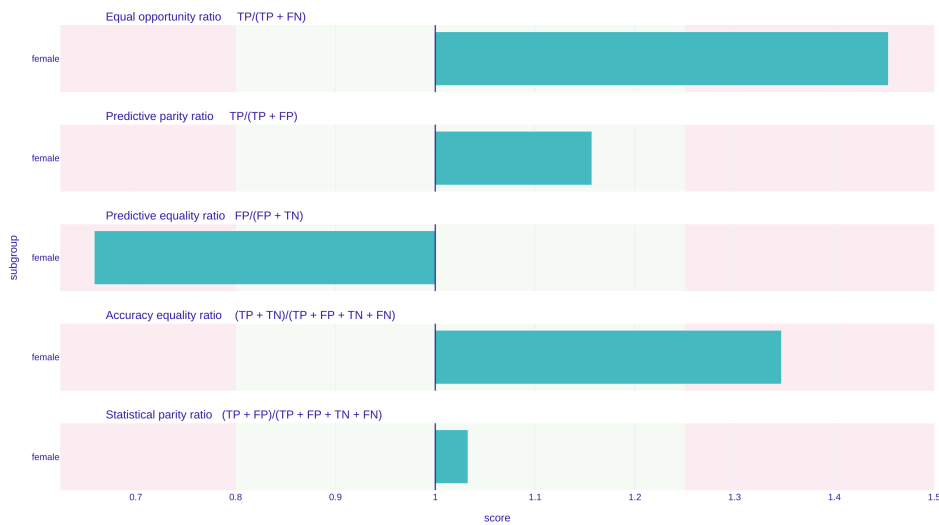


Figure 1: For this model, bias was detected in 3 metrics: TPR, ACC, and FPR. When FPR are too unequal between men and women authors, the fairness criteria Equalised Odds is violated. This is true for 11 of the 20 models. When the TPR is too unequal between men and women authors, the fairness criteria of Equal Opportunity is violated; this holds for 5 of the 20 models. The accuracy for men and women authors is too unequal in 2 out of 20 models.

ensure variability between iterations. Hence, all 166 canon books by women authors are in each run used together with a random subset of canon books by men authors.

When training on a 50-50 gender distribution, the average accuracy on the 20 runs remains approximately the same, 0.71. One potential reason the accuracy is not affected by a smaller data sample is that the balanced gender distribution may enhance the model’s ability to generalise across different author groups, counteracting any potential loss of information from the reduced sample size.

4.1. Fairness

Out of the 20 models trained on a 50-50 gender distribution for both the positive and negative classes, 16 models are unfair according to the fairness criteria. Specifically, for 9 of the models, the FPR is lower for women authors than for men authors, and for 7 models, the TPR is higher for women authors than for men authors. The FPR results indicate that the models have a greater tendency to classify non-canon books by men authors as canon, compared to non-canon books by women authors, violating the equalised odds metric. The higher TPR for women shows that the proportion of correctly recognised canon books is greater for women authors, violating the equality of opportunity metric. To gain insights into these results, in the following section, we summarise the feature distributions in the underlying data and feature importance of the models.

4.2. Feature Importance

4.2.1. Consistent Statistically Different Features

Before examining the predictive models' feature importance, we first tested whether the included features differed between books by women and men authors. To do so, we conducted a Mann-Whitney U test with Bonferroni correction to account for multiple comparisons. This was done for each sample process to ensure that the findings were robust and not related to the random sample. Conducting the test on a 50-50 gender distribution sample rather than the full (imbalanced) dataset minimises the influence of unequal group sizes, providing a clearer understanding of each feature without the confounding effects of gender imbalance. The following features are reported as statistically significant between books by men and women *in the canon set* in more than half of the sampling rounds:

- Narrative features: The mean sentiment of all sentences in the book as well as the mean sentiment of the first and last 10% of the book.
- The normalised frequencies of negation modifiers, auxiliaries, pronouns, verbs, and nominal subjects.
- The ratio between verbs and nouns.

Thus, at least some of the 36 text-intrinsic characteristics differ between the canon books by men and women authors, suggesting that there may be a distinct profile for women and men canon authors. When performing a Mann-Whitney U test with Bonferroni correction for multiple comparisons for the *whole* corpus of 9,000 novels, 31 out of 36 features exhibit a statistically significant difference between books by men vs. books by women authors. Hence, there are larger differences between books written by women and men in the whole corpus than there are in the canon set. Next, we examined each model's feature importance to see if the differences in features between men and women drive the observed biases.

4.3. Feature Importance in Fair and Unfair Models

For each model, we analysed feature importance and counted the presence of each feature in both fair and unfair models, respectively. Using the Dalex Library, which identifies the top 10 most influential features, we counted the presence of these features across all models. Fig. 2 presents an overview of the important features in both fair and unfair models, as well as the features that are reported as statistically significantly different between author genders within the canon set.

The frequency of negation modifiers, type-token-ratio, perplexity and approximate entropy are often reported among the top ten features regardless of whether the classifier is fair or unfair (w.r.t. the considered fairness criteria). Recalling the findings from [8], our results confirm the discriminating power of the textual metric perplexity.

The frequency of negation modifiers and auxiliaries is statistically significant between canon books by men and women authors in *all* sample runs and an important feature in *all* and most fair models, respectively. This might suggest that the canon vs. non-canon signal for these features is stronger than the gender difference. This may also be the case for the type-token

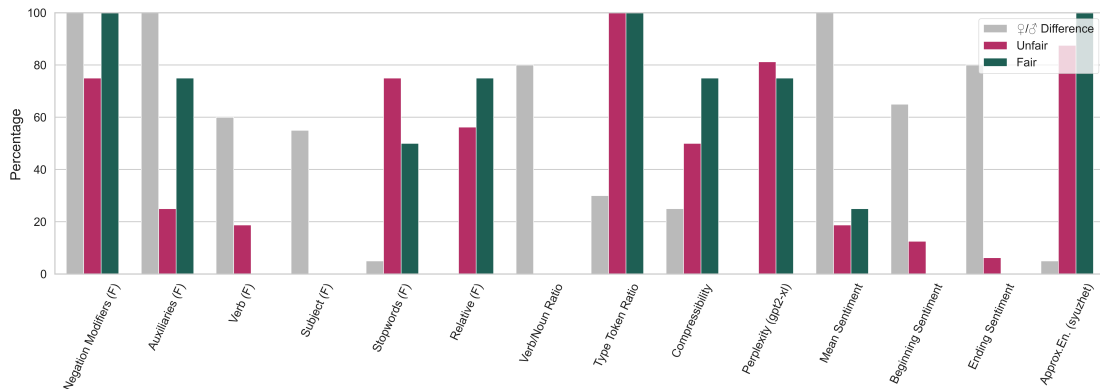


Figure 2: For each feature, the coloured bars show how large a proportion of the fair models (green) and the unfair models (pink) are reporting this feature in their ten most important features. The grey bars show how large a proportion of the sampling rounds the feature is reported as statistically significant between canon books by men and women. Hence, the grey bars are not linked to the classifiers but are descriptive statistics of the underlying data of canon books and can be used to interpret whether the observed biases can be linked to differences in feature distributions. (F) = frequencies, normalised by word count.

ratio, which we report to be different between canon books by men and women in 30% of the sample runs, but important for *all* models.

Furthermore, the frequency of relative clause modifiers and the compressibility of the text are also important features for distinguishing canon books from non-canon books. Both features are reported more often for the fair models, indicating that despite compressibility being reported as different for men and women authors in the canon group (in 25% of the sample runs), this does not explain the observed bias.

For the unfair models specifically, we find that the stop words and verb frequency are more important than in the fair models. Verb frequency is reported as statistically significant between books by men and women canon authors in 60% of the sample runs. It is reported as important only in unfair models, indicating that relying on this feature might contribute to the observed biases. Similarly, although the frequency of stop words is only reported as statistically different in books by women and men canon authors in 5% of the sample runs, it might still add to the observed biases when combined with other features.

For the mean sentiment of the 10% first and last parts of the books, we see that they are only important for the unfair models, while they are reported to be statistically significantly different for books by men and women canon authors. This indicates that these features might contribute to the observed biases. The mean sentiment of all sentences, which is reported as statistically significantly different between men and women authors in *all* runs, is an important feature for 20% of both the fair and the unfair models, and we can, therefore, not conclude how it contributes to biases.

Moreover, the frequencies of nominal subjects and the verb-noun ratio are reported as different between canon books by men and women authors. However, these are not important for the classifiers to tell non-canon from canon books. This suggests that while women canon au-

thors and men canon authors differ in these features, they are not important predictors for the canon category as such. On the other hand, features such as approximate entropy, perplexity, relative clause modifiers, use of stop words, and type-token ratio appear crucial for determining canonicity. Notably there are no substantial differences between men and women authors regarding these features within the canon group, suggesting a shared canon style among men and women canon writers w.r.t. these features.

5. Discussion

The results presented in this paper show that while it is possible to predict canonicity based on text-intrinsic features, it is crucial to consider social biases in these models, such as the effect of author gender. Moreover, our results show that research in DH and computational literary studies can benefit from insights from algorithmic fairness to increase awareness of social biases ingrained into methods and datasets. In the following, we outline our main findings and discuss them in relation to earlier work and fairness considerations.

5.1. Features

Regarding the feature importance results, it is important to note that with the 50-50 gender distribution in this work and the inclusion of PoS frequencies, we do not reproduce the same feature importances as reported in [8]. While perplexity is confirmed as having discriminative power in our results, nominal style, readability, and predictability of sentiment arcs do not appear to be significant predictors of canonicity.

As the experiments in [8] did not take gender into account, their models have been exposed to more men authors than women authors – both of the canon and the non-canon group. In contrast, our bootstrapped sampling process ensures that our models are exposed to an equal number of texts by men and women authors. Predictability and nominal style were reported as statistically significantly different in canon books by men and women authors, but these features did not emerge as predictors of canonicity in our experiments. Therefore, it seems plausible that by highlighting these exact features, the models in [8] might have picked up on a style associated with men (canon) authors rather than canonicity itself. However, keep in mind that our inclusion of PoS features in the analysis may also influence which features are reported as *most important*. It is possible that these features remain important but appear further down the list in our models.

These results underscore the necessity of a careful sampling process when dealing with imbalanced data. Our bootstrapping method, while straightforward, is not without the limitation of reducing datapoints. A more refined approach would involve up-sampling texts by women authors to match the distribution of existing women-authored books.

5.2. False Positives

11 of the 16 unfair models have a higher FPR for men authors than for women authors, showing a tendency to over-include non-canon men in the canon, rather than non-canon women. At the same time, the TPR is higher for women authors, indicating that the models have an

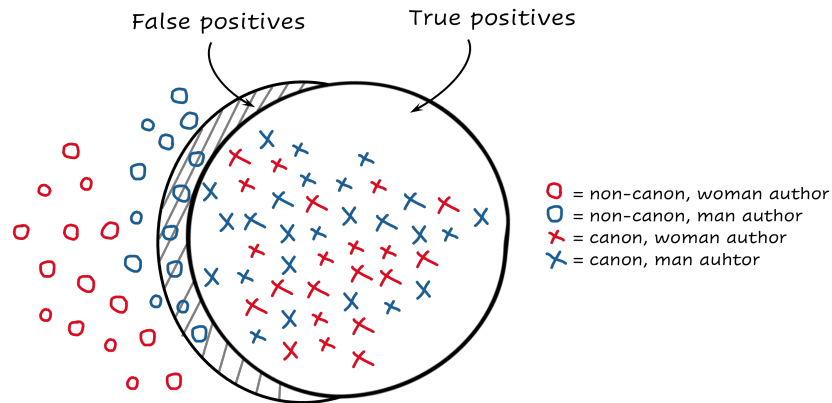


Figure 3: Sketch of *one* potential explanation for the higher FPR for men authors: non-canon books written by women might be more different from the canon group compared to the difference between canon and non-canon books by men.

easier time recognising canon works by women than by men. Overall, this seems to point to a harder divisibility of the men authors’ space between canonical and non-canonical books. One potential explanation for this is that the distance from the canon group might be larger for the non-canon women than for the non-canon men. The hypothesis is sketched in Fig. 3. Further work is needed to test whether this is the case, potentially through techniques like embedding-based clustering of books based on text-intrinsic features used in the present study. A closer examination of how genre plays into the effect observed is also needed, especially as a larger distance between canon and non-canon women authors may be due to other effects related to gender disparity. Women authors are shown to predominantly write in genres such as romance, children’s literature, and young adult fiction [42, 28]. If genres like romance are dominated by women authors and are less represented in canonical compilations [17], and if genres are closely related to writing style[22], the disparity between canon and non-canon books (such as romance novels) by women authors may be larger.

To avoid naturalising these findings, caution is required when speculating that non-canon women authors align less with the canon style; and our study does not draw definitive conclusions about the intrinsic qualities of men’s versus women’s writing. Previous studies have identified gender differences in texts (see Section 2.2), but these findings do not always generalise well [11]. Our analysis reflects the underlying data of the *Chicago Corpus*, which prioritises widely circulated books. If there is a greater disparity between canon and non-canon women authors than men authors, it could result from differential reception [24] and “false labelling” of women’s works. This highlights how seemingly objective text-intrinsic features can embody social biases, as extensively discussed in [11].

5.3. Impossibility considerations

As discussed in Section 3.3, the impossibility theorem of algorithmic fairness [32] shows that different metrics for group fairness are incompatible with each other if the distribution of pos-

itives varies between groups – also known as unequal base rates. In our experiments, we ensured equal base rate through the 50-50 gender distribution for both the positive and negative classes. Despite this, the majority of the models displayed biased predictions based on author gender. Specifically, the FPR is higher for men authors than for women authors in 11 out of the 16 unfair models, leading to a violation of equalised odds.

In real-world scenarios, the base rates are rarely equal. Therefore, addressing such unfairness often involves accepting lower accuracy, a trade-off known as the parity-accuracy trade-off [23]. To balance accuracy and fairness and to choose which fairness metric to prioritise, it is essential to consider the context of use and the intended goals carefully. For a canonicity classifier aimed at understanding canonical literature, it is arguably important to avoid unequal false positives, as this would result in one social group having disproportionate (false) influence over what represents canonical literature. This consideration supports prioritising equalised odds, which addresses fairness in terms of error rates across groups.

The publishing industry might make up another potential use case for binary classifiers predicting categories such as ‘bestseller’ or ‘quality’ [45, 2]. If an ML classifier predicts the success of new manuscripts, it is still preferable to avoid favouring one group over another, thus supporting equalised odds. However, if human experts later sort the manuscripts, over-including false positives is not as harmful as violating equal opportunity (where one group’s positive instances are more likely to be disregarded). In such a use case, the cost of being falsely disregarded is higher than being falsely recognised. Therefore, equality of opportunity is crucial to ensure manuscripts with high potential are equally likely to be recognised, regardless of the author’s group (e.g., gender, ethnicity)

For a binary classifier used in the publishing industry, it is also crucial to consider the fairness criteria *calibration within groups*, as it ensures consistency between predicted probabilities and actual outcomes within each group. Hence, if the classifier consistently predicts a 10% likelihood of bestseller status for manuscripts written by men, then roughly 10% of those manuscripts should indeed turn out to be bestsellers when checked against the actual data, and similarly for other groups. Lack of calibration within groups could lead to systematically overconfident or underconfident predictions for certain groups.

The sections above show how biases can be embedded within ML models used to predict literary phenomena. While the field of algorithmic fairness can help identify and address such skewness, it is worth asking some more fundamental questions about the existing approaches of using imbalanced literary corpora to classify literary works. One thing to keep in mind is that developing predictive ML models relies on an assumption about the existence of classification schema, which can serve as a *ground truth*. In other words, justifying a canonicity classifier through its accuracy relies on an acceptance of the distinction between the canon novels and the non-canon novels and the canon’s historical profile. Such considerations should not be seen as a dismissal of the idea of a literary canon per se; rather, we aim to encourage reflections about what happens when contested classification schema is operationalised into predictive models. Similar points are addressed by Piper [38]: “[W]hile statistical tests can measure the functioning of the model (“the extent to which what we are observing exceeds the boundaries of chance”), they cannot confirm “whether the model is an appropriate approximation of the phenomenon that one is claiming to observe”” (quoted in [11]).

6. Conclusion and Future Works

This study emphasises the critical role of algorithmic fairness in computational literary studies, especially in addressing gender biases in classification models. Despite balanced training data, our findings show that ML models still exhibit significant gender biases, misclassifying non-canon books by men as canon more frequently than those by women, thus violating equalised odds. This suggests that ignoring gender distribution in literary datasets can bias models towards associating men-authored writing styles with canonicity and relatedly can lead to misguided ideas about the textual characteristics of categories like canonic literature.

Our results reveal that seemingly objective text-intrinsic features can harbour social biases, highlighting the need to critically reflect on potential biases in datasets and corpora. By integrating fairness considerations into ML model development and application in computational literary studies, we can not only improve the reliability of the research results but also foster inclusivity by ensuring the representation of all social groups and not just those historically included in established canons.⁶

Further research is needed to understand feature distributions across author genders and their impact on biases. One approach is to create embedding-based clustering to analyse how different author genders are located and distributed within and outside of the canon category.

As pointed out in section 4.3, some features (approximate entropy, perplexity, relative clause modifiers, use of stop words, and type-token ratio) appear crucial for determining canonicity while showing no substantial differences between men and women authors. This suggests a shared canon style among the canon writers, and future work could examine whether these features are consistent across different genres or literary movements within the canon and how this evolves over time.

A limitation of our experiments is that genres were not considered. Future research should incorporate genre distinctions to ensure significant features of canon literature are not conflated with genre-specific ones. This is particularly crucial in the sampling process to avoid comparing canon books against genre literature.

Another limitation is the influence of pressures from the publishing industry. Research has shown that women writers often face constraints from publishers regarding their writing style and subject matter [44, 7]. While this requires further investigation, it highlights the social context shaping how literature is written, published, and distributed - factors that inevitably influence literary data and the resulting predictive models.

References

- [1] M. Algee-Hewitt and M. McGurl. *Between Canon and Corpus: Six Perspectives on 20th-Century Novels*. Literary Lab Pamphlet 8. Stanford Literary Lab, 2015.
- [2] J. Archer and M. L. Jockers. *The bestseller code: Anatomy of the blockbuster novel*. Usa: St. Martin's Press, 2016. DOI: 10.5555/3098683.

⁶Distribution plots for features statistically significant between men and women canon authors are provided in the appendix.

- [3] S. Argamon, M. Koppel, J. Fine, and A. R. Shimoni. “Gender, genre, and writing style in formal written texts”. In: *Text & talk* 23.3 (2003), pp. 321–346. DOI: 10.1515/text.2003.014.
- [4] S. Bagga and A. Piper. “Measuring the effects of bias in training data for literary classification”. In: *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. 2020, pp. 74–84.
- [5] S. Barocas and A. D. Selbst. “Big data’s disparate impact”. In: *Calif. L. Rev.* 104 (2016), p. 671. DOI: 24758720.
- [6] J. Barré, J.-B. Camps, and T. Poibeau. “Operationalizing Canonicity: A Quantitative Study of French 19th and 20th Century Literature”. In: *Journal of Cultural Analytics* 8.3 (2023). DOI: 10.22148/001c.88113.
- [7] I. Berensmeyer. “Authors of Slender Means? Female Authorship in Mid-Twentieth-Century British Fiction”. In: *Zeitschrift für Anglistik und Amerikanistik* 70.4 (2022), pp. 385–402. DOI: 10.1515/zaa-2022-2073.
- [8] Y. Bizzoni, P. Feldkamp, M. Jacobsen, M. R. Thomsen, and K. Nielbo. “Good Books are Complex Matters: Gauging Complexity Profiles Across Diverse Categories of Perceived Literary Quality”. In: *arXiv preprint arXiv:2404.04022* (2024). DOI: 10.48550/arXiv.2404.04022.
- [9] Y. Bizzoni, P. Moreira, M. R. Thomsen, and K. L. Nielbo. “The fractality of sentiment arcs for literary quality assessment: The case of Nobel laureates”. In: *Journal of Data Mining & Digital Humanities* (2023).
- [10] Y. Bizzoni, P. F. Moreira, I. M. S. Lassen, M. R. Thomsen, and K. Nielbo. “A Matter of Perspective: Building a Multi-Perspective Annotated Dataset for the Study of Literary Quality”. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 2024, pp. 789–800.
- [11] K. Bode. “Why you can’t model away bias”. In: *Modern Language Quarterly* 81.1 (2020), pp. 95–124. DOI: /10.1215/00267929-7933102.
- [12] J. Brottrager, A. Stahl, and A. Arslan. “Predicting Canonization: Comparing Canonization Scores Based on Text-Extrinsic and -Intrinsic Features”. In: *CEUR Workshop Proceedings*. Antwerp, Belgium: Ceur, 2021, pp. 195–205.
- [13] J. Burrows. “Textual Analysis”. In: *A Companion to Digital Humanities*. John Wiley & Sons, Ltd, 2004. Chap. 23, pp. 323–347. DOI: 10.1002/9780470999875.
- [14] J. Butler. *Gender Trouble: Feminism and the Subversion of Identity*. Routledge, 2006. DOI: 10.4324/9780203824979.
- [15] P. Chong. “Reading difference: How race and ethnicity function as tools for critical appraisal”. In: *Poetics* 39.1 (2011), pp. 64–84. DOI: <https://doi.org/10.1016/j.poetic.2010.11.003>.
- [16] S. Das, R. Stanton, and N. Wallace. “Algorithmic fairness”. In: *Annual Review of Financial Economics* 15.1 (2023), pp. 565–593. DOI: 10.1146/annurev-financial-110921-125930.
- [17] P. Feldkamp, Y. Bizzoni, M. R. Thomsen, and K. L. Nielbo. *Measuring Literary Quality. Proxies and Perspectives*. Report. Darmstadt, 2024. DOI: 10.26083/tuprints-00027391.

- [18] V. Ganjigunte Ashok, S. Feng, and Y. Choi. “Success with Style: Using Writing Style to Predict the Success of Novels”. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, 2013, pp. 1753–1764.
- [19] M. Hayward. “Are texts recognizably gendered? An experiment and analysis”. In: *Poetics* 31.2 (2003), pp. 87–101. DOI: 10.1016/s0304-422x(03)00005-6.
- [20] M. P. Hiatt. “The feminine style: Theory and fact”. In: *College Composition & Communication* 29.3 (1978), pp. 222–226. DOI: 10.2307/356931.
- [21] S. Jannatus Saba, B. S. Bijoy, H. Gorelick, S. Ismail, M. S. Islam, and M. R. Amin. “A Study on Using Semantic Word Associations to Predict the Success of a Novel”. In: *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*. Online: Association for Computational Linguistics, 2021, pp. 38–51. DOI: 10.18653/v1/2021.stars-em-1.4.
- [22] M. L. Jockers. *Macroanalysis: Digital Methods and Literary History*. Topics in the digital humanities. Urbana: University of Illinois Press, 2013.
- [23] J. Kleinberg, S. Mullainathan, and M. Raghavan. “Inherent Trade-Offs in the Fair Determination of Risk Scores”. In: *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Ed. by C. H. Papadimitriou. Vol. 67. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017, 43:1–43:23. DOI: 10.4230/LIPIcs.ITCS.2017.43.
- [24] C. Koolen. *Women’s books versus books by women*. 2018.
- [25] M. Koppel, S. Argamon, and A. R. Shimoni. “Automatically categorizing written texts by author gender”. In: *Literary and linguistic computing* 17.4 (2002), pp. 401–412. DOI: 10.1093/lc/17.4.401.
- [26] K. Land. “Predicting author gender using machine learning algorithms: Looking beyond the binary”. In: *Digital Studies/Le champ numérique* 10.1 (2020). DOI: 10.16995/dscn.362.
- [27] I. M. S. Lassen, R. D. Kristensen-McLachlan, M. Almasi, K. Enevoldsen, and K. L. Nielbo. “Epistemic consequences of unfair tools”. In: *Digital Scholarship in the Humanities* 39.1 (2024), pp. 198–214. DOI: 10.1093/lc/fqad091.
- [28] I. M. S. Lassen, P. F. Moreira, Y. Bizzoni, M. R. Thomsen, and K. L. Nielbo. “Persistence of Gender Asymmetries in Book Reviews Within and Across Genres”. In: *CEUR Workshop Proceedings*. Vol. 3558. ceur workshop proceedings. 2023, p. 14.
- [29] I. M. S. Lassen, Y. Bizzoni, T. Peura, M. R. Thomsen, and K. L. Nielbo. “Reviewer Preferences and Gender Disparities in Aesthetic Judgments”. In: *CEUR Workshop Proceedings* 3290 (2022), pp. 280–290.
- [30] P. Lauter. “Race and Gender in the Shaping of the American Literary Canon A Case Study from the Twenties”. In: *Canons and Contexts*. Oxford University Press, 1991, pp. 22–47. DOI: 10.1093/oso/9780195055931.003.0007.

- [31] S. Maharjan, J. Arevalo, M. Montes, F. A. González, and T. Solorio. “A Multi-task Approach to Predict Likability of Books”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Ed. by M. Lapata, P. Blunsom, and A. Koller. Valencia, Spain: Association for Computational Linguistics, 2017, pp. 1217–1227.
- [32] T. Miconi. “The impossibility of ”fairness”: a generalized impossibility result for decisions”. In: *arXiv* (2017). DOI: 10.48550/arXiv.1707.01195.
- [33] M. Mohseni, C. Redies, and V. Gast. “Approximate Entropy in Canonical and Non-Canonical Fiction”. In: *Entropy* 24.2 (2022), p. 278. DOI: 10.3390/e24020278.
- [34] P. Moreira, Y. Bizzoni, K. Nielbo, I. M. Lassen, and M. Thomsen. “Modeling Readers’ Appreciation of Literary Narratives Through Sentiment Arcs and Semantic Profiles”. In: *Proceedings of the 5th Workshop on Narrative Understanding*. 2023, pp. 25–35. DOI: 10.18653/v1/2023.wnu-1.5.
- [35] M. L. Newman, C. J. Groom, L. D. Handelman, and J. W. Pennebaker. “Gender differences in language use: An analysis of 14,000 text samples”. In: *Discourse processes* 45.3 (2008), pp. 211–236. DOI: 10.1080/01638530802073712.
- [36] E. L. Overgaard and I. M. Granum. “Kønshierarki i kanonlitteratur: En kvantitativ undersøgelse af køn”. In: *Dansknoter* 2023.3 (2023), pp. 46–49.
- [37] B. G. Pace. “The Textbook Canon: Genre, Gender, and Race in US Literature Anthologies”. In: *The English Journal* 81.5 (1992), pp. 33–38. DOI: 10.2307/819892.
- [38] A. Piper. “Think small: on literary modeling”. In: *Pmla* 132.3 (2017), pp. 651–658.
- [39] P. L. Roth, P. Bobko, and F. S. Switzer III. “Modeling the behavior of the 4/5ths rule for determining adverse impact: Reasons for caution.” In: *Journal of Applied Psychology* 91.3 (2006), p. 507. DOI: 10.1037/0021-9010.91.3.507.
- [40] J. Russ. *How to Suppress Women’s Writing*. Austin, Texas, USA: University of Texas Press, 1983. DOI: 10.7560/316252.
- [41] R. Sarawgi, K. Gajulapalli, and Y. Choi. “Gender Attribution: Tracing Stylometric Evidence Beyond Topic and Genre”. In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*. Ed. by S. Goldwater and C. Manning. Portland, Oregon, USA: Association for Computational Linguistics, 2011, pp. 78–86. DOI: 10.5555/2018936.2018946.
- [42] M. Thelwall. “Book genre and author gender: Romance>Paranormal-Romance to Autobiography>Memoir”. In: *Journal of the Association for Information Science and Technology* 68.5 (2017), pp. 1212–1223. DOI: 10.1002/asi.23768.
- [43] S. Touileb, L. Øvrelid, and E. Vellidal. “Gender and sentiment, critics and authors: a dataset of Norwegian book reviews”. In: *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. Barcelona, Spain (Online): Association for Computational Linguistics, 2020, pp. 125–138.
- [44] G. Tuchman and N. E. Fortin. *Edging women out: Victorian novelists, publishers and social change*. Vol. 13. Oxfordshire, England, UK: Routledge, 2012.

- [45] X. Wang, B. Yucesoy, O. Varol, T. Eliassi-Rad, and A.-L. Barabási. “Success in books: predicting book sales before publication”. In: *EPJ Data Science* 8.1 (2019), p. 31. DOI: 10.1140/epjds/s13688-019-0208-6.

Online Resources

See <https://zenodo.org/records/12699037> for code.

Appendix

Table 2

Number of (sampling) runs where the feature levels are statistically significant between men canon and women canon books. A statistically significant difference is defined as $p < 0.05$, with Bonferroni correction for multiple comparisons. (F) = frequencies, normalised based on the word count.

Feature	Statistically Different n = 20	Unfair models Feature Importance n = 16	Fair models Feature Importance n = 4
Negation Modifiers (F)	20	12	4
Mean Sentiment	20	3	1
Auxiliaries	20	4	3
Pronouns (F)	17	7	1
Verb/Noun Ratio	16	-	-
Ending Sentiment	16	1	-
Beginning Sentiment	13	2	-
Verbs (F)	12	3	-
Subjects (F)	11	-	-
Hurst	8	-	-
Nominal Verb/Noun Ratio	8	-	-
Type Token Ratio	6	16	4
Compressibility	5	8	3
Nouns (F)	4	-	-
Of (F)	3	-	-
Readability (Smog)	3	3	1
Standard deviation, Sentiment	2	8	2
Hurst (Syuzhet)	2	1	1
Approx. En. (Syuzhet)	1	14	4
Stopwords (F)	1	12	2
Perplexity (gpt2-xl_ppl)	-	13	3
Relative clause modifiers (F)	-	9	3
Readability (Dale-Chall New)	-	6	1
Perplexity (gpt2_ppl)	-	6	2
Adjectives (F)	-	4	2
Adverbs (F)	-	4	-
Punctuation (F)	-	4	-
Passive/Active Ratio	-	3	-
Readability (Ari)	-	3	-
Passive (F)	-	3	1
That (F)	-	2	1
Function words (F)	-	2	-
Perplexity(self_model_ppl)	-	2	1
Readability (Flesch Ease)	-	2	1
Readability (Flesch Grade)	-	-	2

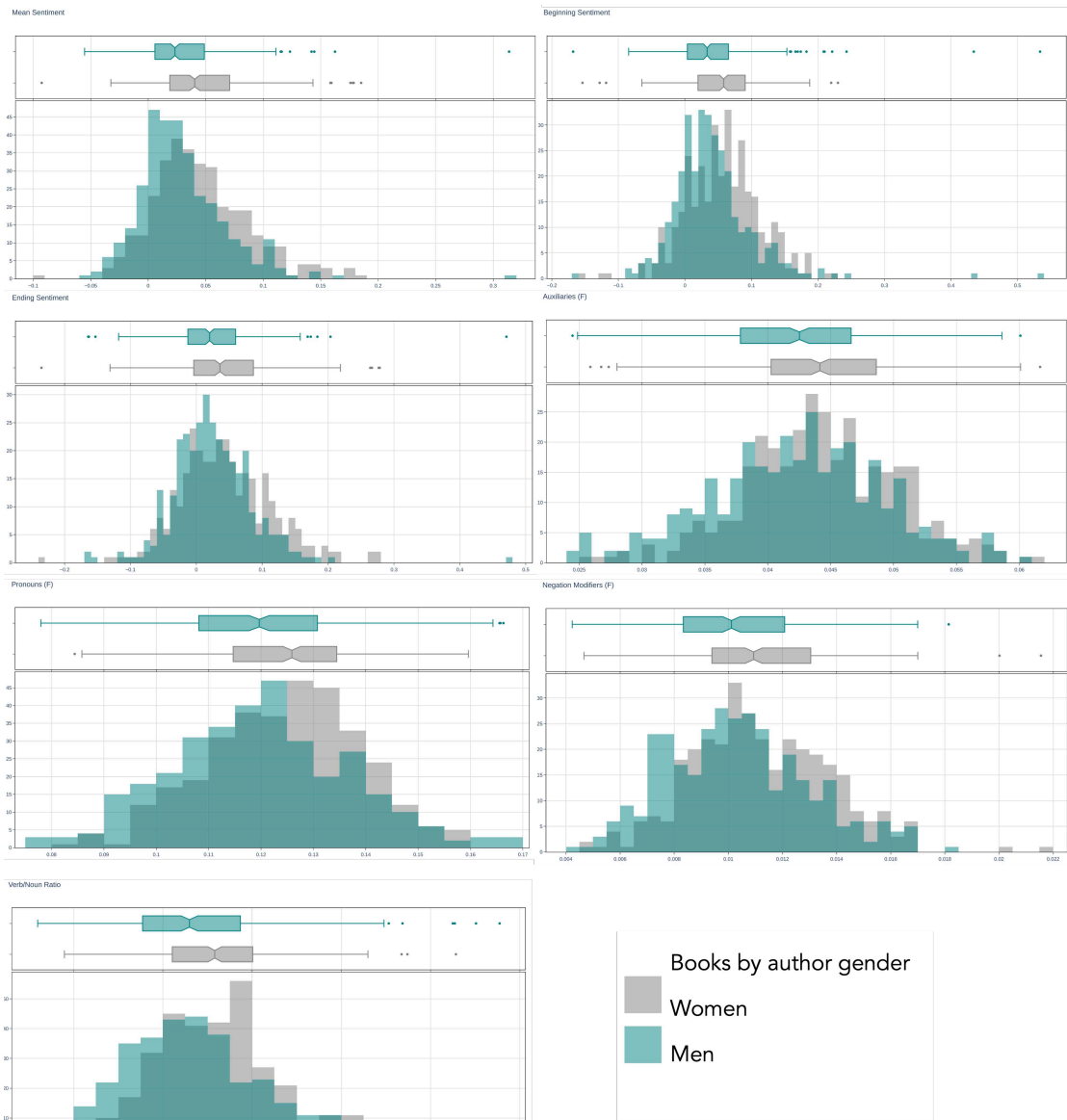


Figure 4: Distribution plots for features reported as statistically significant between men and women canon authors in more than half of the sampling rounds.