These are not the Stereotypes You are Looking For

Koolen, Corina; van Cranenburgh, Andreas

Published in:
Proceedings of the First Ethics in NLP workshop

DOI:
10.18653/v1/W17-1602

These are not the Stereotypes You are Looking For: Bias and Fairness in Authorial Gender Attribution

Corina Koolen
Institute for Logic, Language and Computation, University of Amsterdam
c.w.koolen@uva.nl

Andreas van Cranenburgh
Institut für Sprache und Information
Heinrich Heine University Düsseldorf
cranenburgh@phil.hhu.de

Abstract

Stylometric and text categorization results show that author gender can be discerned in texts with relatively high accuracy. However, it is difficult to explain what gives rise to these results and there are many possible confounding factors, such as the domain, genre, and target audience of a text. More fundamentally, such classification efforts risk invoking stereotyping and essentialism. We explore this issue in two datasets of Dutch literary novels, using commonly used descriptive (LIWC, topic modeling) and predictive (machine learning) methods. Our results show the importance of controlling for variables in the corpus and we argue for taking care not to overgeneralize from the results.

1 Introduction

Women write more about emotions, men use more numbers (Newman et al., 2008). Conclusions such as these, based on Natural Language Processing (NLP) research into gender, are not just compelling to a general audience (Cameron, 1996), they are specific and seem objective, and hence are published regularly.

The ethical problem with this type of research however, is that stressing difference—where there is often considerable overlap—comes with the tendency of enlarging the perceived gap between female and male authors; especially when results are interpreted using gender stereotypes. Moreover, many researchers are not aware of possible confounding variables related to gender, resulting in well-intentioned but unsound research.

But, rather than suggesting not performing research into gender at all, we look into practical solutions to conduct it more soundly.[1] The reason we do not propose to abandon gender analysis in NLP altogether is that female-male differences are quite striking when it comes to cultural production. We focus on literary fiction. Female authors still remain back-benched when it comes to gaining literary prestige: novels by females are still much less likely to be reviewed, or to win a literary award (Berkers et al., 2014; Verboord, 2012). Moreover, literary works by female authors are readily compared to popular bestselling genres typically written by and for women, referred to as 'women's novels,' whereas literary works by male authors are rarely gender-labeled or associated with popular genres (Groos, 2011). If we want to do research into the gender gap in cultural production, we need to investigate the role of author gender in texts without overgeneralizing to effects more properly explained by text-extrinsic perceptions of gender and literary quality.

In other words, NLP research can be very useful in revealing the mechanisms behind the differences, but in order for that to be possible, researchers need to be aware of the issues, and learn how to avoid essentialistic explanations. Thus, our question is: *how can we use NLP tools to research the relationship between gender and text meaningfully, yet without resorting to stereotyping or essentialism?*

Analysis of gender with NLP has roughly two methodological strands, the first *descriptive* and the second *predictive*. First, descriptive, is the technically least complex one. The researcher divides a set of texts into two parts, half written by female and half by male authors, processes these with the same computational tool(s), and tries to explain the

[1] We are not looking to challenge the use of gender as a binary construct in this paper, although this is a position that can be argued as well. Butler (2011) has shown how gender is not simply a biological given, nor a valid dichotomy. We recognize that computational methods may encourage this dichotomy further, but we shall focus on practical steps.

Proceedings of the First Workshop on Ethics in Natural Language Processing, pages 12–22,
Valencia, Spain, April 4th, 2017.

observed differences. Examples are Jockers (2013, pp. 118–153) and Hoover (2013). Olsen (2005) cleverly reinterprets Cixous' notion of *écriture féminine* to validate an examination of female authors separately from male authors (Cixous et al., 1976).

The second, at a first glance more neutral strand of automated gender division, is to use predictive methods such as text categorization: training a machine learning model to automatically recognize texts written by either women or men, and to measure the success of its predictions (e.g., Koppel et al., 2002; Argamon et al., 2009). Johannsen et al. (2015) combines descriptive and predictive approaches and mines a dataset for distinctive features with respect to gender. We will apply both descriptive and predictive methods as well.

The rest of this paper is structured as follows. Section 2 discusses two theoretical issues that should be considered before starting NLP research into gender: preemptive categorization, and the semblance of objectivity. These two theoretical issues are related to two potential practical pitfalls, the ones which we hope to remedy with these paper: dataset bias and interpretation bias (Section 3). In short, if researchers choose to do research into gender (a) they should be much more rigorous in selecting their dataset, i.e., confounding variables need to be given more attention when constructing a dataset; and (b) they need to avoid potential interpretative pitfalls: essentialism and stereotyping. Lastly, we provide computational evidence for our argument, and give handles on how to deal with the practical issues, based on a corpus of Dutch, literary novels (Sections 4 through 6).

Note that none of the gender-related issues we argue are new, nor is the focus on computational analysis (see Baker, 2014). What is novel, however, is the practical application onto contemporary fiction. We want to show how fairly simple, commonly used computational tools can be applied in a way that avoids bias and promotes fairness—in this case with respect to gender, but note that the method is relevant to other categorizations as well.

2 Theoretical issues

Gender research in NLP gives rise to several ethical questions, as argued in for instance Bing and Bergvall (1996) and Nguyen et al. (2016). We discuss two theoretical issues here, which researchers need to consider carefully before performing NLP research into gender.

2.1 Preemptive categorization

Admittedly, categorization is hard to do without. We use it to make sense of the world around us. It is necessary to function properly, for instance to be able to distinguish a police officer from other persons. Gender is not an unproblematic category however, for a number of reasons.

First, feminists have argued that although many people fit into the categories female and male, there are more than two sexes (Bing and Bergvall, 1996, p. 2). Our having to decide how to categorize the novel by the transgender male in our corpus published before his transition is a case in point (we opted for male).

Second, it is problematic because gender is such a powerful categorization. Gender is the primary characteristic that people use for classification, over others like race, age and occupational role, regardless of *actual* importance (Rudman and Glick, 2012, p. 84). Baker (2014) analyzes research that finds gender differences in the spoken section of the British National Corpus (BNC), which indicates gender differences are quite prominent. However, the context also turned out to be different: women were more likely to have been recorded at home, men at work (p. 30). Only when one assumes that gender causes the contextual difference, can we attribute the differences to gender. There is no direct causation, however. Because of the saliency of the category of gender, this 'in-between step' of causation is not always noticed. Cameron (1996) altogether challenges the "notion of gender as a pre-existing demographic correlate which accounts for behavior, rather than as something that requires explanation in its own right" (p. 42).

This does not mean that gender differences do not exist or that we should not research them. But, as Bing and Bergvall (1996) point out: "The issue, of course, is not difference, but oversimplification and stereotyping" (p. 15). Stereotypes can only be built after categorization has taken place at all (Rudman and Glick, 2012). This means that the method of classification itself inherently comes with the potential pitfall of stereotyping.

Although the differences found in a divided corpus are not necessarily meaningful, nor always reproducible with other datasets, an 'intuitive' explanation is a trap easily fallen into: rather than being restricted to the particular dataset, results can

be unjustly ascribed to supposedly innate qualities of all members of that gender, and extrapolated to all members of the gender in trying to motivate a result. This type of bias is called essentialism (Allport, 1979; Gelman, 2003).

Rudman and Glick (2012) argue that stereotypes (which are founded on essentialism) cause harm because they can be used to unfairly discriminate against individuals—even if they are accurate on average differences (p. 95).

On top of that, ideas on how members of each gender act do not remain descriptive, but become prescriptive. This means that based on certain differences, social norms form on how members of a certain gender *should* act, and these are then reinforced, with punishment for deviation. As Baker (2014) notes: "The gender differences paradigm creates expectations that people should speak at the linguistic extremes of their sex in order to be seen as normal and/or acceptable, and thus it problematizes people who do not conform, creating in- and out-groups." (p. 42)

Thus, although categorization in itself can appear unproblematic, actively choosing to apply it has the potential pitfall of reinforcing essentialistic ideas on gender and enlarging stereotypes. This is of course not unique to NLP, but the lure of making sweeping claims with big data, coupled with NLP's semblance of objectivity, makes it a particularly pressing topic for the discipline.

2.2 Semblance of objectivity

An issue which applies to NLP techniques in general, but particularly to machine learning, is the *semblance* of neutrality and objectivity (see Rieder and Röhle, 2012). Machine learning models can make predictions on unseen texts, and this shows that one can indeed automatically identify differences between male and female authors, which are relatively consistent over multiple text types and domains. Note first that the outcome of these machine learning classifiers are different from what many general readers expect: the nature of these differences is often stylistic, rather than content-related (e.g., Flekova et al. 2016; Janssen and Murachver 2005, pp. 211–212). For men they include a higher proportion of determiners, numerical quantifiers (Argamon et al., 2009; Johannsen et al., 2015), and overall verbosity (longer sentences and texts; Newman et al. 2008). For women a higher use of personal pronouns, negative polarity items (Argamon et al., 2009), and verbs stands out (Johannsen et al., 2015; Newman et al., 2008). What these differences mean, or why they are important for literary analysis (other than a functional benefit), is not generally made sufficiently evident.

But while evaluations of out-of-sample predictions provide an objective measure of success, the technique is ultimately not any more neutral than the descriptive method, with its preemptive group selection. Even though the algorithm automatically finds gender differences, the fact remains that the researcher selects the gender as two groups to train for, and the predictive success says nothing about the merits (e.g., explanatory value) of this division. In other words, it starts with the same premise as the descriptive method, and thus needs to keep the same ethical issues in mind.

3 Practical concerns

Although the two theoretical issues are unavoidable, there are two practical issues inextricably linked to them, dataset and interpretation bias, which the researcher should strive to address.

3.1 Dataset bias

Strictly speaking, a corpus is supposed to represent a statistically representative sample, and the conclusions from experiments with corpora are only valid insofar as this assumption is met. In gender research, this assumptions is too often violated, as potential confounding factors are not accounted for, exacerbating the ethical issues discussed.

For example, Johannsen et al. (2015) work with a corpus of online reviews divided by gender and age. However, reflected in the dataset is the types of products that men and women tend to review (e.g., cars vs. makeup). They argue that their use of abstract syntactic features may overcome this domain bias, but this argument is not very convincing. For example, the use of measurement phrases as a distinctive feature for men can also be explained by its higher relevance in automotive products versus makeup, instead of as a gender marker.

Argamon et al. (2009) carefully select texts by men and women from the same domain, French literature, which overcomes this problem. However, since the corpus is largely based on nineteenth century texts, any conclusions are strongly influenced by literary and gender norms from this time period (which evidently differ from contemporary norms).

Koppel et al. (2002) compose a corpus from the

BNC, which has more recent texts from the 1970s, and includes genre classifications which together with gender are balanced in the resulting corpus. Lastly, Sarawgi et al. (2011) present a study that carefully and systematically controls for topic and genre bias. They show that in cross-domain tasks, the performance of gender attribution decreases, and investigate the different characteristics of lexical, syntactic, and character-based features; the latter prove to be most robust.

On the surface the latter two seem to be a reasonable approach of controlling variables where possible. One remaining issue is the potential for publication bias: if for whatever reason women are less likely to be published, it will be reflected in this corpus without being obvious (a hidden variable).

In sum, controlling for author characteristics should not be neglected. Moreover, it is often not clear from the datasets whether text variables are sufficiently controlled for either, such as period, text type, or genre. Freed (1996) has shown that researchers too easily attribute differences to gender, when in fact other intersecting variables are at play. We argue that there is still much to gain in the consideration of author and text type characteristics, but we focus on the latter here. Even within the text type of fictional novels, in a very restricted period of time, as we shall show, there is a variety of subgenres that each have their own characteristics, which might erroneously be attributed to gender.

3.2 Interpretation bias

The acceptance of gender as a cause of difference is not uncommon in computational research (cf. Section 1). Supporting research beyond the chosen dataset is not always sought, because the alignment of results with 'common knowledge' (which is generally based on stereotypes) is seen as sufficient, when in fact this is more aptly described as researcher's bias. Conversely, it is also problematic when counterintuitive results are labeled as deviant and inexplicable (e.g., in Hoover, 2013). This is a form of cherry picking. Another subtle example of this is the choice of visualization in Jockers and Mimno (2013) to illustrate a topic model. They choose to visualize only gender-stereotypical topics, even though they make up a small part of the results, as they do note carefully (Jockers and Mimno, 2013, p. 762). Still, this draws attention to the stereotype-confirming topics.

Regardless of the issue whether differences between men and women are innate and/or socially constructed, such interpretations are not only unsound, they promote the separation of female and male authors in literary judgments. But it can be done differently. A good example of research based on careful gender-related analysis is Muzny et al. (2016) who consider gender as performative language use in its dialogue and social context.

Dataset and interpretation bias are quite hard to avoid with this type of research, because of the theoretical issues discussed in Section 2. We now provide two experiments that show why it is so important to try to avoid these biases, and provide first steps as to how this can be done.

4 Data

To support our argument, we analyze two datasets. The first is the corpus of the Riddle of Literary Quality: 401 Dutch-language (original and translated) novels published between 2007–2012, that were bestsellers or most often lent from libraries in the period 2009–2012 (henceforth: Riddle corpus). It consists mostly of suspense novels (46.4 %) and general fiction (36.9 %), with smaller portions of romantic novels (10.2 %) and other genres (fantasy, horror, etc.; 6.5 %). It contains about the same amount of female authors (48.9 %) as male authors (47.6 %) and 3.5 % of unknown gender, or duo's of mixed gender. In the genre of general fiction however (where the literary works are situated), there are more originally Dutch works by male authors, and more translated work by female authors.

The second corpus (henceforth: Nominee corpus) was compiled because of this skewness; there are few Dutch female literary authors in the Riddle corpus. It is a set of 50 novels that were nominated for one of the two most well-known literary prizes in the Netherlands, the *AKO Literatuurprijs* (currently called ECI Literatuurprijs) and the *Libris Literatuur Prijs*, in the period 2007-2012, but which were not part of the Riddle corpus. Variables controlled for are gender (24 female, 25 male, 1 transgender male who was then still known as a female), country of origin (Belgium and the Netherlands), and whether the novel won a prize or not (2 within each gender group). The corpus is relatively small, because the percentage of female nominees was small (26.2 %).

5 Experiments with LIWC

Newman et al. (2008) relate a descriptive method

of extracting gender differences, using Linguistic Inquiry and Word Count (LIWC; Pennebaker et al., 2007). LIWC is a text analysis tool typically used for sentiment mining. It collects word frequencies based on word lists and calculates the relative frequency per word list in given texts. The word lists, or categories, are of different orders: psychological, linguistic, and personal concerns; see Table 1; LIWC and other word list based methods have been applied to research of fiction (e.g., Nichols et al., 2014; Mohammad, 2011). We use a validated Dutch translation of LIWC (Zijlstra et al., 2005).

5.1 Riddle corpus

We apply LIWC to the Riddle corpus, where we compare the corpus along author gender lines. We also zoom in on the two biggest genres in the corpus, general fiction and suspense. When we compare the results of novels by male authors versus those by female authors, we find that 48 of 66 LIWC categories differ significantly ($p < 0.01$), after a Benjamini-Hochberg False Discovery Rate correction. In addition to significance tests, we report Cohen's d effect size (Cohen, 1988). An effect size $|d| > 0.2$ can be considered non-negligible.

The results coincide with gender stereotypical notions. Gender stereotypes can relate to several attributes: physical characteristics, preferences and interest, social roles and occupations; but psychological research generally focuses on personality. Personality traits related to agency and power are often attributed to men, and nurturing and empathy to women (Rudman and Glick, 2012, pp. 85–86). The results in Table 1 were selected from the categories with the largest effect sizes. These stereotype-affirming effects remain when only a subset of the corpus with general fiction and suspense novels is considered.

In other words, quite some gender stereotype-confirming differences *appear* to be genre independent here, plus there are some characteristics that were also identified by the machine learning experiments mentioned in section 2.2. Novels by female authors for instance score significantly higher overall and within genre in Affect, Pronoun, Home, Body and Social; whereas novels by male authors score significantly higher on Articles, Prepositions, Numbers, and Occupation.

The only result here that counters stereotypes is the higher score for female authors on Cognitive

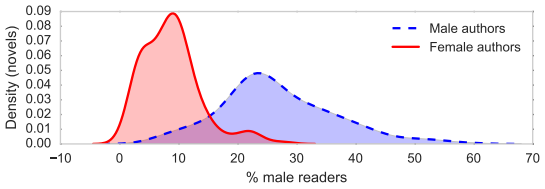

Figure 1: Kernel density estimation of the percentage of male readers with respect to author gender.

Processes, which describes thought processes and has been claimed to be a marker of science fiction—as opposed to fantasy and mystery—because "reasoned decision-making is constitutive of the resolution of typical forms of conflict in science fiction" (Nichols et al., 2014, p. 30). Arguably, reasoned decision-making is stereotypically associated with the male gender.

It is quite possible to leave the results at that, and attempt an explanation. The differences are not just found in the overall corpus, where a reasonable amount of romantic novels (approximately 10 %, almost exclusively by female authors) could be seen as the cause for a gender stereotypical outcome. The results are also found within the traditionally 'male' genre of suspense (although half of the suspense authors are female in this corpus), and within the genre of general fiction.

Nonetheless, there are some elements to the corpus that were not considered. The most important factor not taken into account, is whether the novel has been originally written in Dutch or whether it is a translation. As noted, the general fiction category is skewed along gender lines: there are very few originally Dutch female authors.

Another, more easily overlooked factor is the existence of subgenres which might skew the outcome. Suspense and general fiction are categories that are already considerably more specific than the 'genres' (what we would call text-types) researched in the previously mentioned studies, such as fiction versus non-fiction. For instance, there is a typical subgenre in Dutch suspense novels, the so-called 'literary thriller', which has a very specific content and style (Jautze, 2013). The gender of the author—female—is part of its signature.

Readership might play a role in this as well. The percentage of readers for female and male authors, taken from the Dutch 2013 National Reader Survey (approximately 14,000 respondents) shows how gendered the division of readers is. This distribu-

LIWC category	Examples	Female mean	Female SD	Male mean	Male SD	effect size (d)	sign.
Linguistic							
Prepositions	to, with, above	11.38	0.86	11.92	0.86	-0.63	*
Pronouns	I, them, itself	12.58	1.90	10.14	2.10	1.22	*
Negations	no, not, never	2.02	0.31	1.78	0.35	0.74	*
Article	a, an, the	8.48	1.08	9.71	1.19	-1.08	*
Numbers		0.61	0.15	0.79	0.25	-0.86	*
Psychological							
Social	mate, talk, they, child	10.81	2.00	9.54	1.73	0.68	*
Friends	buddy, friend, neighbor	0.10	0.04	0.09	0.04	0.23	
Humans		0.43	0.16	0.41	0.15	0.11	
Affect	happy, cried, abandon	2.84	0.49	2.35	0.38	1.12	*
Positive emotions	love, nice, sweet	1.38	0.34	1.13	0.23	0.86	*
Cognitive processes	cause, know, ought	5.51	0.67	5.03	0.72	0.69	*
Occupation	work, class, boss	0.54	0.15	0.67	0.20	-0.75	*
Current concerns							
Home	apartment, kitchen, family	0.42	0.13	0.34	0.14	0.57	*
Money	cash, taxes, income	0.20	0.10	0.21	0.10	-0.12	
Body	ache, breast, sleep	1.30	0.41	1.06	0.33	0.63	*

Table 1: A selection of LIWC categories with results on the Riddle corpus. The indented categories are subcategories forming a subset of the preceding category. * indicates a significant result.

tion is visualized in Figure 1, which is a Kernel Density Estimation (KDE). A KDE can be seen as a continuous (smoothed) variant of a histogram, in which the x-axis shows the variable of interest, and y-axis indicates how common instances are for a given value on the x-axis. In this case, the graph indicates the number of novels read by a given proportion of male versus female readers. Male readers barely read the female authors in our corpus, female readers read both genders; there is a selection of novels which is only read by female readers. Hence, the gender of the target reader group differs per genre as well, and this is another possible influence on author style.

In sum, there is no telling whether we are looking purely at author gender, or also at translation and/or subgenre, or even at productions of gendered perceptions of genre.

5.2 Comparison with Nominees corpus

We now consider a corpus of novels that were nominated for the two most well-known literary awards in the Netherlands, the *AKO Literatuurprijs* and *Libris Literatuur Prijs*. This corpus has less confounding variables, as these novels were all originally written in Dutch, and are all of the same genre. They are fewer, however, fifty in total. We hypothesize that there are few differences in LIWC scores between the novels by the female and male authors, as they have been nominated for a literary award, and will not be marked as overtly by a genre. All of them have passed the bar of literary quality—and few female authors have made the cut in this period of time to begin with;[2] thus, we contend, they will be more similar to the male authors in this corpus than in the Riddle corpus containing bestsellers.

However, here we run into the problem that significance tests on this corpus of different size would not be comparable to those on the previous corpus; for example, due to the smaller size, there will be a lower chance of finding a significant effect (and indeed, repeating the procedure of the previous section yields no significant results for this corpus). Moreover, comparing only means is of limited utility. Inspection does reveal that five effect sizes increase: Negations, Positive emotions, Cognitive processes, Friends, and Money; all relate more strongly to female authors. Other effect sizes decrease, mostly mildly.

In light of these problems with the t-test in analyzing LIWC-scores, we offer an alternative. In interpretation, the first step is to note the strengths and weaknesses of the method applied. The largest problem with comparing LIWC scores among two groups with a t-test, is that it only tests means: the mean score for female authors versus the mean score for male authors in our research. A t-test to compare means is restricted to examining the groups as a whole, which, we as we argued, is un-

[2]Note that female authors not being nominated for literary prizes does not say anything about the relationship between gender and literary quality. Perhaps female authors are overlooked, or they write materials of lesser literary quality, or they are simply judged this way because men have set the standard and the standard is biased towards 'male' qualities.

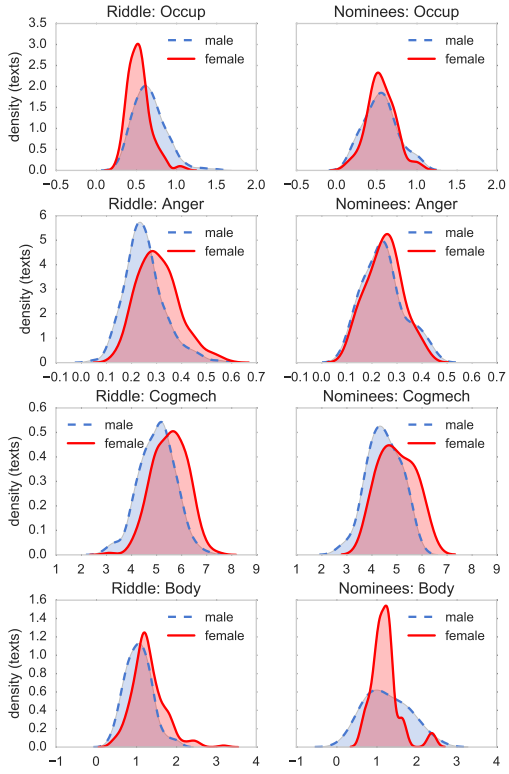

Figure 2: Kernel density estimation of four LIWC categories across the novels of the Riddle (left) and Nominees (right) corpus.

sound to begin with. That is why we only use it as a means to an end. A KDE plot of scores on each category gives better insight into the distribution and differences across the novels; see Figure 2.

Occupation and Anger are two categories of which the difference in means largely disappears with the Nominees corpus, showing an effect size of $d < 0.1$. The plots demonstrate nicely how the overlap has become near perfect with the Nominees corpus, indicating that subgenre and/or translation might have indeed been factors that caused the difference in the Riddle corpus. Cognitive processes (Cogmech) is a category which increases in effect size with the Nominees corpus. We see that the overlap with female and male authors is large, but that a small portion of male authors uses the words in this category less often than other authors and a small portion of the female authors uses it more often than other authors.

While the category Body was found to have a significant difference with the Riddle corpus, in the KDE plot it looks remarkably similar, while in the Nominees corpus, there is a difference not in mean but in variance. It appears that on the one hand, there are quite some male authors who use the words *less* often than female authors, and on the other, there is a similar-sized group of male authors who—and this counters stereotypical explanations—use the words *more* often than female authors. The individual differences between authors appear to be more salient than differences between the means; contrary to what the means indicate, Body apparently is a category and topic worth looking into. This shows how careful one must be in comparing means of groups within a corpus, with respect to (author) gender or otherwise.

Riddle	BoW	char3grams	support
female	83.7	80.8	196
male	82.1	79.9	191
avg / total	**82.9**	80.4	387
Nominees	**BoW**	**char3grams**	**support**
female	63.2	57.9	24
male	77.4	74.2	26
avg / total	**70.6**	66.4	50

Table 2: Gender classification scores (F1) on the Riddle corpus (above) and the Nominees corpus (below).

6 Machine Learning Experiments

In order to confirm the results in the previous section, we now apply machine learning methods that have proved most successful in previous work. Since we want to compare the two corpora, we opt for training and fitting the models on the Riddle corpus, and applying those models to both corpora.

6.1 Predictive: Classification

We replicate the setup of Argamon et al. (2009), which is to use frequencies of lemmas to train a support vector classifier. We restrict the features to the 60 % most common lemmas in the corpus and transform their counts to relative frequencies (i.e., a bag-of-words model; BoW). Because of the robust results reported with character n-grams in Sarawgi et al. (2011), we also run the experiment with character trigrams, in this case without a restriction on the features. We train on the Riddle corpus, and evaluate on both the the Riddle corpus and the Nominees corpus; for the former we use 5-fold cross-validation to ensure an out-of-sample evaluation. We leave out authors of unknown or multiple genders, since this class is too small to learn from.

See Table 2 for the results; Table 4 shows the confusion matrix with the number of correct and in-

female:	toespraak, engel, energie, champagne, gehoorzaam, grendel, drug, tante, echtgenoot, vleug
	$\text{speech}_{\text{NN}}$, angel, energy, champagne, docile, lock, drug, aunt, spouse, tad
male:	wee, datzelfde, hollen, conversatie, plak, kruimel, strijken, gelijk, inpakken, ondergaan
	woe, same, run, conversation, slice, crumble, iron_{VB}, right/just, pack, undergo

Table 3: A sample of 10 distinctive, mid-frequency features.

Riddle	female	male
female	**170**	26
male	40	**151**
Nominees	**female**	**male**
female	**12**	12
male	2	**24**

Table 4: Confusion matrices for the SVM results with BoW. The diagonal indicates the number of correctly classified texts. The rows show the true labels, while the columns show the predictions.

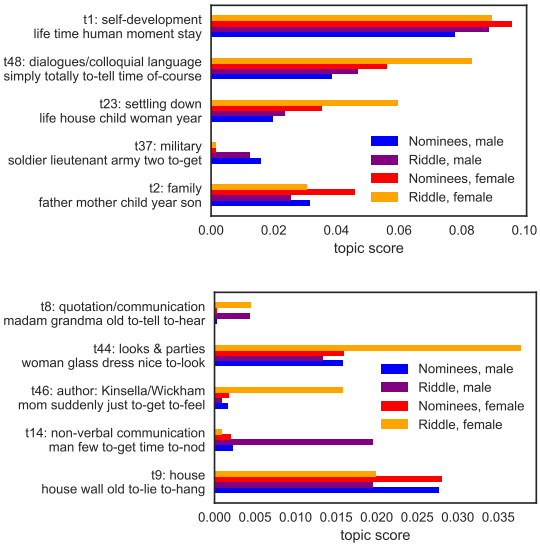

Figure 3: Comparison of mean topic weights w.r.t. gender and corpus, showing largest (above) and smallest (below) male-female differences.

correct classifications. As in the previous section, it appears that gender differences are less pronounced in the Nominees corpus, shown by the substantial difference of almost 10 F1 percentage points. We also see the effect of a different training and test corpus: the classifier reveals a bias for attributing texts to male authors with the Nominees corpus, shown by the distribution of misclassifications in Table 4. On the one hand, the success can be explained by similarities of the corpora; on the other, the male bias reveals that the model is also affected by particularities of the training corpus. Sarawgi et al. (2011) show that with actual cross-domain classification, performance drops more significantly.

A linear model[3] is in principle straightforward to interpret: features make either a positive or a negative contribution to the final prediction. However, due to the fact that thousands of features are involved, and words may be difficult to interpret without context, looking at the features with the highest weight may not give much insight; the tail may be so long that the sign of the prediction still flips multiple times after the contribution of the top 20 features has been taken into account.

Indeed, looking at the features with the highest weight does not show a clear picture: the top 20 consists mostly of pronouns and other function words. We have tried to overcome this by filtering out the most frequent words and sorting words with the largest difference in the Nominees corpus (which helps to focus on the differences that remain in the corpus other than the one on which the model has been trained). As an indication of the sort of differences the classifier exploits, Table 3 shows a selection of features; the results cannot be easily aligned with stereotypes, and it remains difficult to explain the success of the classifier from a small sample as this. We now turn to a different model to analyze the differences between the two corpora in terms of gender.

6.2 Descriptive: Topic Model

We use a topic model of the Riddle corpus presented in Jautze et al. (2016) to infer topic weights for both corpora. This model of 50 topics was derived with Latent Dirichlet Allocation (LDA), based on a lemmatized version of the Riddle corpus without function words or punctuation, divided into chunks of 1000 tokens. We compare the topic weights with respect to gender by taking the mean topic weights of the texts of each gender. From the list of 50 topics we show the top 5 with both

[3]Other models such as decision trees are even more amenable to interpretation. However, in the context of text categorization, bag-of-word models with large numbers of features work best, which do not work well in combination with decision trees.

the largest and the smallest (absolute) difference between the genders (with respect to the Nominees corpus);[4] see Figure 3. Note that the topic labels were assigned by hand, and other interpretations of the topic keys are possible.

The largest differences contain topics that confirm stereotypes: military (male) and settling down (female). This is not unexpected: the choice to examine the largest differences ensures these are the extreme ends of female-male differences.[5] However, the topics that are most similar for the genders in the Nominees corpus contain stereotype-confirming topics as well—i.e., they both score similarly low on 'looks and parties.'

Finally, the large difference on dialogue and colloquial language shows that speech representation forms a fruitful hypothesis for explaining at least part of the gender differences.

7 Discussion and Conclusion

Gender is not a self-explanatory variable. In this paper, we have used fairly simple, commonly applied Natural Language Processing (NLP) techniques to demonstrate how a seemingly 'neutral' corpus—one that consists of only one text-type, fiction, and with a balanced number of male and female authors—can easily be used to produce stereotype-affirming results, while in fact (at least) two other variables were not controlled for properly. Researchers need to be much more careful in selecting their data and interpreting results when performing NLP research into gender, to minimize the ethical issues discussed.

From an ethics point of view, care should be taken with NLP research into gender, due to the unavoidable ethical-theoretical issues we discussed: (1) Preemptive categorization: dividing a dataset in two preemptively invites essentialist or even stereotyping explanations; (2) The semblance of objectivity: because a computer algorithm calculates differences between genders, this lends a sense of objectivity; we are inclined to forget that the researcher has chosen to look or train for these two categories of female and male.

However, we do want to keep doing textual analysis into gender, as we argued we should, in order to analyze gender bias in cultural production. The good news is that we can take practical steps to minimize their effect. We show that we can do this by taking care to avoid two practical problems that are intertwined with the two theoretical issues: dataset bias and interpretation bias.

Dataset bias can be avoided by controlling for more variables than is generally done. We argue that apart from author variables (which we have chosen not to focus on in this paper, but which should be taken into account), text variables should be applied more restrictively. Fiction, even, is too broad as a genre; subgenres as specific as 'literary thriller' can become confounding factors as well, as we have shown in our set of Dutch bestsellers, both in the experiments with LIWC as well as the machine learning experiments.

Interpretation bias stems from considering female and male authors as groups that can be relied upon and taken for granted. We have shown with visualizations that statistically significant differences between genders can be caused by outliers on each end of the spectrum, even though the gender overlap is large on the one hand; and that possibly interesting within-group differences become confounded by solely using means over gender groups on the other hand, missing differences that might be interesting. Taking these extra visualization steps makes for a better basis for analysis that does right by authors, no matter of which gender they are.

This work has focused on standard explanatory and predictive text analysis tools. Recent developments with more advanced techniques, in particular word embeddings, appear to allow gender prejudice in word associations to be isolated, and even eliminated (Schmidt, 2015; Bolukbasi et al., 2016; Caliskan-Islam et al., 2016); applying these methods to literature is an interesting avenue for future work.

The code and results for this paper are available as a notebook at `https://github.com/andreasvc/ethnlpgender`

Acknowledgments

We thank the six (!) reviewers for their insightful and valuable comments.

[4] By comparing absolute differences in topic weights, rarer topics with small but nevertheless consistent differences may be overlooked; using relative differences would remove this bias, but introduces the risk of giving too much weight to rarer topics. We choose the former to focus on the more prominent and representative topics.

[5] Note that the topics were derived from the Riddle corpus, which contains romance and spy novels.

References

Gordon Willard Allport. 1979. *The nature of prejudice*. Basic books.

Shlomo Argamon, Jean-Baptiste Goulain, Russell Horton, and Mark Olsen. 2009. Vive la Différence! Text mining gender difference in French literature. *Digital Humanities Quarterly*, 3(2). http://www.digitalhumanities.org/dhq/vol/3/2/000042/000042.html.

Paul Baker. 2014. *Using corpora to analyze gender*. A&C Black.

Victoria L. Bergvall, Janet M. Bing, and Alice F. Freed, editors. 1996. *Rethinking language and gender research: theory and practice*. Longman, London.

Pauwke Berkers, Marc Verboord, and Frank Weij. 2014. Genderongelijkheid in de dagbladberichtgeving over kunst en cultuur. *Sociologie*, 10(2):124–146. Transl.: *Gender inequality in newspaper coverage of arts and culture*. https://doi.org/10.5117/SOC2014.2.BERK.

Janet M. Bing and Victoria L. Bergvall. 1996. The question of questions: Beyond binary thinking. In Bergvall et al. (1996).

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357. http://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf.

Judith Butler. 2011. *Gender trouble: Feminism and the subversion of identity*. Routledge, New York, NY.

Aylin Caliskan-Islam, Joanna J. Bryson, and Arvind Narayanan. 2016. Semantics derived automatically from language corpora necessarily contain human biases. ArXiv preprint, https://arxiv.org/abs/1608.07187.

Deborah Cameron. 1996. The language-gender interface: challenging co-optation. In Bergvall et al. (1996).

Hélène Cixous, Keith Cohen, and Paula Cohen. 1976. The laugh of the Medusa. *Signs: Journal of Women in Culture and Society*, 1(4):875–893. http://dx.doi.org/10.1086/493306.

Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences*. Routledge Academic, New York, NY.

Lucie Flekova, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. Analyzing biases in human perception of user age and gender from text. In *Proceedings of ACL*, pages 843–854. http://aclweb.org/anthology/P16-1080.

Alice Freed. 1996. Language and gender research in an experimental setting. In Bergvall et al. (1996).

Susan A. Gelman. 2003. *The essential child: Origins of essentialism in everyday thought*. Oxford University Press.

Marije Groos. 2011. Wie schrijft die blijft? schrijfsters in de literaire kritiek van nu. *Tijdschrift voor Genderstudies*, 3(3):31–36. Transl.: *Who writes endures? Women writers in current literary criticism*. http://rjh.ub.rug.nl/genderstudies/article/view/1575.

David Hoover. 2013. Text analysis. In Kenneth Price and Ray Siemens, editors, *Literary Studies in the Digital Age: An Evolving Anthology*. Modern Language Association, New York.

Anna Janssen and Tamar Murachver. 2005. Readers' perceptions of author gender and literary genre. *Journal of Language and Social Psychology*, 24(2):207–219. http://dx.doi.org/10.1177%2F0261927X05275745.

Kim Jautze. 2013. Hoe literair is de literaire thriller? Blog post. Transl.: *How literary is the literary thriller?*, http://kimjautze.blogspot.nl/2013/11/hoe-literair-is-de-literaire-thriller.html.

Kim Jautze, Andreas van Cranenburgh, and Corina Koolen. 2016. Topic modeling literary quality. In *Digital Humanities 2016: Conference Abstracts*, pages 233–237. Krákow, Poland. http://dh2016.adho.org/abstracts/95.

Matthew L. Jockers. 2013. *Macroanalysis: Digital methods and literary history*. University of Illinois Press, Urbana, Chicago, Springfield.

Matthew L. Jockers and David Mimno. 2013. Significant themes in 19th-century literature. *Poetics*, 41(6):750–769. http://dx.doi.org/10.1016/j.poetic.2013.08.005.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of CoNLL*, pages 103–112. http://aclweb.org/anthology/K15-1011.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412. http://llc.oxfordjournals.org/

content/17/4/401.abstract.

Saif Mohammad. 2011. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114. http://aclweb.org/anthology/W11-1514.

Grace Muzny, Mark Algee-Hewitt, and Dan Jurafsky. 2016. The dialogic turn and the performance of gender: the English canon 1782–2011. In *Digital Humanities 2016: Conference Abstracts*, pages 296–299. http://dh2016.adho.org/abstracts/153.

Matthew L. Newman, Carla J. Groom, Lori D. Handelman, and James W. Pennebaker. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3):211–236. http://dx.doi.org/10.1080/01638530802073712.

Dong Nguyen, A. Seza Doğruö, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational Sociolinguistics: A survey. *Computational Linguistics*, 42(3):537–593. http://aclweb.org/anthology/J16-3007.

Ryan Nichols, Justin Lynn, and Benjamin Grant Purzycki. 2014. Toward a science of science fiction: Applying quantitative methods to genre individuation. *Scientific Study of Literature*, 4(1):25–45. http://dx.doi.org/10.1075/ssol.4.1.02nic.

Mark Olsen. 2005. Écriture féminine: searching for an indefinable practice? *Literary and linguistic computing*, 20(Suppl. 1):147–164.

James W. Pennebaker, Roger J. Booth, and Martha E. Francis. 2007. Linguistic inquiry and word count: LIWC [computer software]. www.liwc.net.

Theo Rieder and Bernhard Röhle. 2012. Digital methods: Five challenges. In *Understanding digital humanities*, pages 67–84. Palgrave Macmillan, London.

Laurie A. Rudman and Peter Glick. 2012. *The social psychology of gender: How power and intimacy shape gender relations*. Guilford Press.

Ruchita Sarawgi, Kailash Gajulapalli, and Yejin Choi. 2011. Gender attribution: tracing stylometric evidence beyond topic and genre. In *Proceedings of CoNLL*, pages 78–86. http://aclweb.org/anthology/W11-0310.

Ben Schmidt. 2015. Rejecting the gender binary: a vector-space operation. Blog post, http://bookworm.benschmidt.org/posts/2015-10-30-rejecting-the-gender-binary.html.

Marc Verboord. 2012. Female bestsellers: A cross-national study of gender inequality and the popular–highbrow culture divide in fiction book production, 1960-2009. *European Journal of Communication*, 27(4):395–409. http://dx.doi.org/10.1177%2F0267323112459433.

Hanna Zijlstra, Henriët Van Middendorp, Tanja Van Meerveld, and Rinie Geenen. 2005. Validiteit van de Nederlandse versie van de Linguistic Inquiry and Word Count (LIWC). *Netherlands journal of psychology*, 60(3):50–58. Transl.: *Validity of the Dutch version of LIWC*. http://dx.doi.org/10.1007/BF03062342.