Language differences and accommodation in Amazon book reviews

An analysis of gendered online environments and their impact on linguistic differences between the sexes

Sarah Ball

22 April 2022

1 Introduction

At the heart of many gender attribution attempts in machine learning (ML) is the assumption that gender is a fixed concept with two expressions: female and male (Nguyen et al., 2014). These gender attributions draw on studies suggesting that men and women use different language. Lakoff (1973) was among the first to observe that men and women differ in their general pattern of word usage. Statistical analyses seem to support this finding: Garimella & Mihalcea (2016), for instance, show that broad semantic classes related to sensorial concepts or family are specific to language used by women, whereas classes like sports or job are more related to men's language. While gender attribution studies point to a static understanding of gender, disciplines like sociolinguistics and the social sciences rather consider gender to be "fluid". This means that gender identification and expression are influenced by a variety of variables including the societal context, the culture of a conversation partner, or social roles (Eckert, 2008, 2012). Acknowledging the complexity of gender identification and expression, this paper takes up Ngyuyen et al.'s (2014) call to analyse factors that potentially influence the degree to which a person uses gender-indexical language. Specifically, this paper aims to explore whether the context of gendered book genres influences language con- or divergence between male and female Amazon reviewers. The following three questions are guiding this study:

RQ1: Do men and women use different language when reviewing the same book genre?
RQ2: Does a potential difference in language depend on how "gender-stereotyped" a genre is?
RQ3: Is linguistic adaption, if it exists, equally driven by both genders?

The first research question is based on observations by Sarawgi et al. (2011), who were the first to analyse the performance of ML classifiers in gender attribution tasks after removing the bias introduced by topic and genre. According to the authors, ML techniques which do not account for a text's topic lead to overly optimistic accuracy because men and women tend to be interested in different topics. Their results, however, point to gender-specific language differences that persist even when controlling for a text's topic. The second and third research questions are informed by findings of Boulis & Ostendorf (2005) who show that a ML algorithm can perfectly identify female- and male-only telephone conversations, while the accuracy drops significantly for distinguishing mixed-gender conversations. They conclude that the gender of a telephone partner leads to linguistic accommodation and they also show that this convergence is equally driven by both sexes.

This paper extends the approach of Boulis & Ostendorf (2005) by considering language accommodation in an online context where no simultaneous conversations occur. For studying the research questions, I analyse 767,860 Amazon book reviews in the genres *classics, romance*, and *science fiction*. Since this dataset does not provide the gender of a reviewer directly, I pre-

dict gender based on usernames, employing the Python program *gender-guesser* (Pérez, 2016). Unfortunately, the unavailability of gender information requires to reduce the interpretation of gender to the two categories female and male, although gender is a much broader concept.

Using a Naïve Bayes (NB) classifier with word unigrams as input and Long Short-Term Memory (LSTM) neural networks based on pre-trained fastText word embeddings, this study indicates that men and women write differently about the same book genre. Furthermore, the results imply that gendered book genres are associated with higher levels of language accommodation, which seems to be driven more by men than women. Employing ML techniques to analyse the research questions is based on the idea that a classifier tends to reach higher performance levels the more distinguishable the language between men and women is. However, this methodology has the drawback that the results can be biased by architecture and hyperparameter choices and the quality of the data, which is why the findings of the study are preliminary.

2 Derivation of hypotheses

In line with the outlined findings of Sarawgi et al. (2011), I expect that:

H1: Language differences between men and women persist after controlling for the book genre.

To derive theoretically motivated hypotheses for the second and third research questions, I draw on the Communication Accommodation Theory (CAT, Zhang & Giles, 2018). The theory aims to explain why, when and how speakers in a conversation adapt their language to each other. It generally defines *accommodation* as the ability to adjust one's language to a specific context, characterised by factors like conversation partners and stereotypes (Zhang & Giles, 2018, p.2). Accordingly, *nonaccomodation* describes the effect that speakers refrain from adjusting their language to conversation partners and may furthermore stress linguistic differences. While accommodation is associated with the wish for liking and increased relational closeness, nonaccommodation mainly serves social differentiation and increases relational distance. One contextual factor which influences language accommodation is the *gender* of a conversation parter (Palomares et al., 2016). Studies reveal that many between-gender conversations lead to accommodation (Mulac et al., 1988; Boulis & Ostendorf, 2005; Palomares et al., 2016). However, while Boulis & Ostendorf (2005) show that accommodation is equally driven by both sexes, Hogg (1985), Mulac et al. (1987), and Palomares (2004) find that women tend to adjust their language more than men.
Concerning the Amazon books reviews context, the environments in which the reviews are written are either male-dominated (science fiction), female-dominated (romance) or equally mixed (classic). Following the theoretical outlines, I therefore expect that:

H2: Science fiction and romance reviews show higher levels of language accommodation between men and women compared to classic book reviews.

Furthermore, following the outlined empirical findings, I expect that:

H3: Women adjust their language more than men.

However, according to CAT, between-gender conversations can also be characterised by divergence. This can be explained by another contextual factor which is *gender salience*. Gender salience is a cognitive state where a person implicitly or explicitly counts her- or himself to the corresponding gender group. The level of gender salience depends on the importance of gender in a specific context. It is for instance high if the topic of the conversation is gendered. Gender salience can activate conventions, norms and therefore stereotypical behaviour. Higher levels of gender salience tend to evoke nonaccommodation, thereby strengthening the difference between women and men. Palomares (2004), for example, find that both sexes are using the same level of emotion references in a conversation if gender salience is low, but women use more emotions when gender salience is high.
The genres science fiction and romance are gendered topics, thereby representing high gender salience. This could make women and men behave more in line with classic gender stereotypes, implying nonaccomodation. Therefore, contrary to H2, I expect that:

H4: Science fiction and romance reviews show higher levels of language divergence between men and women compared to classic book reviews.

3 Data

3.1 Amazon books dataset

In order to answer the research questions, I leverage a publicly available Amazon reviews dataset (Ni et al., 2019). The dataset provides information about several characteristics of the reviews, like their rating and helpfulness, next to metadata like the price of a product. The main variables of interest for this paper are the review text, the reviewer's username and the product category. The total number of reviews is 233.1 million, with a date range of May 1996 to October 2018. In order to give the analysis a focus, I choose to only use a book subdataset. Another reason for doing so is that within the books dataset, genres can be both gendered and gender neutral (Thelwall, 2019a; Statista, 2017; Thelwall, 2019b), which serves as an ideal basis to analyse the research questions. Based on prior research and surveys (Thelwall, 2019a; Statista, 2017; Thelwall, 2019b), I choose classic books as the "neutral" comparison category, science fiction as the stereotypically male, and romance as the stereotypical female categories.

Note that all books classified as "classics" entered the analysis apart from classic children books. However, to limit computation time, the specific book categories chosen for romance and science fiction books are "Romance contemporary" and "Science Fiction & Fantasy" respectively. After dropping book duplicates, this selection leads to a preliminary dataset containing around 3.5 million reviews.

3.2 Gender prediction

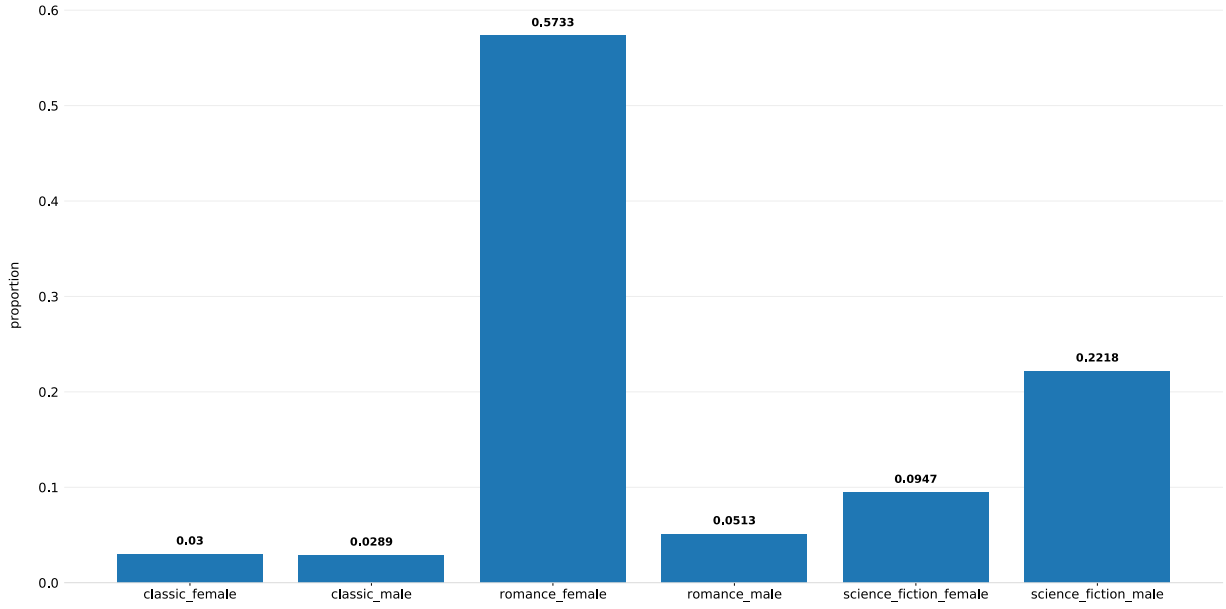
Before preprocessing the review texts, I use an existing gender classifier called *gender-guesser* (GG) (Pérez, 2016 based on Michael, 2008) to predict gender using a reviewer's username. The predictor is based on a dictionary of over 40,000 first names from a variety of countries. It assigns one of the labels *female*, *male*, *mostly female*, *mostly male*, *adrogyn*, and *unknown* to a first name based on country-specific public name statistics. Following Sikdar et al. (2021), I apply the GG to the first token of a reviewer's username and only keep predictions that are "female" or "male". I use the GG because Sikdar et al. (2021) show in an experiment that human annotators' and GG predictions are similar for 98% of the inspected Amazon usernames. Consequently, the GG tool seems reliable and comparable to human perception in the context of Amazon product reviews. Nevertheless, even if the GG predictions are correct, the assigned gender might not correspond with the gender of the actual review writer. This is because I cannot observe who actually wrote the review. While I assume this divergence to be minimal, it might be a source of bias.
When applying the GG, 30.94% of usernames could not be assigned a gender. Of the usernames that were successfully labeled, 70% are female. Figure 1 shows the distribution of predicted male and female observations over the book genres, indicating class imbalance.

3.3 Preprocessing

In the literature on author gender prediction, many different preprocessing steps are used (see e.g. Potthast et al., 2017; Kestemont et al., 2018, 2019). Informed by these authors, I remove all characters that are not in the alphabet or numbers, tokenise each review, use lower case letters, and replace links by the word "url". In a last step, I remove words which are not in the pre-trained fastText word embeddings[1]. For the NB classifier, I additionally delete stop words. Following Rangel et al. (2014), I remove reviews which comprise less than ten words, leading to a final dataset of 767,860 reviews.

[1]Note that I am using a reduced set of the fastText word embeddings.

Figure 1: Distribution of genre-gender labels over the final review dataset



4 Methods

4.1 Model architecture and input choice

The computational task of this paper is to assign genre-gender labels to different reviews and to subsequently assess the performance differences of the classifiers. One comparably simple algorithm that has been employed to conduct this task is the NB algorithm (Calders & Verwer, 2010; Pizarro, 2019). It is a *generative* classifier, meaning that when provided with a review, the algorithm calculates which class most likely generated this review. The NB classifier makes two simplifying assumptions: first, it assumes that the positions of words (or more generally features) in a sentence do not matter and therefore sees the review as a bag-of-words. Second, the probability of a feature to occur in one class is independent of the probability of the other features in the same class, which is called the *Naïve Bayes assumption*. Using these assumptions, the NB classifier returns the predicted class c_{NB} that has the highest posterior probability out of all classes C for having generated a specific review r, which can be formulated as:

$$c_{NB} = argmax_{c \in C} \log(P(c)) + \sum_{f \in F} \log(P(f|c)) \tag{1}$$

where f is a feature. These features can take many forms but I focus on word unigrams for the sake of reducing computational complexity and because one major interest of this paper is lexical differences between the two sexes.

Although the NB assumption is often violated, the NB classifier has been shown to perform well even under violation of this assumption (Domingos & Pazzani, 1997). Furthermore, the

NB can even outperform more complex and hence powerful classifiers (Domingos & Pazzani, 1997). Given these facts and the advantage that few parameters have to be tuned, which reduces calculation time, I choose this classifier as a benchmark.

While the NB classifier in this paper can only account for more "literal" word unigram differences between men and women, a more complex LSTM model can take semantic and syntactic information into account. Therefore, employing this neural network architecture should enable me to investigate a difference in language between the sexes at a more nuanced and detailed level. The improvement over the NB classifier should occur due to more informative input features and due to the more complex architecture of the LSTM.
Concerning the input and following the distributional hypothesis, which states that words occurring in a similar context should have similar meanings (Joos, 1950; Harris, 1954; Firth, 1957), one can use the distribution of words in a text to create numerical representations of word meanings, called embeddings. They capture lexical semantic information such that words with a similar meaning have a similar vector representation. For the LSTM, I employ pretrained fastText embeddings (fastText, 2022), which are based on character n-grams and the skip-gram model.
Concerning the architecture, neural networks like the LSTM (Hochreiter & Schmidhuber, 1997) are *discriminative* classifiers. The idea of discriminative models is that instead of relying on prior class probabilities of words, the model learns which features are more important to distinguish between different classes. Hence, it learns to assign distinguishing features a higher weight and bias via gradient descent and error backpropagation. LSTM models are a special variant of recurrent neural networks (RNN). RNN account for the sequential nature of language by representing prior context, which is why they can capture syntactic information (Shibata et al., 2020). LSTM improve on RNNs by better capturing "long-term dependencies" in a sequence and by resolving the vanishing gradients problem. To achieve this, the LSTM uses a more complex architecture that allows to learn which information to keep and which to forget. For this, the architecture is augmented by a context layer, in addition to the recurrent hidden layers of RNNs, and it uses gates within neural units to influence the information flow into and out of the units. More specifically, there is a "forget gate", which deletes information from the context that has become obsolete, while the "add gate" decides what information from the current input and the previous hidden state should be added to the modified context vector. Lastly, the "output gate" determines the next hidden state. For classification, the last hidden state is passed to a simple feed forward network, which subsequently determines the class via a softmax function.

4.2 Evaluation approach and criteria

To investigate the research questions, I choose two data setups. In the first setup, I train the NB and LSTM classifier on the entire imbalanced dataset and predict six different class labels: *classic_female, classic_male, romance_female, romance_male, science_fiction_female*, and *science_fiction_male*. The performance of the classifiers is measured with F1-scores, which are the harmonic mean of precision and recall values for each class. Especially the macro-averaged F1-score is insightful for this paper, because it allows to detect whether there are performance problems in minority classes. Only considering the overall accuracy could obscure such insights. In the second data setup, I create three genre-specific balanced datasets by upsampling with respect to the *romance_female* class. Note that I choose upsampling over downsampling in order not to lose information. However, I repeat the analysis with downsampled datasets as a robustness check. The resulting classic, romance, and science fiction datasets have 44,436 reviews each, with 50% female and male reviewers. Based on these three datasets, I train three different NB classifiers. I only employ the NB classifier because training three different LSTM classifiers with hyperparameter tuning was not computationally feasible. For evaluating the performance of the classifiers, I consider the overall accuracy.

I choose two different data setups to account for the strengths and weaknesses of both approaches. With the six-label dataset, I can evaluate the performance of the NB and LSTM classifiers on the same test set, which is good practice. However, I cannot distinguish whether performance increases over a random baseline are due to the classifiers learning word differences in book genres or gender or both. Following Boulis & Ostendorf (2005), I hence repeat the analysis on three different genre-specific datasets to explicitly control for the book genre. However, the disadvantage here is that I compare the performance of three different classifiers on three different test sets, which is why the results can be biased by differences in the data and classifiers.

Note that for all the analyses the datasets are split into training, development and test set. For the split on the imbalanced dataset, I use a group shuffle split such that all reviews of a specific book can only occur in one of the three subsets to avoid information leakage. Without this, the accuracy of a classifier could be higher just because it learned book-specific words. The split is conducted such that 80% of the books and their reviews are in the training dataset, and the remaining books are split between the test and development set. Note that I do not use this more complex split for the genre-specific datasets because otherwise, the sizes of the datasets and the class balances would differ, making comparisons difficult. However, this implies that the results can be biased due to information leakage.

4.3 Hyperparameters

Naïve Bayes

The problem with the NB formulation in equation 1 is that if a feature occurs in the test set

but not in the training set, the overall prediction for a class would be zero, thereby ignoring all the information from the other available features. In order to avoid this, I apply smoothing, which basically translates to assigning a non-zero probability to unseen words. There are various forms of smoothing that differ in their complexity (Kaur & Oberai, 2014). In this paper, I use *backoff additive smoothing*, which is a method that adds the parameter α only to the probability of the unseen words. The optimal α is determined by hyperparameter tuning.
In addition to this, I use feature selection. The idea of feature selection is to decrease the dimensionality of the feature space by reducing the number of redundant or noisy data. As a result, the computations should speed up while improving prediction performance. For choosing the subset, I use mutual information (MI) because Novakovic (2010) show that MI increases the accuracy of a NB classifier on different datasets more than other features like the gain ratio. MI is based on entropy (i.e. Shannon information) and measures the uncertainty reduction for one random variable given the value of another random variable. Applied to the Amazon book review context, I am estimating how much information the presence (or absence) of a specific word provides about the different classes. The higher the MI estimate, the more informative is the word for the classes. Using hyperparameter tuning for combinations of α and different numbers of features, I find that for the imbalanced dataset the optimal $\alpha = 0.05$ with 20,000 features and for the upsampled genre-specific datasets $\alpha = 1e^{-10}$ with again 20,000 features.

LSTM

The hyperparameters tuned for the LSTM are the batch size, learning rate, dropout rate, and the number of epochs. Concerning the batch size, this hyperparameter controls how many training examples the classifier sees before updating the weights and biases. If optimally tuned, the advantage of this approach is an increase in computational efficiency because the examples in a batch can be processed in parallel (Jurafsky & Martin, 2022, p.95).
Second, the learning rate influences the magnitude for updating weights and biases at each iteration in the learning process. There is a trade-off: if the rate is too low, the change in the parameters is too small, which is why the learner will take too long to reach the minimum of the loss function or might get stuck in a local minimum. In contrast, if it is too high, the learner will overshoot the minima of the loss function.
Third, the reason for using dropout is based on the observation that ensembles of different neural networks reduce overfitting by combining multiple predictions (Brownlee, 2018). Dropout is a cheap way of simulating the training of "different neural networks", while effectively only one model is trained. This is achieved by dropping edges between hidden layers with a probability that needs to be tuned.
Lastly, the number of epochs determines how often there is a complete pass through the training dataset. More epochs lead to better generalisations but too many epochs result in overfitting and hence bad performance on the test set.
Note that the hyperparameter tuning for this paper only considered a small selection of hyper-

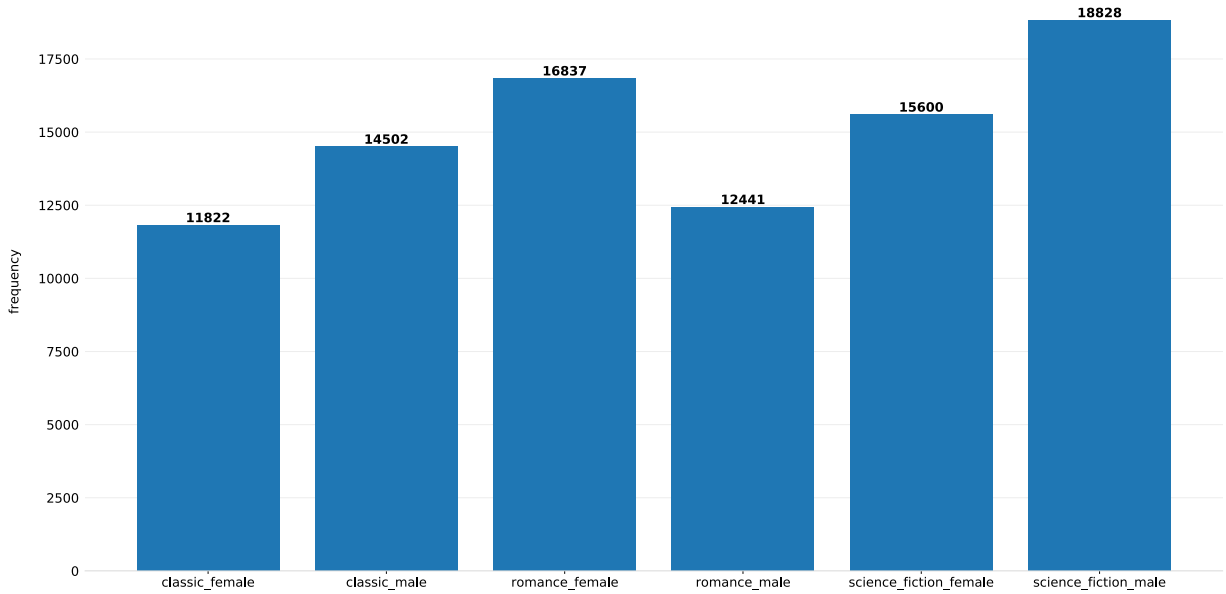
parameters due to limited computational power. The optimal values found are a batch size of 16, a learning rate of 0.0015, a dropout of 0.2, and five epochs.

5 Results

5.1 Preliminary descriptive results

In order to quantitatively assess the similarity of word choice in the different categories, I calculate the Jaccard similarities for the unigrams amongst all six categories. Jaccard similarity is the ratio of the size of two sample set's interaction divided by their union. Table 1 shows that there are very high similarity values for all of the categories. Interestingly, the similarity for gender within the genres is highest for science fiction, with 77.85 % of common unigrams, followed by classics (71.40%), and romance (70.56%). This gives an indication, that women and men use more similar language in the gendered science fiction category while they deviate more in the gendered romance category compared to the neutral classic category. Additionally, the table suggests that some genres are characterised by higher lexical similarities than others. For instance, the class *romance_female* shares more unigrams with *science_fiction_female* and even more with *science_fiction_ male* than with *romance_ male*, which is surprising. These observations could indicate that the classifiers might have problems distinguishing between book genres next to distinguishing gender.

Figure 2: Number of unigrams per class in six-label training dataset



However, the ranking of the similarities in the table can also be influenced by the number

Table 1: Jaccard similarity of word unigrams

	classic_female	classic_male	romance_female	romance_male	science_fiction_female	science_fiction_male
classic_female	1.0000	0.7140	0.6584	0.7041	0.6796	0.6094
classic_male	**0.7140**	1.0000	0.7435	0.6984	0.7542	0.7215
romance_female	0.6584	0.7435	1.0000	0.7056	0.7757	0.7810
romance_male	0.7041	0.6984	**0.7056**	1.0000	0.7013	0.6333
science_fiction_female	0.6796	0.7542	0.7757	0.7013	1.0000	0.7785
science_fiction_male	0.6094	0.7215	0.7810	0.6333	**0.7785**	1.0000

Note: within-genre Jaccard similarity values are highlighted in bold.

of unigrams for the categories, which differ significantly given the imbalance of the dataset (Figure 2). Ignoring the frequency of the n-grams also implies that the measure is biased towards low-frequency words.

5.2 Results of the classifiers

Results imbalanced six-label dataset

Before evaluating the performance of the classifiers, it is helpful to consider an informative baseline.

Table 2: Classification report random baseline, imbalanced dataset

	Precision	Recall	F1-score	Support
classic_female	0.0335	0.0328	0.0331	2318
classic_male	0.0265	0.0267	0.0266	2175
romance_female	0.5538	0.5711	0.5623	42007
romance_male	0.0491	0.0450	0.0470	4218
science_fiction_female	0.1124	0.0938	0.1023	8568
science_fiction_male	0.2148	0.2212	0.2179	16278
accuracy			0.3800	75564
macro avg	0.1650	0.1651	0.1649	75564
weighted avg	0.3714	0.3800	0.3755	75564

Table 3: Classification report correct genre random gender, imbalanced dataset

	Precision	Recall	F1-score	Support
classic_female	0.5073	0.5073	0.5073	2318
classic_male	0.4749	0.4749	0.4749	2175
romance_female	0.9085	0.9085	0.9085	42007
romance_male	0.0891	0.0891	0.0891	4218
science_fiction_female	0.3424	0.3424	0.3424	8568
science_fiction_male	0.6539	0.6539	0.6539	16278
accuracy			0.7190	75564
macro avg	0.4960	0.4960	0.4960	75564
weighted avg	0.7190	0.7190	0.7190	75564

Table 2 shows the classification report for randomly predicting labels according to their class probabilities. This leads to an overall accuracy of ca. 38%, while the macro-averaged F1-score ranges at 16.49%. However, if a classifier beats the F1-baseline-scores for all classes, we cannot directly conclude that this is because the classifier detected linguistic differences between the sexes. Rather, performance improvements could be driven by the classifier only learning how to distinguish the book genres but not the sexes. To see this, Table 3 shows the F1-scores for

correctly predicting the book genres but randomly predicting gender according to their genre-specific probabilities. Correctly predicting the book genres alone already leads to substantial increases in the F1-scores over the original random baseline. Therefore, it is important to also consider the confusion matrices when evaluating the classifiers' performances.

Turning to the NB classifier, Table 4 shows that all the F1-scores are higher than the first random baseline values. The classifier reaches particularly high increases in the F1-score of the classes *science_fiction_male* (+38.29 percentage points, pp) and *classic_male* (+31.07 pp). The improvement is lowest for the class *romance_male* (+5.58 pp). As mentioned before, it is necessary to consider the confusion matrix to assess whether these improvements are due to learning genre and/or gender lexical differences. Figure 6 shows that the classifier has problems distinguishing between the book genres given that a substantial number of predictions are off-diagonal. This indicates that the overall improvement in the classification report (Table 4) was not only due to improvements in distinguishing book genres but also due to distinguishing gender. This notion is supported by the fact that the classifier was able to predict the correct sex (regardless of the genre) for 74.07% of the reviews. This is a performance improvement over randomly predicting a label containing female, which leads to 70% accuracy.

Table 4: Classification report NB classifier, imbalanced dataset

	Precision	Recall	F1-score	Support
classic_female	0.2251	0.3594	0.2768	2318
classic_male	0.3386	0.3361	0.3373	2175
romance_female	0.8026	0.8156	0.8090	42007
romance_male	0.1104	0.0963	0.1028	4218
science_fiction_female	0.3408	0.3681	0.3539	8568
science_fiction_male	0.6475	0.5604	0.6008	16278
accuracy			0.6419	75564
macro avg	0.4108	0.4226	0.4135	75564
weighted avg	0.6471	0.6419	0.6433	75564

Concerning the performance of the LSTM classifier, the overall accuracy increases by 8.38 pp compared to the NB classifier, indicating that the more complex structure is generally better able to detect differences between the reviews. However, as shown by the lower macro average F1-score (Table 5), this increase in performance does not apply to all classes: while the F1-scores for *classic_male*, *romance_female*, and *science_fiction_male* increases compared to the NB F1-scores, there is a decrease for the other three classes. This is especially prevalent for the *science_fiction_female* class for which the F1-score dropped by 17.09 pp. Furthermore, the LSTM is especially bad in detecting the *romance_male* minority class with an F1-score that is even below the first random baseline. When looking at the confusion matrix (Figure 7), we see

that the classifier still has problems in distinguishing between book genres next to gender. Comparing the change in the confusion matrices of the NB and the LSTM classifiers (Tables 6 and 7), the performance loss in the *romance_male* class is entirely due to *romance_male* labels being predicted to be *romance_female*. Similarly, the performance loss in *science_fiction_female* is due to labels being predicted as *science_fiction_male* and *romance_female*. Especially for the *romance_female* class it can be observed that the classifier predicts this label more for all the other classes, which is why the precision value decreased but the recall value increased substantially. Similarly, the classifier predicts more *science_fiction_male*, which increases recall significantly and slightly decreases precision. This indicates that the classifier learned more to predict the majority classes which pushes performance overall at the cost of reduced performance for other labels. However, the percentage of correctly predicted labels only with respect to the sexes increased to 79.43% compared to the 70% of predicting female labels.

Table 5: Classification report LSTM classifier, imbalanced dataset

	Precision	Recall	F1-score	Support
classic_female	0.3520	0.2632	0.3012	2318
classic_male	0.4083	0.3940	0.4010	2175
romance_female	0.7893	0.9594	0.8661	42007
romance_male	0.6667	0.0005	0.0009	4218
science_fiction_female	0.5573	0.1095	0.1830	8568
science_fiction_male	0.6389	0.7452	0.6880	16278
accuracy			0.7257	75564
macro avg	0.5688	0.4120	0.4067	75564
weighted avg	0.6994	0.7257	0.6713	75564

Results balanced two-label datasets

Tables 8, 9 and 10 show the classification reports for the *balanced* classic, romance and science fiction datasets. Note that the random baseline now changes to F1-scores of 50% due to the balance. As can be seen from the tables, all three classifiers reach accuracies higher than 50%, which implies that the classifiers are able to detect language differences when controlling for the book genre. Furthermore, there seem to be performance differences: the highest accuracy (69.15%) and macro-averaged F1-score (68.29%) are reached on the classic dataset, followed by the classifier on the science fiction dataset, which reaches 66.38% accuracy and a macro-averaged F1-score of 65.99%. The worst performance was achieved on the romance dataset with an accuracy of 60.75%, which is almost 10 pp lower than the accuracy for the classic dataset. Note that the same performance ranking is obtained when using a downsampled dataset (Appendix).

Table 6: Confusion matrix NB classifier, imbalanced dataset

	classic_female	classic_male	romance_female	romance_male	science_fiction_female	science_fiction_male
classic_female	833	525	546	22	188	204
classic_male	757	731	310	14	74	289
romance_female	879	173	34259	3012	2143	1541
romance_male	174	83	3028	406	261	266
science_fiction_female	399	168	2070	112	3154	2665
science_fiction_male	659	479	2471	113	3434	9122

Table 7: Confusion matrix LSTM classifier, imbalanced dataset

	classic_female	classic_male	romance_female	romance_male	science_fiction_female	science_fiction_male
classic_female	610	570	785	0.0	18	335
classic_male	438	857	450	1.0	2	427
romance_female	157	47	40302	0.0	131	1370
romance_male	58	32	3876	2.0	20	230
science_fiction_female	198	165	2774	0.0	938	4493
science_fiction_male	272	428	2873	0.0	574	12131

Table 8: Classification report NB classifier, balanced *classic* dataset

	Precision	Recall	F1-score	Support
classic_female	0.6440	0.8563	0.7351	44018
classic_male	0.7857	0.5267	0.6306	44018
accuracy			0.6915	88036
macro avg	0.7148	0.6915	0.6829	88036
weighted avg	0.7148	0.6915	0.6829	88036

Table 9: Classification report NB classifier, balanced *romance* dataset

	Precision	Recall	F1-score	Support
romance_female	0.6044	0.6221	0.6131	44018
romance_male	0.6107	0.5928	0.6016	44018
accuracy			0.6075	88036
macro avg	0.6076	0.6075	0.6074	88036
weighted avg	0.6076	0.6075	0.6074	88036

Table 10: Classification report NB classifier, balanced *science fiction* dataset

	Precision	Recall	F1-score	Support
science_fiction_female	0.6349	0.7707	0.6963	44018
science_fiction_male	0.7083	0.5569	0.6235	44018
accuracy			0.6638	88036
macro avg	0.6716	0.6638	0.6599	88036
weighted avg	0.6716	0.6638	0.6599	88036

6 Discussion and Conclusion

Overall, the NB classifier on the imbalanced six-label dataset shows an improvement over randomly predicting labels according to their class probabilities. In addition, the accuracy of correctly predicting the sexes regardless of the book genre increases over randomly predicting the majority female "sub-label". One possible reason to explain this increase in performance is that women and men indeed show lexical differences when reviewing a book, which supports H1. However, from the chosen setup, it is difficult to see how much of the performance increase was due to correctly predicting the book genre or the gender.

The LSTM classifier shows an improvement in overall accuracy. However, from this I cannot conclude whether men and women have more subtle language differences concerning semantics

and syntax. This is because the improvement over the NB classifier seems mainly driven by the classifier learning to identify the majority classes *romance_female* and *science_fiction_male*. Concerning the setup with three different classifiers on balanced datasets, the results show that all classifiers improve accuracy over randomly predicting female and male labels. This observation supports H1. Furthermore, female and male reviews can best be distinguished in the classic genre, suggesting that men and women use more different language in the "neutral" product environment where gender salience is low. However, when the audiences are more gendered, the accuracy of the classifiers decreases. A substantial difference to the performance of the classic NB classifier can be found for the romance classifier. This implies that the language between men and women is less different in the romance context, supporting the hypothesis of language accommodation to the conversation partner or audience (H2). Similarly, the science fiction classifier performs worse than the classic classifier. However, the difference in the accuracy is not high, meaning that the difference could be random. Overall, these observations support H2 but do *not* support H4, which states that higher gender salience of products leads to language divergence.
Finally, the observation that the classifier for the romance dataset performs worse than the classifier on the science fiction dataset contradicts H3. This is because in the female-dominated romance genre, language adjustment seems to be higher than in the male-dominated science-fiction genre. Assuming that language adjustment happens on the side of the minority, this suggests that men adjust their language more than women.

However, these results have to be interpreted with caution. First, I am comparing the performance of three classifiers which were tested on three separate datasets. Consequently, performance differences could be due to differences in the noisiness of the data. Additionally, using another hyperparameter grid for tuning can potentially result in different performance rankings. Hence, the performance variations found in this study could also be explained by differences in data quality or model set ups and do not necessarily result from language accommodation or divergence.
Second, this analysis is limited with respect to the model inputs and the complexity of the algorithms considered. For instance, most of the papers which reach high performances for classifying gender in social media posts combine several language characteristics like content-based features in the form of word and character n-grams next to style-based features such as character flooding or part-of-speech tagging (Fatima et al., 2017; Rangel et al., 2017; Joo et al., 2019). Concerning more complex architectures, transformer models like BERT (Joo et al., 2019) have been shown to reach accuracy levels of 83%[2]. Compared to the accuracy my model architectures reach, this could imply that my setup is too simplistic and potentially underestimates linguistic differences between men and women.
Furthermore, I only analyse reviewers who revealed their gender through their username. How-

[2]However, the authors did not account for topic differences in social media posts.

ever, if those "revealing" people stress their gender identify significantly more through their language compared to anonymous users, the results are biased. To throughly understand language accomodation, it is necessary to investigate language changes on an individual level. Consequently, one needs text samples of people who have written reviews in all three book genres and evaluate how their style of writing changes given the context.

In summary, this study analysed Amazon book reviews to investigate whether linguistic differences between women and men exist and whether the degree of a potential difference is influenced by the gender stereotype of a book genre. The performances of the NB and LSTM classifiers indicate the presence of language differences even when controlling for the book genres. Additionally, language accommodation seems to be higher in gendered book genres and driven mostly by men. However, these results could also occur due to differences in data quality, model setups and too simplistic algorithm architectures. Hence, further research is necessary to fully address the research questions. Future research should also extend the scope of binary gender studies to acknowledge the complexity of gender identity.

Bibliography

Boulis, C., & Ostendorf, M. (2005). A quantitative analysis of lexical differences between genders in telephone conversations. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, (pp. 435–442).

Brownlee, J. (2018). *A Gentle Introduction to Dropout for Regularizing Deep Neural Networks*. https://machinelearningmastery.com/dropout-for-regularizing-deep-neural-networks/. Accessed March 10, 2022.

Calders, T., & Verwer, S. (2010). Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery*, *21*(2), 277–292.

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple bayesian classifier under zero-one loss. *Machine learning*, *29*(2), 103–130.

Eckert, P. (2008). Variation and the indexical field 1. *Journal of sociolinguistics*, *12*(4), 453–476.

Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology*, *41*, 87–100.

fastText (2022). *English word vectors*. https://fasttext.cc/docs/en/english-vectors.html. Accessed March 10, 2022.

Fatima, M., Hasan, K., Anwar, S., & Nawab, R. M. A. (2017). Multilingual author profiling on facebook. *Information Processing & Management*, *53*(4), 886–904.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Garimella, A., & Mihalcea, R. (2016). Zooming in on gender differences in social media. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, (pp. 1–10).

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

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hogg, M. A. (1985). Masculine and feminine speech in dyads and groups: A study of speech style and gender salience. *Journal of Language and Social Psychology*, *4*(2), 99–112.

Joo, Y., Hwang, I., Cappellato, L., Ferro, N., Losada, D., & Müller, H. (2019). Author profiling on social media: An ensemble learning model using various features. *Notebook for PAN at CLEF*.

Joos, M. (1950). Description of language design. *The Journal of the Acoustical Society of America*, *22*(6), 701–707.

Jurafsky, D., & Martin, J. H. (2022). *Speech and Language Processing. An introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*.

Kaur, G., & Oberai, E. N. (2014). A review article on naive bayes classifier with various smoothing techniques. *International Journal of Computer Science and Mobile Computing*, *3*(10), 864–868.

Kestemont, M., Stamatatos, E., Manjavacas, E., Daelemans, W., Potthast, M., & Stein, B. (2019). Overview of the cross-domain authorship attribution task at {PAN} 2019. In *Working Notes of CLEF 2019-Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019*, (pp. 1–15).

Kestemont, M., Tschuggnall, M., Stamatatos, E., Daelemans, W., Specht, G., Stein, B., & Potthast, M. (2018). Overview of the author identification task at pan-2018: cross-domain authorship attribution and style change detection. In *Working Notes Papers of the CLEF 2018 Evaluation Labs. Avignon, France, September 10-14, 2018/Cappellato, Linda [edit.]; et al.*, (pp. 1–25).

Lakoff, R. (1973). Language and woman's place. *Language in society*, *2*(1), 45–79.

Michael, J. (2008). *List of first names and gender*. `https://raw.githubusercontent.com/lead-ratings/gender-guesser/master/gender_guesser/data/nam_dict.txt` Accessed 2 Nov 2021.

Mulac, A., Studley, L. B., Wiemann, J. M., & Bradac, J. J. (1987). Male/female gaze in same-sex and mixed-sex dyads: Gender-linked differences and mutual influence. *Human communication research*, *13*(3), 323–343.

Mulac, A., Wiemann, J. M., Widenmann, S. J., & Gibson, T. W. (1988). Male/female language differences and effects in same-sex and mixed-sex dyads: The gender-linked language effect. *Communications Monographs*, *55*(4), 315–335.

Nguyen, D., Trieschnigg, D., Doğruöz, A. S., Gravel, R., Theune, M., Meder, T., & de Jong, F. (2014). Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *25th International Conference on Computational Linguistics (COLING 2014)*, (pp. 1950–1961). Dublin City University and Association for Computational Linguistics.

Ni, J., Li, J., & McAuley, J. (2019). Justifying recommendations using distantly-labeled reviews and fined-grained aspects. *Empirical Methods in Natural Language Processing (EMNLP)*.

Novakovic, J. (2010). The impact of feature selection on the accuracy of naïve bayes classifier. In *18th Telecommunications forum TELFOR*, vol. 2, (pp. 1113–1116).

Palomares, N. A. (2004). Gender schematicity, gender identity salience, and gender-linked language use. *Human Communication Research*, *30*(4), 556–588.

Palomares, N. A., Giles, H., Soliz, J., & Gallois, C. (2016). Intergroup accommodation, social categories, and identities. *Communication accommodation theory: Negotiating personal relationships and social identities across contexts*, (pp. 123–151).

Pizarro, J. (2019). Using n-grams to detect bots on twitter. In *CLEF (Working Notes)*.

Potthast, M., Rangel, F., Tschuggnall, M., Stamatatos, E., Rosso, P., & Stein, B. (2017). Overview of pan'17. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, (pp. 275–290). Springer.

Pérez, I. S. (2016). *Gender-guesser 0.4.0*. https://pypi.org/project/gender-guesser/. Accessed 2 Nov 2021.

Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., & Daelemans, W. (2014). Overview of the 2nd author profiling task at pan 2014. In *CLEF 2014 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK, 2014*, (pp. 1–30).

Rangel, F., Rosso, P., Potthast, M., & Stein, B. (2017). Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. *Working notes papers of the CLEF*, (pp. 1613–0073).

Sarawgi, R., Gajulapalli, K., & Choi, Y. (2011). Gender attribution: tracing stylometric evidence beyond topic and genre. In *Proceedings of the fifteenth conference on computational natural language learning*, (pp. 78–86).

Shibata, C., Uchiumi, K., & Mochihashi, D. (2020). How lstm encodes syntax: Exploring context vectors and semi-quantization on natural text. *arXiv preprint arXiv:2010.00363*.

Sikdar, S., Sachdeva, R., Wachs, J., Lemmerich, F., & Strohmaier, M. (2021). The effects of gender signals and performance in online product reviews. *Frontiers in Big Data*, *4*.

Statista (2017). *What genres do you read on a regular basis?* https://www.statista.com/statistics/705758/consumers-genres-regularly-read-by-gender/. Accessed March 10, 2022.

Thelwall, M. (2019a). Reader and author gender and genre in goodreads. *Journal of Librarianship and Information Science*, *51*(2), 403–430.
URL https://doi.org/10.1177/0961000617709061

Thelwall, M. (2019b). Reader and author gender and genre in goodreads. *Journal of Librarianship and Information Science*, *51*(2), 403–430.

Zhang, Y. B., & Giles, H. (2018). Communication accommodation theory. *The international encyclopedia of intercultural communication*, (pp. 95–108).

Appendix

Classification reports of NB classifiers on gender-specific downsampled datasets:

Table 11: Classification report NB classifier, balanced *classic* dataset (downsampled)

	Precision	Recall	F1-score	Support
classic_female	0.6422	0.8263	0.7227	2222
classic_male	0.7565	0.5396	0.6299	2222
accuracy			0.6829	4444
macro avg	0.6993	0.6829	0.6763	4444
weighted avg	0.6993	0.6829	0.6763	4444

Table 12: Classification report NB classifier, balanced *romance* dataset (downsampled)

	Precision	Recall	F1-score	Support
romance_female	0.5891	0.6098	0.5993	2222
romance_male	0.5956	0.5747	0.5850	2222
accuracy			0.5923	4444
macro avg	0.5924	0.5923	0.5921	4444
weighted avg	0.5924	0.5923	0.5921	4444

Table 13: Classification report NB classifier, balanced *science fiction* dataset (downsampled)

	Precision	Recall	F1-score	Support
science_fiction_female	0.6326	0.7408	0.6824	2222
science_fiction_male	0.6873	0.5698	0.6230	2222
accuracy			0.6553	4444
macro avg	0.6599	0.6553	0.6527	4444
weighted avg	0.6599	0.6553	0.6527	4444