Measuring the Effects of Bias in Training Data for Literary Classification

Sunyam Bagga
.txtLAB
McGill University
sunyam.bagga@mcgill.ca

Andrew Piper
Department of Languages, Literatures, and Cultures
McGill University
andrew.piper@mcgill.ca

Abstract

Downstream effects of biased training data have become a major concern of the NLP community. How this may impact the automated curation and annotation of cultural heritage material is currently not well known. In this work, we create an experimental framework to measure the effects of different types of stylistic and social bias within training data for the purposes of literary classification, as one important subclass of cultural material. Because historical collections are often sparsely annotated, much like our knowledge of history is incomplete, researchers often cannot know the underlying distributions of different document types and their various sub-classes. This means that bias is likely to be an *intrinsic* feature of training data when it comes to cultural heritage material. Our aim in this study is to investigate which classification methods may help mitigate the effects of different types of bias within curated samples of training data. We find that machine learning techniques such as BERT or SVM are robust against reproducing certain kinds of social and stylistic bias within our test data, except in the most extreme cases. We hope that this work will spur further research into the potential effects of bias within training data for other cultural heritage material beyond the study of literature.

1 Introduction

One of the challenges facing researchers working with cultural heritage data is the difficulty of producing historically representative samples of data (Bode, 2020). While we have access to very large collections of digitized material (e.g. Hathi Trust, Gale), we often lack knowledge about the distributions of different types of documents and their stylistic qualities within these collections (not to mention within the broader sweep of history more generally). Researchers aiming to build collections for historical study using automated methods are thus faced with a two-part challenge: first, the collection of reliable training data given the absence of annotated data within larger collections; and second, the mitigation of potentially unknown biases within such training data when scaling to the classification of larger historical collections.

In this work, we attempt to measure the effects of such potential *unknown* biases within training data for the purpose of literary classification by testing cases of *known* bias. In essence, we want to simulate the following scenario. A researcher wishes to construct a large sample of historical documents from within a given heritage repository using automated methods. Because there is no consistently annotated data for her purposes, she constructs a small training data sample by hand based on her domain expertise, either by randomly sampling from the larger collection or building around some prior disciplinary consensus. As she moves to implement a process to automatically classify documents based on her training data, she is left with a fundamental uncertainty: Because the underlying distribution of different stylistic and social features of the data within the larger collection are unknown, and given that her sample represents a tiny fraction of all documents, how confident can the researcher be that whatever biases may be present in the training data will (not) be reproduced in the subsequent automated annotations?

Proceedings of LaTeCH-CLfL 2020, pages 74–84
Barcelona, Spain (Online), December 12, 2020.

To model this scenario, we work with a collection of data in which the underlying distributions of stylistic and social features are known and then test a variety of cases of increasing bias within the training data to measure its effects on the test data. We assess three separate forms of stylistic and social bias, which include genre, dialogue, and authorial gender, as well as a variety of classification techniques, including the use of data augmentation, to identify conditions under which the reproduction of bias is minimized. In all cases, the goal of our classification task is the detection of *fiction* from a larger collection of documents. The identification of fictional or "literary" documents within large historical collections is a pressing need for the field of literary studies and has been taken up in several cases (Underwood et al., 2020; Underwood, 2014; Bode, 2020). As yet, however, no assessment has been made of the potential effects of bias within the training data used for such annotation exercises.

As we discuss in detail in Section 5, we surprisingly find that current state-of-the-art techniques in NLP such as BERT (Devlin et al., 2019) (or even SVMs) are robust against reproducing all three kinds of bias within our test data, except in the most extreme cases. One bias in particular, authorial gender, appears to exhibit no effect at all in classification tasks, even under the most extreme circumstances.

2 Related Work

2.1 Bias in NLP

Numerous studies in recent years have analyzed different kinds of biases in NLP systems. They span a wide variety of NLP tasks such as abusive language detection (Sap et al., 2019), language modeling (Lu et al., 2018), learning word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017), and machine translation (Vanmassenhove et al., 2018), among others. For example, Sap et al. (2019) find that African American English tweets are twice as likely to be labelled offensive compared to others, demonstrating racial bias in hate speech detection systems, while Bolukbasi et al. (2016) and Caliskan et al. (2017) show that word embeddings trained on popular datasets strongly exhibit gender stereotypes. In each of these cases, research demonstrates that biases encoded in the training data used for automated detection tasks is reproduced in NLP outputs.

Recent work by Blodgett et al. (2020) has emphasized that in light of the numerous ways "bias" has been interpreted and studied in the literature that researchers explicitly state both their working understanding of bias and also the concrete social harms that can follow from bias in NLP. Our concern in focusing on text classification for cultural heritage materials is designed to address the problem of historical representation and the role that automated systems play in the construction of our understanding of the past. Biases in training data could lead to misleading representations of the past, which could in turn lead to "harms of representation" identified by Blodgett et al. (2020) with respect to different, often historically marginalized social groups. For example, training data that does not adequately reflect women's participation in the production of literature could in turn generate historical samples that severely under-represent women's role in the history of literature. Similar concerns could be raised about racial, ethnic, or regional identities. Moreover, Zadrozny (2004) show that some machine learning classifiers are affected by the problem of sample selection bias, that is, when the training data consist of randomly drawn samples from a different distribution than the test data about which the learned model is expected to make predictions.

2.2 Data Augmentation for NLP

Due to the lack of large-scale and reliably annotated data in historical collections, many researchers will necessarily have to begin with manually collected training data, which in most cases constrains the size and diversity of training data.

Data augmentation has been proposed as a strategy to help train more robust models by improving the quantity and quality of training data. It is commonly and effectively used in the domains of computer vision (Krizhevsky et al., 2017), speech (Cui et al., 2015), and is now being explored for NLP tasks (Wei and Zou, 2019). To date, no work has experimented with data augmentation for the task of literary classification. Here, we experiment with two forms of data augmentation: (1) the Easy Data Augmentation (EDA) model that has been shown to provide performance gains across five classification tasks

(Wei and Zou, 2019), and (2) a hand-engineered model consisting of augmentation techniques such as back-translation (Yu et al., 2018), crossover (FM, 2019), and substituting proper names (Section 3.3).

2.3 Literary Text Classification

Within the larger field of text classification, very few studies have experimented with optimizing classification within the literary domain. Yu (2008) implement naive bayes and support vector machines (SVMs) for two literary classification tasks. Allison et al. (2011) show that compute algorithms and digital methods can be successfully used to build predictive models (if not explanatory models) for literary genres. More recently, Underwood et al. (2020) released a large collection of volumes that were predicted to be fiction through algorithmic modeling. They implement regularized logistic regression using a feature set that consists of unigrams (words) along with a few structural features. Our work provides further understanding of the relationship between classifiers and the prediction of literary documents.

3 Methodology

The goal of our experiments is to test the effects of bias in training data on the classification of test data belonging to the domain of literature. For our purposes, following the work of Underwood et al. (2020), we attempt to predict whether a document is a work of "fiction" within a binary classification task (fiction/non-fiction) and modulate different stylistic features, which we describe in Section 4. In this section, we describe the dataset followed by the classification algorithms used along with the data augmentation techniques applied to the training data.[1]

3.1 Dataset

The data used in this paper consists of 866 digitized works of "bestselling" contemporary writing according to Amazon.com and published between 2000 and 2016 (Piper and Portelance, 2016). The breakdown of works are based on Amazon's genre tags and include: 200 works of non-fiction, comprised of a variety of sub-genres including history, biography, policy, self-help, etc.; 235 works of "Mystery" novels, 220 works of "Science Fiction," and 211 "Romance" novels. All works were selected based on their ranking within the "bestselling" sorting mechanism on Amazon and reviewed by hand for genre appropriateness.

3.2 Classification Algorithms

We experiment with a diverse set of classification algorithms using both traditional machine learning and deep learning for this supervised learning task.

Traditional Machine Learning. We implement the following learning algorithms that are commonly used for classification: Logistic Regression, Support Vector Machines (SVM), Random Forests and Gradient Boosting Classifier (GBC). A crucial aspect of the text classification pipeline is feature representation. We represent the input text as a bag of word n-grams which is one of the most simple yet effective methods for feature vectorization. We experiment with unigrams, bigrams, trigrams and all three word n-grams combined, and pick the one which yields the best performance in a 5-fold cross validation test. Simultaneously, we perform cross-validated hyperparameter tuning for each of our learning algorithms. The algorithms and features are implemented using the scikit-learn library (Pedregosa et al., 2011).

Deep Learning. Deep learning models have achieved state-of-the-art results in many text classification tasks (Minaee et al., 2020). In this work, we implement a number of deep learning models: Convolutional Neural Networks (CNN), Long Short-Term Memory Networks (LSTM), Bi-directional LSTM, and Stacked Bi-directional LSTM. For the input embedding layer to the network, we experiment with: (1) using static pre-trained GloVe embeddings (Pennington et al., 2014), (2) using the dynamic deep contextualized ELMo representations (Peters et al., 2018) which utilizes transfer learning. We use the cross entropy loss function and Adam optimizer (Kingma and Ba, 2014) for learning network weights.

Recently, many transformer-based models have been proposed in the field that have outperformed all other learning models on several NLP tasks including text classification. We implement two such models

[1] We make our code and metadata publicly available at https://github.com/sunyam/bias-literary-classification.

EDA	CDA
1. Synonym Replacement: Replace N random words from the sentence with one of their synonyms.	**1. Crossover**: generates new instances by swapping halves of two instances with the same label (inspired by genetic algorithm's crossover).
2. Random Insertion: Insert a random synonym of a random word in the sentence at a random position in the sentence. Do this N times.	**2. Back-translation**: generates paraphrases of the input (English) text by translating it to another language and translating it back to English. We use four languages – French, Korean, German, Spanish.
3. Random Swap: Randomly choose two words in the sentence and swap their positions. Do this N times.	**3. Proper Names**: We generate two new instances by: (1) deleting all proper names in the input text; (2) substituting all proper names with random names.
4. Random Deletion: Randomly remove each word in the sentence with probability p.	

Table 1: Data augmentation techniques implemented in this work. As recommended by Wei and Zou (2019), we generate 16 instances per training instance for both EDA and CDA ($N = 25$ and $p = 0.05$).

here: Bidirectional Encoder Representations for Transformers (BERT) (Devlin et al., 2019) and XLNet (Yang et al., 2019). More specifically, we first load these large pre-trained models and fine-tune them on our literary classification task. The deep learning models are implemented using PyTorch (Paszke et al., 2019), AllenNLP (Gardner et al., 2017) and Transformers (Wolf et al., 2019) libraries.

3.3 Data Augmentation

The shortage of labeled training data is a major concern for many supervised learning tasks since data annotation is a time-consuming and expensive process. This has led to the idea of data augmentation which refers to increasing the size and diversity of training data without actually collecting new data. In this work, we implement two sets of data augmentation techniques: Easy Data Augmentation (Wei and Zou, 2019) and a set of other techniques we group under the name Custom Data Augmentation (CDA). They are presented in Table 1. Note that we use the Google Translate API[2] for back-translation; for proper names, we use the Stanford Named Entity Recognizer (Finkel et al., 2005) to locate them in the text and NameDatabases[3] for sampling random names. Additionally, we also merge EDA and CDA where 8 augmented instances are generated using the former and 8 using the latter.

4 Experiment Design

In this section, we introduce our experimental setup to test the robustness of the classifiers when presented with different forms of bias in the training data, grouped into the following three categories: genre, dialogue, and gender. Our research question is which classification techniques (if any) mitigate different types of known biases in the training data when it comes to accurately classifying test documents?

In all cases, we hold our test set constant and then manipulate our training data according to different kinds of stylistic and social features described below, and evaluate each model's performance. We begin by establishing a baseline where the distribution of the feature of interest (e.g. genre, dialogue, gender) is the same for both the train and test sets. We then gradually distort the training data and measure observed declines in model performance using the evaluation metrics described below. The size of the training data remains constant throughout: 200 fiction and 200 non-fiction documents.[4] A "document" refers to a 500-word passage randomly sampled from a single work. When sampling passages, we condition on the middle 40% of the work in order to avoid paratext at the beginning or end of a work. Additionally, we

[2]https://py-googletrans.readthedocs.io/

[3]https://github.com/smashew/NameDatabases

[4]With the exception of Uniform-Genre experiment where we use 201 fiction documents, 67 from each genre.

experiment with using longer passages from these volumes (up to 10,000 words) as described in Section 5. Finally, we sample no more than 1-3 passages per work for each scenario.

4.1 Genre

An important characteristic of fictional documents is the way they consist of a variety of sub-types, one aspect of which can be captured through the notion of "genre." While the term genre can be interpreted in different ways, we use it here to mean stylistic distinctions among literary documents that have a strong "thematic" orientation. Genres such as "mystery," "science fiction," or "romance," address significantly different real-world scenarios, which may affect the nature of characterization, narrative voice, or event-types. Research has shown that such generic distinctions exhibit a strong degree of categorical difference from the perspective of machine learning (Underwood, 2016; Piper and Portelance, 2016).

We thus hypothesize that training data that is biased with respect to genre may produce biased representations when it comes to the prediction of test data. If a researcher generated (unknowingly) a training data sample based mostly or only on science fiction, for example, would this training data produce predictions that were 1) less accurate with respect to the broader *fiction* category and 2) less equally distributed among other kinds of genres *within* the category of fiction (i.e. biased towards science fiction)?

Train-Set Scenarios. This experiment is broken down into four training data scenarios listed below.

1. Uniform: We keep the fiction train-set uniformly distributed across the three genres, where the fiction set consists of 1/3 Mystery, 1/3 Romance and 1/3 SciFi passages. This scenario corresponds to the red dot in Figure 1.

2. Genre-Dominated: For each genre, we begin with 50% of the training data being dominated by that genre and increase the genre dominance by 10% until the entire training set consists of a single genre (i.e. 100%). The other two genres are split evenly for each scenario.

Test Set. We keep the test set static across the four scenarios: 99 non-fiction documents and 99 fiction documents equally distributed across the three genres (similar to Scenario 1).

4.2 Dialogue

While there are no systematic studies on the prevalence and distinctiveness of dialogue within fiction (whether over time or by genre), narrative theory suggests that the emphasis or avoidance of dialogue by writers indicates an important quality of fictional narrative (Genette, 1983). As a representation of oral speech, dialogue captures a distinctive stylistic quality of fictional documents that strongly departs from the linguistic norms and codes of narration and description (Bal, 2017). We thus hypothesize that imbalances with respect to the underlying distribution of dialogue within training data will pose challenges for the accurate prediction of fiction and may bias future samples from adequately capturing the underlying distribution of dialogue within a specific sub-domain of fiction.

Sampling. To create our sample of documents, we first process all documents using BookNLP (Bamman et al., 2014) which allows us to identify words that appear in dialogue. We then divide passages into two groups: "with dialogue" and "no dialogue." To select passages with dialogue, we condition on all passages with dialogue from each work and keep the top two passages in terms of percentage of quoted words. Doing so, we observe that the average percentage of quoted words in our "with dialogue" sample is 81.2%. To select passages with no dialogue, we sample passages that have zero quoted words. We only sample from the Mystery novels to control for the effects of genre.

Train-Set Scenarios. For our training scenarios, we begin with a train set consisting of 0 fiction passages with dialogue and gradually increase the number of passages with dialogue to 10%, 20%, 30%, ... up until 100%. As with the genre experiments, non-fiction data is kept constant.

Test Set. The test set is kept static across all train-set scenarios. It contains 100 non-fiction passages and 100 fiction passages where 50 passages are drawn from our "with dialogue" sample and 50 from the "no dialogue" sample.

4.3 Authorial Gender

A great deal of work in the NLP community as well as literary studies has focused on problems of gender bias when it comes to the use of word embeddings (Caliskan et al., 2017; Bolukbasi et al., 2016), classification techniques (Mandell, 2019), and modeling more generally (Bode, 2020). Empirical work has demonstrated that gender represents an important form of social inequality within the literary realm (Underwood et al., 2018; Weinberg and Kapelner, 2018; Kraicer and Piper, 2019). It is thus imperative to develop methods of data curation that do not reproduce or amplify such historical inequalities. In this set of experiments, we explore the effects that biases with respect to authorial gender might have on predictive accuracy and balance when it comes to classifying works of fiction. We manually annotate our gender assignments based on identified gender via the author's public biography. While we found no non-binary authors in our study, it is important to acknowledge that the binary labels we use here are for heuristic purposes of identifying potential bias and are not designed to capture a more diverse understanding of gender. In order to control for the effects of genre, we once again only sample from the Mystery novels.

Train-Set Scenarios. As with our dialogue experiments, we begin with 0% men novelists where all passages of fiction are written by women authors and gradually increase the number of fiction passages written by men to 10%, 20%, 30%, ..., until 100%.

Test Set. Consistent with our other experiments, we use a static test set with 100 non-fiction passages and 100 fiction passages with 50% of fiction written by men and 50% by women.

4.4 Evaluation

In order to assess the effects of bias within our training data, we evaluate our classifiers across the following two dimensions. In all cases, we compare performance with respect to a baseline where the distributions in the train and test sets mirror each other.

- Accuracy: We report standard performance metrics such as F1-score, accuracy, precision, and recall to address the question: did the increase of bias result in a decline of overall predictive accuracy?
- Balance: Our second evaluation goal is to capture distortions in the positive predictions for each of our scenarios. Here we are asking whether an increase in bias in training data leads to an increase in imbalanced sub-classes within our overall class of fiction. When a classifier is trained largely on a single genre or presence of dialogue or gender, is it able to equally identify fiction that does not belong to that genre, level of dialogue or gender? This is in many ways the more important measure for our purposes because it allows us to see how biased training data impacts the underlying distribution of a feature of interest. Will biased training data produce biased samples? To measure this, we report the relative entropy of the true positives for a given classification task. An entropy of 0 would mean that there is no class imbalance produced among the different sub-types of fiction tested, while higher entropy indicates greater skew towards a single sub-type. Table 2 indicates the relationship between entropy scores and class imbalances for the genre experiment.

5 Results and Discussion

5.1 Classifier Performance

Which classifiers perform best at our task of literary classification? In order to systematically compare performance, we start by implementing all classifiers for the *Uniform* Genre experiment setting. The classification metrics and relative entropy for a diverse set of classifiers and augmentation techniques are shown in Table 2.[5] As expected, transformer-based models perform very well on this task with BERT outperforming all other classifiers by at least 2 or more F1 points. Convolutional Neural Networks that utilize transfer learning through ELMo embeddings (F1 = 0.9) perform much better than their GloVe embeddings counterpart (F1 = 0.87).

[5]Due to space constraints, we only present a subset of the top performing classifiers here. Specifically, Random Forest, GBC, and LSTM-based models did not perform well on this task and are not shown in Table 2.

Classifier	F1-score	Precision	Recall	Accuracy	True-Positives Distribution	Relative Entropy
BERT	**0.9394**	0.9394	0.9394	0.9394	{mys: 0.3226, sci: 0.3333, rom: 0.3441}	0.00034
BERT + EDA	0.9333	0.9479	0.9192	0.9343	{rom: 0.3516, mys: 0.3187, sci: 0.3297}	0.00083
XLNet	0.9175	0.9368	0.899	0.9192	{rom: 0.3596, mys: 0.3034, sci: 0.3371}	0.00241
CNN (ELMo) + EDA	0.91	0.8571	0.9697	0.904	{sci: 0.3438, rom: 0.3438, mys: 0.3125}	0.00099
LogReg + EDA	0.9029	0.8692	0.9394	0.899	{sci: 0.3333, mys: 0.3333, rom: 0.3333}	0.0
SVM + EDA and CDA	0.9009	0.8835	0.9192	0.899	{sci: 0.3516, mys: 0.3077, rom: 0.3407}	0.00158
CNN (ELMo)	0.9005	0.8482	0.9596	0.8939	{rom: 0.3474, sci: 0.3368, mys: 0.3158}	0.00078
SVM + EDA	0.8986	0.8611	0.9394	0.8939	{sci: 0.3333, mys: 0.3333, rom: 0.3333}	0.0
BERT + CDA	0.898	0.9072	0.8889	0.899	{rom: 0.3636, mys: 0.3182, sci: 0.3182}	0.00203
LogReg + EDA and CDA	0.8955	0.8824	0.9091	0.8939	{sci: 0.3333, mys: 0.3222, rom: 0.3444}	0.00036
SVM	0.891	0.8393	0.9495	0.8838	{rom: 0.3298, sci: 0.3404, mys: 0.3298}	0.00011
LogReg	0.89	0.8455	0.9394	0.8838	{rom: 0.3333, sci: 0.3333, mys: 0.3333}	0.0
LogReg + CDA	0.8768	0.8558	0.899	0.8737	{mys: 0.3146, rom: 0.3483, sci: 0.3371}	0.00088
SVM + CDA	0.875	0.8349	0.9192	0.8687	{mys: 0.3187, rom: 0.3516, sci: 0.3297}	0.00083
CNN (GloVe)	0.8732	0.8158	0.9394	0.8636	{rom: 0.3548, sci: 0.3226, mys: 0.3226}	0.00102

Table 2: Performance of classifiers across the two dimensions for Genre (Uniform) experiment. Note that both train and test set are uniformly distributed across genres, and Relative Entropy is calculated between the True-Positives distribution and the test-set distribution {mys: 0.333, sci: 0.333, rom: 0.333}.

The traditional learning models – SVM and Logistic Regression – achieve somewhat comparable F1-scores of 89%. Moreover, these algorithms are the only ones to achieve a perfect entropy score of 0. It is worth mentioning, however, that BERT's entropy score of 0.00034 also yields a TP-distribution that is extremely close to the ideal test-set distribution.

Does Data Augmentation Help? We observe marginal performance gains of about 1 F1-point with EDA for Logistic Regression, SVM, and CNN. However, the performance drops by 0.6% for BERT. Our Custom Data Augmentation (CDA) technique does not improve the F1-score for any of the classifiers. From these empirical observations, we conclude that the augmentation techniques implemented in this work do not provide significant performance gains for this classification task on our data.

Passage-Length. In addition to 500-word passages, we experiment with using longer passages – 1,000 to 10,000 words – from these volumes. We find that SVM's F1-score goes up from 0.891 when using 500 words to 0.955 with 10,000 words.[6] The peak is achieved at 5,000 words with an F1-score of 0.959.

Given these findings, we continue our bias analysis only implementing BERT and SVM without any data augmentation for all subsequent experiments using 500-word passages.

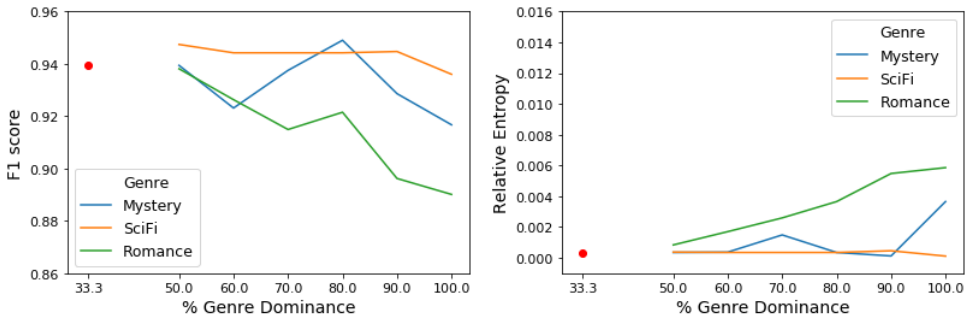

Figure 1: BERT's performance across the two evaluation-dimensions on the Genre experiments.

5.2 Bias Analysis

Genre Bias. Figure 1 shows BERT's performance according to both F1 score (left) and relative entropy (right) for the three different scenarios where a given genre dominates the training data by increasing

[6]Note that BERT's performance stays approximately constant at 0.94 since the pre-trained model's max length is restricted to 512 tokens. This is because the model learns positional-embeddings with sequence lengths of up to 512 tokens only.

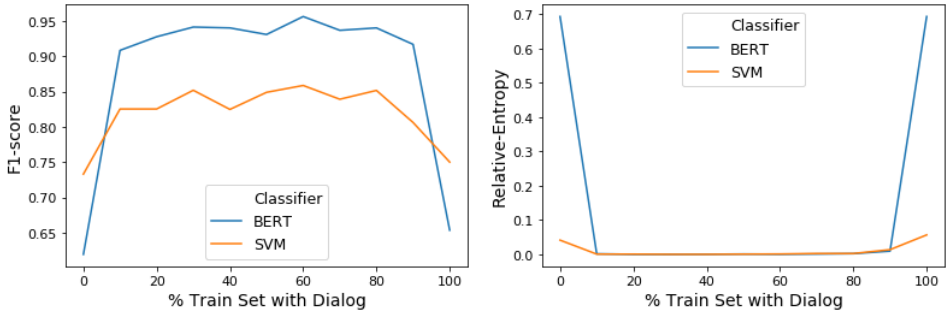

Figure 2: BERT and SVM's performances across the two dimensions on the Dialog experiments.

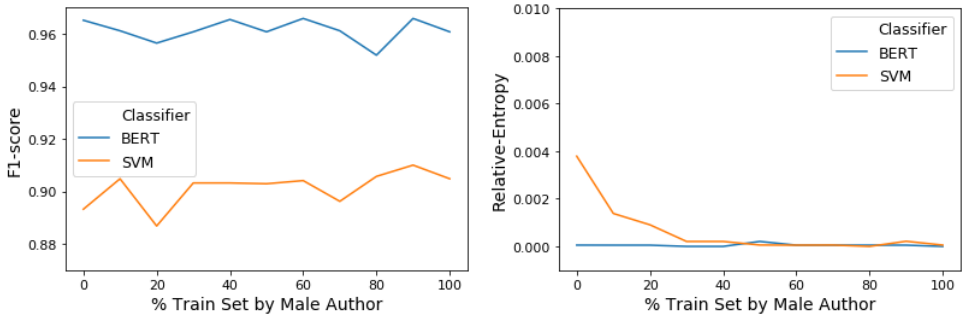

Figure 3: BERT and SVM's performances across the two dimensions on the Gender experiments.

amounts. We can see that when training data and test data mirror each other, BERT achieves an F1-score of 0.939 and an almost-perfect entropy score of 0.00034 (red dot). As we introduce more genre imbalance, we see little performance decrease until 90% imbalance has been achieved. When the train-set is 100% Romance, the F1-score goes down 5 points to 0.89 and the relative entropy goes up to 0.00585 which corresponds to a TP-distribution that is slightly biased towards romance: {sci: 0.296, rom: 0.383, mys: 0.321}. SVM also exhibits a similar trend[7] (not shown here) with its entropy going up to 0.04 where 43% of the True Positives belong to romance. In sum, BERT appears to be robust against genre biases as long as training data is not biased upwards of 90% for a single genre.

Dialogue Bias. Figure 2 shows both SVM and BERT's performance when we change the distribution of fiction passages with dialogue in the training set. Both classifiers' performance is stable across both dimensions except for the most extreme case (i.e. 100% of the train-set has dialogue or no dialogue). Somewhat surprisingly, as long as a classifier does not learn that fiction *only* consists of dialogue (or the opposite), it should not condition on types of fiction with a differential preference for dialogue.

Gender Bias. The findings of this experiment are presented in Figure 3. As can be seen, both BERT and SVM's performance is relatively stable and constant across F1-score and relative entropy. Unlike the other two experiments' findings, this holds true even for the most extreme cases of authorial gender imbalance in the train-set. While there appears to be an asymmetry with respect to gender bias - training data of all women authors will produce more imbalance than training data with all men - the relative entropy (of 0.0037) corresponds to an imbalance of 45.6% men and 54.4% women even in the most extreme case.

6 Conclusion

In this paper, we have tested different classifiers, different data augmentation techniques, and different forms of training data bias to assess their effects on the task of literary classification. Overall, we

[7] In fact, this trend is consistent across all different passage lengths we experimented with. Due to space constraints, the rest of the plots are provided in the GitHub repository.

have found that BERT is the best-performing classifier with SVMs comparable with text-passages above 5,000 words in length. Data augmentation as we have implemented it provides little performance gain. Finally, the stylistic and social biases tested here exhibit little effect except in the most extreme cases ($> 90\%$ bias for a given category) suggesting that at least for the purposes of literary text classification, underlying biases in training data are not as impactful as researchers have initially hypothesized. Nevertheless, our work is limited in its historical scope (different historical periods may exhibit different effects), cultural specificity (our effects have only been observed on English-language documents), classification task (other types of classification may perform differently), and stylistic breadth (stylistic features important to other domains or research questions may behave differently). It is also important to emphasize that while classifiers can mitigate the propagation of bias within training data scenarios (up to a point and under certain conditions), they cannot address biases built into the underlying digital collections from which new collections are created (Bode, 2020). We hope that these experiments provide a useful framework for further refining our understanding of the effects of bias on multiple forms of cultural classification. Future work will want to test different classification scenarios, types of stylistic or social bias, multiple linguistic contexts, as well as further historical document types to better understand how unknown biases in training data may impact our representation of the past using digital collections.

References

Sarah Allison, Ryan Heuser, Matthew Jockers, Franco Moretti, and Michael Witmore. 2011. Quantitative formalism. Stanford Literary Lab.

Mieke Bal. 2017. *Narratology: Introduction to the Theory of Narrative*. University of Toronto Press.

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A Bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379, Baltimore, Maryland, June. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July. Association for Computational Linguistics.

Katherine Bode. 2020. Why You Can't Model Away Bias. *Modern Language Quarterly*, 81(1):95–124.

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 D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Curran Associates, Inc.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

X. Cui, V. Goel, and B. Kingsbury. 2015. Data augmentation for deep neural network acoustic modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(9):1469–1477.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Luque FM. 2019. Atalaya at tass 2019: Data augmentation and robust embeddings for sentiment analysis. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF)*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Gérard Genette. 1983. *Narrative Discourse: An Essay in Method*. Cornell University Press.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Eve Kraicer and Andrew Piper. 2019. Social characters: The hierarchy of gender in contemporary english-language fiction. *Journal of Cultural Analytics*, 3(1).

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, May.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018. Gender bias in neural natural language processing.

Laura Mandell. 2019. Gender and cultural analytics: Finding or making stereotypes? In *Debates in the Digital Humanities*, pages 3–26.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2020. Deep Learning Based Text Classification: A Comprehensive Review. *arXiv e-prints*, page arXiv:2004.03705, April.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Andrew Piper and Eva Portelance. 2016. How cultural capital works: Prizewinning novels, bestsellers, and the time of reading. *Post-45*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy, July. Association for Computational Linguistics.

William E Underwood, David Bamman, and Sabrina Lee. 2018. The transformation of gender in english-language fiction. *Journal of Cultural Analytics*, 2(2).

Ted Underwood, Patrick Kimutis, and Jessica Witte. 2020. Noveltm datasets for english-language fiction, 1700-2009. *Journal of Cultural Analytics*, 4(5).

Ted Underwood. 2014. Understanding genre in a collection of a million volumes, interim report. Technical report, University of Illinois Urbana-Champaign.

William E Underwood. 2016. The life cycles of genres. *Journal of Cultural Analytics*, 1(5).

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium, October-November. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China, November. Association for Computational Linguistics.

Dana B. Weinberg and Adam Kapelner. 2018. Comparing gender discrimination and inequality in indie and traditional publishing. *PLOS ONE*, 13(4):1–20.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv e-prints*, page arXiv:1906.08237, June.

Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. 2018. Fast and accurate reading comprehension by combining self-attention and convolution. In *International Conference on Learning Representations*.

Bei Yu. 2008. An evaluation of text classification methods for literary study. *Literary and Linguistic Computing*, 23(3):327–343, 09.

Bianca Zadrozny. 2004. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 114, New York, NY, USA. Association for Computing Machinery.