

Measuring Intersectional Biases in Historical Documents


WARNING: This paper shows dataset samples that are racist in nature

Nadav Borenstein^{*1} Karolina Stańczak^{*1} Thea Rolskov² Natália da Silva Perez³
Natacha Klein Käfer¹ Isabelle Augenstein¹

¹University of Copenhagen ²Aarhus University ³Erasmus University Rotterdam
nadav.borenstein@di.ku.dk ks@di.ku.dk 201706833@post.au.dk
dasilvaperez@eshcc.eur.nl nkk@teol.ku.dk augenstein@di.ku.dk

Abstract

Data-driven analyses of biases in historical texts can help illuminate the origin and development of biases prevailing in modern society. However, digitised historical documents pose a challenge for NLP practitioners as these corpora suffer from errors introduced by optical character recognition (OCR) and are written in an archaic language. In this paper, we investigate the continuities and transformations of bias in historical newspapers published in the Caribbean during the colonial era (18th to 19th centuries). Our analyses are performed along the axes of gender, race, and their intersection. We examine these biases by conducting a temporal study in which we measure the development of lexical associations using distributional semantics models and word embeddings. Further, we evaluate the effectiveness of techniques designed to process OCR-generated data and assess their stability when trained on and applied to the noisy historical newspapers. We find that there is a trade-off between the stability of the word embeddings and their compatibility with the historical dataset. We provide evidence that gender and racial biases are interdependent, and their intersection triggers distinct effects. These findings align with the theory of intersectionality, which stresses that biases affecting people with multiple marginalised identities compound to more than the sum of their constituents.

 <https://github.com/copenlu/intersectional-bias-pbw>

1 Introduction

The availability of large-scale digitised archives and modern NLP tools has enabled a number of sociological studies of historical trends and cultures (Garg et al., 2018; Kozłowski et al., 2019; Michel et al., 2011). Analyses of historical biases and stereotypes, in particular, can shed light on past

* Equal contribution.

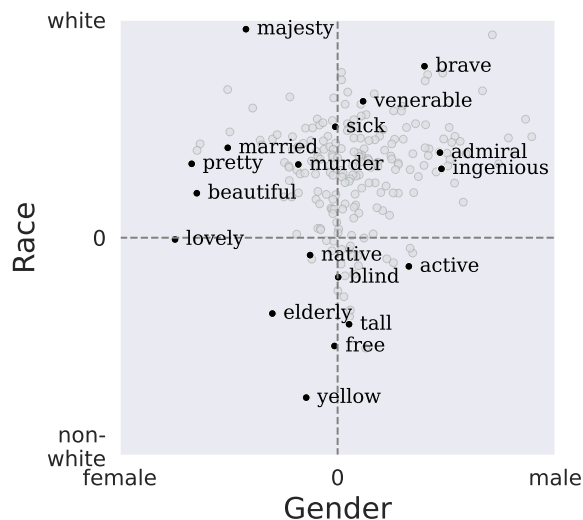


Figure 1: PMI analysis of our historical corpora. Words are placed on the intersectional gender/race plane.

societal dynamics and circumstances (Levis Sulam et al., 2022) and link them to contemporary challenges and biases prevalent in modern societies (Payne et al., 2019). For instance, Payne et al. (2019) consider implicit bias as the cognitive residue of past and present structural inequalities and highlight the critical role of history in shaping modern forms of prejudice.

Thus far, previous research on bias in historical documents focused either on gender (Rios et al., 2020; Wevers, 2019) or ethnic biases (Levis Sulam et al., 2022). While Garg et al. (2018) separately analyse both, their work does not engage with their intersection. Yet, in the words of Crenshaw (1995), intersectional perspective is important because “the intersection of racism and sexism factors into black women’s lives in ways that cannot be captured wholly by looking separately at the race or gender dimensions of those experiences.”

Analysing historical documents poses particular challenges for modern NLP tools (Borenstein et al., 2023; Ehrmann et al., 2020). Misspelt words due

to wrongly recognised characters in the digitisation process, and archaic language unknown to modern NLP models, i.e. historical variant spellings and words that became obsolete in the current language, increase the task’s complexity (Bollmann, 2019; Linhares Pontes et al., 2019; Piotrowski, 2012). However, while most previous work on historical NLP acknowledges the unique nature of the task, only a few address them within their experimental setup.

In this paper, we address the shortcomings of previous work and make the following contributions: (1) To the best of our knowledge, this paper presents the first study of historical language associated with entities at the intersections of two axes of oppression: race and gender. We study biases associated with identified entities on a word level, and to this end, employ distributional models and analyse semantics extracted from word embeddings trained on our historical corpora. (2) We conduct a temporal case study on historical newspapers from the Caribbean in the colonial period between 1770–1870. During this time, the region suffered both the consequences of European wars and political turmoil, as well as several uprisings of the local enslaved populations, which had a significant impact on the Caribbean social relationships and cultures (Migge and Muehleisen, 2010). (3) To address the challenges of analysing historical documents, we probe the applied methods for their stability and ability to comprehend the noisy, archaic corpora.

We find that there is a trade-off between the stability of word embeddings and their compatibility with the historical dataset. Further, our temporal analysis connects changes in biased word associations to historical shifts taking place in the period. For instance, we couple the high association between *Caribbean countries* and “manual labour” prevalent mostly in the earlier time periods to waves of white labour migrants coming to the Caribbean from 1750 onward. Finally, we provide evidence supporting the intersectionality theory by observing conventional manifestations of gender bias solely for white people. While unsurprising, this finding necessitates intersectional bias analysis for historical documents.

2 Related Work

Intersectional Biases. Most prior work has analysed bias along one axis, e.g. race or gender, but not both simultaneously (Field et al., 2021;

Staćzak and Augenstein, 2021). There, research on racial biases is generally centred around the gender majority group, such as Black men, while research on gender bias emphasises the experience of individuals who hold racial privilege, such as white women. Therefore, discrimination towards people with multiple minority identities, such as Black women, remains understudied. Addressing this, the intersectionality framework (Crenshaw, 1989) investigates how different forms of inequality, e.g. gender and race, intersect with and reinforce each other. Drawing on this framework, Tan and Celis (2019a); May et al. (2019); Lepori (2020); Maronikolakis et al. (2022); Guo and Caliskan (2021) analyse the compounding effects of race and gender encoded in contextualised word representations and downstream tasks. Recently, Lalor et al. (2022); Jiang and Fellbaum (2020) show the harmful implications of intersectionality effects in pre-trained language models. Less interest has been dedicated to unveiling intersectional biases prevalent in natural language, with a notable exception of Kim et al. (2020) which provide evidence on intersectional bias in datasets of hate speech and abusive language on social media. As far as we know, this is the first paper on intersectional biases in historical documents.

Bias in Historical Documents. Historical corpora have been employed to study societal phenomena such as language change (Kutuzov et al., 2018; Hamilton et al., 2016) and societal biases. Gender bias has been analysed in biomedical research over a span of 60 years (Rios et al., 2020), in English-language books published between 1520 and 2008 (Hoyle et al., 2019), and in Dutch newspapers from the second half of the 20th century (Wevers, 2019). Levis Sullam et al. (2022) investigate the evolution of the discourse on Jews in France during the 19th century. Garg et al. (2018) study the temporal change in stereotypes and attitudes toward women and ethnic minorities in the 20th and 21st centuries in the US. However, they neglect the emergent intersectionality bias.

When analysing the transformations of biases in historical texts, researchers rely on conventional tools developed for modern language. However, historical texts can be viewed as a separate domain due to their unique challenges of small and idiosyncratic corpora and noisy, archaic text (Piotrowski, 2012). Prior work has attempted to overcome the challenges such documents pose for mod-

Source	#Files	#Sentences
Caribbean Project	7 487	5 224 591
Danish Royal Library	5 661	657 618
Total	13 148	5 882 209

Table 1: Statistics of the newspapers dataset.

Period	Decade	#Issues	Total
International conflicts and slave rebellions	1710–1770	15	1 886
	1770s	747	
	1780s	283	
	1790s	841	
Revolutions and nation building	1800s	604	3 790
	1810s	1 347	
	1820s	1 839	
Abolishment of slavery	1830s	1 838	7 453
	1840s	1 197	
	1850s	1 111	
	1860s	1 521	
	1870s	1 786	

Table 2: Total number of articles in each period and decade.

ern tools, including recognition of spelling variations (Bollmann, 2019) and misspelt words (Boros et al., 2020), and ensuring the stability of the applied methods (Antoniak and Mimno, 2018).

We study the dynamics of intersectional biases and their manifestations in language while addressing the challenges of historical data.

3 Datasets

Newspapers are considered an excellent source for the study of societal phenomena since they function as transceivers – both producing and demonstrating public discourse (Wevers, 2019). As part of this study, we collect newspapers written in English from the “Caribbean Newspapers, 1718–1876” database,¹ the largest collection of Caribbean newspapers from the 18th–19th century available online. We extend this dataset with English-Danish newspapers published between 1770–1850 in the Danish colony of Santa Cruz (Saint Croix) downloaded from Danish Royal Library’s website.² See Tab 1 and Fig 8 (in App A.1) for details.

As mentioned in §1, the Caribbean islands experienced significant changes and turmoils during the 18th–19th century. Although chronologies

¹<https://www.readex.com/products/caribbean-newspapers-series-1-1718-1876-american-antiquarian-society>

²<https://www2.statsbiblioteket.dk/mediestream/>

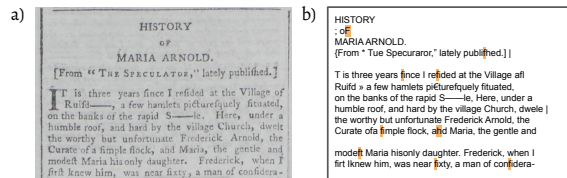


Figure 2: An example of a scanned newspaper (a) and the output of the OCR tool Tesseract (b). We fix simple OCR errors (highlighted) using a rule-based approach.

can change from island to island, key moments in Caribbean history can be divided into roughly four periods (Higman, 2021; Heuman, 2018): 1) colonial trade and plantation system (1718 to 1750); 2) international conflicts and slave rebellions (1751 to 1790); 3) revolutions and nation building (1791 to 1825); 4) end of slavery and decline of European dominance (1826 to 1876). In our experimental setup, we conduct a temporal study on data split into these periods (see Tab 2 for the number of articles in each period). As the resulting number of newspapers for the first period is very small (< 10), we focus on the three latter periods.

Data Preprocessing. Starting with scans of entire newspaper issues (Fig 2.a), we first OCR them using the popular software Tesseract³ with default parameters and settings. We then clean the dataset by applying the DataMunging package,⁴ which uses a simple rule-based approach to fix basic OCR errors (e.g. long s’ being OCRed as f’, (Fig 2.b)). As some of the newspapers downloaded from the Danish royal library contain Danish text, we use spaCy⁵ to tokenise the OCRed newspapers into sentences and the python package langdetect⁶ to filter out non-English sentences.

4 Bias and its Measures

Biases can manifest themselves in natural language in many ways (see the surveys by Stańczak and Augenstein (2021); Field et al. (2021); Lalor et al. (2022)). In the following, we state the definition of bias we follow and describe the measures we use to quantify it.

³<https://github.com/tesseract-ocr/tesseract>

⁴<https://github.com/tedunderwood/DataMunging>

⁵<https://spacy.io/>

⁶<https://github.com/Mimino666/langdetect>

4.1 Definition

Language is known to reflect common perceptions of the world (Hitti et al., 2019) and differences in its usage have been shown to reflect societal biases (Hoyle et al., 2019; Marjanovic et al., 2022). In this paper, we define bias in a text as the use of words or syntactic constructs that connote or imply an inclination or prejudice against a certain sensitive group, following the bias definition as in Hitti et al. (2019). To quantify bias under this definition, we analyse word embeddings trained on our historical corpora. These representations are assumed to carry lexical semantic meaning signals from the data and encode information about language usage in the proximity of entities. However, even words that are not used as direct descriptors of an entity influence its embedding, and thus its learnt meaning. Therefore, we further conduct an analysis focusing exclusively on words that describe identified entities.

4.2 Measures

WEAT The Word Embedding Association Test (Caliskan et al., 2017) is arguably the most popular benchmark to assess bias in word embeddings and has been adapted in numerous research (May et al., 2019; Rios et al., 2020). WEAT employs cosine similarity to measure the association between two sets of attribute words and two sets of target concepts. Here, the attribute words relate to a sensitive attribute (e.g. male and female), whereas the target concepts are composed of words in a category of a specific domain of bias (e.g. career- and family-related words). For instance, the WEAT statistic informs us whether the learned embeddings representing the concept of *family* are more associated with females compared to males. According to Caliskan et al. (2017), the differential association between two sets of target concept embeddings, denoted X and Y , with two sets of attribute embeddings, denoted as A and B , can be calculated as:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

where $s(w, A, B)$ measures the embedding association between one target word w and each of the sensitive attributes:

$$s(w, A, B) = \text{mean}[\cos(w, a)] - \text{mean}[\cos(w, b)]$$

The resulting effect size is then a normalised measure of association:

$$d = \frac{\text{mean}_{x \in X}[s(x, A, B)] - \text{mean}_{y \in Y}[s(y, A, B)]}{\text{std}_{w \in X \cup Y}[s(w, A, B)]}$$

As a result, larger effect sizes imply a more biased word embedding. Furthermore, concept-related words should be equally associated with either sensitive attribute group assuming an unbiased word embedding.

PMI We use point-wise mutual information (PMI; Church and Hanks 1990) as a measure of association between a descriptive word and a sensitive attribute (gender or race). In particular, PMI measures the difference between the probability of the co-occurrence of a word and an attribute, and their joint probability if they were independent as:

$$\text{PMI}(a, w) = \log \frac{p(a, w)}{p(a)p(w)} \quad (1)$$

A strong association with a specific gender or race leads to a high PMI. For example, a high value for $\text{PMI}(\textit{female}, \textit{wife})$ is expected due to their co-occurrence probability being higher than the independent probabilities of *female* and *wife*. Accordingly, in an ideal unbiased world, words such as *honourable* would have a PMI of approximately zero for all gender and racial identities.

5 Experimental Setup

We perform two sets of experiments on our historical newspaper corpus. First, before we employ word embeddings to measure bias, we investigate the stability of the word embeddings trained on our dataset and evaluate their understanding of the noisy nature of the corpora. Second, we assess gender and racial biases using tools defined in §4.2.

5.1 Embedding Stability Evaluation

We use word embeddings as a tool to quantify historical trends and word associations in our data. However, prior work has called attention to the lack of stability of word embeddings trained on small and potentially idiosyncratic corpora (Antoniak and Mimno, 2018; Gonen et al., 2020). We compare these different embeddings setups by testing them with regard to their stability and capturing meaning while controlling for the tokenisation algorithm, embedding size and the minimum number of occurrences.

We construct the word embeddings employing the continuous skip-gram negative sampling model from Word2vec (Mikolov et al., 2013b) using `gensim`.⁷ Following prior work (Antoniak and Mimno, 2018; Gonen et al., 2020), we test two common vector dimension sizes of 100 and 300, and two minimum numbers of occurrences of 20 and 100. The rest of the hyperparameters are set to their default value. We use two different methods for tokenising documents, the `spaCy` tokeniser and a subword-based tokeniser, Byte-Pair Encoding (BPE, Gage (1994)). We train the BPE tokeniser on our dataset using the Hugging Face tokeniser implementation.⁸

For each word in the vocabulary, we identify its 20 nearest neighbours and calculate the Jaccard similarity across five algorithm runs. Next, we test how well the word embeddings deal with the noisy nature of our documents. We create a list of 110 frequently misspelt words (See App A.2). We construct the list by first tokenising our dataset using `spaCy` and filtering out proper nouns and tokens that appear in the English dictionary. We then order the remaining tokens by frequency and manually scan the top 1 000 tokens for misspelt words. We calculate the percentage of words (averaged across 5 runs) for which the misspelt word is in immediate proximity to the correct word (top 5 nearest neighbours in terms of cosine similarity).

Based on the results of the stability and compatibility study, we select the most suitable model with which we conduct the following bias evaluation.

5.2 Bias Estimation

5.2.1 WEAT Evaluation

As discussed in §4.2, WEAT is used to evaluate how two attributes are associated with two target concepts in an embedding space, here of the model that was selected by the method described in §5.1.

In this work, we focus on the attribute pairs (*female, male*)⁹ and (*white, non-white*). Usually, comparing the sensitive attributes (*white, non-white*) is done by collecting the embedding of popular white names and popular non-white names (Tan and Celis, 2019b). However, this approach can introduce noise when applied to our dataset (Handler

and Jacoby, 1996). First, non-whites are less likely to be mentioned by name in historical newspapers compared to whites. Second, popular non-white names of the 18th and 19th centuries differ substantially from popular non-white names of modern times, and, to the best of our knowledge, there is no list of common historical non-white names. For these reasons, instead of comparing the pair (*white, non-white*), we compare the pairs (*African countries, European countries*) and (*Caribbean countries, European countries*).

Following Rios et al. (2020), we analyse the association of the above-mentioned attributes to the target concepts (*career, family*), (*strong, weak*), (*intelligence, appearance*), and (*physical illness, mental illness*). Following a consultation with a historian, we add further target concepts relevant to this period (*manual labour, non-manual labour*) and (*crime, lawfulness*). Tab 6 (in App A.3) lists the target and attribute words we use for our analysis.

We also train a separate word embedding model on each of the dataset splits defined in §3 and run WEAT on the resulting three models. Comparing the obtained WEAT scores allows us to visualise temporal changes in the bias associated with the attributes and understand its dynamics.

5.2.2 PMI Evaluation

Different from WEAT, calculating PMI requires first identifying entities in the OCRed historical newspapers and then classifying them into pre-defined attribute groups. The next step is collecting descriptors, i.e. words that are used to describe the entities. Finally, we use PMI to measure the association strength of the collected descriptors with each attribute group.

Entity Extraction. We apply `F-coref` (Otmazgin et al., 2022), a model for English coreference resolution that simultaneously performs entity extraction and coreference resolution on the extracted entities. The model’s output is a set of entities, each represented as a list of all the references to that entity in the text. We filter out non-human entities by using `nltk`’s WordNet package,¹⁰ retaining only entities for which the synset “`person.n1`” is a hypernym of one of their references.

Entity Classification. We use a keyword-based approach (Lepori, 2020) to classify the entities into groups corresponding to the gender and race axes

⁷<https://radimrehurek.com/gensim/models/word2vec.html>

⁸<https://huggingface.co/docs/tokenizers>

⁹As we deal with historical documents from the 18th–19th centuries, other genders are unlikely to be found in the data.

¹⁰<https://www.nltk.org/howto/wordnet.html>

#Entities	#Males	#Females	#Non-whites	#Non-white males	#Non-white females
601 468	387 292	78 821	8 525	4 543	1 548

Table 3: The entities in our Caribbean newspapers dataset. Notice that #males and #females do not sum to #entities as some entities could not be classified. Similarly, #non-white males and #non-white females do not sum to #non-whites.

and their intersection. Specifically, we classify each entity as being a member of *male vs female*, and *white vs non-white*. Additionally, entities are classified into intersectional groups (e.g. we classify an entity into the group *non-white females* if it belongs to both *female* and *non-white*).

Formally, we classify an entity e with references $\{r_e^1, \dots, r_e^m\}$ to attribute group G with keyword-set $K_G = \{k_1, \dots, k_n\}$ if $\exists i$ such that $r_e^i \in K_G$. See App A.3 for listing the keyword sets of the different groups. In Tab 3, we present the number of entities classified into each group. We note here the unbalanced representation of the groups in the dataset. Further, it is important to state, that because it is highly unlikely that an entity in our dataset would be explicitly described as white, we classify an entity into the *whites* group if it was not classified as *non-white*. See the [Limitations](#) section for a discussion of the limitations of using a keyword-based classification approach.

To evaluate our classification scheme, an author of this paper manually labelled a random sample of 56 entities. The keyword-based approach assigned the correct gender and race label for $\sim 80\%$ of the entities. See additional details in Tab 7 in App B. From a preliminary inspection, it appears that many of the entities that were wrongly classified as *female* were actually ships or other vessels (traditionally “ship” has been referred to using female gender). As F-coref was developed and trained using modern corpora, we evaluate its accuracy on the same set of 56 entities. Two authors of this paper validated its performance on the historical data to be satisfactory, with especially impressive results on shorter texts with fewer amount of OCR errors.

Descriptors Collection. Finally, we use spaCy to collect descriptors for each classified entity. Here, we define the descriptors as the lemmatised form of tokens that share a dependency arc labelled “amod” (i.e. adjectives that describe the tokens) to one of the entity’s references. Every target group G_j is then assigned with descriptors list $D_j = [d_1, \dots, d_k]$.

To calculate PMI according to Eq (1), we estimate the joint distribution of a target group and a descriptor using a simple plug-in estimator:

$$\hat{p}(G_j, d_i) \propto \text{count}(G_j, d_i) \quad (2)$$

Now, we can assign every word d_i two continuous values representing its bias in the gender and race dimensions by calculating $\text{PMI}(\text{female}, d_i) - \text{PMI}(\text{males}, d_i)$ and $\text{PMI}(\text{non-white}, d_i) - \text{PMI}(\text{white}, d_i)$. These two continuous values can be seen as d_i ’s coordinates on the intersectional gender/race plane.

Tokenisation	Embedding Size	Min Freq	Mean JS Top 20	Correct Word in Top 5 (all words)	% Misspelling in vocabulary
BPE	100	20	0.66	37.04	94.44
	100	100	0.66	37.04	94.44
	300	20	0.63	40.74	94.44
	300	100	0.64	39.81	94.44
SpaCy	100	20	0.59	63.89	74.07
	100	100	0.65	48.15	56.48
	300	20	0.55	63.89	74.07
	300	100	0.61	50.00	56.48

Table 4: Results of the stability analysis of different word embedding methods (measured with Jaccard similarity) and their compatibility with the historical corpora (ability to recognise misspelt words).

5.2.3 Lexicon Evaluation

Another popular approach for quantifying different aspects of bias is the application of specialised lexica (Staćzak and Augenstein, 2021). These lexica assign words a continuous value that represents how well the word aligns with a specific dimension of bias. We use NRC-VAD lexicon (Mohammad, 2018) to compare word usage associated with the sensitive attributes *race* and *gender* in three dimensions: *dominance* (strength/weakness), *valence* (goodness/badness), and *arousal* (activeness/passiveness of an identity). Specifically, given a bias dimension \mathcal{B} with lexicon $L_{\mathcal{B}} =$

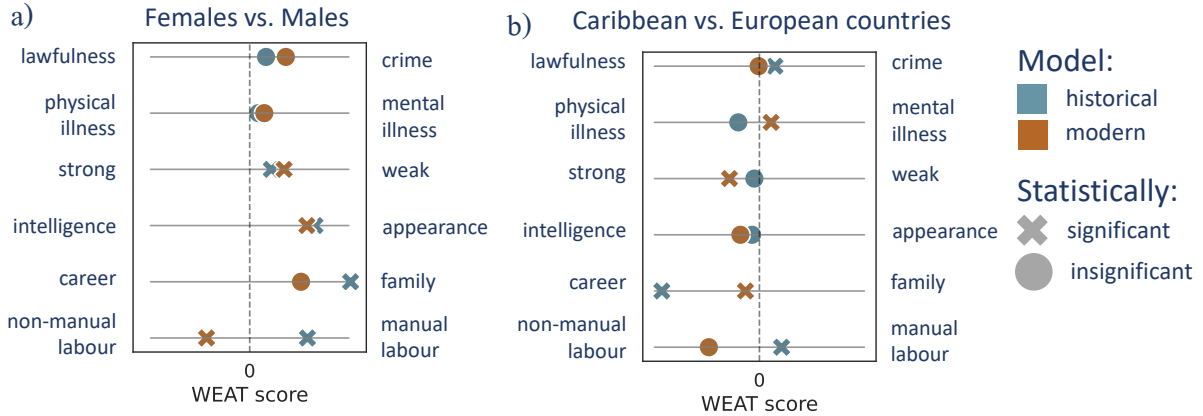


Figure 3: a) WEAT results of *females vs males*. The location of a marker measures the association strength of *females* with the concept (compared to *males*). For example, according to the modern model, *females* are associated with “weak” and *non-manual labour* while *males* are associated with “strong” and *manual labour*. b) WEAT results of *Caribbean countries vs European countries*. The location of a marker measures the association strength of *Caribbean countries* with the concept (compared to *European countries*).

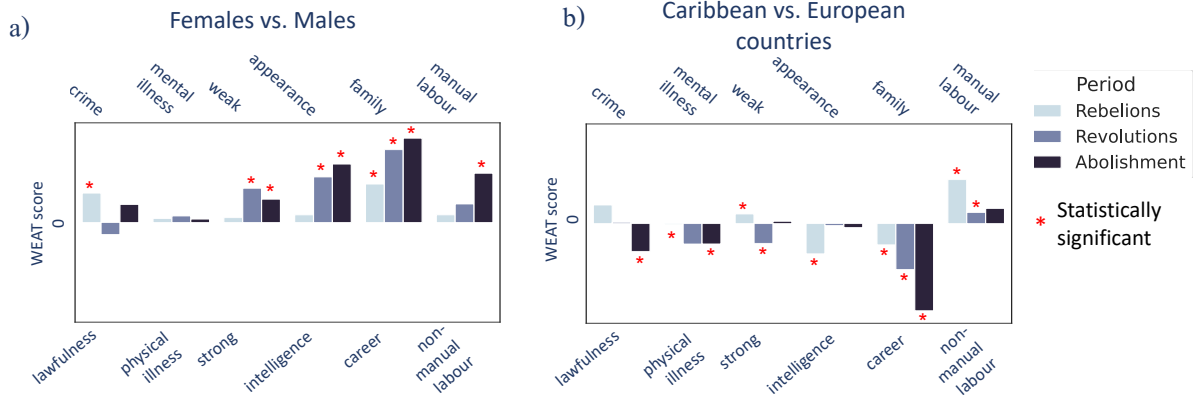


Figure 4: Temporal WEAT analysis conducted for the periods 1751–1790 (rebellions), 1791–1825 (revolutions) and 1826–1876 (abolishment). Similar to Fig 3, the height of each bar represents how strong the association of the attribute is with each concept.

$\{(w_1, a_1), \dots, (w_n, a_n)\}$, where (w_i, a_i) are word-value pairs, we calculate the association of \mathcal{B} with a sensitive attribute G_j using:

$$A(\mathcal{B}, G_j) = \frac{\sum_i^n a_i \cdot \text{count}(w_i, D_j)}{\sum_i^n \text{count}(w_i, D_j)} \quad (3)$$

where $\text{count}(w_i, D_j)$ is the number of times the word w_i appears in the descriptors list D_j .

6 Results

First, we investigate which training strategies of word embeddings optimise their stability and compatibility on historical corpora (§6.1). Next, we analyse how bias is manifested along the gender and racial axes and whether there are any notice-

able differences in bias across different periods of the Caribbean history (§6.2).

6.1 Embedding Stability Evaluation

In Tab 4, we present the results of the study on the influence of training strategies of word embeddings. We find that there is a trade-off between the stability of word embeddings and their compatibility with the dataset. While BPE achieves a higher Jaccard similarity across the top 20 nearest neighbours for each word across all runs, it loses the meaning of misspelt words. Interestingly, this phenomenon arises, despite the misspelt words occurring frequently enough to be included in the BPE model’s vocabulary.

For the remainder of the experiments, we aim to select a model which effectively manages this

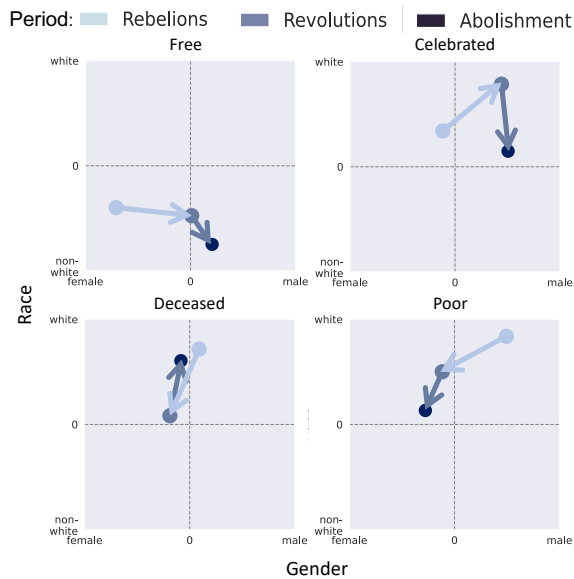


Figure 5: Intersectional PMI analysis of “free”, “celebrated”, “deceased” and “poor” across the periods.

trade-off achieving both high stability and captures meaning despite the noisy nature of the underlying data. Thus, we opt to use a `spacy`-based embedding with a minimum number of occurrences of 20 and an embedding size of 100 which achieves competitive results in both of these aspects. Finally, we note that our results remain stable across different algorithm runs and do not suffer from substantial variations which corroborates the reliability of the findings we make henceforth.

6.2 Bias Estimation

6.2.1 WEAT Analysis

Fig 3 displays the results of performing a WEAT analysis for measuring the association of the six targets described in §5.2 with the attributes (*females*, *males*) and (*Caribbean countries*, *European countries*), respectively.¹¹ We calculate the WEAT score using the embedding model from §6.1 and compare it with an embedding model trained on modern news corpora (`word2vec-google-news-300`, Mikolov et al. (2013a)). We notice interesting differences between the historical and modern embeddings. For example, while in our dataset *females* are associated with the target concept of *manual labour*, this notion is more aligned with *males* in the modern corpora. A likely cause is that during this period, womens’ intellectual and

¹¹See Fig 9 in App B for analysis of the attributes (*African countries*, *European countries*).

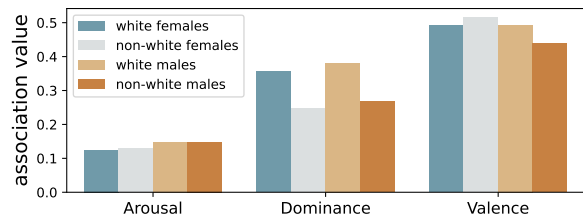


Figure 6: Association of attributes with the lexicon of dominance, valence, and arousal.

administrative work was not commonly recognised (Wayne, 2020). It is also interesting to note that the attribute *Caribbean countries* has a much stronger association in the historical embedding with the target *career* (as opposed to *family*) compared to the modern embeddings. A possible explanation is that Caribbean newspapers referred to locals by profession or similar titles, while Europeans were referred to as relatives of the Caribbean population.

In Fig 4 and Fig 10 (in App B), we present a dynamic WEAT analysis that unveils trends on a temporal axis. In particular, we see an increase in the magnitude of association between the target of *family vs career* and the attributes (*females*, *males*) and (*Caribbean countries*, *European countries*) over time. It is especially interesting to compare Fig 3 with Fig 4. One intriguing result is that the high association between *Caribbean countries* and *manual labour* can be attributed to the earlier periods. This finding is potentially related to several historical shifts taking place in the period. For instance, while in the earlier years, it was normal for plantation owners to be absentees and continue to live in Europe, from 1750 onward, waves of white migrants with varied professional backgrounds came to the Caribbean.

6.2.2 PMI Analysis

We report the results of the intersectional PMI analysis in Fig 1. As can be seen, an intersectional analysis can shed a unique light on the biased nature of some words in a way that single-dimensional analysis cannot. *White males* are “brave” and “ingenious”, and *non-white males* are described as “active” and “tall”. Interestingly, while words such as “pretty” and “beautiful” (and peculiarly, “murdered”) are biased towards *white* as opposed to *non-white females*, the word “lovely” is not, whereas “elderly” is strongly aligned with *non-white females*. Another intriguing dichotomy is the word pair “sick” and “blind” which are both independent

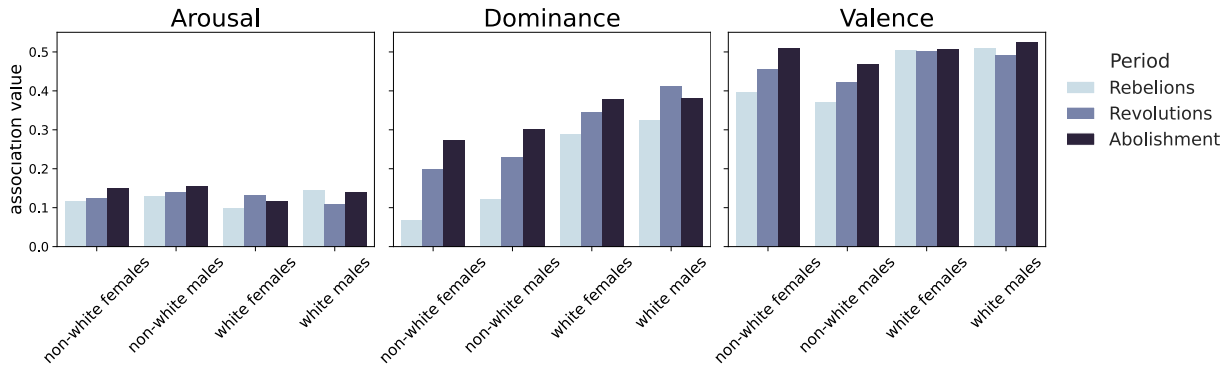


Figure 7: Association of attributes with the lexicon of dominance, valence, and value done on the periods 1751–1790 (rebellions), 1791–1825 (revolutions) and 1826–1876 (abolishment).

along the gender axis but manifest a polar racial bias. In Tab 8 in App B, we list some examples from our dataset featuring those words.

Similarly to §6.2.1, we perform a temporal PMI analysis by comparing results obtained from separately analysing the three dataset splits. In Fig 5, we follow the trajectory over time of the biased words “free”, “celebrated”, “deceased” and “poor”. Each word displays different temporal dynamics. For example, while the word “free” moved towards the *male* attribute, “poor” transitioned to become more associated with the attributes *female* and *non-white* over time (potentially due to its meaning change from an association with poverty to a pity).

These results provide evidence for the claims of the intersectionality theory. We observe conventional manifestations of gender bias, i.e. “beautiful” and “pretty” for *white females*, and “ingenious” and “brave” for *white males*. While unsurprising due to the societal status of non-white people in that period, this finding necessitates intersectional bias analysis for historical documents in particular.

6.2.3 Lexicon Evaluation

Finally, we report the lexicon-based evaluation results in Fig 6 and Fig 7. Unsurprisingly, we observe lower dominance levels for the *non-white* and *female* attributes compared to *white* and *male*, a finding previously uncovered in modern texts (Field and Tsvetkov, 2019; Rabinovich et al., 2020). While Fig 7 indicates that the level of dominance associated with these attributes raised over time, a noticeable disparity to white males remains. Perhaps more surprising is the valence dimension. We see the highest and lowest levels of associations with the intersectional attributes *non-white female* and *non-white male*, respectively. We hypothesise that this connects to the nature of advertisements

for lending the services of or selling non-white women where being agreeable is a valuable asset.

7 Conclusions

In this paper, we examine biases present in historical newspapers published in the Caribbean during the colonial era by conducting a temporal analysis of biases along the axes of gender, race, and their intersection. We evaluate the effectiveness of different embedding strategies and find a trade-off between the stability and compatibility of word representations on historical data. We link changes in biased word usage to historical shifts, coupling the development of the association between *manual labour* and *Caribbean countries* to waves of white labour migrants coming to the Caribbean from 1750 onward. Finally, we provide evidence to corroborate the intersectionality theory by observing conventional manifestations of gender bias solely for white people.

Limitations

We see several limitations regarding our work. First, we focus on documents in the English language only, neglecting many Caribbean newspapers and islands with other official languages. While some of our methods can be easily extended to non-English material (e.g. WEAT analysis), methods that rely on the pre-trained English model F-coref (i.e. PMI, lexicon-based analysis) can not.

On the same note, F-coref and spaCy were developed and trained using modern corpora, and their capabilities when applied to the noisy historical newspapers dataset, are noticeably lower compared to modern texts. Contributing to this issue is the unique, sometimes archaic language in

which the newspapers were written. While we validate F-coref performance on a random sample (§5.2), this is a significant limitation of our work. Similarly, increased attention is required to adapt the keyword sets used by our methods to historical settings.

Moreover, our historical newspaper dataset is inherently imbalanced and skewed. As can be seen in Tab 2 and Fig 8, there is an over-representation of a handful of specific islands and time periods. While it is likely that in different regions and periods, less source material survived to modern times, part of the imbalance (e.g. the prevalence of the US Virgin Islands) can also be attributed to current research funding and policies.¹² Compounding this further, minority groups are traditionally under-represented in news sources. This introduces noise and imbalance into our results, which rely on a large amount of textual material referring to each attribute on the gender/race plane that we analyse.

Relating to that, our keyword-based method of classifying entities into groups corresponding to the gender and race axes is limited. While we devise a specialised keyword set targeting the attributes *female*, *male* and *non-white*, we classify an entity into the *white* group if it was not classified as *non-white*. This discrepancy is likely to introduce noise into our evaluation, as can also be observed in Tab 7. This tendency may be intensified by the NLP systems that we use, as many tend to perform worse on gender- and race-minority groups (Field et al., 2021).

Finally, in this work, we explore intersectional bias only along the race and gender axes. Thus, we neglect the effects of other confounding factors (e.g. societal position, occupation) that affect asymmetries in language.

Ethical Considerations

Studying historical texts from the era of colonisation and slavery poses ethical issues to historians and computer scientists alike since vulnerable groups still suffer the consequences of this history in the present. Indeed, racist and sexist language is not only a historical artefact of bygone days but has a real impact on people’s lives (Alim et al., 2020).

We note that the newspapers we consider for this analysis were written foremost by the European

¹²The Danish government has recently funded a campaign for the digitisation of historical newspapers published in the Danish colonies; <https://stcroixsource.com/2017/03/01/>.

oppressors. Moreover, only a limited number of affluent people (white males) could afford to place advertisements in those newspapers (which constitute a large portion of the raw material). This skews our study toward language used by privileged individuals and their perceptions.

This work aims to investigate racial and gender biases, as well as their intersection. Both race and gender are considered social constructs and can encompass a range of perspectives, including one’s reflected, observed, or self-perceived identity. In this paper, we classify entities as observed by the author of an article and infer their gender and race based on the pronouns and descriptors used in relation to this entity. We follow this approach in an absence of explicit demographic information. However, we warn that this method poses a risk of misclassification. Although the people referred to in the newspapers are no longer among the living, we should be considerate when conducting studies addressing vulnerable groups.

Finally, we use the mutually exclusive *white* and *non-white* race categories as well as *male* and *female* gender categories. We acknowledge that these groupings do not fully capture the nuanced nature of bias. This decision was made due to limited data discussing minorities in our corpus. While gender identities beyond the binary are unlikely to be found in the historical newspapers from the 18th-19th century, future work will aim to explore a wider range of racial identities.

Acknowledgements

This work is funded by Independent Research Fund Denmark under grant agreement number 9130-00092B, as well as the Danish National Research Foundation (DNRF 138). Isabelle Augenstein is further supported by the Pioneer Centre for AI, DNRF grant number P1.

References

- H. Samy Alim, Angela Reyes, and Paul V. Kroskrity, editors. 2020. *The Oxford Handbook of Language and Race*. Oxford University Press.
- Maria Antoniak and David Mimno. 2018. [Evaluating the stability of embedding-based word similarities](#). *Transactions of the Association for Computational Linguistics*, 6:107–119.
- Marcel Bollmann. 2019. [A large-scale comparison of historical text normalization systems](#). 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 3885–3898, Minneapolis, Minnesota. Association for Computational Linguistics.
- Nadav Borenstein, Natalia da Silva Perez, and Isabelle Augenstein. 2023. **Multilingual event extraction from historical newspaper adverts**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada. Association for Computational Linguistics.
- Emanuela Boros, Ahmed Hamdi, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose G. Moreno, Nicolas Sidere, and Antoine Doucet. 2020. **Alleviating digitization errors in named entity recognition for historical documents**. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 431–441, Online. Association for Computational Linguistics.
- Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. **Semantics derived automatically from language corpora contain human-like biases**. *Science*, 356(6334):183–186.
- Kenneth Ward Church and Patrick Hanks. 1990. **Word association norms, mutual information, and lexicography**. *Computational Linguistics*, 16(1):22–29.
- Kimberle Crenshaw. 1989. **Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics**. *The University of Chicago Legal Forum*, 140:139–167.
- Kimberlé Crenshaw. 1995. **Mapping the Margins: Intersectionality, Identity Politics, and Violence Against Women of Color**. In *Critical race theory: the key writings that formed the movement*. New Press, New York.
- Maud Ehrmann, Matteo Romanello, Simon Clematide, Phillip Benjamin Ströbel, and Raphaël Barman. 2020. **Language resources for historical newspapers: the impresso collection**. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 958–968, Marseille, France. European Language Resources Association.
- Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. **A survey of race, racism, and anti-racism in NLP**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925, Online. Association for Computational Linguistics.
- Anjalie Field and Yulia Tsvetkov. 2019. **Entity-centric contextual affective analysis**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2550–2560, Florence, Italy. Association for Computational Linguistics.
- Philip Gage. 1994. **A new algorithm for data compression**. *C Users Journal*, 12(2):23–38.
- Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. **Word embeddings quantify 100 years of gender and ethnic stereotypes**. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.
- Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. 2020. **Simple, interpretable and stable method for detecting words with usage change across corpora**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555, Online. Association for Computational Linguistics.
- Wei Guo and Aylin Caliskan. 2021. **Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases**. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 122–133, New York, NY, USA. Association for Computing Machinery.
- William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. **Diachronic word embeddings reveal statistical laws of semantic change**. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.
- Jerome S. Handler and JoAnn Jacoby. 1996. **Slave names and naming in barbados, 1650-1830**. *The William and Mary Quarterly*, 53(4):685–728.
- Gad Heuman. 2018. *The Caribbean: A Brief History*, 3 edition. Bloomsbury Academic, London, England.
- B. W. Higman. 2021. *A Concise History of the Caribbean*, 2 edition. Cambridge Concise Histories. Cambridge University Press.
- Yasmeen Hitti, Eunbee Jang, Ines Moreno, and Carlyne Pelletier. 2019. **Proposed taxonomy for gender bias in text; a filtering methodology for the gender generalization subtype**. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 8–17, Florence, Italy. Association for Computational Linguistics.
- Alexander Miserlis Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, Isabelle Augenstein, and Ryan Cotterell. 2019. **Unsupervised discovery of gendered language through latent-variable modeling**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1706–1716, Florence, Italy. Association for Computational Linguistics.
- May Jiang and Christiane Fellbaum. 2020. **Interdependencies of gender and race in contextualized word embeddings**. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 17–25, Barcelona, Spain (Online). Association for Computational Linguistics.

- Jae Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. 2020. [Intersectional bias in hate speech and abusive language datasets](#). *arXiv:2005.05921 [cs]*.
- Austin C. Kozlowski, Matt Taddy, and James A. Evans. 2019. [The geometry of culture: Analyzing the meanings of class through word embeddings](#). *American Sociological Review*, 84(5):905–949.
- Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Veldal. 2018. [Diachronic word embeddings and semantic shifts: a survey](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- John Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. 2022. [Benchmarking intersectional biases in NLP](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3598–3609, Seattle, United States. Association for Computational Linguistics.
- Michael Lepori. 2020. [Unequal representations: Analyzing intersectional biases in word embeddings using representational similarity analysis](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1720–1728, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Simon Levis Sullam, Giorgia Minello, Rocco Tripodi, and Massimo Warglien. 2022. [Representation of jews and anti-jewish bias in 19th century french public discourse: Distant and close reading](#). *Frontiers in Big Data*, 4.
- Elvys Linhares Pontes, Ahmed Hamdi, Nicolas Sidère, and Antoine Doucet. 2019. [Impact of OCR Quality on Named Entity Linking](#). In *International Conference on Asia-Pacific Digital Libraries 2019*, Kuala Lumpur, Malaysia.
- Sara Marjanovic, Karolina Stańczak, and Isabelle Augenstein. 2022. [Quantifying gender biases towards politicians on Reddit](#). *PLOS ONE*, 17(10):1–36.
- Antonis Maronikolakis, Philip Baader, and Hinrich Schütze. 2022. [Analyzing hate speech data along racial, gender and intersectional axes](#). In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 1–7, Seattle, Washington. Association for Computational Linguistics.
- Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. [On measuring social biases in sentence encoders](#). 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 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. [Quantitative analysis of culture using millions of digitized books](#). *Science*, 331(6014):176–182.
- Bettina M Migge and Susanne Muehleisen. 2010. [Earlier Caribbean English and Creole in Writing](#). In Raymond Hickey, editor, *Varieties in writing: The written word as linguistic evidence*, pages 223–244. John Benjamins.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. [Efficient estimation of word representations in vector space](#). *arXiv:1301.3781 [cs]*.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. [Distributed representations of words and phrases and their compositionality](#). In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS’13*, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.
- Saif Mohammad. 2018. [Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.
- Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2022. [F-coref: Fast, accurate and easy to use coreference resolution](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 48–56, Taipei, Taiwan. Association for Computational Linguistics.
- B. Keith Payne, Heidi A. Vuletic, and Jazmin L. Brown-Iannuzzi. 2019. [Historical roots of implicit bias in slavery](#). *Proceedings of the National Academy of Sciences*, 116(24):11693–11698.
- Michael Piotrowski. 2012. *Natural Language Processing for Historical Texts*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.
- Ella Rabinovich, Hila Gonen, and Suzanne Stevenson. 2020. [Pick a fight or bite your tongue: Investigation of gender differences in idiomatic language usage](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5181–5192, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Anthony Rios, Reenam Joshi, and Hejin Shin. 2020. [Quantifying 60 years of gender bias in biomedical research with word embeddings](#). In *Proceedings of*

the 19th SIGBioMed Workshop on Biomedical Language Processing, pages 1–13, Online. Association for Computational Linguistics.

Karolina Stańczak and Isabelle Augenstein. 2021. *A survey on gender bias in natural language processing*. *arXiv:2112.14168 [cs]*.

Yi Chern Tan and L. Elisa Celis. 2019a. *Assessing Social and Intersectional Biases in Contextualized Word Representations*, chapter 1. Curran Associates Inc., Red Hook, NY, USA.

Yi Chern Tan and L Elisa Celis. 2019b. *Assessing social and intersectional biases in contextualized word representations*. *Advances in Neural Information Processing Systems*, 32.

Valerie Wayne. 2020. *Women’s labour and the history of the book in early modern England*. Bloomsbury Publishing.

Melvin Wevers. 2019. *Using word embeddings to examine gender bias in Dutch newspapers, 1950-1990*. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 92–97, Florence, Italy. Association for Computational Linguistics.

A Additional Material

A.1 Dataset Statistics

In Fig 8, we present the geographical distribution of the newspapers in the curated dataset.

A.2 Misspelt Words

Here we list 110 frequently misspelt words and their correct spelling, which was used for the embedding evaluation described in Sec 5.1.

hon'ble - honorable, honble - honorable, majetty - majesty, mujesty - majesty, mojesty - majesty, houfe - house, calied - called, upen - upon, cailed - called, reeeived - received, betore - before, kaow - know, reecived - received, bope - hope, fonnd - found, dificult - difficult, qnite - quite, convi-need - convinced, satistied - satisfied, intinate - intimate, demandcd - demanded, snuccessful - successful, abie - able, impossibie - impossible, although - although, foreed - forced, giad - glad, preper - proper, understocd - understood, fuund - found, almest - almost, nore - more, atter - after, oceupied - occupied, understuod - understood, satis'y - satisfy, impofible - impossible, imppoible - impossible, inseusible - insensible, accessary - accessory, contident - confident, koown - known, receiv - receive, calied - calles, appellunt - appellant, Eniperor - emperor, auxious - anxious, ofien - often, lawiul - lawful, posstble - possible, Svanish - Spanish, fuffictent - sufficient, furcher - further, yery - very, uader - under, ayreeable - agreeable, ylad - glad, egreed - agreed, unabile - unable, giyen - given, uecessary - necessary, alrendy - already, entitied - entitled, cffered - offered, pesitive - positive, creator - creator, prefound - profound, exam-ived - examined, successiul - successful, pablic - public, propor - proper, cousiderable - considerable, lvely - lovely, fold - sold, seeond - second, huuse - house, excellen - excellent, auetion - auction, Engiand - England, peopie - people, govero-ment - government, yeurs - years, exceliency - excellency, generel - general, foliowing - following, gonal - general, preperty - property, wondertul - wonderful, o'clock - o'clock, exeellency - excellency, tollowing - following, Eugland - England, gentieman - gentleman, colontal - colonial, govern-ment - government, excelleney - excellency, gover-ament - government, Lendon - London, Bermupa - Bermuda, goverument - government, himeelf - himself, entlemen - gentlemen, sublcriber - subscriber, majelijy - majesty, Weduesday - Wednesday, o'cleck - o'clock, o'cluck - o'clock, colonics -

colonies, sngar - sugar.

A.3 Keyword Sets

Tab 5 and Tab 6 describe the various keyword sets that we used for entity classification (Section 5.2.2) and for performing the WEAT tests (Section 5.2.1).

B Supplementary Results

In Tab 7, we report the accuracy of the classified entities using the keyword-based approach. In Tab 8, we list examples of sentences from our newspaper dataset. Fig 9 presents the WEAT results of the attributes *African countries vs European countries*. Fig 10 presents temporal WEAT analysis conducted for the attributes *African countries vs European countries*.

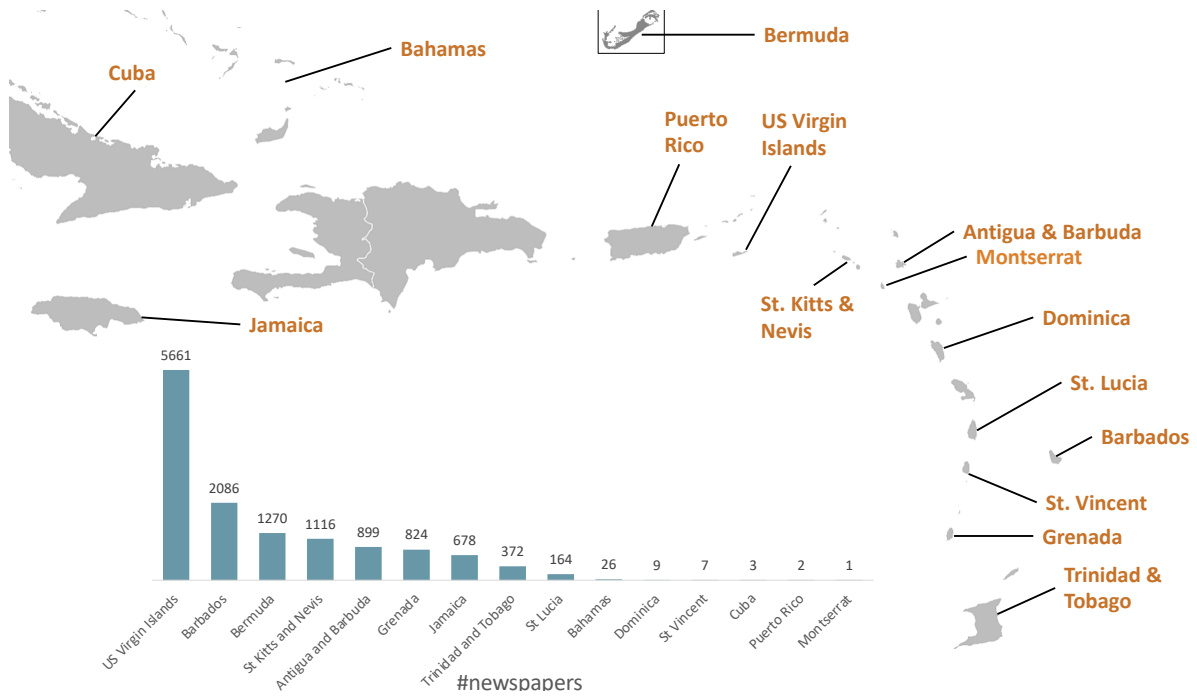


Figure 8: The geographical distribution of the curated Caribbean newspapers dataset.

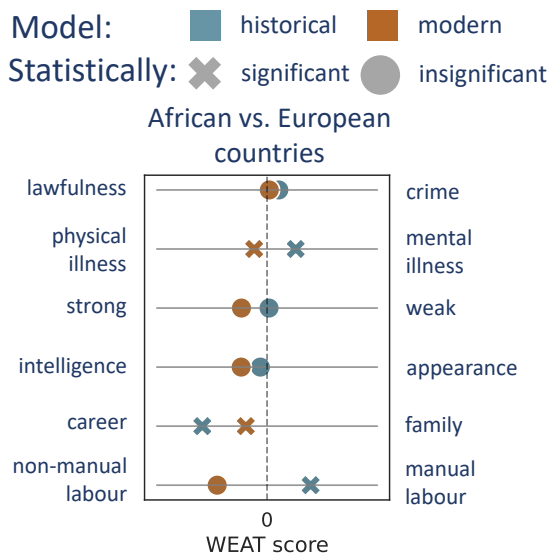


Figure 9: WEAT results of African countries vs European countries.

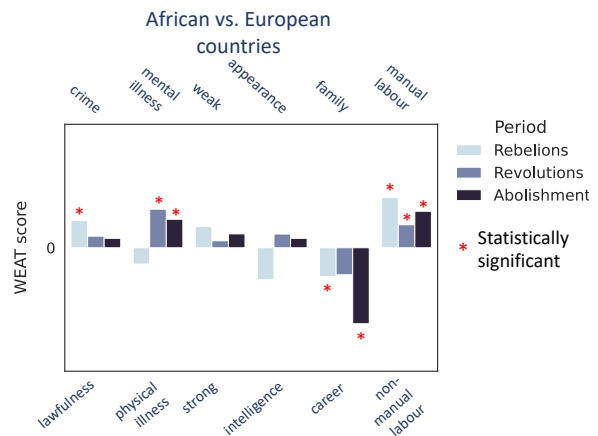


Figure 10: Temporal WEAT analysis conducted for the periods 1751–1790 (rebellions), 1791–1825 (revolutions) and 1826–1876 (abolishment). Similar to Fig 3, the height of each bar represents how strong the association of the attribute of African countries is with each concept.

Subgroup	Wordlist
Males	husband, suitor, brother, boyhood, beau, salesman, daddy, man, spokesman, chairman, lad, mister, men, sperm, dad, gelding, gentleman, boy, sir, horsemen, paternity, statesman, prince, sons, countryman, pa, suitors, stallion, fella, monks, fiance, chap, uncles, godfather, bulls, males, grandfather, penis, lions, nephew, monk, countrymen, grandsons, beards, schoolboy, councilmen, dads, fellow, colts, mr, king, father, fraternal, baritone, gentlemen, fathers, husbands, guy, semen, brotherhood, nephews, lion, lads, grandson, widower, bachelor, kings, male, son, brothers, uncle, brethren, boys, councilman, czar, beard, bull, salesmen, fraternity, dude, colt, john, he, himself, his
Females	sisters, queen, ladies, princess, witch, mother, nun, aunt, princes, housewife, women, convent, gals, witches, stepmother, wife, granddaughter, mis, widows, nieces, studs, niece, actresses, wives, sister, dowry, hens, daughters, womb, monastery, ms, misses, mama, mrs, fillies, woman, aunts, girl, actress, wench, brides, grandmother, stud, lady, female, maid, gal, queens, hostess, daughter, grandmothers, girls, heiress, moms, maids, mistress, mothers, mom, mare, filly, maternal, bride, widow, goddess, diva, maiden, hen, housewives, heroine, nuns, females', she, herself, hers, her
Non-whites	negro, negros, creole, indian, negroes, colored, mulatto, mulattos, negresse, munding, brown, browns, african, congo, black, blacks, dark, creoles
Whites	(any entity that was not classified as Non-white)

Table 5: Keywords used for classification entities into subgroups.

Attribute	Wordlist
Males	husband, man, mister, gentleman, boy, sir, prince, countryman, fiance, godfather, grandfather, nephew, fellow, mr, king, father, guy, grandson, widower, bachelor, male, son, brother, uncle, brethren
Females	sister, queen, lady, witch, mother, aunt, princes, housewife, stepmother, wife, granddaughter, mis, niece, ms, misses, mrs, woman, girl, wench, bride, grandmother, female, maid, daughter, mistress, bride, widow, maiden
European countries	ireland, georgia, france, monaco, poland, cyprus, greece, hungary, norway, portugal, belgium, luxembourg, finland, albania, germany, netherlands, montenegro, scotland, spain, europe, russia, vatican, switzerland, lithuania, bulgaria, wales, ukraine, romania, denmark, england, italy, bosnia, turkey, malta, iceland, austria, croatia, sweden, macedonia
African countries	liberia, mozambique, gambia, ghana, morocco, chad, senegal, togo, algeria, egypt, benin, ethiopia, niger, madagascar, guinea, mauritius, africa, mali, congo, angola
Caribbean countries	barbuda, bahamas, jamaica, dominica, haiti, antigua, grenada, caribbean, barbados, cuba, trinidad, dominican, nevis, kitts, lucia, croix, tobago, grenadines, puerto, rico
Target	Wordlist
Appearance	apt, discerning, judicious, imaginative, inquiring, intelligent, inquisitive, wise, shrewd, logical, astute, intuitive, precocious, analytical, smart, ingenious, reflective, inventive, venerable, genius, brilliant, clever, thoughtful
Intelligence	bald, strong, muscular, thin, voluptuous, blushing, athletic, gorgeous, handsome, homely, feeble, fashionable, attractive, weak, plump, ugly, slim, stout, pretty, fat, sensual, beautiful, healthy, alluring, slender
Weak	failure, loser, weak, timid, withdraw, follow, fragile, afraid, weakness, shy, lose, surrender, vulnerable, yield
Strong	strong, potent, succeed, loud, assert, leader, winner, dominant, command, confident, power, triumph, shout, bold
Family	loved, sisters, mother, reunited, estranged, aunt, relatives, grandchildren, godmother, kin, grandsons, sons, son, parents, stepmother, childless, paramour, nieces, children, niece, father, twins, sister, fiance, daughters, youngest, uncle, uncles, aunts, eldest, cousins, grandmother, children, loving, daughter, paternal, girls, nephews, friends, mothers, grandfather, cousin, maternal, married, nephew, wedding, grandson
Career	branch, managers, usurping, subsidiary, engineering, performs, fiscal, personnel, duties, offices, clerical, engineer, executive, functions, revenues, entity, competitive, competitor, employing, chairman, director, commissions, audit, promotion, professional, assistant, company, auditors, oversight, departments, comptroller, president, manager, operations, marketing, directors, shareholder, engineers, corporate, salaries, internal, management, salaried, corporation, revenue, salary, usurpation, managing, delegated, operating
Manual labour	sailor, bricklayer, server, butcher, gardener, cook, repairer, maid, guard, farmer, fisher, carpenter, paver, cleaner, cabinetmaker, barber, breeder, washer, miner, builder, baker, fisherman, plumber, labourer, servant
Non-manual labour	teacher, judge, manager, lawyer, director, mathematician, physician, medic, designer, bookkeeper, nurse, librarian, doctor, educator, auditor, clerk, midwife, translator, inspector, surgeon
Mental illness	sleep, pica, disorders, nightmare, personality, histrionic, stress, dependence, anxiety, terror, emotional, delusion, depression, panic, abuse, disorder, mania, hysteria
Physical illness	scurvy, sciatica, asthma, gangrene, gerd, cowpox, lice, rickets, malaria, epilepsy, sars, diphtheria, smallpox, bronchitis, thrush, leprosy, typhus, sids, watkins, measles, jaundice, shingles, cholera, boil, pneumonia, mumps, rheumatism, rabies, abscess, warts, plague, dysentery, syphilis, cancer, influenza, ulcers, tetanus
Crime	arrested, unreliable, detained, arrest, detain, murder, murdered, criminal, criminally, thug, theft, thief, mugger, mugging, suspicious, executed, illegal, unjust, jailed, jail, prison, arson, arsonist, kidnap, kidnapped, assaulted, assault, released, custody, police, sheriff, bailed, bail
lawfulness	loyal, charming, friendly, respectful, dutiful, grateful, amiable, honourable, honourably, good, faithfully, faithful, pleasant, praised, just, dignified, approving, approve, compliment, generous, faithful, intelligent, appreciative, delighted, appreciate

Table 6: Keywords used for performing WEAT evaluation.

Attribute	Ratio of correctly classified entities	Ratio of incorrectly classified entities	Ratio of unable to classify
Non-whites	0.89	0.036	0.07
Whites	0.75	0.18	0.07
Males	0.89	0.036	0.07
Females	0.79	0.21	0

Table 7: Performance of the keyword-based classification approach.

Word	Sentence
ingenious	This comprehensive piece of clockwork cost the ingenious and indefatigable artist (one Jacob Lovelace, of Exeter,) 34 years' labour.
elderly	yun away for upwards of 16 Months past; elderly NEGRO WOMAN named LOUISA, belonging to the Estate of the late Ancup.
active	FOR SALE, STRONG active NEGRO GIRL, about 24 Years of Age, she is a good Cook, can Wash, Iron, and is well acquainted with Housework in general.
beautiful	and the young husband was hurried away, being scarcely permitted to take a parting kiss from his blooming and beautiful bride.
blind	Dick, of the Mundingo Country, blind mark, about 18 years of age, says he belongs to the estate of Nicholas, dec. of Mantego bay.
sick	The young wife had snatched up,; few of her own and her baby's clothes; the husband, I Openig Chorus, though sick , had attended to his duty to the last, and escaped Song caped penniless with the clothes on his back.
free	A free black girl JOSEPHINE, detained by the Police as being diseased; Proprietors and Managers in the Country are kindly requested to have the said Josephine apprehended 'and lodged in the Towa Prison, the usual reward will be paid
brave	From that moment the brave Lopez Lara was only occupied in devising means for delivering this notorious criminal into the hands of justice.

Table 8: Examples from our dataset that contain biased words. Notice the high levels of noise and OCR errors.

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
Limitations
- A2. Did you discuss any potential risks of your work?
Ethical considerations
- A3. Do the abstract and introduction summarize the paper’s main claims?
abstract, 1
- A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

3, 5

- B1. Did you cite the creators of artifacts you used?
3, 5
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Limitations, ethical considerations
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Ethical considerations
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3, limitations, appendix
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3, appendix

C Did you run computational experiments?

5

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
5

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

5

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

6

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

3, 5

D Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

Not applicable. Left blank.

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

Not applicable. Left blank.

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

Not applicable. Left blank.

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

Not applicable. Left blank.

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

Not applicable. Left blank.