Towards Gender Biased Language Classification: A Case Study with British English Archival Metadata Descriptions

Lucy Havens
School of Informatics / University of Edinburgh
Informatics Forum / 10 Crichton Street
Edinburgh, United Kingdom / EH8 9AB
lucy.havens@ed.ac.uk

Abstract

This thesis-in-progress summarizes the work completed and potential directions for a Ph.D. project researching the classification of gender biased language. Recognizing bias as inherent in language and thus inevitable in natural language processing systems, the project aims to make bias transparent. An interdisciplinary methodology is applied to define gender bias, annotate documents according to that definition, and train classification models on the annotated dataset to identify types of gender bias. Having created a gender biased language taxonomy and an annotated dataset, the project now moves towards the development of document classification models. There are several directions the classifier development could follow. The project would benefit from participation in the Student Research Workshop to discuss which direction would add the most valuable contribution to computational linguistics.

1 Introduction and Background

The need to mitigate bias in data has become urgent as evidence of harms from such data grows (Perez, 2019; Noble, 2018). Due to the complexities of bias often overlooked in natural language processing (NLP) bias research (Devinney et al., 2022; Stańczak and Augenstein, 2021), Blodgett et al. (2020) and Crawford (2017) call for greater interdisciplinary collaboration and stakeholder involvement in NLP and machine learning (ML) research. The gallery, library, archive, and museum (GLAM) sector has made similar calls for interdisciplinary engagement, looking to applications of data science and ML to better understand and mitigate bias in GLAM collections (Padilla, 2017, 2019; Geraci, 2019). Supporting the NLP and GLAM communities' shared aim of of mitigating the minoritization[1] of certain social groups that biased language causes, this project aims to develop a classification model that categorizes biased language in GLAM documentation. The project uses the term *biased language* to refer to "written or spoken language that creates or reinforces inequitable power relations among people, harming certain people through simplified, dehumanizing, or judgmental words or phrases that restrict their identity; and privileging other people through words or phrases that favor their identity" (Havens et al., 2020). The project uses the term GLAM *documentation* to refer to the descriptions of cultural heritage collection items written in catalogs of galleries, libraries, archives, and museums. Figure 3 in A.9 shows an example of GLAM documentation online.

Studying GLAM documentation provides an opportunity to study the evolution of biased language, because descriptions in contemporary GLAM catalogs contain both historical and contemporary language. To provide a record of the past, GLAM continually acquire and describe heritage items, structuring descriptions of the items according to metadata standards (such as Research Description and Access (RDA Steering Committee, 2022)) and subject authorities (such as Library of Congress Subject Headings (Library of Congress, 2021)). The heritage items included in GLAM, along with the language used to describe them in catalogs, have a continual influence on society (Benjamin, 2019; Cook, 2011; Smith, 2006). The processes of selecting which items to bring into GLAM, and organizing those items according to standards and authorities, privilege particular perspectives (Adler, 2017; de Jong and Koevoets, 2013; Furner, 2007; Tanselle, 2002; Olson, 2001; Bowker and Star, 1999). These processes shape society's understanding of the present and can either reinforce or challenge existing power relationships among people (Benjamin, 2019; Noble, 2018; Yale, 2015; de Jong and Koevoets, 2013; Cook, 2011; Smith, 2006).

Through case studies of free-text descriptions in

[1] D'Ignazio and Klein (2020) propose the term "minoritization" to describe a group of people's *experience* of oppression, in place of "minority" which defines people as oppressed.

many GLAM catalogs, variations in biased language over time and across locations could be better understood. Should patterns in the evolution of biased language emerge, language technology could one day be trained to identify newly-emerging types of bias that it has not yet seen. This project takes the first step in that direction, with a case study of biased language classification for GLAM documentation.

To create biased language classifiers, the project defined three objectives:

O1. Define types of bias for GLAM.

O2. Measure the prevalence of biased language in GLAM documentation.

O3. Build and evaluate classifiers to detect bias.

O1 has been achieved and O2 is in progress (§4). As the project proceeds, several approaches are under consideration for building and evaluating classifiers (§5). Recently passing the halfway point of a three-and-half-year Ph.D., the project would benefit from feedback at the Student Research Workshop to discuss approaches to O3.

2 Related Work

Awareness of limitations in approaches to bias mitigation in Natural Language Processing (NLP) and the wider Machine Learning (ML) community is growing. Publications about NLP bias research now include not only efforts to debias datasets and algorithms (Webster et al., 2018; Zhao et al., 2018), but also recommendations to address the complexity of bias that debiasing efforts often miss (Goldfarb-Tarrant et al., 2021; Blodgett et al., 2021; Jo and Gebru, 2020; Havens et al., 2020; Gonen and Goldberg, 2019; Mitchell et al., 2019; Bender and Friedman, 2018). Recognizing the harmful impacts of deep learning models trained on datasets too large to be adequately interrogated (Birhane and Prabhu, 2021; Bender et al., 2021; Noble, 2018), this project will train supervised NLP models on a dataset small enough to be interrogated (399,957 words, 24,474 sentences). Moreover, collaborators include archivists who manage the collections described in the project's data and have expert knowledge to inform annotation and analysis processes.

Recognizing the subjective nature of certain NLP tasks, such as detecting hate speech and bias, Davani et al. (2022), Sang and Stanton (2022), and Basile et al. (2021) have questioned annotation approaches that create a single gold standard or ground truth dataset. The "perspectivist" approach to NLP this inspired, which incorporates multiple annotators' perspectives in published datasets (Basile, 2022), aligns with the data feminist approach that D'Ignazio and Klein (2020) put forth. Data feminism views data as situated and partial, drawing on intersectional feminism's view of knowledge as particular to a specific time, place, and people (Harding, 1995; Crenshaw, 1991; Haraway, 1988). Feminist theories argue that the standpoint (perspective) of a person impacts knowledge and understanding, and that a universal standpoint cannot exist (Harding, 1995; Haraway, 1988). Indigenous epistemologies, such as the Lakota concept of *waȟkàŋ*, further the notion of the impossibility of a universal truth (Lewis et al., 2018). Translated as "that which cannot be understood," waȟkàŋ communicates that knowledge may come from a place beyond what we are capable of imagining (ibid.). To create a dataset of GLAM documentation annotated for gender biased language, this project creates an annotation taxonomy that allows for gender information to be labeled as uncertain or excluded, and incorporates multiple annotators' perspectives in the model training data.

To practically apply theories and approaches from perspectivist NLP, data feminism, and indigenous epistemologies, the project applies the case study method common to social sciences and design research. Case studies use a combination of data and information gathering approaches to study particular phenomena in context, focusing on "consideration of the whole, covering interrelationships," which provides a "depth [that] compensates for any shortcomings in breadth and the ability to generalize" (Martin and Hanington, 2012, 28). Furthermore, case studies report and reflect upon outliers discovered in the research process (ibid.), useful for this project's effort to create space for the perspectives of minoritized people. This project provides a case study for NLP bias research with the long-term aim of building a collection of case studies, which would enable the NLP community to determine the aspects of bias mitigation approaches that can and cannot be generalized across contexts.

3 Methodology

The interdisciplinary nature of the Ph.D. project warrants a combination of methods and frameworks from several disciplines. Adopting the bias-aware methodology of Havens et al. (2020), case study and participatory action research methods complement NLP methods for creating the project's annotation taxonomy, annotated datasets, and classification models. Critical discourse analysis, feminist theories, queer theory, and indigenous epistemologies provide frameworks through which to analyze the project's metadata descriptions and annotated datasets. To begin, the author defined gender bias using the definition of biased language of Havens et al. (2020) (quoted in §1) narrowed to *gender* bias. This definition informs the annotation taxonomy, which in turn will influence classifiers created with the annotated data.

Participatory action research methods are used to incorporate stakeholder perspectives, necessary for situating a study of gender bias in a particular time, place, and people. Situated in the United Kingdom, the project works with archival documentation written in British English from the Centre for Research Collections at the University of Edinburgh (CRC).[2] Due to the numerous characteristics on which bias may be based, such as racialized ethnicities, economic class, gender, and sexuality, a focus on *gender* bias was chosen. This focus supports the CRC's existing effort to mitigate gender bias in its collections. A person's gender is considered to be self-described, changeable, and capable of falling anywhere along a spectrum of femininity to masculinity (Scheuerman et al., 2020; Keyes, 2018). Archivists provided feedback during the development of the project's annotation taxonomy (§A.8), and will provide feedback in future work analyzing the data annotated with the taxonomy.

The annotation taxonomy and instructions for applying the taxonomy focus on documenting information explicit in the text to avoid misgendering (Scheuerman et al., 2020). The annotations do not infer a person's gender from the person's name, occupation, or other descriptive information, nor do the annotations assign a particular gender to a person. Rather, the annotation process records whether the terms used to describe a person are "feminine," "non-binary," "masculine," or, if only gender-neutral terms are used, "unknown." Annotators were instructed to read the metadata descriptions from their contemporary perspective. That being said, as the historian Shopland writes, "when writing of historic LGBTQIA+ people, we use a definition which simply did not exist in their lifetimes" (2020, 1). Consequently, the project acknowledges that the perspectives documented in the annotation process are situated not only geographically and culturally, but also temporally, in the 21st century.

Following Smith's (2006) approach, the project views heritage as a process. Smith writes, "what makes certain activities 'heritage' are those activities that actively engage with thinking about and acting out not only 'where we have come from' in terms of the past, but also 'where we are going' in terms of the present and future" (ibid., 84). The annotation process of this project visits, interprets, and negotiates with heritage (ibid.) in the form of archival documentation, directing NLP technology towards trans-inclusive conceptualizations of gender, and making gender biases in archival documentation transparent. Smith's approach to heritage draws on Fairclough's (2003) approach to critical discourse analysis (CDA).

Discourse consists of language and its production, interpretation, and social context (Fairclough, 2003; Marston, 2000). CDA thus provides a valuable lens through which to study the heritage material of this project: descriptions from an archival catalog. Considering language in its context of use, CDA offers an approach to studying how language legitimizes, maintains, and challenges power (Smith, 2006; Fairclough, 2003; Marston, 2000). The project uses CDA to follow the data feminism principles of examining and challenging power (D'Ignazio and Klein, 2020). Through annotations of gender biased language, the project examines and challenges the dominant perspective of men in the archival metadata descriptions, making this perspective explicit and identifying opportunities for perspectives of additional genders to be incorporated into the descriptions.

4 Work Achieved

The project has accomplished O1, defining and categorizing types of gender biased language for archives, through the creation of an annotation taxonomy. The taxonomy defines types of gender bias to label in a corpus of archival documentation. Currently the project is progressing on O2 and O3, which are interrelated: the manual annotation process allowed for calculations of the prevalence of gender biased language on a subset of archival documentation, and the classifiers, once built, will enable more complete calculations of gender biased language on the remainder of the descriptions in the archive's catalog. This section summarizes the

[2] archives.collections.ed.ac.uk

	Title	Biographical/Historical	Scope & Contents	Processing Information	Total
Count	4,834	576	6198	280	11,888
Words	51,904	75,032	269,892	3,129	399,957
Sentences	5,932	3,829	14,412	301	24,474

Table 1: Total counts, words and sentences for descriptive metadata fields in the aggregated dataset. Calculations were made using Punkt tokenizers in the Natural Language Toolkit Python library (Loper and Bird, 2002).

work achieved on O1, O2, and O3; the next section (§5) outlines potential directions for completing O3. Havens et al. (2022) contains a detailed discussion of the annotation taxonomy and its application to create an annotated dataset.

The project's annotation taxonomy builds on literature from ML (Hitti et al., 2019), human-computer interaction (Keyes, 2018; Scheuerman et al., 2020), gender studies (Butler, 1990), archival science (Tanselle, 2002), and linguistics (Fairclough, 2003; Bucholtz, 2003, 1999). Group interviews and workshops (participatory action research methods (Moore, 2018; Swantz, 2008; Reid and Frisby, 2008)) further informed the annotation taxonomy. The final annotation taxonomy consists of three categories. Each category contains subcategories with the labels that the annotators applied to archival metadata descriptions. §A.8 contains the complete taxonomy with definitions and examples.

The first two categories of labels, Person Name and Linguistic, annotate vocabulary choices and lexical relations that are explicit in the descriptions, providing a record of the "internal' relations of texts" (Fairclough, 2003, 36-7). The third category of labels, Contextual, annotates according to the descriptions' relationship with a social context (for example, events, behaviors, and power structures), providing a record of the "'external' relations of texts" (Fairclough, 2003, 36). Approaching the archival documentation as discourse, the annotations make the connections between the internal and external relations of language transparent.

Annotating heritage in the form of archival metadata descriptions adds to the process that is heritage, evolving the meaning of the descriptions (Smith, 2006). Applying annotations to archival metadata descriptions from a 21st century perspective recontextualizes the descriptions, adding to the genre chain, or network, of archival documentation that begins with the archival items and continued with catalogers' descriptions of the items (Fairclough, 2003). The taxonomy permits annotators to record uncertainty and absence of information (Shopland, 2020; Lewis et al., 2018), deviating from past NLP documentation approaches (i.e., Garnerin et al. (2020); Dinan et al. (2020)).[3] Participatory action research found that archivists view archival documentation as incomplete. The primary purpose of describing archival items is to enable their discoverability, but this must be balanced with the need to describe a backlog of new archival items perpetually being acquired.

The corpus of archival documentation for annotation were created by harvesting metadata descriptions from an online catalog, reformatting the descriptions for annotation, and manually labeling the descriptions according to the annotation taxonomy. The archival documentation comes from four metadata descriptions in the online archival catalog of the CRC: Title, Biographical / Historical, Scope and Contents, and Processing Information. Though not all descriptions have a date recording when they were written, the earliest recorded date of a description's writing is 1896 and the latest, 2020. The CRC's Archives include a variety of material, such as photographs, letters, manuscripts (letters, lecture notes, and other handwritten documents), and instruments; and cover a range of topics, including town planning, research and teaching, and Scottish Presbyterianism. The language of the Archives' materials are mostly English (1,018 out of 1,315 collections, about 77%), though over 80 languages total are present across the collections. The descriptions that were annotated account for about 20% of the text in the entire online catalog of the CRC's Archives. Table 1 provides summary statistics about the data. For further detail on the size, contents, and organization of the annotation corpus, please refer to the paper by Havens et al. (2022) and the data statement (Bender and Friedman, 2018) in Appendix A.9.

The project received grants to hire four anno-

[3] Domains beyond GLAM also face the challenge of uncertain and absent information (Andrus et al., 2021).

tators, who were Ph.D. students selected for their experience in gender studies or archives. The total cost of the annotation work amounted to circa 400 hours of work and £5,333.76. The four hired annotators each worked 72 hours over eight weeks, receiving £18.52 per hour. The author spent 86 hours annotating over 16 weeks for her Ph.D. project. Though all annotators identify as women, due to the historical dominance of men's perspective in the English language and the pejoration of terms describing women (Spencer, 2000; Schulz, 2000; Lakoff, 1989),[4] the project's annotated dataset does challenge dominant perspectives in archival discourse to advance gender equity (D'Ignazio and Klein, 2020; Fairclough, 2003).

Inter-annotator agreement (IAA) calculations reflect the subjectivity of gender bias (see §B, tables 2, 3, 4, and 5). Annotating *gendered* language proved to be more straightforward than annotating gender *biased* language. We report IAA with F_1 as our metric due to the limitations of coefficients' assumptions and interpretability as Artstein and Poesio (2008) discuss. F_1 scores for the gendered language labels "Gendered Role" and "Gendered Pronoun" fall between 0.71 and 0.99. F_1 scores for annotating gender biased language are relatively low, with the greatest agreement on the "Generalization" label at only 0.56, on the "Omission" label at 0.48, and on the "Stereotype" label at 0.57. Manual analysis of disagreements among annotators demonstrated the value of a perspectivist approach to disagreements (Davani et al., 2022; Sang and Stanton, 2022; Basile et al., 2021), as multiple annotators' labels were often deemed correct for the same text span.

The five annotators' datasets were merged into one aggregated dataset, which will be divided into training, development, and test sets for creating classifiers. Aggregation began with a one-hour manual review of each annotator's labels to identify patterns and common mistakes, which informed the subsequent aggregation steps. Disagreeing labels for the same text span were then manually reviewed, with either a combination or an individual label being chosen for each text span to include in the aggregated dataset.

Next, for annotations with overlapping text spans and the same label (considered to be in agreement), the annotation with the largest text span was added to the aggregated dataset. All remaining annotations were then added to the aggregated dataset, with the exception of one annotator's Person Name labels, as these were applied with great inconsistency relative to other annotators. §A.1 details the review and aggregation of the annotated datasets. Figures 2 illustrates the prevalence of each of the taxonomy's labels in the aggregated dataset and Figure 3 illustrates how many annotations from each annotator are in the aggregated dataset. The annotated datasets are a starting point to identify gender bias in GLAM documentation in the UK; they are not intended to comprehensively cover of all gender biases that may come through in GLAM documentation. They will serve as training, development, and test data for developing classifiers, and will be published alongside the classifiers in future work.

5 Discussion and Conclusion

Now passing the halfway point of a Ph.D. degree, with a year and six months remaining, the project would benefit from feedback on possible approaches to the project's last objectives. Several approaches are under consideration for building classification models (O3).

Four algorithms are under consideration for building a gender biased document classifier: (1) logistic regression (LR), as a classification baseline (Jurafsky and Martin, 2000); (2) decision tree or (3) random forest (a combination of randomized decision trees), as the decision trees are the most transparent algorithm for classification (Bird et al., 2019); and (4) support vector machines, as Adhikari et al. (2019) found this outperformed LR and neural models on document classification for select datasets. The document classifiers could be developed as single task or multitask; the project would like to investigate correlations between labels. As the perspectivist approach to disagreements in NLP encourages (Basile, 2022), classifiers could be trained on individual annotators' datasets in addition to the aggregated dataset. Document classifiers could also be pre-trained on a deep learning model such as DocBERT (Adhikari et al., 2020) to see if pre-training improves their performance.

The focus on document classification comes from the intended use case of the classification models: to support archivists in identifying descriptions with gender biases in their catalogs. Such identification would support the efficient prioriti-

[4] I.e., in the 16th century, grammarians instructed that *man* precede *woman* in writing; in the 18th century, *man* and *he* began to be used in place of *human* and *their* (Spencer, 2000).

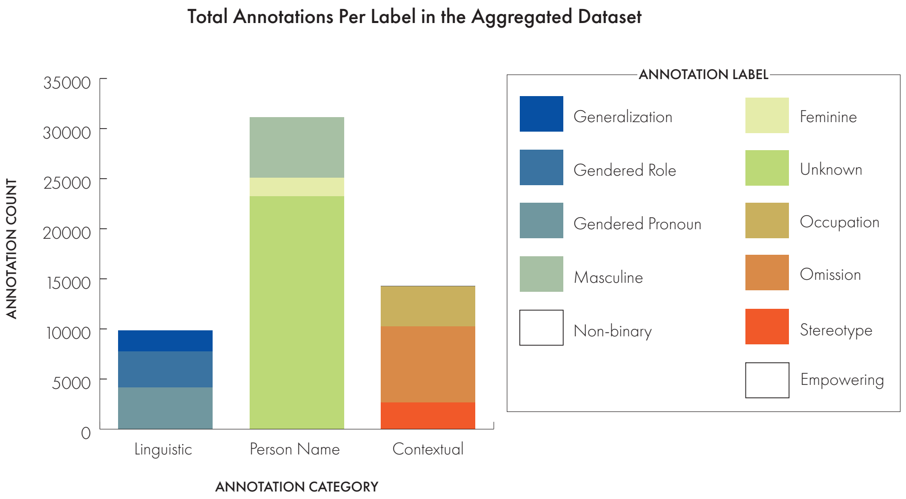

Figure 1: A stacked bar chart of counts of annotations per label across all annotators in the aggregated dataset of 55,260 total annotations, organized into the three categories of labels: Linguistic, Person Name, and Contextual. “Non-binary” (a Person Name label) and “Empowering” (a Contextual label) both have a count of zero.

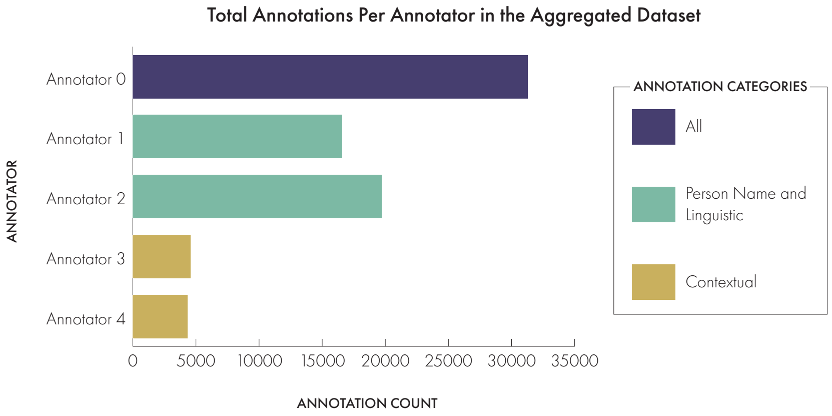

Figure 2: A bar chart of the total annotations from each annotator included in the aggregated dataset, with colors indicating the category of labels each annotator used. For annotations that matched or overlapped, only one was added to the aggregated dataset, so the total number of annotations in the aggregated dataset (55,260) is 21,283 less than the sum of the annotators’ annotations in this chart (76,543).

zation of reparative description practices that add context to or reword harmful descriptions. That being said, annotators applied labels to text spans, not documents, so sentence classification could also be pursued. Though all approaches have the potential to contribute to NLP and GLAM’s efforts to mitigate bias, the 18 months remaining in the Ph.D. provides only enough time for select approaches to be pursued. The project would appreciate feedback at the Student Research Workshop on approaches under consideration to build and evaluate classifiers that detect gender biased documents.

Acknowledgements

Thank you to the organizers of the Student Research Workshop at NAACL-HLT 2022 for the opportunity to submit this in-progress work, and to the reviewers of this paper for their feedback. Additional thanks go to my Ph.D. supervisory team, Bea Alex, Ben Bach, and Melissa Terras; as well as Rachel Hosker and her team at the CRC for their enthusiastic collaboration. Thank you also to the organizations who have funded this Ph.D.: the UK’s Engineering and Physical Sciences Research Council and the University of Edinburgh’s School of Informatics Graduate School. The annotation work in this project was made possible by generous funding from the Institute for Language, Cognition, and Computation in the School of Informatics and the Edinburgh Futures Institute, both at the University of Edinburgh.

References

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, William L. Hamilton, and Jimmy Lin. 2020. Exploring the Limits of Simple Learners in Knowledge Distillation for Document Classification with DocBERT. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 72–77, Online. Association for Computational Linguistics.

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. Rethinking Complex Neural Network Architectures for Document Classification. 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 4046–4051, Minneapolis, Minnesota. Association for Computational Linguistics.

Melissa Adler. 2017. Introduction: A Book is Being Cataloged. In *Cruising the Library: Perversities in the Organization of Knowledge*, pages 1–26. Fordham University Press.

McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 249–260, New York, US. Association for Computing Machinery.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Valerio Basile. 2022. The Perspectivist Data Manifesto. [Online; accessed March 21, 2022].

Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. *CoRR*, abs/2109.04270.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, US. Association for Computing Machinery.

Ruha Benjamin. 2019. *Race after technology : abolitionist tools for the new Jim code*. Polity, Cambridge, UK.

Steven Bird, Ewan Klein, and Edward Loper. 2019. *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision? *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546.

Su 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.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. 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 1004–1015, Online. Association for Computational Linguistics.

Geoffrey C. Bowker and SUSn Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences*. Inside technology. MIT Press, Cambridge, US.

Mary Bucholtz. 1999. Gender. *Journal of linguistic anthropology*, 9(1-2):80–83.

Mary Bucholtz. 2003. Theories of Discourse as Theories of Gender: Discourse Analysis in Language and Gender Studies. In *The Handbook of Language and Gender*, pages 43–68, Oxford, UK. Blackwell Publishing Ltd.

Judith Butler. 1990. *Gender trouble: feminism and the subversion of identity*. Thinking gender. Routledge, New York, US.

Terry Cook. 2011. 'We Are What We Keep; We Keep What We Are': Archival Appraisal Past, Present and Future. *Journal of the Society of Archivists*, 32(2):173–189.

Kate Crawford. 2017. The Trouble with Bias. In *Neural Information Processing Systems Conference Keynote*. [Online; accessed 10-July-2020].

Kimberlé Crenshaw. 1991. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review*, 43(6):1241–1299.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Sara de Jong and Sanne Koevoets. 2013. Introduction. In *Teaching Gender with Libraries and Archives The Power of Information*, Teaching with gender. European women's studies in international and interdisciplinary classrooms, vol. 10, Utrecht. ATGENDER, the European Association for Gender Research, Education and Documentation.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "Gender" in NLP Bias Research. *Computing Research Repository*, arXiv:2205.02526.

Catherine D'Ignazio and Lauren F. Klein. 2020. *Data Feminism*. Strong ideas series. The MIT Press, Cambridge, US.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-dimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online. Association for Computational Linguistics.

Norman Fairclough. 2003. *Analysing Discourse: Textual Analysis for Social Research*. Routledge, London, UK.

Jonathan Furner. 2007. Dewey Deracialized: A Critical Race-Theoretic Perspective. *KNOWLEDGE ORGANIZATION*, 34(3):144–168.

Mahault Garnerin, Solange Rossato, and Laurent Besacier. 2020. Gender representation in open source speech resources. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6599–6605, Marseille, France. European Language Resources Association.

Noah Geraci. 2019. Programmatic approaches to bias in descriptive metadata. In *Code4Lib Conference 2019*. [Online; accessed 28-May-2020].

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. 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 1926–1940, Online. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. *NAACL 2019*, arXiv:1903.03862v2.

Donna Haraway. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, 14(3):575.

Sandra Harding. 1995. "Strong objectivity": A response to the new objectivity question. *Synthese*, 104(3).

Lucy Havens, Benjamin Bach, Melissa Terras, and Beatrice Alex. 2022. Uncertainty and Inclusivity in Gender Bias Annotation: An Annotation Taxonomy and Annotated Datasets of British English Text. In *Proceedings of the Fourth Workshop on Gender Bias in Natural Language Processing*, Seattle, WA, USA. Association for Computational Linguistics. [forthcoming].

Lucy Havens, Melissa Terras, Benjamin Bach, and Beatrice Alex. 2020. Situated data, situated systems: A methodology to engage with power relations in natural language processing research. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 107–124, Barcelona, Spain (Online). Association for Computational Linguistics.

Yasmeen Hitti, Eunbee Jang, Ines Moreno, and Carolyne 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, IT. Association for Computational Linguistics.

Eun Seo Jo and Timnit Gebru. 2020. Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 306–316, New York, US. Association for Computing Machinery.

Dan Jurafsky and James H. Martin. 2000. *Speech & Language Processing*. Pearson Education India.

Os Keyes. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW).

Robin Lakoff. 1989. *Language and Woman's Place*. Harper & Row, New York, US.

Jason Edward Lewis, Nick Philip, Noelani Arista, Archer Pechawis, and Suzanne Kite. 2018. Making Kin with the Machines. *Journal of Design and Science*.

Library of Congress. 2021. Library of Congress Subject Headings PDF Files.

Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, US. Association for Computational Linguistics.

Greg Marston. 2000. Metaphor, morality and myth: a critical discourse analysis of public housing policy in Queensland. *Critical Social Policy*, 20(3):349–373.

Bella Martin and Bruce Hanington. 2012. 11 Case studies. In *Universal Methods of Design: 100 Ways to Research Complex Problems, Develop Innovative Ideas, and Design Effective Solutions*, Beverly, US. Rockport Publishers.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT*'19*.

Niamh Moore. 2018. A cat's cradle of feminist and other critical approaches to participatory research. In *Connected Communities Foundation Series*, Bristol, UK. University of Bristol/AHRC Connected Communities Programme. [Online; accessed 24-July-2020].

Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, New York, US.

Hope A. Olson. 2001. The Power to Name: Representation in Library Catalogs. *Signs: Journal of Women in Culture and Society*, 26(3):639–668.

Thomas Padilla. 2017. On a Collections as Data Imperative. *UC Santa Barbara Previously Published Works*.

Thomas Padilla. 2019. Responsible Operations: Data Science, Machine Learning, and AI in Libraries. *OCLC Research*, page 38.

Caroline Criado Perez. 2019. *Invisible Women: Exposing Data Bias in a World Designed for Men*. Vintage, London, UK.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, Jun'ichi Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Session at EACL 2012*. Association for Computational Linguistics.

RDA Steering Committee. 2022. About RDA.

Colleen Reid and Wendy Frisby. 2008. 6 Continuing the Journey: Articulating Dimensions of Feminist Participatory Action Research (FPAR). In *The SAGE Handbook of Action Research*, pages 93–105. SAGE Publications Ltd.

Yisi Sang and Jeffrey Stanton. 2022. The Origin and Value of Disagreement Among Data Labelers: A Case Study of Individual Differences in Hate Speech Annotation. In *Information for a Better World: Shaping the Global Future*, Lecture Notes in Computer Science, pages 425–444. Springer International Publishing, Cham.

Morgan Klaus Scheuerman, Katta Spiel, Oliver L. Haimson, Foad Hamidi, and Stacy M. Branham. 2020. HCI Guidelines for Gender Equity and Inclusion: Misgendering.

Muriel R. Schulz. 2000. The Semantic Derogation of Women. In Lucy Burke, Tony Crowley, and Alan Girvin, editors, *The Routledge language and cultural theory reader*. Routledge, London, UK.

Norena Shopland. 2020. *A Practical Guide to Searching LGBTQIA Historical Records*. Taylor & Francis Group, Milton.

Laurajane Smith. 2006. *Uses of Heritage*. Routledge, London, UK.

Dale Spencer. 2000. Language and reality: Who made the world? (1980). In Lucy Burke, Tony Crowley, and Alan Girvin, editors, *The Routledge language and cultural theory reader*. Routledge, London, UK.

Karolina Stańczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *CoRR*, abs/2112.14168.

Marja Liisa Swantz. 2008. 2 Participatory Action Research as Practice. In *The SAGE Handbook of Action Research*, pages 31–48. SAGE Publications Ltd.

G. Thomas Tanselle. 2002. The World as Archive. *Common Knowledge*, 8(2):402–406.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *Computing Research Repository*, arXiv:1810.05201.

Anne Welsh. 2016. The Rare Books Catalog and the Scholarly Database. *Cataloging & Classification Quarterly*, 54(5–6):317–337.

Elizabeth Yale. 2015. The History of Archives: The State of the Discipline. *Book History*, 18(1):332–359.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, US. Association for Computational Linguistics.

A Data Statement: Annotated Datasets of Archival Documentation

A.1 Curation Rationale

These datasets were created from a corpus of 1,460 files of archival metadata descriptions totaling circa 15,419 sentences and 255,943 words. That corpus is the first 20% of text from the corpus described in the Provenance Appendix (§A.9), annotated for gender bias according the the taxonomy in Other (§A.8). 73 of files (10% of the text) were triply annotated; the remaining 1,387 files (90% of the text) were doubly annotated. There are six instances of the annotated corpus: one for each of the five annotators and one that aggregates all annotators' labels. Participatory action research with archivists led the project to choose four metadata fields were chosen in the archival catalog to extract for annotation: Title, Scope and Contents, Biographical / Historical, and Processing Information.

The five annotated datasets were merged into a single aggregated dataset for classifier training and evaluation, so comparisons could be made on classifiers' performances after training on an individual annotator's dataset versus on the aggregated dataset. The merging process began with a one-hour manual review of each annotator's labels to identify patterns and common mistakes in their labeling, which informed the subsequent steps for merging the five annotated datasets.

The second step of the merging process was to manually review disagreeing labels for the same text span and add the correct label to the aggregated dataset. Disagreeing labels for the same text span were reviewed for all *Person Name*, *Linguistic*, and *Contextual* categories of labels. For *Person Name* and *Linguistic* labels, where three annotators labeled the same span of text, majority voting determined the correct label: if two out of the three annotators used one label and the other annotator used a different label, the label used by the two annotators was deemed correct and added to the aggregated dataset. For *Contextual* labels, unless an obvious mistake was made, the union of all three annotators' labels was included in the aggregated dataset.

Thirdly, the "Occupation" and "Gendered Pronoun" labels were reviewed. A unique list of the text spans with these labels was generated and incorrect text spans were removed from this list. The "Occupation" and "Gendered Pronoun" labels in the annotated datasets with text spans in the unique lists of valid text spans were added to the aggregated dataset. Fourthly, the remaining *Linguistic* labels ("Gendered Pronoun," "Gendered Role," and "Generalization") not deemed incorrect in the annotated datasets were added to the aggregated dataset. Due to common mistakes in annotating *Person Name* labels with one annotator, only data from the other two annotators who annotated with *Person Name* labels was added to the aggregated dataset. Fifthly, for annotations with overlapping text spans and the same label, the annotation with the longer text span was added to the aggregated dataset. The sixth and final step to constructing the aggregated dataset was to take the union of the remaining *Contextual* labels ("Stereotype," "Omission," "Occupation," and "Empowering") not deemed incorrect in the three annotated datasets with these labels and add them to the aggregated dataset.

A.2 Language Variety

The metadata descriptions extracted from the Archive's catalog are written primarily in British English, with the occasional word in another language such as French or Latin.

A.3 Producer Demographic

The producing research team are of American, German, and Scots nationalities, and are three women and one man. We all work primarily as academic researchers in the disciplines of natural language processing, data science, data visualization, human-computer interaction, digital humanities, and digital cultural heritage. Additionally, one of us is audited an online course on feminist and social justice studies.

A.4 Annotator Demographic

The five annotators are of American and European nationalities and identify as women. Four annotators were hired by the lead annotator for their experience in gender studies and archives. The four annotators worked 72 hours each over eight weeks in 2022, receiving £1,333.44 each (£18.52 per hour). The lead annotator completed the work for her Ph.D. project, which totaled to 86 hours of work over 16 weeks.

A.5 Speech or Publication Situation

The archival metadata descriptions describe material about a range of topics, such as teaching, research, town planning, music, and religion. The materials described also vary, from letters and journals to photographs and audio recordings. The descriptions in this project's dataset with a known date (which describe 38.5% of the archives' records) were written from 1896 through 2020.

The annotated dataset will be published with a forthcoming paper detailing the methodology and theoretical framework that guided the development of the annotation taxonomy and the annotation process, accompanied by analysis of patterns and outliers in the annotated dataset.

A.6 Data Characteristics

The datasets were organized for annotation in a web-based annotation paltform, the brat rapid annotation tool (Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, Jun'ichi Tsujii, 2012). Consequently, the data formats conform to the brat formats: plain text files that end in '.txt' contain the original text and plain text files that end in '.ann' contain the annotations. The annotation files include the starting and ending text span of a label, the actual text contained in that span, the label name, and any notes annotators recorded about the rationale for applying the label they did. The names of all the files consist of the name of the fonds (the archival term for a collection) and a number indicating the starting line number of the descriptions. Descriptions from a single fonds were split across files so that no file contained more than 100 lines, because brat could not handle the extensive length of certian fonds' descriptions.

A.7 Data Quality

A subset of annotations were applied automatically with a grep script and then corrected during the manual annotation process. All three categories of the annotation taxonomy were manually applied by the annotators. The lead annotator then manually checked the labels for accuracy. That being said, due to time constraints, mistakes are likely to remain in the application of labels (for example, the starting letter may be missing from a labeled text span or a punctuation mark may have accidentally been included in a labeled text span).

A.8 Other: Annotation Schema

The detailed schema that guided the annotation process is listed below with examples for each label. In each example, the labeled text is underlined. All examples are taken from the dataset except for labels 1.1, "Non-binary," and 3.4, "Empowering," as the annotators did not find any text to which the provided label definitions applied. The annotation instructions permitted labels to overlap as each annotator saw fit, and asked annotators to read and annotate from their contemporary perspective. The categories of labels from the annotation taxonomy were divided among annotators: two hired annotators labeled with categories 1 and 2, two hired annotators labeled with category 3, and the lead annotator labeled with all categories.

The annotation taxonomy includes labels for *gendered* language, rather than only explicitly gender-biased language, because measuring the use of gendered words across an entire archives' collection provides information about gender bias at the overall collections' level. For example, using a gendered pronoun such as "he" is not inherently biased, but if the use of this masculine gendered pronoun far outnumbers the use of other gendered pronouns in our dataset, we can observe that the masculine is over-represented, indicating a masculine bias in the archives' collections overall. Labeling gender-biased language focuses on the individual description level. For example, the stereotype of a wife playing only or primarily a supporting role to her husband comes through in the following description:

> *Jewel took an active interest in her husband's work, accompanying him when he travelled, sitting on charitable committees, looking after missionary furlough houses and much more. She also wrote a preface to his Baptism and Conversion and a foreward [sic] to his A Reasoned Faith.*. (Fonds Identifier: Coll-1036)

1. **Person Name:** the name of a person, including any pre-nominal titles (i.e., Professor, Mrs., Sir, Queen), when the person is the primary entity being described (rather than a location named after a person, for example)

   1.1 **Non-binary:*** the pronouns or roles of the named person within the descriptive field in which this instance of the name appears (either Title, Scope and Contents, Biographical / Historical, or Processing Information) are non-binary

   Example 1.1: Francis McDonald went to the University of Edinburgh where they studied law.
   Note: the annotation process did not find suitable text on which to apply this label in the dataset.

   1.2 **Feminine:** the pronouns, titles, or roles of the named person within the descriptive field in which this instance of the name appears (either Title, Scope and Contents, Biographical / Historical, or Processing Information) are feminine

Example 1.2: "Jewel took an active interest in her husband's work..." (Fonds Identifier: Coll-1036)

1.3 **Masculine:** the pronouns, titles, or roles of the named person within the descriptive field in which this instance of the name appears (either Title, Scope and Contents, Biographical / Historical, or Processing Information) are masculine

Example 1.3: "Martin Luther, the man and his work." (Fonds Identifier: BAI)

1.4 **Unknown:** any pronouns, titles, or roles of the named person within the descriptive field in which this instance of the name appears (either Title, Scope and Contents, Biographical / Historical, or Processing Information) are gender neutral, or no such pronouns or roles are provided within the descriptive field

Example 1.4: "Testimonials and additional testimonials in favour of Niecks, candidacy for the Chair of Music, 1891" (Fonds Identifier: Coll-1086)

2. **Linguistic:** gender marked in the way a word, phrase or sentence references a person or people, assigning them a specific gender that does not account for all genders possible for that person or people

2.1 **Generalization:** use of a gender-specific term (i.e. roles, titles) to refer to a group of people that could identify as more than the specified gender

Example 2.1: "His classes included Anatomy, Practical Anatomy, ... Midwifery and Diseases of Women, Therapeutics, Neurology, ... Public Health, and Diseases of the Skin." (Fonds Identifier: Coll-1118)

2.2 **Gendered Roles:** use of a title or word denoting a person's role that marks either a non-binary, feminine, or masculine gender

Example 2.2: "New map of Scotland for Ladies Needlework, 1797" (Fonds Identifier: Coll-1111)

2.3 **Gendered Pronouns:** explicitly marking the gender of a person or people through the use of pronouns (e.g., he, him, himself, his, her, herself, and she)

Example 2.3: "He obtained surgical qualifications from Edinburgh University in 1873 ([M.B.])." (Fonds Identifier: Coll-1096)

3. **Contextual:** expectations about a gender or genders that comes from knowledge about the time and place in which language is used, rather than from linguistic patterns alone (i.e., sentence structure or word choice)

3.1 **Stereotype:** a word, phrase, or sentence that communicates an expectation of a person or group of people's behaviors or preferences that does not reflect the reality of all their possible behaviors or preferences; or a word, phrase, or sentence that focuses on a particular aspect of a person that doesn't represent that person holistically

Example 3.1: "The engraving depicts a walking figure (female) set against sunlight, and holding/releasing a bird." (Fonds Identifier: Coll-1116)

3.2 **Omission:** focusing on the presence, responsibility, or contribution of a single gender in a situation in which more than one gender has a presence, responsibility or contribution; or defining one person's identity in terms of their relation to another person

Example 3.2: "This group portrait of Laurencin, Apollinaire, and Picasso and his mistress became the theme of a larger version in 1909 entitledApollinaire [sic] and his friends." (Fonds Identifier: Coll-1090).

3.3 **Occupation:** a word or phrase that refers to a person or people's job title (singular or plural) for which the person or people received payment; do not annotate occupations used as a pre-nominal title (for example, "Colonel Sir Thomas Francis Fremantle" should not have an occupation label)

Example 3.3: "He became a surgeon with the Indian Medical Service." (Fonds Identifier: Coll-1096).

3.4 **Empowering:** reclaiming derogatory words or phrases to empower a minoritized person or people

Example 3.4: a person describing themself as queer in a self-affirming, positive manner
Note: the annotation process did not find enough text on which to apply this label in the dataset to include it when training a classifier.

*The "Non-binary" label was not used by the annotators. That being said, this does not mean there were not people who would identify as non-binary represented in the text of the annotation corpus. When relying only on descriptions written by people other than those represented in the descriptions, knowledge about people's gender identity remains incomplete (Shopland, 2020). Additional linguistic research informed by a knowledge of terminology for the relevant time period may identify people who were likely to identify as non-binary in the corpus of archival metadata descriptions. For example, Shopland (2020) finds that focusing on actions that people were described doing can help to locate people of minoritized genders (and sexualities) in historical texts, but also cautions researchers against assuming too much. A full understanding of a person's gender often remains unattainable from the documentation that exists about them.

A.9 Provenance Appendix

Data Statement: Corpus of Archival Documentation

A.9.1 Curation Rationale

We (the research team) will use the extracted metadata descriptions to create a gold standard dataset annotated for contextual gender bias. We adopt Hitti et al.'s definition of contextual gender bias in text: written language that connotes or implies an inclination or prejudice against a gender through the use of gender-marked keywords and their context (2019, p. 10-11).

A member of our research team has extracted text from four descriptive metadata fields for all collections, subcollections, and items in the Archive's online catalog. The first field is a title field. The second field provides information about the people, time period, and places associated with the collection, subcollection, or item to which the field belongs. The third field summarizes the contents of the collection, subcollection, or item to which the field belongs. The last field records the person who wrote the text for the collection, subcollection, or item's descriptive metadata fields, and the date the person wrote the text (although not all of this information is available in each description; some are empty).

Using the dataset of extracted text, we will experiment with training a discriminative classification algorithm to identify types of contextual gender bias. Additionally, the dataset will serve as a source of annotated, historical text to complement datasets composed of contemporary texts (i.e. from social media, Wikipedia, news articles).
We chose to use archival metadata descriptions as a data source because:

1. Metadata descriptions in the Archive's catalog (and most GLAM catalogs) are freely, publicly available online

2. GLAM metadata descriptions have yet to be analyzed at large scale using natural language processing (NLP) methods and, as records of cultural heritage, the descriptions have the potential to provide historical insights on changes in language and society (Welsh, 2016)

3. GLAM metadata standards are freely, publicly available, often online, meaning we can use historical changes in metadata standards used in the Archive to guide large-scale text analysis of changes in the language of the metadata descriptions over time

4. The Archive's policy acknowledges its responsibility to address legacy descriptions in its catalogs that use language considered biased or otherwise inappropriate today[5]

A.9.2 Language Variety

The metadata descriptions extracted from the Archive's catalog are written in British English.

A.9.3 Producer Demographic

We (the research team) are of American, German, and Scots nationalities, and are three females and one male. We all work primarily as academic researchers in the disciplines of natural language processing, data science, data visualization, human-

[5] The Archive is not alone; across the GLAM sector, institutions acknowledge and are exploring ways to address legacy language in their catalogs' descriptions. The "Note" in We Are What We Steal provides one example: https://dxlab.sl.nsw.gov.au/we-are-what-we-steal/notes/.

computer interaction, digital humanities, and digital cultural heritage. Additionally, one of us has been auditing a feminism and social justice course, and reading literature on feminist theories, queer theory, and indigenous epistemologies.

A.9.4 Annotator Demographic

Not applicable

A.9.5 Speech or Publication Situation

The metadata descriptions extracted from the Archive's online catalog using Open Access Initiative - Protocol for Metadata Harvesting (OAI-PMH). For OAI-PMH, an institution (in this case, the Archive) provides a URL to its catalog that displays its catalog metadata in XML format. A member of our research team wrote scripts in Python to extract three descriptive metadata fields for every collection, subcollection, and item in the Archive's online catalog (the metadata is organized hierarchically). Using Python and its Natural Language Toolkit library (Loper and Bird, 2002), the researcher removed duplicate sentences and calculated that the extracted metadata descriptions consist of a total of 966,763 words and 68,448 sentences across 1,231 collections. The minimum number of words in a collection is 7 and the maximum, 156,747, with an average of 1,306 words per collection and standard deviation of 7,784 words. The archival items described in resulting corpus consist of a variety of material, from photographs and manuscripts (letters, lecture notes, and other handwritten documents) to instruments and tweets.

A.9.6 Data Characteristics

Upon extracting the metadata descriptions using OAI-PMH, the XML tags were removed so that the total words and sentences of the metadata descriptions could be calculated to ensure the text source provided a sufficiently large dataset. A member of our research team has grouped all the extracted metadata descriptions by their collection (the "fonds" level in the XML data), preserving the context in which the metadata descriptions were written and will be read by visitors to the Archive's online catalog.

A.9.7 Data Quality

As a member of our research team extracts and filters metadata descriptions from the Archive's online catalog, they write assertions and tests to ensure as best as possible that metadata is not lost or unintentionally changed.

A.9.8 Other

Not applicable

A.9.9 Provenance Appendix

Not applicable

B Inter-Annotator Agreement

The following pages display four tables of inter-annotator agreement (IAA) scores: among annotators, table 2 for the Person Name and Linguistic categories of labels, and table 3 for the Contextual category of labels; annotators versus the aggregated dataset, table 4 for the Person Name and Linguistic categories of labels, and table 5 for the Contextual category of labels. IAA was calculated such that overlapping text spans with the same label were considered to be in agreement, in addition to matching text spans with the same label. Due to the aim of training document classification models on the annotated datasets, the existence of a particular type of gender bias in an archival metadata description was deemed more important than the precise words that communicate gender bias.

exp	pred	label	true pos	false pos	false neg	precision	recall	F_1	files
0	1	Unknown	5031	1524	4268	0.76751	0.54103	0.63467	584
0	2	Unknown	2776	537	432	0.83791	0.86534	0.85140	170
1	2	Unknown	1048	1421	315	0.42446	0.76889	0.54697	72
0	1	Masculine	2367	2372	1079	0.49947	0.68688	0.57838	584
0	2	Masculine	728	111	146	0.86770	0.83295	0.84997	170
1	2	Masculine	380	169	411	0.69217	0.48040	0.56716	72
0	1	Feminine	627	427	642	0.59488	0.49409	0.53982	584
0	2	Feminine	724	128	178	0.84977	0.80266	0.82554	170
1	2	Feminine	287	496	279	0.36654	0.50707	0.42550	72
0	1	Non-binary	0	0	0	-	-	-	584
0	2	Non-binary	0	0	0	-	-	-	170
1	2	Non-binary	0	0	0	-	-	-	72
0	1	Gendered Role	1802	306	882	0.85484	0.67139	0.75209	584
0	2	Gendered Role	1404	162	257	0.89655	0.84527	0.87016	170
1	2	Gendered Role	438	292	52	0.60000	0.89388	0.71803	72
0	1	Gendered Pronoun	3398	101	190	0.97113	0.94705	0.95894	584
0	2	Gendered Pronoun	869	70	60	0.92545	0.93541	0.93041	170
1	2	Gendered Pronoun	518	7	11	0.98667	0.97921	0.98292	72
0	1	Generalization	37	35	262	0.51389	0.12375	0.19946	584
0	2	Generalization	74	51	63	0.59200	0.54015	0.56489	170
1	2	Generalization	2	50	7	0.03846	0.22222	0.06557	72

Table 2: Inter-annotator agreement (IAA) measures for annotators who used the *Person Name* and *Linguistic* categories of labels to annotate archival documentation. The first two columns note the annotator whose labels were considered expected or predicted, respectively. The abbreviation “pos” is for “positive;” “neg,” for “negative.” The last column lists the number of files with annotations by both annotators for that row. No annotators applied the “Non-binary” label.

exp	pred	label	true pos	false pos	false neg	precision	recall	F_1	files
0	3	Occupation	1988	613	724	0.76432	0.73303	0.74835	485
0	4	Occupation	738	396	240	0.65079	0.75460	0.69886	149
3	4	Occupation	422	327	134	0.56341	0.75899	0.64674	57
0	3	Omission	1376	914	3259	0.60087	0.29687	0.39740	485
0	4	Omission	416	317	875	0.56753	0.32223	0.41106	149
3	4	Omission	215	315	155	0.40566	0.58108	0.47777	57
0	3	Stereotype	505	539	227	0.48371	0.68989	0.56869	485
0	4	Stereotype	507	525	600	0.49127	0.45799	0.47405	149
3	4	Stereotype	34	60	161	0.36170	0.17435	0.23529	57
0	3	Empowering	0	80	0	-	-	-	485
0	4	Empowering	0	0	0	-	-	-	149
3	4	Empowering	0	0	80	-	-	-	57

Table 3: Inter-annotator agreement (IAA) measures for annotators who used the *Contextual* category of labels to annotate archival metadata descriptions. The first two columns note the annotator whose labels were considered expected or predicted, respectively. The abbreviation “pos” is for “positive;” “neg,” for “negative.” The last column lists the number of files with annotations by both annotators for that row. Only annotator 3 applied the “Empowering” label.

exp	pred	label	true pos	false pos	false neg	precision	recall	F_1	files
Agg	0	Unknown	10561	36	1900	0.99660	0.84752	0.91604	714
Agg	1	Unknown	6608	0	4511	1.00000	0.59430	0.74553	597
Agg	2	Unknown	15140	117	679	0.99233	0.95708	0.97439	444
Agg	0	Masculine	3963	18	2446	0.99548	0.61835	0.76285	714
Agg	1	Masculine	4749	1	1099	0.99979	0.81207	0.89621	597
Agg	2	Masculine	1007	5	525	0.99506	0.65731	0.79167	444
Agg	0	Feminine	1454	19	523	0.98710	0.73546	0.84290	714
Agg	1	Feminine	1076	0	707	1.00000	0.60348	0.75271	597
Agg	2	Feminine	994	12	410	0.98807	0.70798	0.82490	444
Agg	0	Nonbinary	0	0	0	-	-	-	714
Agg	1	Nonbinary	0	0	0	-	-	-	597
Agg	2	Nonbinary	0	0	0	-	-	-	444
Agg	0	Gendered-Role	3108	697	330	0.81682	0.90401	0.85821	714
Agg	1	Gendered-Role	1924	218	716	0.89823	0.72879	0.80468	597
Agg	2	Gendered-Role	1471	652	230	0.69289	0.86479	0.76935	444
Agg	0	Gendered-Pronoun	3933	160	165	0.96091	0.95974	0.96032	714
Agg	1	Gendered-Pronoun	3498	3	190	0.99914	0.94848	0.97315	597
Agg	2	Gendered-Pronoun	1016	1	41	0.99902	0.96121	0.97975	444
Agg	0	Generalization	405	1	1370	0.99754	0.22817	0.37139	714
Agg	1	Generalization	69	4	1123	0.94521	0.05789	0.10909	597
Agg	2	Generalization	127	0	862	1.00000	0.12841	0.22760	444

Table 4: Inter-annotator agreement (IAA) between the aggregated dataset and annotators for the *Person Name* and *Linguistic* categories of labels to annotate archival documentation. The first two columns note the annotator whose labels were considered expected or predicted, respectively. The abbreviation “pos” is for “positive;” “neg,” for “negative.” The last column lists the number of files with annotations by both annotators for that row. No annotators applied the “Non-binary” label.

exp	pred	label	true pos	false pos	false neg	precision	recall	F_1	files
Agg	0	Occupation	2725	23	571	0.99163	0.82676	0.90172	631
Agg	3	Occupation	2320	290	873	0.88889	0.72659	0.79959	508
Agg	4	Occupation	1746	147	253	0.92235	0.87344	0.89723	450
Agg	0	Omission	5916	12	1187	0.99798	0.83289	0.90799	631
Agg	3	Omission	2310	13	3475	0.99440	0.39931	0.56981	508
Agg	4	Omission	1876	5	967	0.99734	0.65987	0.79424	450
Agg	0	Stereotype	1748	11	1058	0.99375	0.62295	0.76583	631
Agg	3	Stereotype	1089	9	279	0.99180	0.79605	0.88321	508
Agg	4	Stereotype	1400	2	715	0.99857	0.66194	0.79613	450
Agg	0	Empowering	0	0	0	-	-	-	631
Agg	3	Empowering	0	80	0	0.0	-	0.0	508
Agg	4	Empowering	0	0	0	-	-	-	450

Table 5: Inter-annotator agreement (IAA) between the aggregated dataset and annotators for the *Contextual* category of labels to annotate archival metadata descriptions. The first two columns note the annotator whose labels were considered expected or predicted, respectively. The abbreviation “pos” is for “positive;” “neg,” for “negative.” The last column lists the number of files with annotations by both annotators for that row. Only annotator 3 applied the “Empowering” label.