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 mitigating the minoritization 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
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, Duvani 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 comple-
ment 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). 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
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)).

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), 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 priorit-

---

4I.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


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].


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


RDA Steering Committee. 2022. About RDA.


### 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 to 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 platform, 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 certain 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 entitled Apollinaire [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.

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...
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.
### 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.

<table>
<thead>
<tr>
<th>exp</th>
<th>pred</th>
<th>label</th>
<th>true pos</th>
<th>false pos</th>
<th>false neg</th>
<th>precision</th>
<th>recall</th>
<th>F₁</th>
<th>files</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>Unknown</td>
<td>5031</td>
<td>1524</td>
<td>4268</td>
<td>0.76751</td>
<td>0.54103</td>
<td>0.63467</td>
<td>584</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>Unknown</td>
<td>2776</td>
<td>537</td>
<td>432</td>
<td>0.83791</td>
<td>0.86534</td>
<td>0.85140</td>
<td>170</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>Unknown</td>
<td>1048</td>
<td>1421</td>
<td>315</td>
<td>0.42446</td>
<td>0.76889</td>
<td>0.54697</td>
<td>72</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>Masculine</td>
<td>2367</td>
<td>2372</td>
<td>1079</td>
<td>0.49947</td>
<td>0.68688</td>
<td>0.57838</td>
<td>584</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>Masculine</td>
<td>728</td>
<td>111</td>
<td>146</td>
<td>0.86770</td>
<td>0.83295</td>
<td>0.84997</td>
<td>170</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>Masculine</td>
<td>380</td>
<td>169</td>
<td>411</td>
<td>0.69217</td>
<td>0.48040</td>
<td>0.56716</td>
<td>72</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>Feminine</td>
<td>627</td>
<td>427</td>
<td>642</td>
<td>0.59488</td>
<td>0.49409</td>
<td>0.53982</td>
<td>584</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>Feminine</td>
<td>724</td>
<td>128</td>
<td>178</td>
<td>0.84977</td>
<td>0.80266</td>
<td>0.82554</td>
<td>170</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>Feminine</td>
<td>287</td>
<td>496</td>
<td>279</td>
<td>0.36654</td>
<td>0.50707</td>
<td>0.42550</td>
<td>72</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>Non-binary</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>584</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>Non-binary</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>170</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>Non-binary</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>72</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>Gendered Role</td>
<td>1802</td>
<td>306</td>
<td>882</td>
<td>0.85484</td>
<td>0.67139</td>
<td>0.75209</td>
<td>584</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>Gendered Role</td>
<td>1404</td>
<td>162</td>
<td>257</td>
<td>0.89655</td>
<td>0.84527</td>
<td>0.87016</td>
<td>170</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>Gendered Role</td>
<td>438</td>
<td>292</td>
<td>52</td>
<td>0.60000</td>
<td>0.89388</td>
<td>0.71803</td>
<td>72</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>Gendered Pronoun</td>
<td>3398</td>
<td>101</td>
<td>190</td>
<td>0.97113</td>
<td>0.94705</td>
<td>0.95894</td>
<td>584</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>Gendered Pronoun</td>
<td>869</td>
<td>70</td>
<td>60</td>
<td>0.92545</td>
<td>0.93541</td>
<td>0.93041</td>
<td>170</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>Gendered Pronoun</td>
<td>518</td>
<td>7</td>
<td>11</td>
<td>0.98667</td>
<td>0.97921</td>
<td>0.98292</td>
<td>72</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>Generalization</td>
<td>37</td>
<td>35</td>
<td>262</td>
<td>0.51389</td>
<td>0.12375</td>
<td>0.19946</td>
<td>584</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>Generalization</td>
<td>74</td>
<td>51</td>
<td>63</td>
<td>0.59200</td>
<td>0.54015</td>
<td>0.56489</td>
<td>170</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>Generalization</td>
<td>2</td>
<td>50</td>
<td>7</td>
<td>0.03846</td>
<td>0.22222</td>
<td>0.06557</td>
<td>72</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th>exp</th>
<th>pred</th>
<th>label</th>
<th>true pos</th>
<th>false pos</th>
<th>false neg</th>
<th>precision</th>
<th>recall</th>
<th>F₁</th>
<th>files</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3</td>
<td>Occupation</td>
<td>1988</td>
<td>613</td>
<td>724</td>
<td>0.76432</td>
<td>0.73303</td>
<td>0.74835</td>
<td>485</td>
</tr>
<tr>
<td>0</td>
<td>4</td>
<td>Occupation</td>
<td>738</td>
<td>396</td>
<td>240</td>
<td>0.65079</td>
<td>0.75460</td>
<td>0.69886</td>
<td>149</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>Occupation</td>
<td>422</td>
<td>327</td>
<td>134</td>
<td>0.56341</td>
<td>0.75899</td>
<td>0.64674</td>
<td>57</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>Omission</td>
<td>1376</td>
<td>914</td>
<td>3259</td>
<td>0.60087</td>
<td>0.29687</td>
<td>0.39740</td>
<td>485</td>
</tr>
<tr>
<td>0</td>
<td>4</td>
<td>Omission</td>
<td>416</td>
<td>317</td>
<td>875</td>
<td>0.56753</td>
<td>0.32223</td>
<td>0.41106</td>
<td>149</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>Omission</td>
<td>215</td>
<td>315</td>
<td>155</td>
<td>0.40566</td>
<td>0.58108</td>
<td>0.47777</td>
<td>57</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>Stereotype</td>
<td>505</td>
<td>539</td>
<td>227</td>
<td>0.48371</td>
<td>0.68989</td>
<td>0.56869</td>
<td>485</td>
</tr>
<tr>
<td>0</td>
<td>4</td>
<td>Stereotype</td>
<td>507</td>
<td>525</td>
<td>600</td>
<td>0.49127</td>
<td>0.45799</td>
<td>0.47405</td>
<td>149</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>Stereotype</td>
<td>34</td>
<td>60</td>
<td>161</td>
<td>0.36170</td>
<td>0.17435</td>
<td>0.23529</td>
<td>57</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>Empowering</td>
<td>0</td>
<td>80</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>485</td>
</tr>
<tr>
<td>0</td>
<td>4</td>
<td>Empowering</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>149</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>Empowering</td>
<td>0</td>
<td>80</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>57</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>exp</th>
<th>pred</th>
<th>label</th>
<th>true pos</th>
<th>false pos</th>
<th>false neg</th>
<th>precision</th>
<th>recall</th>
<th>F₁</th>
<th>files</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agg 0</td>
<td>Unknown</td>
<td>10561</td>
<td>36</td>
<td>1900</td>
<td>0.99660</td>
<td>0.84752</td>
<td>0.91604</td>
<td>714</td>
<td></td>
</tr>
<tr>
<td>Agg 1</td>
<td>Unknown</td>
<td>6608</td>
<td>0</td>
<td>4511</td>
<td>1.00000</td>
<td>0.59430</td>
<td>0.74553</td>
<td>597</td>
<td></td>
</tr>
<tr>
<td>Agg 2</td>
<td>Unknown</td>
<td>15140</td>
<td>117</td>
<td>679</td>
<td>0.99233</td>
<td>0.95708</td>
<td>0.97439</td>
<td>444</td>
<td></td>
</tr>
<tr>
<td>Agg 0</td>
<td>Masculine</td>
<td>3963</td>
<td>18</td>
<td>2446</td>
<td>0.99548</td>
<td>0.61835</td>
<td>0.76285</td>
<td>714</td>
<td></td>
</tr>
<tr>
<td>Agg 1</td>
<td>Masculine</td>
<td>4749</td>
<td>1</td>
<td>1099</td>
<td>0.99979</td>
<td>0.81207</td>
<td>0.89621</td>
<td>597</td>
<td></td>
</tr>
<tr>
<td>Agg 2</td>
<td>Masculine</td>
<td>1007</td>
<td>5</td>
<td>525</td>
<td>0.99506</td>
<td>0.65731</td>
<td>0.79167</td>
<td>444</td>
<td></td>
</tr>
<tr>
<td>Agg 0</td>
<td>Feminine</td>
<td>1454</td>
<td>19</td>
<td>523</td>
<td>0.98710</td>
<td>0.73546</td>
<td>0.84290</td>
<td>714</td>
<td></td>
</tr>
<tr>
<td>Agg 1</td>
<td>Feminine</td>
<td>1076</td>
<td>0</td>
<td>707</td>
<td>1.00000</td>
<td>0.60348</td>
<td>0.75271</td>
<td>597</td>
<td></td>
</tr>
<tr>
<td>Agg 2</td>
<td>Feminine</td>
<td>994</td>
<td>12</td>
<td>410</td>
<td>0.98807</td>
<td>0.70798</td>
<td>0.82490</td>
<td>444</td>
<td></td>
</tr>
<tr>
<td>Agg 0</td>
<td>Nonbinary</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>714</td>
<td></td>
</tr>
<tr>
<td>Agg 1</td>
<td>Nonbinary</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>597</td>
<td></td>
</tr>
<tr>
<td>Agg 2</td>
<td>Nonbinary</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>444</td>
<td></td>
</tr>
<tr>
<td>Agg 0</td>
<td>Gendered-Role</td>
<td>3108</td>
<td>697</td>
<td>330</td>
<td>0.81682</td>
<td>0.90401</td>
<td>0.85821</td>
<td>714</td>
<td></td>
</tr>
<tr>
<td>Agg 1</td>
<td>Gendered-Role</td>
<td>1924</td>
<td>218</td>
<td>716</td>
<td>0.89823</td>
<td>0.72879</td>
<td>0.80468</td>
<td>597</td>
<td></td>
</tr>
<tr>
<td>Agg 2</td>
<td>Gendered-Role</td>
<td>1471</td>
<td>652</td>
<td>230</td>
<td>0.69289</td>
<td>0.86479</td>
<td>0.76935</td>
<td>444</td>
<td></td>
</tr>
<tr>
<td>Agg 0</td>
<td>Gendered-Pronoun</td>
<td>3933</td>
<td>160</td>
<td>190</td>
<td>0.99914</td>
<td>0.94848</td>
<td>0.97315</td>
<td>597</td>
<td></td>
</tr>
<tr>
<td>Agg 1</td>
<td>Gendered-Pronoun</td>
<td>3498</td>
<td>3</td>
<td>190</td>
<td>0.99914</td>
<td>0.94848</td>
<td>0.97315</td>
<td>597</td>
<td></td>
</tr>
<tr>
<td>Agg 2</td>
<td>Gendered-Pronoun</td>
<td>1016</td>
<td>1</td>
<td>41</td>
<td>0.99902</td>
<td>0.96121</td>
<td>0.97975</td>
<td>444</td>
<td></td>
</tr>
<tr>
<td>Agg 0</td>
<td>Generalization</td>
<td>405</td>
<td>1</td>
<td>1370</td>
<td>0.99754</td>
<td>0.22817</td>
<td>0.37139</td>
<td>714</td>
<td></td>
</tr>
<tr>
<td>Agg 1</td>
<td>Generalization</td>
<td>69</td>
<td>4</td>
<td>1123</td>
<td>0.94521</td>
<td>0.05789</td>
<td>0.10909</td>
<td>597</td>
<td></td>
</tr>
<tr>
<td>Agg 2</td>
<td>Generalization</td>
<td>127</td>
<td>0</td>
<td>862</td>
<td>1.00000</td>
<td>0.12841</td>
<td>0.22760</td>
<td>444</td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>exp</th>
<th>pred</th>
<th>label</th>
<th>true pos</th>
<th>false pos</th>
<th>false neg</th>
<th>precision</th>
<th>recall</th>
<th>F₁</th>
<th>files</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agg 0</td>
<td>Occupation</td>
<td>2725</td>
<td>23</td>
<td>571</td>
<td>0.99163</td>
<td>0.82676</td>
<td>0.90172</td>
<td>631</td>
<td></td>
</tr>
<tr>
<td>Agg 3</td>
<td>Occupation</td>
<td>2320</td>
<td>290</td>
<td>873</td>
<td>0.88889</td>
<td>0.72659</td>
<td>0.79959</td>
<td>508</td>
<td></td>
</tr>
<tr>
<td>Agg 4</td>
<td>Occupation</td>
<td>1746</td>
<td>147</td>
<td>253</td>
<td>0.92235</td>
<td>0.87344</td>
<td>0.89723</td>
<td>450</td>
<td></td>
</tr>
<tr>
<td>Agg 0</td>
<td>Omission</td>
<td>5916</td>
<td>12</td>
<td>1187</td>
<td>0.99798</td>
<td>0.83289</td>
<td>0.90799</td>
<td>631</td>
<td></td>
</tr>
<tr>
<td>Agg 3</td>
<td>Omission</td>
<td>2310</td>
<td>13</td>
<td>3475</td>
<td>0.99440</td>
<td>0.39931</td>
<td>0.56981</td>
<td>508</td>
<td></td>
</tr>
<tr>
<td>Agg 4</td>
<td>Omission</td>
<td>1876</td>
<td>5</td>
<td>967</td>
<td>0.99734</td>
<td>0.65987</td>
<td>0.79424</td>
<td>450</td>
<td></td>
</tr>
<tr>
<td>Agg 0</td>
<td>Stereotype</td>
<td>1748</td>
<td>11</td>
<td>1058</td>
<td>0.99375</td>
<td>0.62295</td>
<td>0.76583</td>
<td>631</td>
<td></td>
</tr>
<tr>
<td>Agg 3</td>
<td>Stereotype</td>
<td>1089</td>
<td>9</td>
<td>279</td>
<td>0.99180</td>
<td>0.79605</td>
<td>0.88321</td>
<td>508</td>
<td></td>
</tr>
<tr>
<td>Agg 4</td>
<td>Stereotype</td>
<td>1400</td>
<td>2</td>
<td>715</td>
<td>0.99857</td>
<td>0.66194</td>
<td>0.79613</td>
<td>450</td>
<td></td>
</tr>
<tr>
<td>Agg 0</td>
<td>Empowering</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>631</td>
<td></td>
</tr>
<tr>
<td>Agg 3</td>
<td>Empowering</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>508</td>
<td></td>
</tr>
<tr>
<td>Agg 4</td>
<td>Empowering</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>450</td>
<td></td>
</tr>
</tbody>
</table>