LET'S TALK DATA, BIAS, AND MENSTRUAL CRAMPS: VOICING GERWOMANNESS IN THE NINETEENTH CENTURY AND TODAY

Jana Keck
GERMAN HISTORICAL INSTITUTE

An increasing amount of research in the humanities involves analyzing machine-readable data with computer software. The creation of born digitals such as social media, digital art, or e-books as well as the digitization of maps, literary works, images, government census data or historical newspapers allows scholars to study digital representations of cultural heritage data at scale and to combine close and distant, macro- and microscopic, quantitative and qualitative approaches. My PhD project "Text-Mining America's German-Language Newspapers, 1830-1914: Processing Ger(wo)manness" (Working Title) connects methods from text mining, machine learning, network analysis, and close reading in a German-American studies context. In my project, I aim to examine reprinting practices in German-language newspapers in the United States by analyzing transnational and transcultural text production and circulation. The nineteenth century was marked by scissors-and-paste journalism, that is, editors and journalists (re-) produced material from other venues, often without attribution to the original authors or sources.[1] What kind of texts were reprinted in the German immigrant papers across states and decades? How did information of a political, scientific, economic, literary or religious nature circulate in the public sphere? To answer these questions, I develop and use a statistical language model to identify and cluster texts that were printed several times in the same and different newspapers. Due to the abundance and diversity of reprinted texts, or unstructured data, machine learning techniques are applied to classify texts into genres such as, for instance, hard news, ads, poems or lists. This approach has theoretical and pragmatic reasons: first, computationally classifying texts seeks to better understand the nature and role of genres, their similarities and anomalies, and how they relate to migrant groups that differ in sex, gender, age, or class. Second, the objective is to create an archive of reprinted texts and make it available for further research in the field.

This article, however, is not about "measuring" text similarity in order to find these reprinted texts, nor about explaining artificial

1 My PhD project developed out of a larger international Digital Humanities research project, which brought together a research team of scholars from seven countries in Europe and the Americas to study structural, textual, and conceptual network systems of digitized historical newspaper collections. See Oceanic Exchanges Project Team, *Oceanic Exchanges: Tracing Global Information Networks In Historical Newspaper Repositories, 1840-1914* (2017), DOI 10.17605/OSF.IO/WA94S.

intelligence – trying to teach the properties of genres to the computer – to automatically classify texts into newspaper genres.[2] Instead, this article is a data feminist critique of the nineteenth-century German-language press in the United States and prioritizes that which has been systematically neglected or remained hidden: migrant women's lives and their experiences. Since the nineteenth century, scholars have emphasized the role of the immigrant newspaper as the voice and mirror of German ethnic life while predominantly focusing on the HIStories of the German-language press. Did women, their perspectives, concerns, issues or problems never figure in the pages of the immigrant newspapers? Writing HIStories may be valid when we examine the editors, but it appears to be unjustified when we dig into the newspapers' textual realm. This article takes up the challenge of understanding feminism as a political as well as a methodological question by quantitatively and qualitatively uncovering inequalities in a structured manner in order to determine decision-making processes and to (re-)create datasets of minoritized groups.

Researchers in data science use scientific methods, processes, algorithms and systems to extract knowledge and insights from many structured and unstructured data. As data scientists Catherine D'Ignazio and Lauren F. Klein illustrate in *Data Feminism,* "[m]any people think of data as numbers alone, but data can also consist of words or stories, colors or sounds, or any type of information that is systematically collected, organized, and analyzed. The term science in data science simply implies a commitment to systematic methods of observation and experiment."[3] Structured data is usually held in a database in which all values have identifiers and clear relations. They resemble clearly defined categories, such as, for instance, "sex: female" and "age: 30." Since such data is highly organized and formatted, it is easily searchable in relational databases. However, every categorization is an act of interpretation and subject to change as the category "gender" nowadays most clearly illustrates. Plain text can be seen as an example of unstructured data because the boundaries of individual items, the relations between them, and their meaning, are mostly implicit. Unstructured data is information that either does not have a predefined data model or is not organized in a predefined manner. Both types of data – structured and unstructured – are not free from bias because "data are never neutral; they are always the biased output of unequal social, historical, and economic conditions."[4] From historical case studies about women as

2 My thesis will consist of two parts: the first will describe the data and computational models developed and used for this research of computationally detecting and clustering text reuse and classifying genres. The second part will deal with specific viral events to show how sexist, racist, and nationalistic ideas spread across states and decades.

3 Catherine D'Ignazio and Lauren F. Klein, *Data Feminism* (Cambridge, 2020), 14.

4 D'Ignazio and Klein, 39.

computers to the latest Amazon scandals regarding racist AI models, *Data Feminism* offers a bracing and pithy set of intersectional feminist provocations around the political aesthetic of data science to distill and dismantle the race/class/sex and other normative parameters of data discourses. The authors' work raises attention to increase research foci on data ethics based on the observation that data reflects the same old oppression. Data can help communicate intersectional feminism as diverse projects which name, challenge and change forces of oppression and are not only concerned with women and gender, but also with questions of power.

Bias can be introduced by the source material, the collecting of the source material and, additionally, by methods of analysis. In her work, *Scholarship in the Digital Age*, information scientist Christine Borgmann emphasizes that the potential of using data for humanistic inquiry is not only a technical, but especially a theoretical, methodological and a social issue.[5] What scholars produce and label as bias is predominantly shaped by their own mental machinery and less by the cost of information or limited computing power. Since bias is a component of the human thought process, data collected from humans therefore inherently reflects that bias. Who gets to be remembered and historicized by ways of (digital) record creation and analysis? Who is forgotten or silenced in history by way of omission or even destruction of records?[6] The first section of this article lays out the opportunities presented by working with a digitized edition of historical newspapers, describes some challenges of bias from data to analysis as well as limitations of keyword searching the (digital) archive. The second part briefly describes the objective of creating a digitized archive of reprinted texts to demonstrate that studying what people shared across states and decades can be seen as focusing on diverse migrant groups. Simultaneously, this part shows that counting is not enough to both unveil normative forces and recover women's concerns. The article does not describe the algorithmic procedures and models in detail, but rather provides a selective methodological and, thus, theoretical zooming in and out of my project using the example of the genre of advertisements. In order to foreground women's representations in the immigrant newspaper and to expose the biases embedded in the GerMANness of earlier accounts, the third part is framed around a case study of GerWOMANness, a viral event about menstrual cramps, female weakness, and drug abuse. I use GerWOMANness as a category to describe the collection of reprinted texts for, by, and

5 Christine Borgmann, *Scholarship in the Digital Age: Information, Infrastructure, and the Internet* (Cambridge and London, 2010).

6 See Sam Winn, "The Hubris of Neutrality in Archives," *On Archivy* (Blog), April 24, 2017, https://medium.com/on-archivy/the-hubris-of-neutrality-in-archives-8df6b523fe9f. Issues of representation and power are fundamentally rooted in archival work.

about women from different genres. This article seeks to illustrate how virality and genre through close and distant reading can tell us more about collective experiences of women as well as gender roles, stereotypes, and sexism.

I. Naming and challenging GerMANness: What do data say about inequality?

German-language newspapers in the United States flourished during the era of nineteenth-century transatlantic mass migration. These papers became the most widely read and influential non-English newspapers in the United States and, by disseminating local, national, and transnational news and information, helped immigrants both to assimilate into the new environment and to maintain ties with their country of origin. In this sense, they functioned as powerful tools for retaining language and preserving culture while influencing public opinion and setting the agenda for what (story) became news and how it was constructed (discourse). So far, research on the German-language press in the United States can be summarized as relatively outdated and limited not necessarily in terms of scope, but in terms of the connections between the individual newspapers, the diversity of genres and their relation to migrant groups that differ in sex, gender, class or age. [7] The majority of studies represent overwhelmingly white and male migration experiences by placing emphasis on the editors and writers and their influence on U.S. politics and economy. Research on the topic (1) was mostly published in the twentieth century, (2) focuses on individual newspapers, or (3) covers periods before the 1860s or towards the beginning of the 1900s.[8] Women and their issues seem to occupy comparatively little space in newspapers, or at least in the histories of these newspapers.[9]

As an object of analysis, media can become a powerful instrument for highlighting the inequalities in the balance of power that have historically characterized the relationships between women and men. According to historian Karen Offen, the history of feminism is political history, more specifically, it is "a more expansive history of politics that incorporates women and analyzes gender politics."[10] Yet, the task of recovering women's voices is complicated by normative forces that shape our access to women's histories since archives are built upon and reflect systems of imperial and patriarchal power. "Working with data from a feminist perspective requires knowing

7 The only study pertaining to multiple states to date is Christopher Dolmetsch, "Locations of German Language Newspaper and Periodical Printing in the United States: 1732-1976," *Monatshefte* 68:2 (1976): 188-195, which uses the statistical material in Karl John Richard Arndt and May E. Olson, *German-American Newspapers and Periodicals, 1732-1955* (Heidelberg, 1961) to create maps of the periodicals' geographic distribution as comparative material about German settlement to the maps provided by the U.S. Bureau of the Census.

8 See, for instance, (1) Hermann Boeschenstein, "Zum Studium der deutschen Zeitungen in Amerika," *Monatshefte für Deutschen Unterricht* 26:4 (1934): 103-08. JSTOR, www.jstor.org/stable/30168861. Accessed January 13, 2019; Carl Wittke, *The German-Language Press in America* (Lexington, 1957); James M. Bergquist, "The German-American Press," in *The Ethnic Press in the United States: A Historical Analysis and Handbook,* ed. Sally M. Miller (Westport, 1987), 131-160; Elliott Shore et al. *The German-American Radical Press* (Chicago, 1992); (2) Steven Rowan, "The German press in St. Louis and Missouri in the Nineteenth Century: The establishment of a tradition," *The Papers of the Bibliographical Society of America* 99:3 (2005): 459-467; Frank Boles, "Michigan Newspapers: A Two-Hundred-Year Review," *Michigan Historical Review* 36:1 (2010): 31-69; Kathleen Condray, "Arkansas's Bloody German-Language Newspaper War of 1892," *The Arkansas Historical Quarterly* 74:4 (2015): 327-51; (3) Henry John Groen, "A Note on the German-American Newspapers of Cincinnati

and acknowledging this history. (…) Data are part of the problem, to be sure. But they are also part of the solution."[11] Does the digital newspaper archive allow scholars access to and analysis of different stories?[12] How do we find textual evidence, especially since the practice of searching for information in both digital and physical archives remains a mostly "undertheorized" practice in the humanities.[13] Where are the women in the newspapers? They are there and you do not really have to dig very deeply. How are they represented as writers, readers, users, and news themselves? Answering these questions is a rather cumbersome task. This article is both descriptive and theoretical, starting with a keyword search in a digital archive to describe selection options and increase scholarly awareness of representational biases introduced by the labor "behind the archive." This step is essential for a data ethical framework because documenting and reflecting on these decisions lead to clarity and to defining what we will or will not do, can and cannot do. This article therefore illustrates how we can use data as "a check-in, (…), a resource to begin and continue dialogue"[14] in gender studies and beyond.

Searching in a physical or a digital newspaper archive for "data" is like going on a treasure hunt. Both "searches" depend on the prior knowledge of the scholar, who decides which keywords to enter as well as that of the (digital) archivist, who collected and structured the material and entered information about the source – structured data – into the database. Even though the scholar's as well as the archivist's labor play a role in the analogue and the digital archive, the digitized repositories open up a different perspective on representation, comparison, and search methods. A digital newspaper archive such as Chronicling America brings together sources from different decades (1777-1963), places (52 states), ethnicities (33) and languages (22), and offers keyword searches of the textual content

10 Karen Offen, "The History of Feminism is Political History," *Perspectives on History,* May 1, 2011.

11 D'Ignazio and Klein, 17.

12 For current projects working with digitized historic newspaper archives in Europe, see Clemens Neudecker and Gregory Markus, eds., "Newspapers," *Europeana Pro* 16 (2020), https://pro.europeana.eu/page/issue-16-newspapers.

13 Ted Underwood, "Theorizing Research Practices We Forgot to Theorize Twenty Years Ago," *Representations*, 127:1 (2014): 64-72.

14 M. Cifor et al., "Feminist Data Manifest-No," 2019, https://www.manifestno.com.

15 Chronicling America, Library of Congress, accessed January 2, 2021, https://chroniclingamerica.loc.gov.

before 1860," *Monatshefte für deutschen Unterricht* (1945): 67-71; Peter Conolly-Smith, *Translating America: An Immigrant Press Visualizes American Popular Culture, 1895–1918* (Washington, 2004); Daniel Stein, "Transatlantic Politics as Serial Networks in the German-American Mystery Novel, 1850–1855," in *Traveling Traditions: Nineteenth-Century Cultural Concepts and Transatlantic Intellectual Networks,* ed. Erik Redling (Berlin, 2016), 247-265.

9 There are few works on the role of women in German-language newspapers in the United States, even though there were papers published by women such as, for instance, *Deutsche Frauen Zeitung* (starting 1851) by Mathilde Franziska Anneke. "No copies of *Deutsche Frauen Zeitung* are known to exist so we cannot know what its substance was. But our records of Frau Anneke document her understanding of women's unique collective situation and her desire for a genuine people's movement in which women would be equal to men." See Charles Cantrell, "Wisconsin's First Newspaper … by Women," *Quixote* 8:3 (1974): 6.

to produce one or more documents that contain those words or phrases.[15] Each hit corresponds to one newspaper page and typically you need to read a document to decide if is relevant or not. Thereby, Chronicling America allows access to multiple digitized newspapers, that is, data (plain text as unstructured data) and metadata (date of publication, location or title). As literary scholar Ryan Cordell argues, the digitized "newspaper is not simply a surrogate for the material object, but instead constitutes a new edition of text with both new affordances – such as the ability to analyze patterns and text strings across the corpus – and limitations – such as leveling out of scale among objects that, in their printed editions, vary widely in size and format."[16]

Even though using digital repositories sounds promising in terms of quantity, these new formats of nineteenth-century newspapers pose a number of additional questions related to biases ranging from investigation, analysis, and synthesis to the presentation of information in electronic form: Which methods have been used to convert the printed text into machine-encoded text? What is the quality of the data?[17] What has not been digitized? Which algorithms assist in extracting, processing, mining and presenting data in the digital collection? Who is represented in the data and the metadata? An ethical framework of research with digital collections refers to the process that scholars execute to decide on and document for the creation, curation and analysis of data. Even though data ethics builds on the foundation provided by computer and information ethics, it refines the approach by shifting the level of abstraction of ethical enquiries, from being information-centric to being data-centric. Information-centric approaches focus on what is being digitized and how we can search and analyze it. Data-centric approaches expand these objectives by examining who is – and who is not – represented in the corpus. Both approaches start with examining metadata to receive information about the corpus.

Different metadata such as "language" (categorical data which takes quantitative values with qualitative characteristics) and "page numbers" (numerical or quantitative data) can be used to gather information about the corpus by examining the relationship of different categories. As illustrated in figure 1, according to the metadata, the German-language newspapers make up the majority of non-English newspapers in Chronicling America. Earlier scholarship has already used numeric data such as sales and circulation figures to argue

16 Ryan Cordell, "Reprinting, Circulation, and the Network Author in Antebellum Newspapers," *American Literary History*, 27:3 (2015): 417-445.

17 This step refers to Optical Character Recognition (OCR). OCR is a field of research in pattern recognition, artificial intelligence and computer vision. When records are digitized, scanning is only the first step. The software creates an image of the document, but that image, and the data that composes it, is neither editable nor searchable. OCR is a type of software that converts those scanned images into structured data that is extractable, editable and searchable. Thus, OCR involves two steps: converting print text into machine-encoded text as well as creating information about this data: metadata. In short, metadata is data about data. Many distinct types of metadata exist, including descriptive metadata, structural metadata, administrative metadata, reference metadata and statistical metadata.

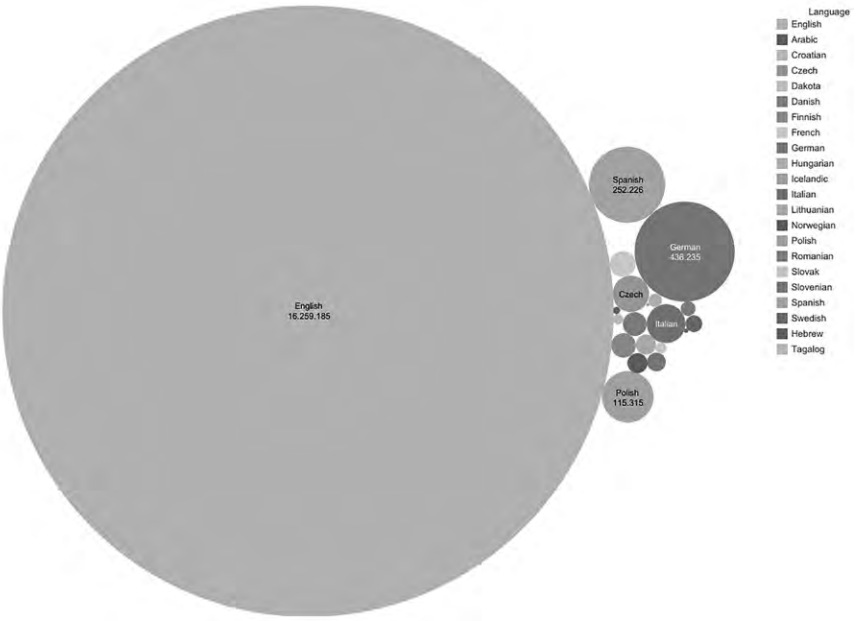


Figure 1. Image of interactive visualization using metadata about "language" and "page numbers" of the corpus in *Chronicling America*. The visualization shows the distribution by language of newspaper pages published between 1690-1963. Data as of 09/12/2020.

that they were the most read and influential non-English newspapers. But influential for whom? Data are not just simple statements of fact. Means of visual representations, government census data or sales figures are prone to leaving out important context and meaning. In 1961, Karl J. R. Arndt and May E. Olsen published *German-American Newspapers and Periodicals 1732-1955* (1965), which is a detailed record of about five thousand newspapers and periodicals, with exact dates of changes and titles of names of editors and publishers, if possible.[18] This dataset provided a first conceptual framework across time and space and emphasized the value of rich statistical material. It encouraged scholarly investigation in the following decades. The majority of these studies, however, have only used this dataset and the rich archival material of newspapers (in non-digital form) for site-related and genealogical work. As historian Wolfgang Helbich wrote in 2009, "[l]arge segments of sources like German-American newspapers and other periodicals are almost terra incognita."[19] While Arndt and Olson's works provided metadata about the male-run newspaper business, Chronicling America allows the user to move beyond the socio-historical perspective of sales figures, individual newspapers' publication periods or their editorship and into the newspapers' textual realm. Although Chronicling America has transformed access to these newspapers, search

18 Arndt and Olson.

19 Wolfgang Helbich, "German Research on German Migration to the United States," *Amerikastudien/American Studies* 54.3 (2009): 383-404.

and analysis within these big sets of structured and unstructured data still remain undertheorized.

As D'Ignazio and Klein have written, a "key way that power and privilege operate in the world today has to do with the word data itself."[20] Data is predominantly not seen as information, but rather as fact or evidence to serve a rhetorical purpose. Visual theorist Johanna Drucker has argued that the etymological root of the term data, "that which is given," does not adequately represent its meaning, and prefers to speak of "capta," which literally means "that which has been captured or gathered."[21] Every act of capturing data, both textual and numeric, is already oriented toward certain goals, performed with specific instruments and driven by a specific attention to a small part of what could have been captured, given different goals and instruments. To some degree, we can trace these processes by examining, for instance, the metadata of the corpus. Using a digital archive does not only imply the investigation of data but presupposes that we reflect on algorithmic performance as well. On the page "All Digitized Newspapers 1777-1963," Chronicling America allows users to select "language" and "ethnicity" in order to examine which newspapers the project has digitized and made available on the site: (1) if we choose "location: all states," "ethnicity: German," and "languages: all languages," we receive the following information: "59 newspapers filtered on German are available for viewing on this site."[22] (2) If we select "all states," "all ethnicities," and "languages: German," the returned results are 74 newspapers and (3) if we combine the two approaches by selecting "all states," "German," and "German," the result lists 59. Why do the first and the last selection option return the "same" result even though each of the three steps is different? While in general, algorithm refers to any special method of solving a certain kind of problem, in computer science it stands for a computable set of steps to achieve a desired result. The sequence of steps is based on performing decisions, which are dependent on order. The reason for (1) and (3) returning the same results even though we have three distinct selections is hierarchical organization of procedures in programs. The command does what it is documented to do, which is always prioritizing "ethnicity" when selecting several features. Moreover, there can be missing metadata in digital archives such as, for instance, information about the language of a newspaper title.[23] Since there is the potential of unreliable results due to missing or

20 D'Ignazio and Klein, 10.

21 Johanna Drucker, "Humanities Approaches to Graphical Display," *DHQ: Digital Humanities Quarterly* 5:1 (2011).

22 Chronicling America, accessed December 23, 2020.

23 Chronicling America's data and metadata can be accessed through the site's API. The Library of Congress also provides access to bulk data, which can be retrieved through web crawling tools.

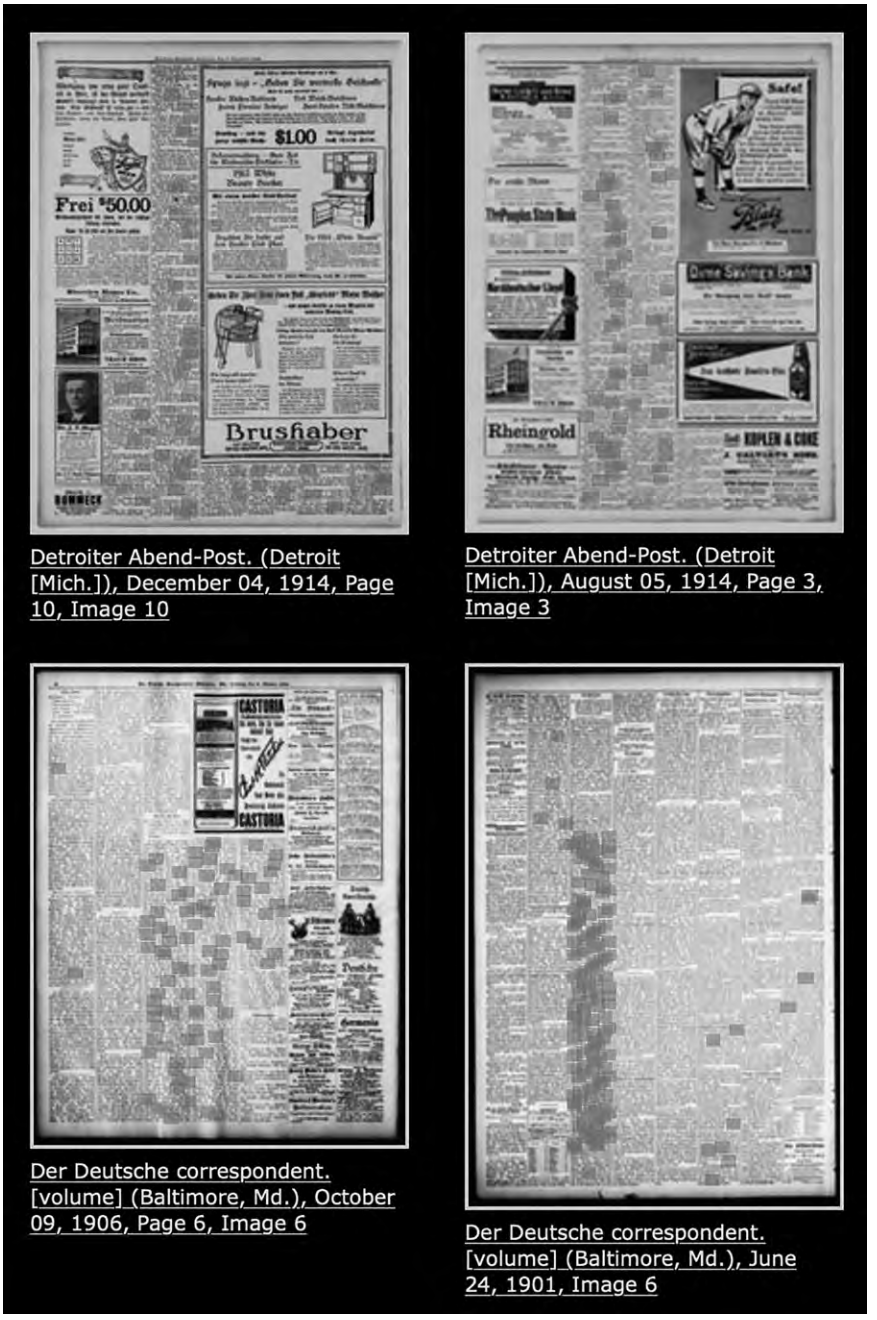
incorrect data, the dataset for my project was preprocessed using a language classifier to extract only German-language texts.

Apart from examining the metadata, users can use full-text search by entering keywords or phrases and select some features (time frames, newspaper titles or languages), but no longer ethnicity. This shows a further example of hierarchical organization in Chronicling America. For the full-text search, the machine-readable text of the newspapers (OCR-derived, page-level text data is provided in both TXT and XML formats) is used. Additionally, users have the possibility to study the page images (PDF and JPG2). Even though researchers have the possibility to study several representations of the newspaper page, the first results, after entering a word or phrase, that the user sees are the digitized images, even though the latter is used for information retrieval. This prioritization can be seen as an act of manipulation because the image recalls memories of the user reading the paper in the physical archive. When keyword searching the archive, all representation options result from a specific data structure and offer both advantages and disadvantages. The image, for instance, can be very useful for a first analysis of word hits in documents and the location of their occurrences. If we select "language: German," "time frame: 1830-1914" and "search term: *Frau*," it returns 257,107 hits, where one hit refers to the number of newspaper pages, in which the word occurs at least once.[24]

As the red spots in figure 2 illustrate, the word does not occur once, but several times on one newspaper page. If I use the same filters, but enter *Mann,* it returns 298,547 hits. As regards hits, the search terms only have a difference of 13.88%. *Mutter* (142,839) to *Vater* (147,807) has a reduction of 0.82%. While the machine-readable textual depiction provides the "numbers" of occurrences (all documents), the digital images provide a first overview into the textual realm and where the terms are embedded in the newspaper page. This analysis shows that, as regards occurrences in the corpus and even the individual newspaper page, there is not a huge gap between the "binary" sexes and there seems to be quite a massive amount of evidence available to be studied.

In *History in the Age of Abundance*? the historian Ian Milligan has argued that we must be prepared to understand and evaluate the tools and platforms we use and adapt to a research environment

24 Currently, there are 436,235 newspaper pages from 1834 to 1954 according to the metadata "language: German." See Chronicling America, accessed December 23, 2020.



Detroiter Abend-Post. (Detroit [Mich.]), December 04, 1914, Page 10, Image 10

Detroiter Abend-Post. (Detroit [Mich.]), August 05, 1914, Page 3, Image 3

Der Deutsche correspondent. [volume] (Baltimore, Md.), October 09, 1906, Page 6, Image 6

Der Deutsche correspondent. [volume] (Baltimore, Md.), June 24, 1901, Image 6

Figure 2. First images displayed of 257,107 hits when entering *Frau* in Chronicling America, sorted by "relevance" (accessed December 31, 2020).

characterized not by the scarcity of primary sources but by quantities of data too vast for any human to read.[25] Keyword searches possess the seemingly contradictory weaknesses of finding too few documents (under-inclusion) and finding too many documents (over-inclusion). Using the search term *Frau,* for instance, results in a case of over-inclusion. If we assume that one needs ten minutes in order to read a newspaper page, to read all documents that contain the word *Frau*, one needs approximately 42,851 hours (5 years of reading without breaks). Apart from these "time issues," there are several more obstacles that can exacerbate corpus creation and analysis when keyword searching the digital archive: there is, for instance, the likelihood of missing some documents because of OCR errors, the problem of selecting the "right" terms due to ambiguity, or diachronic language change. Most importantly, reading the plain text can be tedious because there is no separation between articles due to page-level OCR.

The first steps of searching in and analyzing the digitized newspaper as a historical source already involve a variety of quantitative methods such as using search algorithms and filtering options. The objective of documenting and reflecting on the datasets and diverse decision-making processes is to investigate biases – even our own – in order to conceptualize or rethink the next (re-)search steps. This analysis shows that women – in terms of the occurrences of the word that represents this sex – are visible in the newspapers' textual realm and not really underrepresented compared to men. However, the figures do not say anything about how women are represented: are they mentioned or described in hard news about crimes, advertisements of beer products, nationalistic poems about Germania or lists of newly arrived immigrants? As this example has shown, keyword search algorithms can become problematic in the case of over-inclusion and complicate finding relevant texts for, by and about women.

II. Creating an archive of GerWO/MANness

Many scholars in nineteenth-century periodical studies emphasize the high level of exchange between different publications and advocate for studying what was reprinted in order to understand what editors and readers of the period valued.[26] Scholars have not, however, until very recently, explored the nature of these reprinting practices due to simple pragmatism. Research in periodical studies

25 Ian Milligan, *History in the Age of Abundance?* (Montreal, 2019).

26 See Meredith L. McGill, *American Literature and the Culture of Reprinting, 1834-1853* (Philadelphia, 2007).

has started to use computational methods because no one can read every newspaper ever printed along with the personal insight of the editors or printers. [27] Since I do not have five years to read (without breaks) all pages that contain the word "woman," how can I use quantitative methods (which does not exclude qualitative ones) using the data from Chronicling America and Arndt and Olson to both examine the reprinting practices in these newspapers and focus on multiple social categories? To trace such a circulation system is to map a migrant social network.[28] In my project, I conceptualize historical phenomena in contemporary contexts by developing a model that seeks to study "viral events" in the German immigrant community, applying Ryan Cordell's terminology for describing reprinting practices in the nineteenth century.[29] A text becomes viral since editors do not only reprint it but also change and modify textual properties for various political, economic or social purposes. This section will briefly describe the underlying mechanisms for identifying and clustering reprinted texts and elaborate on the need for an automatic genre classification of these clusters. Additionally, this section is followed by an exploratory data analysis – a distant feminist reading – designed to recover texts for, by, and about women. A data feminist approach seeks to expose and challenge unequal power by looking at many axes of inequality and focuses on (algorithmic) subjectivity and the role of the human and non-human in data science. In my project, I am not only interested in detecting which texts had the longest journeys, but especially if and how these traveling texts of political, economic, scientific, or literary nature can provide information about a rather comprehensive network of immigrants that differed in sex, gender, class, age or (national) identities. Therefore, the digital archive of reprints will consist of texts for, by and about women, men, children, young and old, fiction writers, business owners, and many more.

I use passim software to curate the dataset of German-language newspapers from Chronicling America into a dataset that only contains reprinted texts. The program aligns similar passages of text by detecting passages of similar textual content.[30] Basically, the algorithm looks into each document and if it finds that similar sequences of characters occur in different documents, it takes out these different "textual snippets" and combines them into a cluster. When it clusters these texts, it also copies the metadata, that is, information about the text's source such as the name or place of publication, date or the link to Chronicling America. The reprinting

27 In 2014, the Viral Texts Project, which at the time was the largest-scale study of how content spread through the news networks of the nineteenth century, started to develop a text reuse detection algorithm to analyze 500 digitized English-language newspapers in the U.S. They found that about 650 articles were reprinted 50 times or more, which is a working definition of "viral" in the industrial age. See Ryan Cordell and David Smith, *Viral Texts: Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines* (2017), http://viraltexts.org.

28 See Ruth Ahnert et al., *The Network Turn: Changing Perspectives in the Humanities* (Cambridge, 2020), https://doi.org/10.1017/9781108866804. Networks are a category of study uniting diverse disciplines through a shared understanding of complexity in our world.

29 Ryan Cordell et al., *Going the Rounds: Virality in Nineteenth-Century American Newspapers* (Minneapolis, 2020), https://manifold.umn.edu/projects/going-the-rounds.

30 The model uses shingling techniques, based on indexing n-grams to find document pairs that share a large number of n-grams. See David A. Smith, Ryan Cordell, and Elisabeth Maddock Dillon, "Infectious Texts: Modelling Text Reuse in Nineteenth-Century Newspapers," *Proceedings of the Workshop on Big Humanities* (2013), 86–94.

dataset is now smaller than the Chronicling America corpus and, more importantly, the non-existent complete corpus of German-language newspapers, which would make up approximately 2,000 newspapers for this time frame.[31] Even though Chronicling America is still digitizing more newspapers, the current corpus (approx. 500,000 texts) can be considered representative for the analysis of viral content because it encompasses several states (17) and decades (1830-1914).

Why is one of my goals to create a digital archive of reprinted texts? As it turns out, German-language newspapers shared a huge amount of information even though publication locations were far apart. The sheer quantity of reprinted material makes it a daunting task to do empirical research and to close and distant read reprinted texts. As Cordell identifies the challenge: "When every nineteenth-century newspaper brims with original and reprinted content of all kinds, it is difficult to know where to even begin studying that content."[32] Moreover, examining texts that have the highest number of reprints or the longest reprinting period reveals that these are not necessarily texts about women's concerns. This comes as no surprise, because the majority of reprinted texts belong to the genre of hard news: in other words, predominantly texts by, for, and about men. However, when digging further, there are also reprint clusters that I would classify as ads, poems, jokes, factual texts, or lists to study women's representations. In order to recover these texts and add more metadata to the reprinting clusters, I develop and use a classification model to automatically categorize texts into genres.[33] As mentioned earlier, this approach has a theoretical implication because genre is a key category in the construction of social reality through the news, thereby providing ways of seeing how communicants perform identities and mediate situations and to analyze power and difference. Genre is not a fixed or clearly defined idea but a reflection of an audience's assumptions and wants at a certain point in time. Some genres presuppose a stronger sense of factuality, while others are expected to be entertaining instead. Additionally, there is a pragmatic reason for analyzing style and classifying texts into genres using machine learning techniques. Similar to streaming platforms such as Netflix, this additional metadata will offer users an expanded search through genre filtering.

This article does not explain the details of the genre classification model using different unsupervised and supervised machine

31 Arndt and Olson.

32 Ryan Cordell, "Reprinting, Circulation, and the Network Author in Antebellum Newspapers," *American Literary History*, 27:3 (2015): 417-445.

33 I am developing the genre classifier in collaboration with André Blessing, Institute for Natural Language Processing, University of Stuttgart.

learning techniques, which means that I am switching between annotating and allowing the model to work on its own to discover patterns. However, I want to emphasize that computationally classifying genres does not mean that a text will be 100% categorized as belonging to one type.[34] The model does not provide binary results such as text A is or is not hard news, but rather describes that a text may have the likelihood of belonging 0.1 to genre X, 0.2 to genre Y and 0.7 to genre Z. Investigating these properties is essential to comparing genres and to understanding how they develop over time because genres change and fall out of fashion while new ones emerge. The genre that receives the highest score will become decisive for the publication process of the reprinting archive. Similar to digital film, book, or music stores, which offer genre categories that define the entire search and selection process, this classification has to be seen as an act of my interpretation of genre because "learning algorithms rely on examples rather than fixed definitions."[35] With this method, I have identified ten different genres: hard news, ads, factual texts, lists, notifications, religious texts, jokes, poems, short stories, and novels. Once digitally published, this new edition of German-American literature can be read by the public or used for further research in the field.

Do we expect genres to favor the representation of, for instance, a specific gender? To illustrate a distant feminist reading, let us look at the frequencies of the terms *Frau*, *Dame*, *Mann,* and *Herr* as well as some adjectival complements to these nouns in the non-genre classified reprinting dataset first. The most frequent occurrences of the four nouns and their adjectives (figure 3, row no. 1) show a similar balance between woman and man as to the one found in keyword searching Chronicling America. However, examining the 29 next most frequent occurrences unveils that women prevail. In terms of adjectival complements, both sexes share characteristics of the categories age, manners, status, appearances, nationality (*deutsche/r*) as well as race (*weiß/e*). However, there are many differences as to how they are described within those categories. As regards appearances, while women are described as beautiful or pretty, small, veiled, men are depicted as strong and powerful. When examining the manners, the sexes seem to share attributes such as being kind, educated, or unhappy, but unlike men, women are not accompanied by adjectives relating to honesty and goodness. Additionally, there are three categories, which are not found in

34 This reflects Derrida's approach, who argues that texts do not belong to a genre, but they participate in at least one genre. See Jacques Derrida, "The Law of Genre," *Critical Inquiry* 7:1, trans. Avital Ronell (1980): 212.

35 Ted Underwood, "Machine Learning and Human Perspective," *Varieties of Digital Humanities* 135.1 (2020): 92-109. I will also let other scholars manually classify genres to see how they differ from my conceptualization.

Word Frequencies of the Terms Frau, Dame, Mann and Herr in the Unclassified Reprinting Dataset (pos. 1-30 of 4085 in total)

No	Search Result Frau	No. of occurrences	Search Result Dame	No. of occurrences	Search Result Mann		Search Result Herr	
1	junge Frau	3171	junge Dame	3108	junge Mann	4350	„Aû Herr	3142
2	gn√§dige Frau	2264	jungen Dame	1236	junger Mann	3442	alte Herr	2172
3	alte Frau	1479	alte Dame	1086	jungen Mann	2020	lieber Herr	1037
4	jungen Frau	1247	alten Dame	327	alte Mann	1131	werther Herr	503
5	feine Frau	930	√§ltere Dame	152	alter Mann	617	Gn√§diger Herr	417
6	feiner Frau	717	gekleidete Dame	129	alten Mann	516	Geehrter Herr	396
7	arme Frau	653	vornehme Dame	125	reicher Mann	460	junge Herr	388
8	sch√∂ne Frau	555	sch√∂ne Dame	118	lieber Mann	369	alter Herr	387
9	Frau	505	reiche Dame	66	Ter Mann	278	junger Herr	263
10	liebe Frau	498	vornehmen Dame	62	gro√üer Mann	247	gn√§dige Herr	232
11	alten Frau	474	√§lteren Dame	56	ehrlicher Mann	238	verehrter Herr	171
12	gute Frau	417	verschleierte Dame	55	Armer Mann	237	Ter Herr	163
13	kleine Frau	365	gro√üe Dame	54	gute Mann	235	bester Herr	144
14	ungl√ºckliche Frau	309	feine Dame	47	arme Mann	230	gekleideter Herr	143
15	gn√§digen Frau	303	h√ºbsche Dame	47	sch√∂ner Mann	191	hohe Herr	135
16	sch√∂nen Frau	231	√§ltliche Dame	47	braver Mann	186	ber Herr	118
17	armen Frau	222	bie Dame	45	ber Mann	152	„Aï Herr	113
18	erste Frau	211	erste Dame	44	armen Mann	148	fremde Herr	103
19	deutsche Frau	188	betreffende Dame	44	wohlhabender Mann	147	!"" Herr	101
20	zweite Frau	184	hochgeachteten Dame	43	ungl√ºckliche Mann	146	guter Herr	97
21	bie Frau	182	elegante Dame	42	kr√§ftiger Mann	146	dicke Herr	91
22	ungl√ücklichen Frau	168	fremde Dame	42	freier Mann	139	√§lterer Herr	90
23	ersten Frau	165	„Aû Dame	40	gekleideter Mann	136	genannter Herr	81
24	kranke Frau	158	sch√∂nen Dame	39	gebildeter Mann	135	gestrenge Herr	75
25	zweiten Frau	144	gebildete Dame	37	„Aû Mann	134	o Herr	72
26	wohnende Frau	141	hohe Dame	36	stattlicher Mann	131	gestrenger Herr	72
27	h√ºbsche Frau	140	Notre Dame	35	guter Mann	130	eigener Herr	62
28	geschiedene Frau	124	reichen Dame	34	reichen Mann	119	fremder Herr	60
29	erwiderte Frau	123	wei√üe Dame	34	kleine Mann	117	erwiderte Herr	56
30	ber Frau	123	j√ºngere Dame	34	wei√üer Mann	106	kleine Herr	53

Figure 3. Word frequencies of the terms *Frau, Dame, Mann* and *Herr* in the unclassified reprinting archive (pos. 1-30 of 4085 in total). As one can see, there are quite a few OCR errors, which shows how questionable analyses are when the search depends on individual words, rather than phrases.

the *Mann/Herr* datasets: sickness (*krank)*, marriage status (*geschieden*), and affiliation (*seine Frau/Dame*).[36]

Now what to do with these results? While the number of occurrences shows the visibility of women in the German-language press, their adjectival complements indicate that they are represented in different social contexts. Frequency analyses of the entire dataset without looking closely into the "data" are questionable because newspapers include diverse information: from political reports to nationalistic poems. The dataset is asymmetrical because it consists of different genres that vary in length and proportion within the dataset. When using the entire non-genre classified reprinting archive for such an analysis, representational bias plays an important role because the genre prediction model reveals that the majority of texts are hard news. There is a high probability that women, for instance, are not the topic in hard news of political events. As D'Ignazio and Klein have noted, "as is true in so many cases of data

36 I want to thank Martin Krupp, who was an intern at the GHI (2020/2021) and assisted me with analyzing and interpreting these datasets.

collected (or not) about women and other minoritized groups, the collection environment is compromised by imbalances of power."[37] I consider such results as starting points that guide me to ask further questions: Do I receive different results when taking genre into consideration?

III. GerWOMANness makes the viral ads

Studying genre provides an interesting lens for investigating asymmetrical power by focusing on the intersection between gender and other dimensions of identity such as sexuality, geography, and ability and how they are linked to economic profit. Linking virality, that is the publication, circulation, and modification of a text, to genre in periodicals implies overcoming the boundaries between the textual and the social to investigate representational bias. Sociologist Patricia Hill Collins calls the realm of culture and media, where oppressive ideas circulate, the hegemonic domain.[38] This is taken from her matrix of domination, consisting of four domains – the structural, the disciplinary, the hegemonic, and the interpersonal. The matrix is a concept used for studying the several intersections between power, social categories, and realms. Even though the analysis presented in the previous section is based on the archive of reprinted texts, it does not reveal anything about viral texts per se nor does it show how virality functions as a system to circulate oppressive ideas and how they are constructed and experienced. How does viral GerWOMANness in ads, poems, lists or factual texts provide insight into questions of the legal and social equality of women, democracy and nationalism, violence, remembering and forgetting, and above all, migration or the arrival in a new culture and a new everyday life? Reading topologically, as literary scholar Andrew Piper frames it, implies the use of computational methods to map relationships among multiple elements and categories such as genre and publication information of multiple texts.[39] My goal is not so much to find the "original" version of a text, or the source, but primarily to find, collect, and read texts that have been printed several times (and in different locations), and analyze how these texts, and their modifications, relate to different social categories. This section of my article provides one example of GerWOMANness, which is quite prevalent in the viral texts belonging to the genre of advertisements, and it shows how zooming in on "ads" allows us to better understand the relation between genre and gender. I use the concept of GerWOMANness both as an argument against dominant foci on men's bodies, con-

37 D'Ignazio and Klein, 35.

38 Patricia Hill Collins, *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment* (New York, 2008).

39 Andrew Piper, "Reading's Refrain," *ELH* 80 (2013): 373–99.

Term	Genre	Hits in Genre	Frequency per million words in genre	Selection of adjectival complements to term X (pos. no. 1-100)
Frau	Advertisement	6624	506.45	Junge, seine, leidende, meine schwächliche, kränkliche, zarteste, arme, sonderbare, sterbende, zitternde X
Mann	Advertisement	2096	160.25	Junger, gesunder, weiser, solider, geschwächter, ehrlicher, lieber, glücklicher, starker, kränklicher, reicher X

Figure 4. Number of hits for the terms *Frau* and *Mann* in the genre-classified subset "advertisements" as well as a selection of adjectival complements to both

cerns, and perspectives in earlier accounts and as an umbrella term for collecting data: stories about girls, women, and mothers.

As figure 4 illustrates, according to word hits, women appear more often than men using all reprint clusters (6624 > 2096). Additionally, a selection here of some of the complements to the nouns provides a first representation of the sexes within this genre. While *Frau* is preceded by modifiers such as feeble, tender, strange, or dying as well as (male) possessive pronouns, *Mann* is accompanied by adjectives such as healthy, wise, honest or happy.[40] Zooming in—close reading—on some of the reprint clusters with the highest numbers and widest circulation reveals that these ads are for patent medicines marketed in a way that was not gender specific. However, even though these products target both sexes, women are dominant in these ads because they are used as marketing strategies, often with the underlying storyline: woman A was suffering from B; subsequently she took medicine C; now she feels reborn; you need to buy product C or visit doctor D. The narratives usually end with listing the name, price, and seller of the product, thereby featuring the men as the ones to cure male and female suffering. In their few lines, these ads demonstrate how widely reprinted periodical texts reflect the values of the larger culture. They depict women as the ones who need to be protected, thereby reinforcing their status as second-class citizens. Advertisements like these, as information designed to be spread for scientific and economic purposes, reinforced preexisting societal views about the place of women in society.

40 A similar representation arises when I use the terms *Dirn*, *Kerl*, *Fräulein*, *weiblich*, or *männlich*.

Even though most of the ads refer to products targeting both women and men, there are also quite a number of reprinted ads of products for women, specifically pain relief medicine for menstrual cramps. Even though these texts were written and published by men, creating and analyzing these datasets reveals typical clichés about femininity and masculinity and to some degree allows access to women's voices. Even nowadays, misrepresentations of the menstrual cycle are quite prevalent in the media because many ads about pantyliners, for instance, show narratives about how women are suffering from a certain feminine condition, while they suddenly feel happy once they use the "right" product.[41] In the nineteenth century, one of the most frequently reprinted texts about regulating the menstrual function deals with a product to cure "female weakness" produced by Ray Vaughn Pierce (1840-1914), a Buffalo physician and U.S. Representative from New York (1879-1880). Pierce manufactured several medicines – many of them were little more than alcohol or opium solutions – and nearly one million bottles of his elixirs were shipped from Buffalo annually, but his product "Dr. Pierce's Favorite Prescription" marketed to women to cure female weakness became his best seller. As he wrote in *The People's Common Sense Medical Adviser in Plain English: or, Medicine Simplified*: "In all diseases involving the female reproductive organs, with which there is usually associated an irritable condition of the nervous system, it is unsurpassed as a remedy."[42] The product was a tonic to quiet nervous irritation and strengthen the enfeebled nervous system. Pierce was a master of the media, using his medical works, broadsides, billboards, and above all newspapers, to saturate the country with word of his success. "Favorite Prescription" can be seen as a prime example of a viral event and, specifically, viral marketing in the industrial age, a business strategy that uses existing social networks such as newspapers to promote a product.

The ad for the tonic was not only featured in thousands of English-language newspapers in the U.S., but also in hundreds of the German-language ones (1877-99).[43] The creative nature of viral marketing enables an endless variety of potential forms: there is not one single story about Pierce's product, but many different textual and visual representations in different but also in the same newspaper titles. Even though the majority of distinct reprinted ads became front-page material, they differed in headlines ("Die Dame," "Mütter," "Ein heruntergekommenes" or "Ne sonderbare Frau muss es sein") and with regard to the information about the female body, related symp-

41 See, for instance, "Sofy Pantyliner - How's your Day?" June 19, 2018, video, https://www.youtube.com/watch?v=p28-5c1s_c8.

42 R. V. Pierce, *The People's Common Sense Medical Adviser in Plain English: or, Medicine Simplified* (Buffalo, 1895), Project Gutenberg, May 28, 2006.

43 Many thanks to Moritz Knabben from the Oceanic Exchanges team, Institute for Visualization and Interactive Systems, University of Stuttgart, who extracted "Favorite Prescription" (1840-1914) from the English-language dataset (United States, Chronicling America; United Kingdom, British Library 19th Century Newspapers and Times Digital Archive; New Zealand, Papers Past; and Australia, Trove). For more information about the databases, see M.H. Beals and Emily Bells, *The Atlas of Digitized Newspapers and Metadata: Reports from Oceanic Exchanges* (Loughborough, 2020).

toms, and causes for "female weakness," but also in terms of storyline and accompanying images, which illustrated diverse marketing strategies. This viral event showcases GerWOMANness because it is a story about the misrepresentation and abuse of the female body in the nineteenth century and it is a missing dataset because, so far, scholars' narratives have prioritized representations of Pierce, his success, and of the individual products.[44] Studying such reprinted ads, their textuality and circulation, gives us an insight into the lives of women at that time.

One of the most frequently reprinted ads for "Dr. Pierce's Favorite Prescription" in the corpus is titled "Ein Frauenantlitz" and was predominantly published in the *Freie Presse für Texas* (San Antonio, Texas), a weekly newspaper run by Robert Hanschke (1879-1906).[45] Hanschke's company claimed that the *Freie Presse* had "undoubtedly the largest circulation of all the German papers in the State" and was "the best advertising medium."[46] "Ein Frauenantlitz" (70 reprints from 1897-99), does not simply provide information about Pierce's flawless reputation and the product's cure for insomnia, fatigue, and all maladies caused by disorders of the female organs.[47] The ad begins with a romantic depiction of women, whose age-related transformations are metaphorically imagined as the natural fading of flowers. The color red, compared to roses, disappears from the female glowing *Antlitz* (face) because the organs of the female sex start to malfunction as the years go by. These viral texts "in the best advertising medium" represent not only information about ways to relieve pain: these ads are stories about beauty or rather societal expectations of beauty, gender injustice and how the economy creates images about women, ranging from vulnerable to dysfunctional, in order to sell their products. They mirror and (re-) produce society's understandings of making the condition of female weakness – and age – responsible for every complaint of mind and body that a woman might have experienced.

This viral event represents one of the nineteenth century's biggest drug abuses perpetrated on women. Although advertised as "botanicals," such tonics promising to cure female weakness, recommended for young girls as well as nursing mothers, contained alcohol in double-digit percentages and even opium. Prior to the 1906 Federal Pure Food and Drug Act, the product labels of patent medicines did not have to list the ingredients that were addictive or dangerous such as morphine, opium or alcohol. Six years ear-

44 The "Nickell Collection of Dr. R.V. Pierce Medical Artifacts" (*New York Heritage*) includes prescription bottles and packaging, advertisements, and a book of medical advice by R.V. Pierce. Many of the materials are undated. The rest date between 1885-1890 and 1928-1938.

45 Arndt and Olsen, 630.

46 "Freie Presse für Texas," The Portal to Texas History, accessed January 29, 2021, https://texashistory.unt.edu/explore/collections/FRPRTX/

47 See, for instance, *Freie Presse für Texas*, April 17, 1899, Page 2, CA.

lier, in a reprinted ad in the U.S. English-language press that was published in DC, Texas, and North Dakota (1898-99), the text even negated the alcoholic substance: "'Favorite Prescription' contains no alcohol or whisky." I have not found this information in the German-language newspapers, but there is another common marketing strategy in both languages, which shows that advertising is personal. Along with the causes and symptoms, these texts include personal stories by women who testify to feeling so much better after drinking more than one bottle at once while they were pregnant or even while they were breastfeeding. These ads provide material and informational content about the female body, but they are directly accompanied by sociopolitical rhetoric and strategies embedded in their content: you need to get healthy as soon as possible to be able to fulfill your duties as a wife, housewife, and mother and no longer be a burden to your husband. These textual snippets – data about women – did not derive from their individual creators – Pierce and the editors – but from their passage through the exchange system. These ads illustrate how powerful virality, durability, and statistics of narrating "personal" experience are for marketing strategies: the product "ist seit vielen Jahren, wie Tausende bezeugen, mit völligem Erfolge gebraucht worden"[48] or "90,000 women have testified, over their own signatures, to its wonderful merits."[49]

These different viral ads about "Dr. Pierce's Favorite Prescription" not only provide a case study of successful viral marketing in nineteenth-century newspapers by placing women as witnesses in the center of attention, but they also show how bias against these women was practiced in reporting their issues and how they are directly connected to profit. While the distant readings documented in figures 3 and 4 have shown that women were predominantly represented in terms of their physical appearances (*schön*, *hübsch* or *zarteste*), examples such as "Ein Frauenantlitz" demonstrate how medical products were promoted in a powerful way to restore these features. In other publications, targeted at a "scientific," in other words, male audience, Pierce provided further information about the female weakness as being predominantly caused by "rheumatic affections, constipation, a morbid state of the blood, suppression of the menstrual function, uterine difficulties, masturbation, or self-abuse, or blows."[50] These detailed descriptions, which link female sexuality to violence inflicted by men, do not appear in the newspapers: a medium that was read by both sexes.

48 See, for instance, *Der Nordstern*, September 29, 1880, Page 4, CA.

49 See, for instance, *Evening star*, January 11, 1898, Page 8, CA.

50 Pierce.

Der Deutsche Beobachter.

Jahrgang 26. New Philadelphia, O., Mittwoch, den 9. Mai, 1894. Nummer 1.

Investigating the textual content and circulation figures is one way to analyze the representation – and abuse – of women. The majority of ads about "Dr. Pierce's Favorite Prescription" come with images. While the image in "Ein Frauenantlitz" depicts a young woman holding a flower in her hand, another ad titled "Zu Hülfe" (figure 5), which had a longer circulation period (1883-1894) and wider distribution (from Texas to Ohio), shows – textually and visually – a more sinister depiction of the female condition.

Figure 5. One example of "Zu Hülfe" (top right) in *Der Deutsche Beobachter*, New Philadelphia, Ohio, May 9, 1894, p. 1. https://chroniclingamerica.loc.gov/lccn/sn86063815/1894-05-09/ed-1/seq-1/.

While "Ein Frauenantlitz" starts with the metaphorical relation between the female sex and springtime, "Zu Hülfe" begins directly with a narrative about a woman's cry for help whose savior can only be a man: Dr. Pierce. Such textual and visual representations of women and Pierce's product as a restoring stimulant for the nervous system mirror contemporary fictional writings such as Charlotte Perkins Gilman's "The Yellow Wallpaper" (1892). In Gilman's short story, which is regarded as an important early work of American feminist literature for its illustration of the attitudes towards women's mental and physical health, the narrator suffers from post-partum depression and is therefore locked up in the attic. Similar to the woman behind the wallpaper, the "black" woman without any facial expression in "Zu Hülfe" (figure 6, right) seems to be trapped – or locked up – and in need of help, suffering from mania, delirium or madness. In contrast to other ads, the woman in "Zu Hülfe" does not even have a face. These images produced public stigma in the hegemonic domain such as stereotypes, prejudices, and discrimination of women, who are depicted as sad and hysterical, holding their malfunctioning bodies responsible for their "female weakness." While the number of reprints as well as their wider circulation in both the German- and English-language newspapers clearly shows Favorite's Prescription's popularity as well as Pierce's and the editors' HIStories, the representation of women in these ads reflect the dominant narratives of the relation between the menstrual cycle, mood fluctuations, and mental illness. Writings such as Gilman's short story, which was also widely reprinted in the newspapers, not only represent fictional challenges to the patriarchal diagnoses of women's condition, but also embody a public critique of a real medical treatment that was popularized through these viral ads. They created a sense of an imagined community – within the imagined immigrant and the larger American republic and offered a sense of participation.[51] These observations about patriarchal biases, misrepresentations of the female body, the rational man and the imaginative woman, lead me to further investigate the representation of GerWOMANness in other genres as well as their relation to one another.

The current analysis of the genre classification assumes that the three "dominating" genres that have the highest number of reprint clusters in the archive of reprinted texts are news, ads, and short stories. Examining advertisements as well as the correlation to other genres gives insight into different representations and shows

51 Benedict Anderson, *Imagined Communities: Reflections on the Origin and Spread of Nationalism* (London, 2016).

the blurred lines between fact and fiction, information and data, dominant and minoritized groups. Seriality and virality enable continuing movement and rely on the commercial affordances of capitalist production such as low-cost printing technology and advertising. The patterns of textuality, circulation, and categorization point us back toward the (digital) archive, suggesting new theoretical and descriptive questions about visibility, the body, and women's rights when using data-rich, "computer-assisted" approaches.

In this article, I have shown an analysis of women as the minoritized group in contrast to men in order to, as D'Ignazio and Klein have formulated it, "describe groups of people who are positioned in opposition to a more powerful social group."[52] Studying virality historically by computationally analyzing and enriching data and metadata about migrants sheds light on the social, political, economic, and personal aspirations of readers and writers. The resulting data in this project can be considered a substantial set of German-American literature to provide insight into the reprint history of texts as well as editorial and reading practices. A more comprehensive study of Ger(wo)manness will include examining the representations of women and men in other genres and the interaction between different social categories. Data feminism, that is, applying intersectional feminism to data science, implies changing perspectives by investigating the matrix of domination in terms of women as migrants, subordinated to the dominant group of "American" women or within the group of migrants included in the dominant group due to their whiteness. There are many more examples to investigate the intersection of sex and race, for instance: in a viral factual text about the true nature of a woman's beauty ("Die Wahre Schönheit der Frau"), the author mansplains that women's most fertile years in terms of beauty are between 30-45.[53] While providing details about how to be the best version of yourself through contentment and physical activity in order to retain one's beauty until the age of 50, the text emphasizes the imperative of pure blood – race – visible through a woman's skin color. In the beginning, however, the author lists Cleopatra – an Egyptian woman – as an argument for his age thesis because she was over 30 when she met Antony. Used in this way, intersectionality affords a necessary optic on the uneven ways in which power operates across social groups as well as a set of practices to collectively contest these distinct forms of domination.

52 D'Ignazio and Klein, 26.

53 See, for instance, *Der Deutsche Beobachter,* August 15, 1894, Page 2, CA.

IV. Conclusion

A recent blog post titled "Whose History? AI Uncovers Who Gets Attention in High School Textbooks" discusses a new study of American history textbooks used in Texas. The project conducted by Stanford University researchers who also use machine learning techniques revealed that "high school history textbooks pay much more attention "to white men than to Blacks, ethnic minorities, and women."[54] As regards quantity, "five of the 50 most-mentioned individuals were white men" and "[o]nly one woman made that list – Eleanor Roosevelt."[55] While men were more likely to be associated with words denoting power, women were more likely to be associated with marriage and family. The conclusion is, at this point, that many textbooks focus on formal political events and (male) leaders rather than on the lives of people. Only accessible datasets can be passed on to future generations. Uncovering biases through computational methods has the potential to name, challenge, and change underrepresentation in textbooks and beyond. As D'Ignazio and Klein note, "addressing bias in a dataset is a tiny technological Band-Aid for a much larger problem."[56]

My work calls for increased consciousness of bias in historical research and simultaneously challenges the analysis of bias on different levels: from historical or archival labor to data science practices. This addresses the fourth part of the matrix of domination, which deals with the interpersonal, that is, the domain that concentrates on the influences of the everyday experience of individuals in the world. How would you feel if you were a woman, with monthly cramps, who read these ads? Using the concept of virality and genre, the matrix of domination and the distinction between dominant and minoritized groups, we can begin to examine how power unfolds in and around data. Asking these data-centric questions – such as who is represented by whom, where, and how – allows us to see how discrimination is baked into our data practices and products. By illustrating biases in the genre of information literature, specifically the quantitative and qualitative analysis of texts representing this genre, this article has sought to revise our understanding of the role of women as writers and readers in the German immigrant communities. As D'Ignazio and Klein have argued, "machine learning algorithms don't just predict the past; they also reflect current social inequities."[57] By becoming aware of these different levels as well as by statistically and non-statistically analyzing them, I want to show how to use humanities data science to diagnose those

54 Edmund L. Andrews, "Whose History? AI Uncovers Who Gets Attention in High School Textbooks," *Stanford University Human-Centered Artificial Intelligence* (Blog), November 17, 2020.

55 Ibid.

56 D'Ignazio and Klein, *Data Feminism*, 60.

57 Ibid, 55.

problems and suggest solutions by offering a new window into the hidden histories and mysterious mechanisms of human cultures. In charting this process of data ethics when using digitized collections as well as methods for text mining and classification, I hope to offer contributions to (German-)American Studies, Gender Studies as well as Digital Humanities.

Jean Lee Cole, editor of *American Periodicals*, has convincingly argued that if we rethink periodical studies transnationally as vessels of national ideologies, we have to come back to questions concerning people and thus, inclusion and exclusion, migration and xenophobia.[58] Data science provides analytical leverage for studying these phenomena of discrimination and likewise opens up new ways of addressing and answering these questions because it provides tools to interpret the beliefs and behaviors of people, groups, and organizations at large. Both the data itself and scholarship to make sense of it are critical for advances across disciplines. Methods of statistical reasoning, natural language processing, classification, textual analysis, machine learning, and other data science approaches that developed largely yet not exclusively in the computer science professions have all become essential tools for scholars across the disciplines. Nowadays, this interdisciplinarity goes in both directions: data science has benefited from the complex, critical and, thus, consequential research questions targeting the rich history of human cultures, societies, and histories. Likewise, research in the humanities, not only under the umbrella of digital humanities, has benefited from computer and data science as regards digitization, content management, processing, and information retrieval and extraction.

Even though there is no commonly agreed definition of the term digital history, it generally seems to subsume two main orientations: it is concerned with the "constitution, management, and processing of digitized archives" and thus the transformation and preservation of cultural heritage.[59] And yet, as Ted Underwood writes in *Distant Horizons*, a digital collection "doesn't replace one version of the past with another," but rather "provides an enlarged repertoire of options."[60] These options consist in combining state-of-the-art computational methods, which focus on mathematical abstraction and the development of numerical and formal models, with state-of-the-art theories in the humanities. "Digital" does not simply mean algorithms, statistics or computing power or presupposes

58 Jean Lee Cole, "What's so American about American Periodicals," (presentation, Transnational Periodical Cultures – Interdisciplinary Perspectives, Johannes Gutenberg University, Mainz, Germany, January 30, 2020).

59 See C. Roth, "Digital, digitized, and numerical humanities," *Digital Scholarship in the Humanities*, 34.3 (2019): 616-632.

60 Ted Underwood, *Distant Horizons* (Chicago, 2019), 177.

the absence of theory, as data feminism illustrates, but helps translate humanistic thinking into computational practice. Would I have found these texts without computational methods? The answer is no.

Jana Keck is a Research Fellow in Digital History at the GHI. Among other projects, she is developing the digital research infrastracture "Migrant Connections," which will bring together diverse sources of German-American migration: letters, newspapers, lists, and more. Before joining the GHI in September 2020, she was based at the University of Stuttgart as a member of the international DH project "Oceanic Exchanges Tracing Global Information Network In Historic Newspaper Repositories, 1840-1914)," which developed computational methods – from text mining, deep learning to interactive visualization – in order to examine the global multilingual transfer of information in historic newspapers (DFG). For more information see https://orcid.org/0000-0002-9416-1256.