Computational Humanities Research

www.cambridge.org/chr

Whose news? Critical methods for assessing bias in large historical datasets

Kaspar Beelen[1,2], Jon Lawrence[3], Katherine McDonough[2,4] and Daniel C.S. Wilson[2,5]

[1]School of Advanced Study, University of London, London, UK; [2]The Alan Turing Institute, London, UK; [3]Department of Archaeology and History, University of Exeter, Exeter, UK; [4]Department of History, Lancaster University, Lancaster, UK and [5]Department of Information Studies, University College London, London, UK

Research Article

Cite this article: Beelen Kaspar, Jon Lawrence, Katherine McDonough and Daniel C.S. Wilson. 2025. "Whose news? Critical methods for assessing bias in large historical datasets" *Computational Humanities Research*, **1**:e8, https://doi.org/10.1017/chr.2025.10007

Received: 12 December 2024
Revised: 22 May 2025
Accepted: 1 July 2025

Keywords:
collections as data; data bias; digitization; historical newspapers; metadata; missingness

Corresponding author:
Kaspar Beelen;
Email: kaspar.beelen@sas.ac.uk

Abstract

This article implements a critical method for assessing bias in large historical datasets that we term the "Environmental Scan." The Environmental Scan sheds new light on newspaper collections by linking newly available "reference metadata" gathered from historical sources to existing full-text and catalogue metadata. The rise of computational methods in history and the social sciences, in tandem with newly "datafied" source materials, creates a challenge for researchers to adapt their existing critical practices to the increasing scale and complexity of computational research. To help address this challenge, the Environmental Scan situates big historical datasets in much greater context, including estimating what materials are missing, thereby revealing the ways digital collections can be "oligoptic" in nature. Using the British Newspaper Archive (BNA) as a case study, we diagnose the biases and imbalances in the digitised Victorian press. We determine which voices are under- or over-represented in relation to the political composition of the collection as well as its content and we trace the origins of these biases in the digitisation process. This article informs future interdisciplinary discussions about data bias and offers a conceptual model adaptable to diverse historical datasets. The Environmental Scan provides a more nuanced and accurate understanding of how newspaper data reflects past societies, making it a valuable tool for researchers.

Plain Language Summary

This article showcases a method we call the Environmental Scan, designed to help researchers identify, explore and analyse biases in large historical datasets. As historians and other humanities scholars increasingly use digital data and tools, it becomes harder to use traditional methods to analyse our sources critically. The Environmental Scan puts data into context and estimates gaps in digital collections. Using the British Newspaper Archive (BNA) and new, enriched newspaper metadata about this collection from the historical Mitchell's *Newspaper Press Directory* volumes, we examine Victorian-era newspapers and identify which political viewpoints are over- or under-represented in the digitised collections available on different platforms. Early digitisation initiatives targeted overtly political newspapers (e.g., *Conservative* and *Liberal*), with less emphasis on the *Neutral* or *Independent* press, but this trend shifted in later digitisation phases, led by different partners. Changing digitisation priorities stem from a variety of interests driving heritage digitisation, and with the Environmental Scan and the Press Directories metadata, it is finally possible to see how digital British newspaper collections relate to what was printed in the 19th century. Having identified these collection-level attributes, we also examine the language of the press. We measure how language relates to newspapers' political viewpoints: i.e., what is typically Liberal or Conservative language? Using an algorithm for detecting words that distinguishes between such categories, we outline the extent to which newspapers focus on different themes, such as religion, politics or sports. Used here with historical newspaper collections, the Environmental Scan offers a flexible model that can be used with different types of digitised historical data and metadata to contextualise those collections and enable archivally-aware computational research.

Introduction

Historical sources are increasingly digitised and analysed using computational methods. For decades, libraries and archives have pursued vast digitisation programmes to make their rich holdings available to researchers and the public alike. Historians are an important audience for the creation of these "datafied" collections (Bode and Goodlad 2023). They increasingly turn to big datasets to understand the past. However, the careful methods of source criticism,

This article was awarded the Open Data badge for transparent practices. See the Data availability statement for details.

which are the hallmark of historical research, have yet to find their equivalent in new forms of computational history.

Currently, humanities researchers, as well as public users of this data, are entirely dependent on the unfortunately opaque search interfaces provided by libraries and their commercial partners, which give the impression of comprehensiveness, where none exists (Milligan 2013; Putnam 2016).

This article develops a new critical approach to understanding big historical datasets. We investigate one of the biggest and perhaps most significant historical datasets yet created: the *British Newspaper Archive* (BNA). This dataset is in a constant state of flux, growing at a remarkable speed: to date, it contains over 92 million pages of historical newspapers, or about 150 billion words. Yet it still represents less than 20% of the newspapers held at the British Library (BL), which has one of the largest collections of print newspapers in the world.

In Britain, as elsewhere, big historical datasets have often been created through commercial partnerships between public institutions and private companies. Insofar as there is big data in the domain of cultural heritage – in the digitised or born-digital collections of libraries, archives and museums – it is usually a hybrid product: open to the public, but only under limited access conditions, with re-use constrained by legal and copyright issues arising from its complex provenance. Moreover, the digitisation arrangements entered into by libraries can allow private companies free rein in selecting material for digitisation, thereby shaping the datasets according to unknown criteria, but most likely in line with commercial rather than curatorial considerations. This process is inscrutable to the general public and researchers alike, including those working computationally. The inevitable preponderance of certain features and omission of others within such historical datasets is likely to be perpetuated and exaggerated by algorithmic approaches (such as using datasets like this as training data, or using models to predict some feature of a dataset), which do not take into account their composition. As Bender et al. (2021) argue, "[i]n accepting large amounts of web text as 'representative' of 'all' of humanity, we risk perpetuating dominant viewpoints, increasing power imbalances and further reifying inequality" (614). Emphasizing the link between privilege and power, D'Ignazio and Klein term this source of bias "privilege hazard" (2023, 28). Their lesson is clear: whether dealing with born-digital text or digitised 19th-century newspapers, researchers need to unpack and understand big data before using it, especially as part of a machine learning pipeline.

In light of such calls for critical approaches to large datasets, we develop the "Environmental Scan" (henceforth, ES): an approach to working with big historical data, which helps locate digitised newspaper data within the broader landscape of historical sources from which it was created. Historians generally acknowledge that digitised historical newspapers fail to straightforwardly reflect the past, but understanding in what way these collections shape and potentially skew our view remains hard and is not explicitly or empirically addressed at scale. The difficult work of contextualisation we perform with ES is the stock-in-trade of historians working with analogue sources, but it has yet to find its parallel in an age of historical big data.

To further the empirical analysis of bias for heritage data, we examine three forms of bias:

1. *Bias-as-missingness* investigates gaps in the data due to under-digitisation. After outlining the complete collection or population, this approach estimates what proportion of the originally printed press has been digitised;
2. *Bias-as-divergence* scrutinizes what missingness implies in terms "representativeness," and what voices tend to be over- or under-sampled as in the digitised corpus;
3. *Bias-as-partisanship* considers the implicit political bias in the data itself, and how the text reflects political ideologies.

This typology is not intended to be exhaustive. But by focusing on these biases, we demonstrate how missingness shapes collections in fundamental ways and risks producing datasets that are skewed in their composition and partisan in their content. The principal goal of this article is to offer concepts and methodologies that help researchers monitor how missing data might influence their analysis and findings.

Besides charting what has not been digitised, we point to the ramifications of such gaps and silences in historical collections. We make explicit whose voices are included (and so excluded) in digitised collections, helping to produce a more robust interpretation of past societies and their complex social and political realities. As an example of this methodology, we investigate political biases present in digitised British newspapers. Firstly, we inspect missingness and skews in the composition of the collection over time (*bias-as-missingness* and *bias-as-divergence*). We assess the extent to which the digitised sample (e.g., the BNA collection) is representative of the larger population of historical newspapers printed in the long 19th century. Secondly, we perform a content analysis of this newspaper collection's text, investigating political partisanship in the language of the press. Lastly, we explore how the observed biases arose and evolved throughout the process of digitisation, and discuss what this tells us about the lengthy processes of creating such datasets and its consequences. In the context of British newspapers, we hope that our findings will be used not only to inform further research but also to shape future digitisation programmes. We propose that the model underlying the ES method is a general approach that could also be used for other (inevitably) partially digitised historical sources.

This article extends previous work on the ES as outlined in Beelen et al. (2023) in four ways:

- We propose a more elaborate conceptual framework for contextualizing collections, enriching existing descriptive and administrative metadata with contemporary reference works.
- We complicate the notion of representativeness. In previous work, we principally understood representativeness as the proportional similarity of the part to the whole: the sample being a smaller set of the larger information landscape. In this article, we explore "equal" and "reweighted" forms of representation as ways to deal with bias and missingness.
- Beelen et al. (2023) considered only a small and early collection of digitised newspapers (JISC 1 and 2). Here, we analyse the complete BNA as it was shared with the Living with Machines (LwM) project in 2021. (The JISC newspapers constitute just a small subset of the BNA (Westerling et al. 2025)).
- Finally, we scrutinize the content of digitised newspapers by analysing patterns of partisanship in newspaper language.

The Victorian press as oligoptic data

The rhetoric around big data often implies impossibly comprehensive or panoptic view of phenomena. Any claim to a totalising "view from nowhere" can, according to Kitchin 2014, never truly provide more than an "oligoptic" view. "Oligoptic" emphasises the fact that big data necessarily offer fragmented, partial views on the past; views that remain constrained by the limits of historical sources. With its nod to the concept of oligarchy, the term reflects the reality

https://doi.org/10.1017/chr.2025.10007 Published online by Cambridge University Press

that datasets are constructed by a small number of actors from sources that themselves give precedence to some actors' perspectives over others.

Oligoptics is a useful conceptual lens in developing the ES for two reasons. As a visual metaphor, it emphasizes the need to look and to see the characteristics of a dataset, as well as to read its contents. Moreover, in its original sense in Science Studies, it did not imply a normative claim. Donna Haraway's work is instrumental in showing how knowledge is always situated, while Bruno Latour sets out "oligopticons" as a research programme for sociology that embraces partial perspectives (Haraway 2013; Latour 2005). The ES offers a fresh way to assess the oligoptic perspective that digitised newspapers offer us on past society, and thereby for assessing knowledge claims made using big newspaper data as evidence.

The national and regional press grew enormously across the 19th century, in Britain as elsewhere. As such, it constitutes one of the largest, richest and diverse sources of historical information that has been preserved. However, if users of digitised newspaper collections want to make claims about the social reality of the past, they face a double oligoptic bind which is rarely acknowledged. As noted, they are working with sampled data that bears an unknown relationship to the physical material from which it is drawn. But even if every single surviving newspaper issue were digitised, it would remain an open question as to how representative the newspapers are of historical reality. Newspapers reflect the political and socio-economic biases of their owners, editors and advertisers, even before we arrive at the biases of selection and omission introduced by collecting and digitisation practices. We must therefore acknowledge that the historical newspaper collections in widespread use today relate to the past much like Russian dolls: the working sample is nestled inside, removed from social reality by several hermeneutic layers.

However, the marketing language of historical big data tends to obfuscate these complex realities by portraying newspapers as "mirrors" of society or "a truly comprehensive record of times past."[1] Invoking metaphors like "mirrors" or "windows" fails to acknowledge, let alone accurately describe, the oligoptic nature of such data. Patterns in newspaper data – for example, measures of happiness (Hills et al. 2019) or morality (Lansdall-Welfare et al. 2017) – are being extrapolated to whole societies, ignoring the fact that many voices are excluded or radically underrepresented in digitised corpora and also that the socio-economic profile of newspaper readership changed dramatically over time. Data-driven research which fails to contextualize its datasets properly risks conclusions which are likely to be biased, unconvincing and, at worst, irresponsible. Historical big data in particular risks reproducing and legitimizing the omissions, exclusions and injustices already encoded in the archival record (Thylstrup et al. 2021). Instead, we aim to move beyond established cataloging traditions through the addition of rich, contemporaneous information about digitised sources we call *reference metadata*. In so doing, we make visible the politics of knowledge (Noble 2018) and the archive (Stoler 2002).

Contextualising big data

The principal contribution of the ES is to harness different types of metadata as diagnostic tools for gauging latent imbalances and biases in big historical datasets. In this sense, we reimagine the possibilities presented by "the essentially unlimited space available for documentation and metadata" in the digital era (Hauswedell et al. 2020) and demonstrate new ways of critically documenting data collections (Padilla 2017).

Our approach represents a multidisciplinary effort combining historical understanding of how newspaper collections were created, with statistical measures to gauge and analyse hidden biases in a collection. This requires broadening traditional notions of metadata as collected in library catalogues to incorporate terminologies and classifications gathered from contemporary (19th-century) reference works. We propose a method that combines well-known types of descriptive metadata (e.g., titles and dates of publication) and administrative metadata (related to digitisation or digital formatting) with a new level of information that we term contemporary **reference metadata**. This is metadata that:

- has its origins outside the object of inquiry and is derived from a reference publication;
- provides contextual information in the language of past actors;
- sheds light on how the dataset represents only a sample from a wider potentially unpreserved population.

For the BNA, we introduce reference metadata for important variables, such as price, ownership and declared political allegiance. Where such information exists (as it does for newspapers in multiple countries), the ES allows us to see a digital sample used by researchers in relation to the whole of which it is merely a part: as it were, the entire landscape of newspapers published at the time.

The reference metadata for the research presented here is derived from contemporary newspaper press directories. The longest-running of these was Charles Mitchell's *Newspaper Press Directory*, first published in 1846, and issued annually from 1856. It existed both as a reference work for publishers and as a guide for would-be advertisers. Like rival directories, which proliferated from the 1870s, the Mitchell directories are recognised as invaluable tools for surveying the extent and character of the Victorian press. For generations, the hard-bound volumes of these directories have been historians' go-to source for understanding the broad character of provincial newspaper titles when planning their research. Laurel Brake observes: "Annual listings in the press directories are of immense value to scholars: not only do they provide a trajectory of prices hard to find elsewhere except through examinations of issues, but they also represent the industry diachronically and synchronically, offering information about changes in titles, readerships, publishers, illustrations and geographical distribution" (Brake 2015).

Even for contemporaries, Mitchell's directory was considered an essential guide to the country's rapidly changing newspaper landscape. In 1861, the Post Office relied on it as the authoritative list of London and provincial newspapers (Gliserman 1969). Other epithets, such as "Whitaker of the press,"[2] buttressed its status as a respected publication (O'Malley 2015).

However useful they may be, these directories are not a ground truth for the historical press. O'Malley (2015) points out that Mitchell did not simply provide an impartial record of the press but should be considered as an actor who shaped and influenced the newspaper industry. The political categories, for example, reflect top-down attempts to organise the press along particular classifications. On the other hand, the directories also contain traces of bottom-up manipulations by publishers, who tried to exaggerate the geographical reach or inflate the seniority of their paper. In

[1] Find a wealth of newspaper stories from the past. 2023. https://web.archive.org/web/20230817205215/https://www.britishnewspaperarchive.co.uk/content/how_you_can_use_it.

[2] Whitaker was a yearly publication that covered the intricate details of cricket (O'Malley 2015).

https://doi.org/10.1017/chr.2025.10007 Published online by Cambridge University Press

other words, the codification of the press resulted from the interaction of multiple actors and vested interests.

In terms of completeness, Mitchell's directories do not exactly overlap with other records of the press. When processing this data, we consulted the newspaper metadata derived from the BL catalogue (Ryan and McKernan 2021). While a systematic comparison is beyond the scope of this article, we do have evidence of slight differences. For a few paper titles that occur in Mitchell's directories, we could not trace a corresponding entry in the BL catalogue and vice versa. However, these occurrences were very rare. What Mitchell considered a newspaper does not therefore always align with our current understanding. Moreover, politically, the directories could exhibit restrictive, conservative definitions of what was included. For example, multiple suffragette publications that now appear in the BNA as newspapers were not listed as such by Mitchell at the time.

Unpacking bias at scale

When working with individual newspapers, it would be second nature for a historian to triangulate each newspaper's reporting by supplementing the library catalogue entry with the press directories' richer information about each newspaper's price, declared politics and ownership. It is also well-understood that digitised collections are biased in the sense that they are not in any way a perfect replica of a physical collection, just as materials in a library or archive are in no way a "complete" set of what was produced in the past. Furthermore, it is then clear that selection bias reflected in the content of a digitised collection puts new limits on research, ones which are different from the experience of viewing documents in person (Putnam 2016; Zaagsma 2023). Understanding bias in the context of research with big data, is, however, relatively new territory for historians (see, however, (Jo and Gebru 2020)).

Accounts of large digitised newspaper collections provide insights into the history of digitisation, reflections on different provider interfaces and critiques of the commercial agreements that restrict access to full text and image files for out-of-copyright print collections (Hobbs 2013; King 2005; Fyfe 2016; Tolfo et al. 2021). Fyfe's (2016) "archaeology" of British newspapers studies how these collections have emerged through a long process of multiple (re)mediations, from print, to microfilm, and now to digital image. Fyfe, moreover, brings attention to the discursive practices and context in which such technologies were embedded and how they shaped digital resources. Where Fyfe aims at a historiography of research objects, we attempt to understand the implications of these past decisions and contexts for today's (digital) historians working with these collections at scale (Wilson 2023).

Within the field of machine learning, concerns about data bias and quality have become prevalent, despite value systems and incentive structures which reward data modelling rather than data wrangling and documentation (Sambasivan et al. 2021). Recent papers scrutinise practices of (and attitudes towards) data creation, use and management, and have proposed various frameworks, such as "data sheets" (Gebru et al. 2021), "data statements" (Bender and Friedman 2018) and "nutrition labels" (Holland et al. 2020) for improving transparency.

Documentation and contextualisation are critical tools for building ethical, socially responsible AI. Because data is almost always about people, assessing representativeness is imperative. Disregarding latent cultural or demographic characteristics introduces representation bias, risk over-exposing the language of some groups at the expense of others (Hovy 2018). Following Hovy and Prabhumoye (2021) and Shah, Schwartz, and Hovy (2020), our principal definition of representation bias concerns the mismatch, or divergence distributions of cultural and demographic variables (see Section "Representation Bias as Divergence" for more details).

Failure to delineate demographic and cultural limitations could lead to emergent bias, usually described as a demographic mismatch between the source (a model's training data) and target (users of this model) (Hovy and Prabhumoye 2021; Bender and Friedman 2018), which may affect humanities research using digitised collections. The ES, therefore, aims to elucidate the cultural and demographic characteristics of a dataset, thereby helping users to reason critically about the data. Adding *reference metadata* makes it possible for users to understand how partial or incomplete data may skew analysis.

Data and methods

The ES contextualises big data by broadening the concept of metadata (Figure 1).

Reference Metadata: As shared by FMP, the BL's digitised newspapers are accompanied by minimal descriptive metadata. To situate the digitised press in its historical landscape, we therefore enriched the collection with information derived from press directories. Directories collated information about many aspects of the newspaper press that would otherwise be difficult to recover, including not just price, ownership, and place of publication, but also declared political allegiance, claimed areas of circulation, and even principal subject matter. Though still not definitive, as discussed above, the directories offer a close approximation of the publishing landscape from which today's digitised collections have been drawn.

In 2019, LwM digitised all copies of Mitchell's directories printed between 1846 and 1920 held by the British Library. These were processed and enriched by Beelen et al. (2023).[3] Here, we use a curated subsample of the directories, for which we manually corrected the data.[4] Table 1 shows an example of the structured data extracted from the scans.

Newspaper Data and Metadata: The digitised newspaper files we analyse were provided by Findmypast (FMP) and contain the data behind their BNA: this search interface has been one of the main access points for historians searching newspaper content since 2011, when the BL consolidated previous and ongoing digitisation efforts so that they would be available to FMP subscribers and BL on-site users. BNA therefore incorporates earlier phases of digitisation, including the Gale collection and the initial JISC 1 & 2 projects, which scanned approximately two million pages of 19th-century titles between October 2004 and 2009 (Shaw 2005; Beals et al. 2020). Our analysis also includes titles recently digitised by the BL's Heritage Made Digital project and by LwM and subsequently incorporated into BNA (Tolfo et al. 2021).[5]

[3] The digitisation workflow will be described and evaluated in a forthcoming paper.

[4] We used directories published in 1846, 1847, 1851, 1856–1858, 1860–1875, 1877–1886, 1888–1891, 1893–1896, 1898–1900, 1902, 1903, 1905, 1907, 1908, 1910, 1912, 1914, 1915 and 1920.

[5] British Library blogpost, 5 September 2022, https://blog.britishnewspaperarchive.co.uk/2022/09/05/one-million-more-free-to-view-pages-added-to-the-archive/

https://doi.org/10.1017/chr.2025.10007 Published online by Cambridge University Press

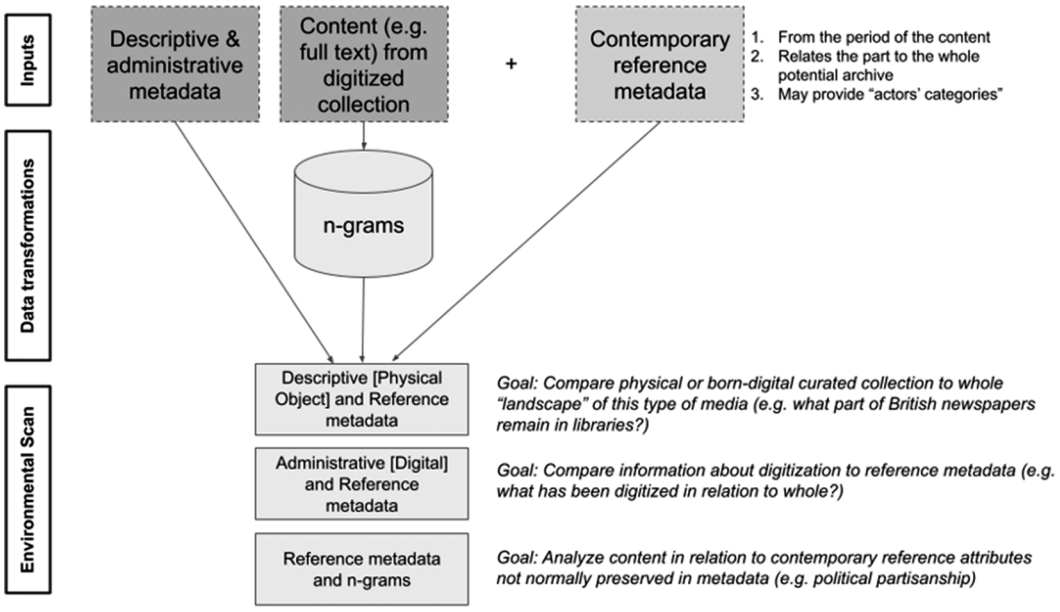

Figure 1. Overview of the environmental scan method and data.

Table 1. The structured press directories showing the extracted data for the *Brackley Observer* and *Altrincham and Bowdon Guardian*

Title	Politics	District	Price	Date	Description	Link_ID
Altrincham and Bowdon Guardian	Neutral	Bowdon	2d; 3d	1862	Takes no part in politics, [. . .]	null
Brackley Observer	Liberal	Brackley	1½d	1858	Taking a liberal view of politics. [. . .]	CID_00570

The BNA data, as obtained in 2021, contains ca. 1600 titles, spanning from 1780 until 1920 (Figure 2). We converted the structured newspaper content to plain text using the *alto2text* tool.[6] Subsequently, we converted the plain text documents to word counts: for each newspaper, we collected monthly word frequencies. This work takes inspiration from the "History Playground" (Lansdall-Welfare and Cristianini 2020), but extends it in crucial ways, allowing for more granular exploration of newspaper content. Instead of aggregating counts by year ("How often was the word 'machine' used in 1845?"), we enable more fine-grained queries ("How often does the *Aberdeen Free Press* mention the word 'machine' in June 1834?").

We linked the counts to the extensive metadata derived from the press directories, thereby creating a dataset similar to Google n-grams (Michel et al. 2011), but notably richer and more tailored to the needs of (digital) humanists. Abstract counts are contextualised by situating them not just in time but also in space (place of publication) and in social context (a newspaper's price, political leaning recorded in the press directories, etc.). In tandem with this article, we therefore publish the *NewsWords* datasets, one containing only word counts, and the other "contextualized" version with associated Mitchell's reference metadata (Beelen 2025). The non-contextualised *NewsWords* in a sparse matrix format contains 321,888 rows (representing monthly word counts by newspaper title) and 245,750 columns (the total vocabulary size, e.g., one column per word). The dataset contains close to 150 billion words (149,899,967,183 to be exact).

[6]https://github.com/Living-with-machines/alto2txt

Representation bias as divergence

Adding reference metadata allows us to investigate the perspectivist nature of big historical data, and critically assess the extent to which missingness affects representativeness because it skews the digitised collection in particular ways. We can compare how observed aspects of a dataset diverge from a hypothetical scenario that encapsulates the properties and values set by the researcher. We focus on the political composition of the collections, analysing the distribution of newspaper allegiances, such as *Liberal* and *Conservative*.

Concretely, for a given attribute like the political leaning of a newspaper, we assess whether the distribution of these labels in sample data resembles a hypothetical distribution whose shape is anchored in a specific interpretation of representativeness. Imagine newspapers are classified as either *Liberal* or *Conservative*. If we adhere to the principle of equal representation – a scenario which requires each political orientation to appear with equal probability – we compute the divergence between the distribution of categories in the sample to one in which each category appears with equal probability. If the digitised collection only contains a few liberal papers, let us say 10%, we compute the bias-as-divergence by comparing the sample distribution $\vec{p}$ ([0.1, 0,9]) with a uniform one $\vec{q}$ ([0.5 0.5]). $\vec{p}$ and $\vec{q}$ are vectors with the first number indicating the probability of a label being *Liberal*, the second number shows the probability for the *Conservative* label.

To compare these distributions, we calculate the Jensen–Shannon Divergence (JSD) between two vectors $\vec{p}$ and $\vec{q}$ that represent the probability distribution of a variable in the sample

https://doi.org/10.1017/chr.2025.10007 Published online by Cambridge University Press

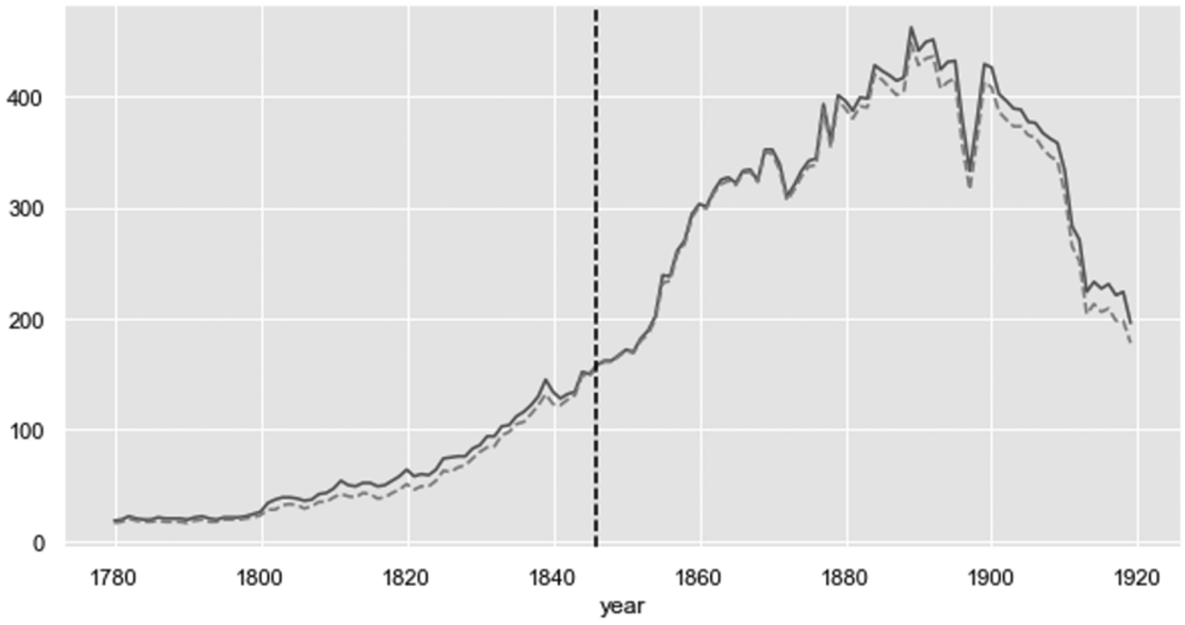

Figure 2. Number of digitised newspaper titles in BNA by year in the BNA corpus (2021).
Note: The solid line shows the total number of digitised titles in BNA, the dashed line shows the number of digitised titles successfully linked to press directory data. The vertical line shows the first year the press directories were published. Newspaper issues published before 1846 were linked to their earliest post-1846 appearance in the press directories, where possible. (The dip in the late 1890s is due to a damaged edition of the directories.)

and the hypothetical scenario, respectively,

$$JSD(\vec{p},\vec{q}) = \frac{1}{2}\left[KL\left(\vec{p}\middle\|\frac{1}{2}(\vec{p}+\vec{q})\right) + KL\left(\vec{q}\middle\|\frac{1}{2}(\vec{p}+\vec{q})\right)\right].$$

Because we model bias as a comparison between two probability distributions, JSD is generally a better choice than other common metrics such as cosine similarity or Euclidean distance. Unlike cosine or Euclidean metrics, JSD accounts for both the support (the number of non-zero values) and the distribution of probability mass. In other words, it captures the amount of information and uncertainty in the distribution. Moreover, JSD is symmetric, always finite, and bounded between 0 and 1. This ensures that it is more stable, interpretable, and robust, especially when distributions can have zero entries. JSD arose from information theory and reflects subtle shifts in distribution shape or entropy, making it an adequate and sensitive measure of bias.

In this article, the hypothetical distribution $\vec{q}$ can take different forms. We distinguish between three models of representativeness: *proportional*, *equal* or *weighted*.

Proportional representation assesses whether the sample is a proportional reflection of the population. In technical terms, we compute the divergence between the distribution of (political) labels in the digitised press and the historical press directories.

Equal representation computes the extent to which a variable's probability distribution diverges from a uniform one (where each value occurs with equal probability).[7] In practice, we only apply this method to numerically or historically significant categories (not all labels are equal).

Weighted representation: Both proportional and equal representation can be misleading. Proportionality seems intuitively fair, but it ignores biases built into the newspaper corpus itself. The 19th-century press had strong commercial incentives to reflect the worldview of the propertied classes, and only a small number of newspapers were aimed specifically at working-class audiences even though they constituted a large majority of the Victorian population. By contrast, adhering to equal representation may radically over-represent some minority perspectives (depending on the categories incorporated). Such representation may be only hypothetical, practically impossible to achieve. Even new digitisation does not provide a solution as some perspectives will be absent from the archive. The weighted representation is a compromise between these two approaches. It attaches value to a substantive presence of minority voices but does not seek to put them on an equal footing with the dominant categories of Victorian newspapers. We use this measure only for diagnostic purposes, to see how far the corpus varies from this hypothetical distribution, but we also suggest that it could be a more practical way to guide future digitisation and rebalance skewed collections.

More formally, for this model of representation, we reweight the count c_i of label l_i (e.g., *Radical*) from a fixed set of labels L=*Radical*, *Conservative*, etc., using N, the total number of observations, and p, a penaliser value:

$$c_{reweighted} = \frac{c_i}{N + (c_i * p)}.$$

We then convert these reweighted scores to a probability distribution. The penaliser value p determines the extent to which we oversample minority (and downsample majority) categories.

To give a concrete example, Figure 3 shows how each model or definition of representativeness produces different hypothetical scenarios (captured by the outcome vectors $\vec{q}$).[8] The black bars

[7] Continuing the leading example of political bias, assume that we observe n unique categories in the press directories. To compute bias, we would then compare the probability distribution of labels in the digitised press to a vector of size n in which each element is equal to $1/n$.

[8] In this example, we reduce L to six labels, which is a simplification of the actual set of political allegiances.

https://doi.org/10.1017/chr.2025.10007 Published online by Cambridge University Press

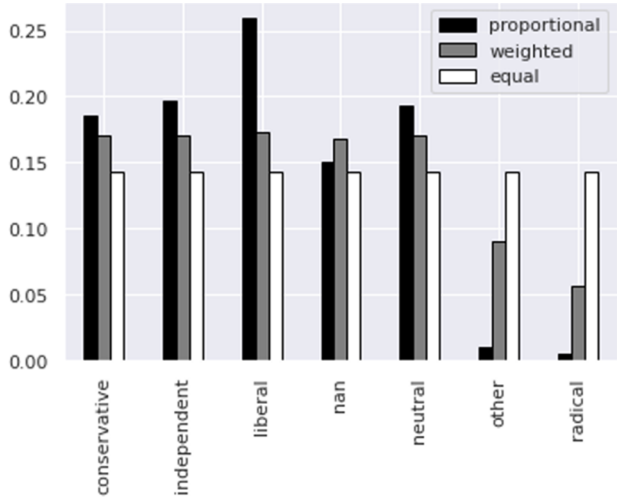

Figure 3. Comparing hypothetical distributions.
Note: The figure shows how definitions of representativeness correspond with hypothetical distributions of political labels (*nan* = no category recorded).

capture the distribution of the political labels in the press directories, used to assess whether a sample is proportionally representative. In this scenario, a sample *without* any *Radical* newspapers could still appear as statistically unbiased, notwithstanding the fact that it fails to include important strands of political argument *within society* that were marginalised because they were under- (or un-)represented in print. The white bars visualise a scenario where each category is equally represented. Situated between these two, in grey, is the re-weighted distribution, which tends to follow the proportional definition, but puts more probability on the minority categories by reducing the preponderance of the dominant categories.

Representativeness is a complex notion, and reducing it to a single, definite number infuses it with a certain rigidity and precision that is problematic. By introducing multiple hypothetical distributions, we offer a more elaborate and flexible diagnostic framework for establishing representatives (and bias) in historical collections based on varying assumptions. We developed (and implemented) the reweighted model to enable a certain parametrisation: it enables the adjustment of probabilities to match expectations about (the distribution of) the data. The choice of the penaliser depends, therefore, on the assumptions and expectations of the researchers. For our experiments, we gauge how putting more probability on categories such as *Radical* (which are historically important, but less present) might alter the results.

We treat hypothetical distributions mainly as a diagnostic tool to understand how our estimate of bias varies depending on our assumption of what a representative sample ought to look like. But it can serve different functions in other scenarios. For example, in the case of digitisation, the hypothetical distribution might serve as an ideal baseline against which we measure how adding data shapes the corpus in anticipated ways. Although bias often appears as a scourge, there is no single or definitive cure: it cannot be solved, and there is no single metric to guide us. Instead, we propose a versatile diagnostic framework which adopts different conceptual and ethical perspectives depending on the context, and which draws on the notion of "representativeness" (Chasalow and Levy 2021).

Content analysis: Discriminating words and partisan bias

Historical biases also percolate through the language of the press. News, as it is recorded in the headlines and articles, is a representation of events, and not a direct reflection of reality itself. It involves curation instead of covering every possible topic. In this sense, it carries signals of its producers' belief systems through the selection and representation of events. We therefore examine how demographic variables – primarily political allegiances – are associated with differences in language use: to what extent can we associate groups of newspapers with distinct priorities and patterns in language use?

For this content analysis, we use the *NewsWords Data "Contextualized Word Counts"*. *NewsWords* contains word count data derived from all types of newspaper items: articles, advertisements, notices, etc. This bag-of-words representation involves losing important distinctions when reducing texts to word counts; however, as an initial step toward outlining patterns of language use, these lexical patterns can nonetheless be revealing. To discover discursive biases in newspaper content, we estimate which words are discriminative and so distinguish, for example, the Conservative from the Liberal press, or cheap from expensive newspapers. By contrasting lexical patterns, we aim to uncover how newspapers created distinct representations of the world, often signalling strongly partisan or ideological stances.

Methods for modelling discriminative words often build on multinomial language models (Grimmer, Roberts, and Stewart 2022). Monroe, Colaresi, and Quinn (2008) proposed the *Fightin' Words* (*FW*) algorithm, a probabilistic model for feature selection that identifies the content of partisan conflict. Gentzkow, Shapiro, and Taddy (2019) extended this approach harnessing the power of Distributed Multinomial Regression (distrom) developed by (Taddy 2013). More recently, Kelly, Manela, and Moreira (2021) proposed a two-step approach. Their n-gram model (n-gramDMR) first estimates the inclusion of a phrase, and only subsequently models its repetition.

For the purposes of this article, we experimented extensively with the three variations on the multinomial language model, but ultimately opted for an adaptation of the *FW* algorithm. The main reasons were scalability and stability: running a full model on the complete dataset was significantly beyond the capacity of our (rather powerful) virtual machines. Secondly, we noted that the distrom and HurdleDMR models proved less reliable at identifying stable political differences in language use over time.

Tailoring *FW* to our research questions required significant adaptations. *FW* proposes a probabilistic model that factors in the varying levels of uncertainty associated with the different frequencies by which words appear. To obtain better estimates of a word's distinctiveness Monroe, Colaresi, and Quinn (2008) propose a regularisation method that shrinks coefficients to a common value (Grimmer, Roberts, and Stewart 2022), only words that are truly distinctive can deviate from this value. Rare words shrink towards the average.

Before explaining the *FW* algorithm in more detail, we point out some of its limitations. The principal aim is to foreground historically stable, partisan words by contrasting the content of different political categories of newspapers. Such a discriminative approach unfortunately reduces the complexity of the political press and removes important nuances by grouping newspapers from different places, price-points and interests together under one label, namely, *Liberal*, *Conservative*, *Independent* and *Neutral*. Moreover, the method assumes that frequencies point to meaningful (in our

https://doi.org/10.1017/chr.2025.10007 Published online by Cambridge University Press

context "political" or "ideological") differences between groups. If one group mentions X often, while another largely ignores X, our algorithm tends to attribute higher partisanship to X. However, because we rely on simple unigram frequencies (single tokens), we ignore the context in which words appear. If both groups use X often but disagree on its meaning and apply it to different contexts, X will not rank highly as a partisan feature.

While data-driven methods are useful for capturing macroscopic trends, we acknowledge that our processing of historical data is fundamentally different from how past audiences consumed information and the quantitative differences we observe are not self-evidently historically meaningful without further analysis and contextualisation.

We apply the *FW* algorithm in a binary setting, either contrasting two categories (e.g., Conservative versus Liberal) or one versus the rest (e.g., Conservative versus not Conservative). The regularised estimate of the probability that word j appears in document belonging to category k is:

$$\hat{u}_{jk} = \frac{W^*_{jk} + \alpha_j}{n_k + \sum_{j=1}^{J} \alpha_j}.$$

α_j is used to smooth the probabilities. After computing this estimate of $\hat{u}_{jk}$, we calculate the standardised log odds ratio, which compares the log odds between groups. For example, in the one versus all scenario, Liberal papers versus everything else.

$$\text{log odd ratio}_{kj} = \log\left(\frac{\mu_{kl}}{1-\mu_{kj}}\right) - \log\left(\frac{\mu_{-kl}}{1-\mu_{-kj}}\right).$$

We use the standardised log odds (Θ_{kj}) as the measure of separation, which is obtained by dividing the log odds ratio by the square root of its variance, but adapt the procedure in important ways to align with our understanding of partisanship, which we (intuitively) defined as words that are:

1. discriminative across categories,
2. prevalent within categories,
3. and exhibit both properties (i.e., 1 and 2) persistently over time.

To match the *FW* algorithm with this understanding of historical partisanship, we computed *FW* over time by only contrasting publications that appear within the same year. This produces a Θ_{kjt} score for each time step t. In some experiments, we would block selection on more attributes, for example, only comparing publications within a specific year and place, to account for systematic regional variation. This effectively introduces one more dimension in the equation.

Secondly, for inference, we computed Θ_{kjt} on subsamples of the data. Effectively, we create a sample containing 10% of the observations, in both kt and $-kj$ (i.e., Liberal papers in 1846 versus all other papers published in that year), and calculate the standardised log odds ratio Θ_{kjts} for this section. We repeat this procedure n times, producing n different estimates. This helps us prioritise words that are prevalent within a category (i.e., those that tend to reappear as distinctive tokens in multiple subsamples).

Lastly, we compute a final score p_{jk} for each word j, and use this to rank the complete vocabulary from most to least partisan. To foreground stable lexical markers of separation, we calculated p_{jk} as

$$p_{jk} = \frac{\text{mean}(\Theta_{kjts})}{1 + \text{var}(\Theta_{kjts})}.$$

This approach confers a higher score to words that, on average, have high distinctiveness values (across years and samples) but also low variance. This method is fast and straightforward to apply to a corpus with a lexicon containing 245,750 distinct words (including variants due to OCR errors) and more than 321,888 observations.

Analysis

Missingness: What has been digitised?

Before we dig deeper, we sketch the BNA collection's contours in broad strokes. The solid line in Figures 2 and 4 (top) shows the number of digitised titles by year, foregrounding the considerable temporal bias in the corpus: more data comes from the latter half of the 19th century. However, as a proportion of all published newspapers, the digitised corpus falls almost continuously from c.1860.

Overall, we estimate that around 20% of once printed British newspaper titles had been digitised by 2021. But this estimate hides quite substantive variation over time. The digital sample does not follow the trends of the larger newspaper landscape. It fails to reflect the steep rise in titles across the 1860s and 1870s, when the growth of the digitised sample is outpaced by the actual expansion of the press. Moreover, the digitised sample declines more sharply after 1900 than the number of titles listed in the directories. Press consolidation was accelerating in this period, and probably explains the blip in both series around 1897 (Lee 1976, 133, 176), but the digital sample may also be registering the effect of tighter copyright restrictions for 20th-century publications. Generally, measured as a proportion of the landscape (in relative terms), the sample becomes smaller over time. This pattern clearly emerges in Figure 4 (bottom left), which exhibits a steep decline in the percentage of digitised titles over time. This trend holds for both the metropolitan (London) and provincial (including Scotland and Wales) newspapers. In general, a smaller portion of the metropolitan press has been digitised.[9]

Divergence: Political biases in the digitised Victorian press

We now move our attention to representation bias with respect to the political orientation of newspaper titles. The directories record approximately 120 unique political allegiances, reflecting in part the freedom newspapers had to self-identify politically (one title, the *London Morning Advertiser*, returned five different political identifiers in just a few years, each time combining the term "independent" with other political terms). In addition, party names were themselves mutable, and many newspapers adjusted their allegiance over time. The Liberal Party was a coalition of pro-reform groupings, from Whigs to radicals, offering newspaper editors' many possible qualifying labels and nuances. Similarly, after the Liberal split over Irish Home Rule in 1886, many Conservatives adopted the term "Unionist" to signal their alliance with the anti-Home Rule "Liberal Unionists," while others chose to call themselves "Constitutionalist." In the analyses that follow, we work with composite allegiances that take account of these historical specificities.

[9] It is important to note that our analysis is necessarily restricted to the title level, as this is the only information we have about the newspapers at the population level (in other words, we know what titles were printed in a year, but not how many issues).

https://doi.org/10.1017/chr.2025.10007 Published online by Cambridge University Press

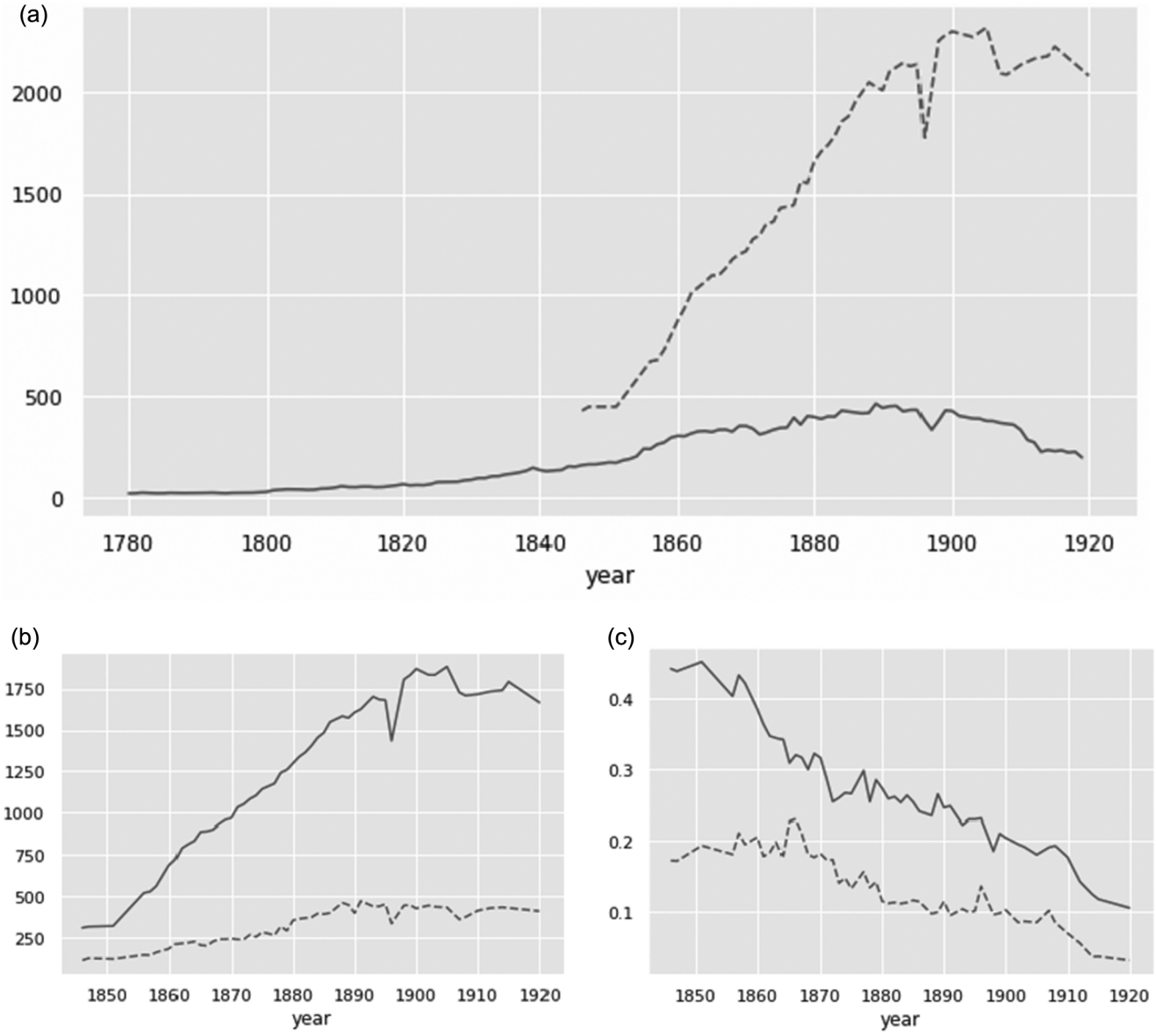

Figure 4. What has been digitised?
Note: a) **Top**: Total number of titles reported in the Newspaper Press Directories (dashed line) and number of digitised newspaper titles in the BNA (solid line); b) **Bottom Left**: The number of titles of Provincial (solid) and Metropolitan (dashed) reported in the Press Directories; c) **Bottom Right** show the digitised press as a proportion of the total number of titles for Provincial (solid) and Metropolitan (dashed) newspapers.

In broad terms, newspapers can be classified into five political blocks: *Conservative*, *Liberal*, *Radical*, *Independent* and *Neutral*. Alone or in combination, these five account for more than 83% of the returns (with a further 15% recording no allegiance). To measure bias, we simplified newspapers' political allegiances to seven categories: *Conservative*, *Liberal*, *Radical*, *Independent*, and *Neutral*, *Other*,[10] and *nan* ("no label"). These collapse some of the fine-grained distinctions into larger categories, while retaining the main political distinctions (see Figure 5).[11]

Turning to the distribution of these labels over time, Figure 6 visualises the shifting politics of the newspaper landscape. In terms of their share of the 19th-century press, Liberal newspapers show a steady decline, from almost 40% in the 1840s to only 25% in 1920. Conservative newspapers decline sharply in the 1850s, stabilise at around 20% of titles until 1890 and then register a modest increase. The category that gained the most ground over this period is the Independent press, from a rarity in the mid-19th century to the largest category of newspapers by the 1920s. If we amalgamate *Independent* and *Neutral* titles as the non-partisan press, they represent a third of titles from 1860, and were the largest grouping by the 1890s. This conflicts with the usual emphasis on party politics in accounts of the Victorian press. (See Lee (1976), Koss (1981), Jones (2016), and Hampton (2004); with Hobbs (2018) as an important exception.) That said, as the preceding discussion suggests, there were also significant differences between

[10]Includes national(ist) and religious newspapers in the remaining analysis.
[11]The mapping is explained in Appendix A. We relied on historical domain expertise within the team to group the labels. Using a data-driven approach could be another option, as one reviewer pointed out. In this scenario, we could assign to a category based on the similarity of the lexicon selection, e.g., assess if their content resembles Liberal or Conservative word use.

https://doi.org/10.1017/chr.2025.10007 Published online by Cambridge University Press

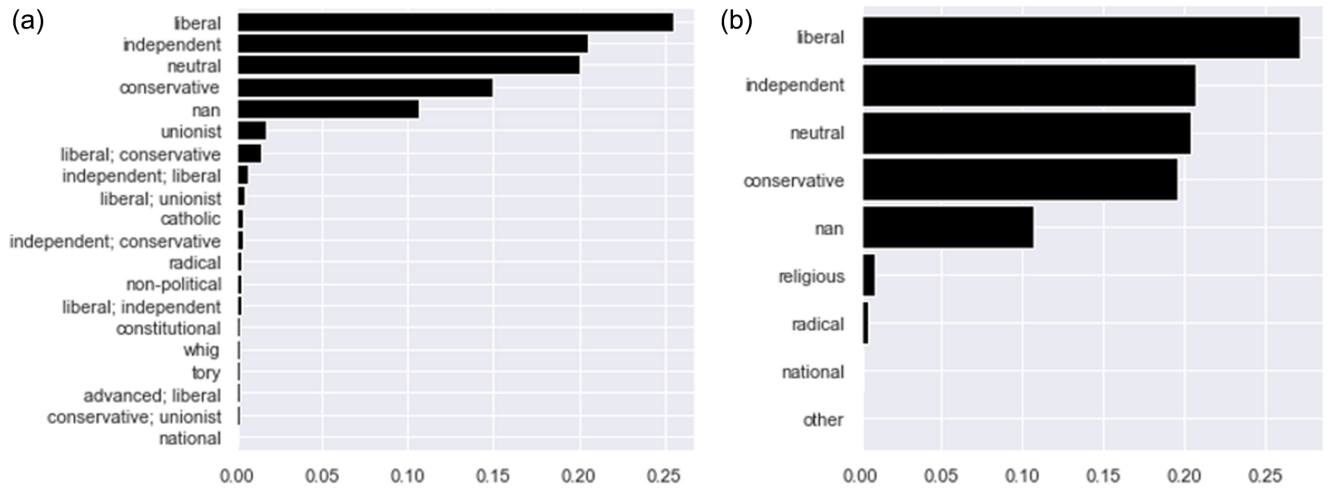

Figure 5. Distribution of Political Labels: proportional distribution of political allegiances in the digitised press directories. **Left**: before reclassification (top 20 labels). **Right**: after simplification of the labels.

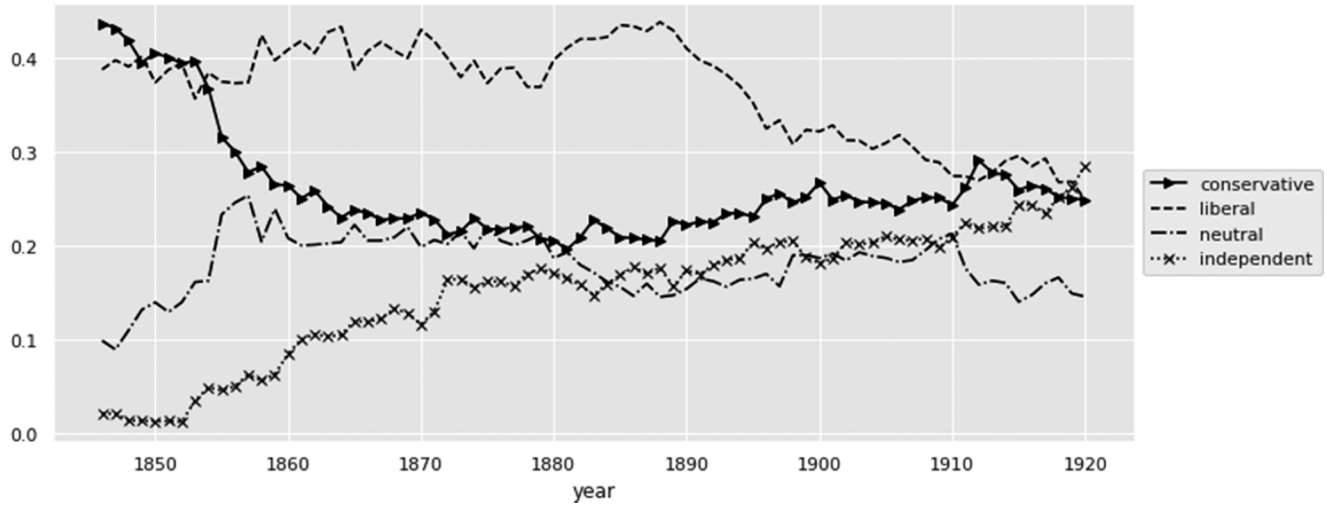

Figure 6. Proportion of newspaper titles by press directory political label over time.

Independent and *Neutral* newspapers, with the latter more resolutely non-political as well as non-party (an aspect of the contemporary distinction that is too often missed, e.g., Lee (1976)).

When comparing the digitised press to the wider newspaper landscape, it appears that overtly political newspapers tend to be over-represented (in proportional terms). For almost the complete period, Liberal newspapers are over-represented, comprising approximately 10% more of the sample than the population (Figure 7). Conservative titles are also over-represented, but less markedly. For the neutral and independent press, the historical pattern generally points in the opposite direction, even though the differences between the digitised sample and the newspaper landscape widen or shrink at different moments. The neutral press is under-represented for the third quarter of the 19th century, while for independent titles, under-representation is greatest in the fourth quarter. Although in these cases the differences remain smaller, generally around 5 percentage points.

As noted, the *proportional* scenario in which we measure how closely the sample is a proportional reflection of the larger landscape, is only one model of representativeness. Using the JSD, we can assess how the digitised press diverges from the other two models (*equal* and *reweighted*). To demonstrate how interpretations of bias-as-divergence may vary, we applied all three approaches to two different categorisations of the British political landscape:

1. *Coarse-grained*, which comprises the four principal categories (*Liberal*, *Conservative*, *Independent* and *Neutral*).
2. *Fine-grained*, including the minority category of *Radical* titles as well as papers which did not report any affiliation (*nan*).

By diagnosing a corpus using different categories and bias metrics, we avoid reducing bias to a single score. Instead, we treat it as a way of investigating or interrogating big data, which requires us to look at a collection from multiple angles. If data are oligoptic, so too are our bias measures.

The definitions of representation yield slightly different interpretations of how bias in the data changes over the 19th century. Emphasising proportionality, the timelines suggest a drop in representativeness in the 1860s that is reversed from 1870, climbing to a higher level again in the early 20th century (Figure 8). In other words, it suggests that during this period of rapid press expansion, the BNA becomes less representative of the total titles indexed in the press directories.

https://doi.org/10.1017/chr.2025.10007 Published online by Cambridge University Press

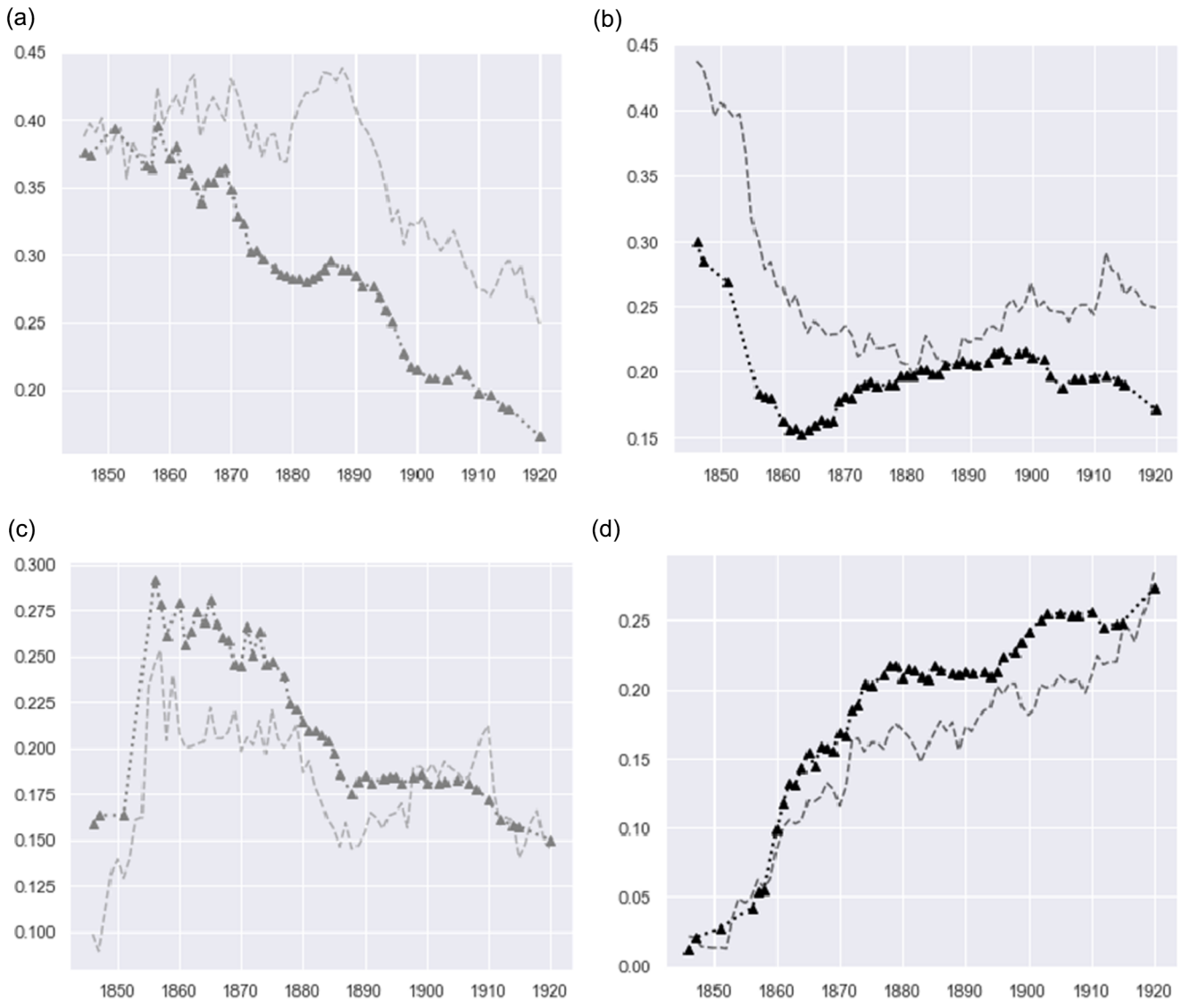

Figure 7. A timeline showing the proportion of newspapers by political leaning-based press directories and BNA.
Note: Dashed lines show the proportional presence of each leaning in the sample (BNA), while the dotted line represents the population (press directories). The top-left figure shows trends for *Liberal* newspapers (grey); the top-right *Conservative* newspapers (black). The bottom-left figure shows *Neutral* newspapers (grey); the bottom-right *Independent* titles (black).

However, when prioritising *equal* or *reweighted* representation, the scores suggest an overall reduction of bias; attention is more equally distributed over the principal political categories at the end of our period. Especially after 1900, the corpus is closer to an equal than a proportional representativeness. However, looking at the fine-grained taxonomy challenges this interpretation, and underlines the importance of including a wider array of partisan and non-partisan voices, not just the dominant ones. There is still evidence of declining bias, but this is now modest. The BNA is far from representative once we include minority voices such as the *Radical* press.

We inspect changes in the bias scores systematically by computing the contributions of each coarse-grained label to the divergence, and repeating this for the different types of bias (Figure 9). Again, the narrative tends to depend on what constitutes representativeness. Also in this case, we need to underline the importance of tackling bias from multiple perspectives.

All measures largely agree on the over-representation of the partisan press at the expense of *Independent* and *Neutral* titles (regardless of whether we base our calculation on coarse or fine-grained classifications). But simultaneously, these figures reveal crucial distinctions. For example, they clearly disagree in their timing: from a proportional perspective, the over-representation of the *Liberal* press grows with time, while it declines in the other figure. Moreover, these bias measures give different weight to the under-representation of the non-partisan press. Equal and reweighted representation measures clearly signal the radical under-representation of *Independent* titles in the middle of the 19th century, and *Neutral* titles later on. The preceding analysis demonstrates how the ES enables scholars to switch between different models of representation. Each of these postulate a shape to which a sample ought to adhere, and thereby offer different measures of the representativeness of the digitised collection. Again, the goal here is not reduce bias to a number, but to better understand its oligoptic view.

https://doi.org/10.1017/chr.2025.10007 Published online by Cambridge University Press

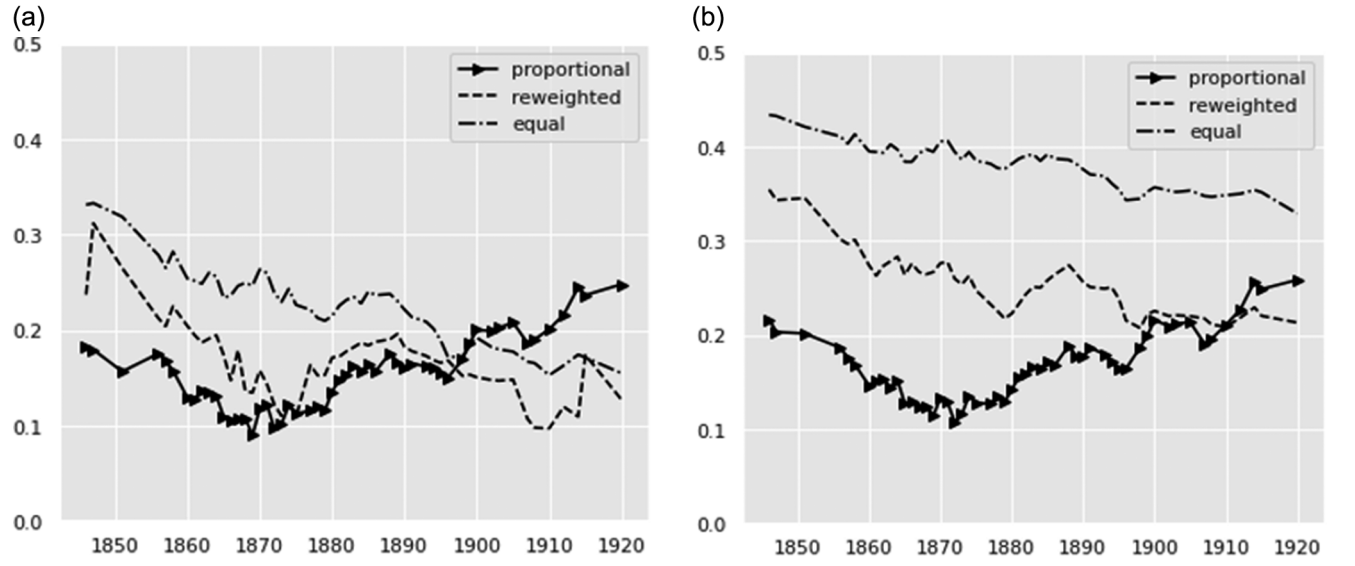

Figure 8. JSD scores over time. The *Y*-axis measures the divergence between the observed distribution of political labels and a hypothetical distribution, i.e., the extent to which the sample achieves proportional, equal or "reweighted" representation.
Note: The left figure focuses on the coarse-grained taxonomy, while the right uses the fine-grained, original categories.

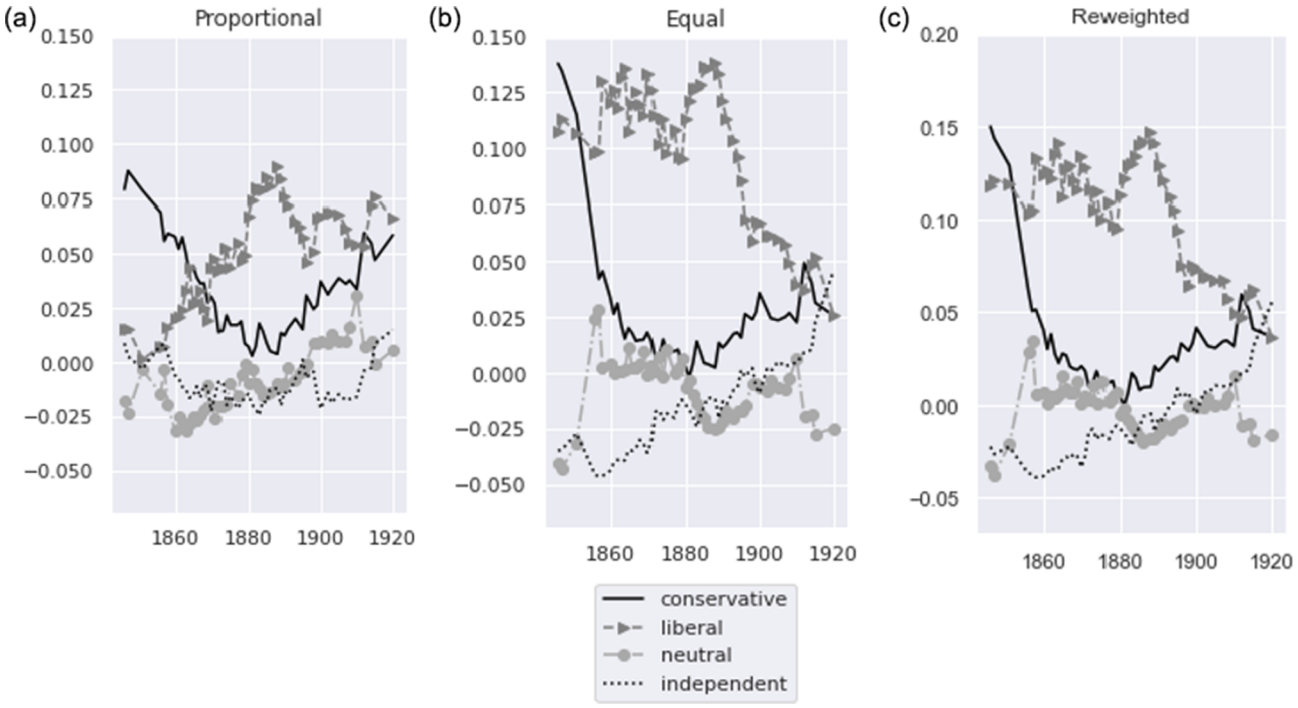

Figure 9. Contribution of the four principal political labels to bias scores for each measure of representativeness (coarse-grained classification).

We also looked more closely at the representation of minority voices, especially interested in the oppositional categories, such as *Radical*,*Labour* and *Socialist*, newspapers aimed to represent working-class, plebeian readers. Proportionally speaking, there is no overall anti-radical or plebian bias, but other models of representation suggest an under-representation of radical titles. In times of rapid democratisation of the press, the BNA increasingly underrepresents radical voices. Titles without an explicit political allegiance in the press directories, and often with minimal contextual information in the BL catalogue, are also less likely to be digitised. A detailed analysis is available in Appendix B.

Partisanship: Content bias in the Victorian press

While we have been discussing the *Liberal* or *Conservative* press, it remains unclear to what extent these categories reflect different discursive repertoires (in terms of newspapers' printed content). In this section, we investigate differences in language by political allegiance to assess the extent and nature of Victorian newspapers' partisanship. This helps us to assess what, for example, a Conservative bias implies in terms of content, but can also help us to understand whether, as suggested from their directory entries, nominally non-party-aligned newspapers could still be partisan on specific topics.

https://doi.org/10.1017/chr.2025.10007 Published online by Cambridge University Press

Table 2. Overview of the most distinctive words by newspapers' political allegiances documented in the press directories and thematic clusters

Cluster	Conservative	Liberal	Neutral	Independent
1. Partisan	Radicals, radical, conservative, conservatives, agitators	Tory, tories, liberal, liberals, toryism		
2. Politics	Property, proposed, undersigned, institution, appointment	Reform, reformers, independent, reforms, ballot		
3. Religion	Diocese, rector, bishop, curate, diocesan	Baptist, congregational, methodist, chapel, unitarian	Wesleyan, abbots, abbot, preached, hymns	
4. Economy	Valuable, auctioneers, auction, offices, hotel	Shop, drapery, shoe, trade, trades	Wood, beam, bone, teak, salt	Potteries, advertiser, turnpike, earthenware, pottery
5. Legal and crime	Pursuant, sessions, executors, solicitor, statute	Charged, police		Belonging, attorneys, stealing, attorney, sentenced
6. Morality and social behaviour	Desirable, highly, pleasure, regret, liberality	Temperance, liquor, moral, social, evils	Eery, mood	death, respectfully, health
7. Education and young people	College, university, scholarship, grammar, scholarships	Youth, teachers	Alma	Schoolroom
8. Social groups	Bart, gentry, society, honorary, earl	People, public, joiner, miner, miners	Moms, mums, mama, woodman, mamma	Yeomanry, woman, father, stationer, inhabitants
9. Place	County, residence, situate, premises, parish	Burgh, street, borough, houe, dwelling	Hill, mews, heath, alba, manor	Messuage, parish, messauges, neighbourhood, outbuildings
10. Agriculture	Agricultural, acres, agriculture, farm, acre		Loam, mows, mule, sows, beet	Dairy, wethers, cowhouse, stirks, wether
11. Military	Colonel, major		Ammo	Brigade, officer, volunteer, musketry, volunteers
12. Royalty	Loyal, majesty, anthem, toast, royal			
13. Time and periodisation	Late, classical, yearly, perpetual	Right, yesterday, daily, yeerday, dail		Forenoon
14. Sport			Wing, away, bottom, bowl, fifa	

We discuss the main themes that emerge from a contrastive analysis. We first computed the most distinctive words for each category, and subsequently grouped these discriminative features in a bottom-up fashion. Clustering words ignores some of the complexity and ambiguity of language, but it clarifies strong lexical patterns in the data. We refrained from assigning words to a category if we deemed them too ambiguous or multi-valent.[12]

We grouped the *FW* into 14 categories (see Appendix C). Clustering by semantic theme makes it easier to compare patterns across types of newspapers. Table 2 shows the top five most distinctive words by category in the top 200 most distinctive words (some clusters have no occurrences in the top 200, others fewer than five). The partisan cluster (1) confirms that Liberal newspapers used variants of "tory" to describe their Conservative opponents, while Conservative newspapers retaliated by talking about "radicals" and "agitators" (see Figure 10).

A party's official name was also a distinguishing feature of partisan newspapers' vocabulary (Figure 10). Other clear patterns include the monopoly of religious words associated with nonconformity in the *Liberal* press and with the Church of England for *Conservative* papers (cluster 3); and the stark differences in partisan newspapers' distinctive languages of social description (cluster 8). The distinctive social vocabulary of *Conservative* newspapers leans

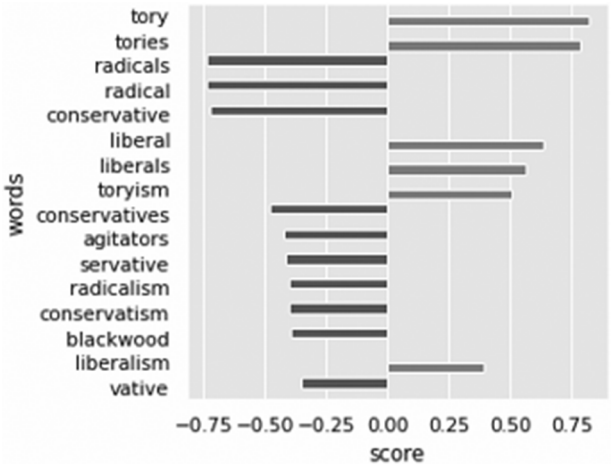

Figure 10. Contrastive plot of distinctive *Liberal* and *Conservative* partisan words (cluster 1).

[12]We stress that the clustering was intended as a heuristic device to help make sense of the various *FW* returned by the algorithm. This is not meant to provide a taxonomy of newspaper content. Two of the authors inspected the top 100 words for each political orientation, discussed principal themes and their content, and then set about annotating each word by its principal theme. Disagreements were resolved through discussion.

https://doi.org/10.1017/chr.2025.10007 Published online by Cambridge University Press

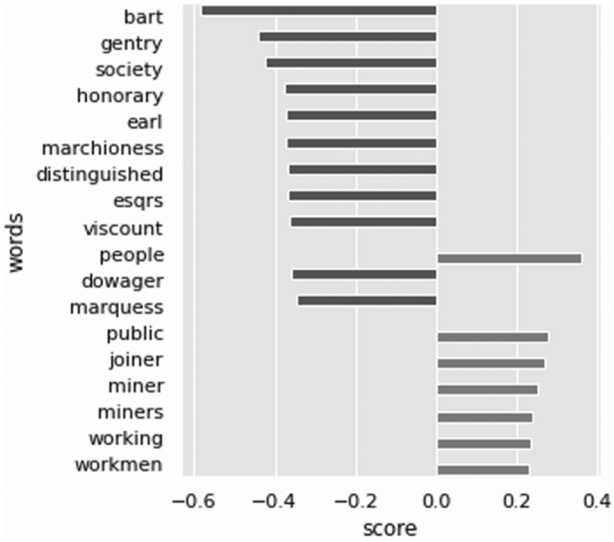

Figure 11. Contrastive plot of distinctive *Liberal* and *Conservative* social group words (cluster 8).

heavily on Burke's *Peerage*, quite probably including its distinctive use of the term "Society" to refer to the social elite, whereas the distinctive language of *Liberal* newspapers focused on working-class occupations, and on demotic abstractions ("people" and "public," see Figure 11). The strong association between party allegiance and religious denomination ("Church" or "chapel") is expected and confirms the method's validity, but the strong class polarisation in newspapers' social language is less obvious in an age when Conservative leaders were seeking to cultivate the "tory working man" (Lawrence 2017; Quinault 1979). It suggests that their efforts may not have been widely amplified in the provincial *Conservative* press.

One sees a similar pattern for cluster 4 (Economy), where *Liberal* newspapers' vocabulary is considerably more quotidian than that of their *Conservative* rivals. The most striking patterns for *Independent* newspapers are probably the absence of partisan and political words, the mix of nonconformist and Anglican words in their religious lexicon (cluster 3), and the importance of economy, agriculture and military words (4, 10 and 11). *Independent* newspapers' distinctive agriculture words are more vernacular than those associated with the conservative press ("dairy," "wethers" and "cowhouse"), and military terms are strongly associated with the local Volunteer movement. Both may offer clues to the distinctive readership of *Independent* newspapers. The same can be said of the *Neutral* newspapers, which are also notable for their large-scale and distinctive use of sporting terms (cluster 14, see Figure 12).

Only the *Conservative* newspapers' use of (Anglican) religious words rivals this distinctive feature of the *Neutral* press. Figure 12 also underlines the importance of political process words for *Liberal* newspapers and the distinctive economic language of *Independent* newspapers.

We can also harness this method to track changing patterns of word use over time. With *FW*, we can investigate how the distinctiveness of a set of words changes over time. Thus, we can explore the extent to which "tory," "liberal," etc. are discriminative of *Liberal* newspapers, and how this tends to vary over the 19th century. In Figure 13, we computed how discriminative the partisan words were for *Liberal* and *Conservative* newspapers, respectively. For each year, we aggregated the scores for all partisan words (cluster 1). Focusing specifically on the use of these words by these two classes of party-aligned newspapers, we see two periods of high and broadly equal emphasis on partisan words: before 1850 and between 1885 and 1895. There are also two periods of more unequal usage: 1850–1885, when *Conservative* newspapers put more emphasis on partisan words, and 1895–1920 when *Liberal* newspapers did so (Figure 13). Both register a significant drop in traditional partisan language during World War I, but for *Conservative* newspapers, this appears as the continuation of a trend begun in the 1900s. This decline suggests that partisan newspapers may have been becoming less overtly political in this period despite considerable polarisation within society (Lawrence 2009). This trend requires further investigation, but may point to the commercial pressures that would ultimately lead most local newspapers to disavow party political allegiances by the mid-20th century.

The discursive impact of partisanship was not constant and can be measured in different ways. Bias in sample composition and newspaper content may interact, adding to the complexity of understanding the effects of bias and missingness for users of digitised corpora. When conducting the analysis of newspapers' distinctive words, it was striking that OCR errors occurred much more frequently in lists generated for *Liberal* and *Neutral* newspapers than for *Conservative* or *Independent* ones. The same pattern existed when we compared cheap newspapers with more expensive ones, suggesting that paper and ink quality may help explain these differences (see Appendix D). Further work is needed to understand the significance of this pattern, but it is a salutary reminder that digitised newspaper collections offer a highly skewed view of the past.

Digitisation, divergence and the origins of bias

Up to this point, we have mainly scrutinised how the British digitised newspaper collection changed over time, where "time" has meant year of publication. While the digitised press allows us to study the past, we should also remember it carries its own history, having emerged through a long and complex digitisation process now spanning more than two decades. In this section, we take a closer look at digitisation as a process, and ask: how did (the BNA sample) biases change over time as digitisation progressed? We show that selection bias has fluctuated and argue that researchers therefore need to understand its extent and nature for the specific version of the corpus they are accessing.

In the following analysis, we sorted the newspaper titles by their position in the digitisation sequence: from the first digitised title to the most recent.[13] This step assigns a rank to each title, ranging from 1 (first digitised newspaper) to n (the last). After sorting all titles by their digitisation order (or rank), we iterate through the collection, and traverse the data from 1 to n with stride s (here 25 titles). We start by computing the bias on the first "digitisation batch," i.e., newspapers with rank between 1 and s (or 25 in our case); in the second iteration, we expand this batch to the papers with rank between 1 and $2s$ (or 50), etc. We thus increment the batch size at each step with stride s and compute the bias by comparing the distribution of newspapers' political allegiance labels in the digitisation batch to a target distribution (see Figure 14). The latter distribution remains static at each iteration. Put simply, we

[13]To sort titles by date, we used the BL NLP identifiers which, with minor caveats, were issued in broadly chronological order and therefore serve as a reasonable proxy for order of digitisation.

https://doi.org/10.1017/chr.2025.10007 Published online by Cambridge University Press

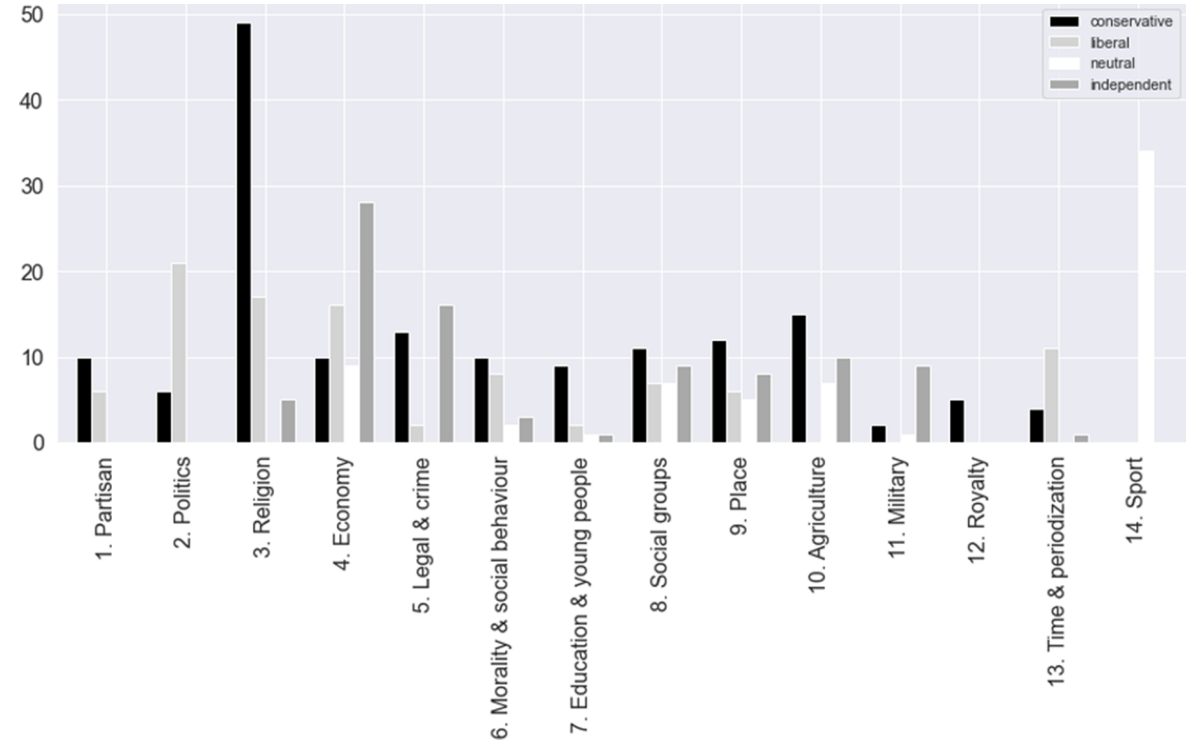

Figure 12. Distribution of thematic word clusters by newspapers' political leaning.

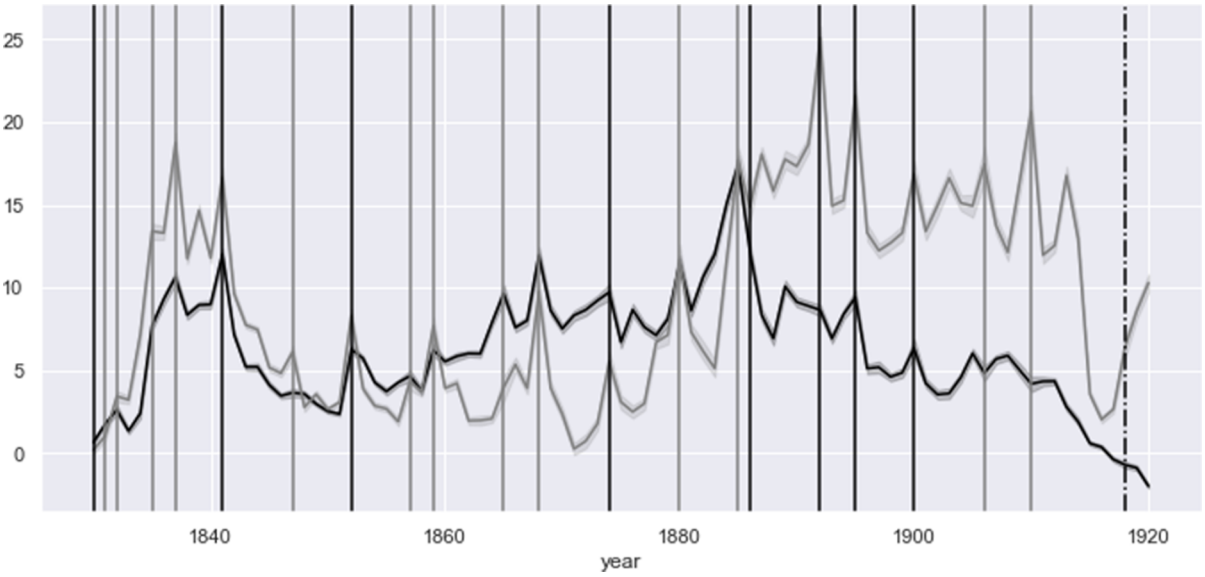

Figure 13. Mean *FW* scores by year for all partisan words in the *Liberal* (grey) and *Conservative* (black) press.
Note: Vertical bars denote General Election years, and the dashed lines indicate a coalition government.

measure how bias has changed as a function of incremental corpus growth (as new titles get added). The X-axis on Figure 14 shows the highest rank of each digitisation batch (in other words, 200 means that the corpus at this step comprises the first 200 newspapers).

Overall, the growth of the corpus has reduced political bias, and especially from a proportional perspective, the data has become more representative (less so according to the equality measure). However, this is not a linear process in all cases. The digitisation timeline suggests that at some point in the process, stagnation in bias reduction appears. For example, if digitisation is assessed using the model of equal representation – giving each group equal presence at each stage – scores show a starker decline during the earlier phase of digitisation, slowing after around 550 titles. A similar pattern occurs for the proportional timeline, albeit the stagnation occurs somewhat later in the process.

The right panel of Figure 14 partly explains these rather abstract patterns, by plotting the political labels' contribution to overall divergence as a function of corpus growth. We used the proportional representation as the target distribution. The initial over-representation of *Conservative* and subsequently *Liberal*

https://doi.org/10.1017/chr.2025.10007 Published online by Cambridge University Press

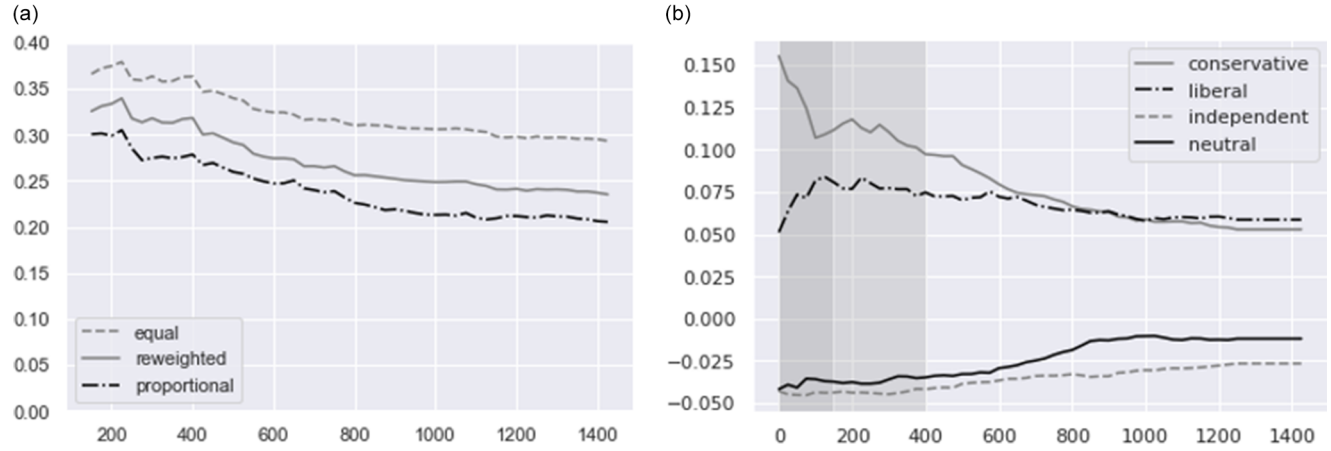

Figure 14. Political bias as a function of digitisation order.
Note: Left: Overall bias measure, lines relate to different interpretations of representativeness. Right: Proportional bias measure for the four principal categories of newspaper affiliation. Shading broadly corresponds with JISC (dark grey) and Gale (light grey). Stride size is equal to 25 titles. Bias was computed on the finer-grained classification.

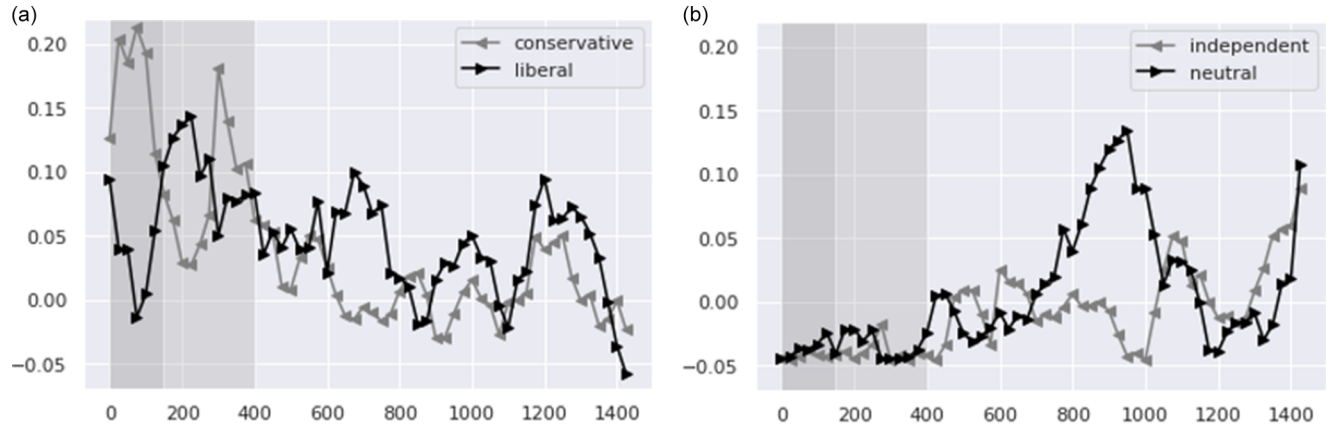

Figure 15. Partial Kullback–Leibler values for each batch of 200 newspapers, looping through the corpus in digitisation order, with a step size of 25 newspapers.
Note: The shading corresponds with JISC (dark grey) and Gale (light grey). No shading reflects the Findmypast period.

newspapers declines over time but persists. The opposite applies to the *Neutral* and *Independent* press.

For Figure 15, we computed the contribution of each political label to the divergence for each "digitisation batch." This analysis is slightly different from the preceding one. Instead of incrementally adding data, we defined a digitisation batch as the set of newspapers that fall within a specific sliding time window w. For example, if we set s to 25 and w to 100, the first digitisation batch comprises all titles with rank 1 to 100, the second contains titles with rank 25 to 125, etc. We then compared the distribution of political labels in each batch to the proportions of those observed in the directories.

Visualising the creation of the BNA in this way underscores how digitisation priorities have changed over time. Earlier efforts in particular seem to have focused on the partisan press, alternating between over-representing first *Conservative* and then *Liberal* newspapers. Something drastic changes after approximately one third of the corpus has been digitised. We observe a steep increase in preference given to *Neutral* newspapers, which were largely ignored in earlier stages. This contributes to driving bias scores lower, as seen in Figure 15. It also corresponds with our previous intuitions about how organisational priorities shape the contours of research data. Here, we can clearly trace the consequences of a transition from digital collections being constructed as research resources for academic users to a focus on the needs of a broader public concerned primarily with local and family histories (Shaw 2005).

These decisions have important implications for the political and other biases encoded in the data; biases that by their nature have been largely undocumented. Therefore, at different stages in the digitisation process, the partisan profile, and hence the social perspective, of the newspapers we can access, changes. Newspaper data are not oligoptic in a static way, but dynamically: their view on the past will always be partial, but shifts with the characteristics of the newspapers selected for scanning.

It is critical to point out that many of these sub-corpora live their own lives. The BNA contains the results of multiple digitisation projects led first by JISC and later by Gale, all of which continue to survive as separate (structured text) datasets that are available to researchers. However, even though the data has changed significantly in its political and other biases, these variations are not flagged to users and each provider tends to emphasise the scale and comprehensiveness of their data (Tolfo et al. 2021). As more researchers depend on these datasets – not simply to identify specific pieces of evidence from a very large collection but also to analyse macro trends, and train language models – the importance of providing clear guidance about their content and biases only increases.

https://doi.org/10.1017/chr.2025.10007 Published online by Cambridge University Press

Conclusion

The principles behind the digital ES could be applied to any big data collection where researchers are able to secure open access to *reference metadata*, or contextual information rich enough to shed light on its hidden biases. With historical big data, we argue that the key is to identify, digitise, and structure high-quality information collated at the time of data creation. Here, we use annually published reference works about British newspapers (the press directories) to enrich the metadata associated with digitised newspaper collections and thereby model the relationship between the contours of our digital corpus and the characteristics of the wider publishing landscape; the vast majority of which remains undigitised. We have focused primarily on the issue of newspapers' declared political leanings (their "partisanship," because this seemed likely to be an attribute that would in turn influence newspaper content, and hence the composition of the big bag of words that constitutes any large text corpus. In the process, we show both that politically-aligned newspapers are over-represented in every iteration of the digitised British press, and that this matters because the content of partisan newspapers differed significantly from that of non-partisan newspapers. Our analysis also strongly suggests that OCR errors are far from randomly distributed across the corpus (see Appendix D). Regardless of whether scholars are using a search interface or conducting large-scale text analysis, they will be disproportionately analysing the words printed by more expensive and conservative newspapers given the uneven occurrence of OCR errors across the collection.

Few would argue with the claim that individual newspapers offer an oligoptic rather than a panoptic view of society, but we insist that this is also true of any digital newspaper *collection*, and indeed of newspapers as a *genre*. Although data providers like to deploy metaphors to suggest that their collections provide a window or mirror through which we can see past society, in practice, they necessarily offer partial perspectives shaped by the worldview of editors, advertisers, and influential readers. We also need to recognise that newspapers were not equal in importance, although given the absence of robust circulation data after the 1850s, it is very hard to take account of readership figures, let alone more intangible factors such as influence. The strong bias towards partisan newspapers in early digitisation efforts almost certainly represents an unintended consequence of the curators' laudable aim to prioritise the selection of long-run, "serious" newspapers from different regions.

Taken together, these factors underscore our central point that the ES is not designed to enable researchers to *correct* for bias in a corpus. Rather, it is a tool for understanding the composition of a corpus and the different biases it embodies. Indeed, our finding that, given its radically uneven distribution, OCR quality is an additional, hidden source of bias within the newspaper corpus, underlines the futility of seeking to create an ideal, "unbiased" sample (just as there can be no "ideal" literary sample, see Bode (2020)). But it remains vital to scale-up the conventional techniques of source criticism to map the vast new collections of data, containing billions of words, that are reshaping scholarship. Only by embedding such procedures across all branches of the Digital Humanities can we avoid the pitfalls of bold overgeneralisation that continue to feed digital scepticism within the academy.

Acknowledgements. Many Living with Machines (LwM) colleagues supported this research. We thank Ruth Ahnert as PI of LwM; David Beavan (an LwM Co-I), who played an important role in leading and supporting the earlier phases of the Environmental Scan research, and contributed to its first publication (Beelen et al. 2023); and Tim Hobson (also an LwM Co-I), Sarah Gibson, and Federico Nanni, who were crucial in transferring, processing, and storing LwM newspaper data. Nanni assisted with the creation of the *NewsWords* data in Turing's Safe Haven environment, where the digitised newspapers were stored. Hobson took the lead in documenting and monitoring the newspaper transfer. These data were fundamental to the analysis presented in this article. We also thank Griffith Rees for his assistance with running and testing HurdleDMR models and providing feedback on the early drafts of this article. Finally, we thank *FindMyPast* for providing access to BNA data.

Data availability statement. Code for reproducing the results is available on GitHub at https://github.com/kasparvonbeelen/whose_news. This article is built on two principal sources: 1) a structured dataset of Mitchell's Newspaper Press Directories from the British Library and 2) the *NewsWords* dataset. The Press Directories are available at https://zenodo.org/uploads/16413810. An older version, used for Beelen et al. (2023) remains accessible on the British Library Research Repository at https://bl.iro.bl.uk/concern/datasets/adcef12a-bb3d-40d9-871d-5784022a77e8*NewsWords* datasets are available on Zenodo as well: The version of unigrams linked with the press directories metadata is accessible at https://zenodo.org/records/14996278 and the un-linked or "raw" version is accessible at https://zenodo.org/records/14826348.

Author contributions. Conceptualisation: K.B., J.L., K.M. and D.W. Data curation: K.B. Analysis and Interpretation: K.B., J.L., K.M. and D.W. Methodology: K.B. and J.L. Project administration: K.B. Resource Management: K.B. Code: K.B. Data Cleaning and Validation: K.B. Visualisation: K.B. and KM. Writing – original draft: K.B., J.L., K.M. and D.W. Writing – review & editing: K.B., J.L., K.M. and D.W.

Funding statement. This article began as a part of the Living with Machines project (AH/S01179X/1), funded by the UK Research and Innovation (UKRI) Strategic Priority Fund, and delivered by the Arts and Humanities Research Council (AHRC), with The Alan Turing Institute, the British Library, the University of Cambridge, King's College London, University of East Anglia, University of Exeter and Queen Mary University of London. It has been completed thanks to Data/Culture: Building Sustainable Communities around Arts and Humanities Datasets and Tools (AH/Y00745X/1), a collaborative project between The Alan Turing Institute, Queen Mary University London, Lancaster University and the Complexity Science Hub funded by the AHRC.

Competing interests. No competing interests to declare.

Ethical standards. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

References

Beals, Melodee, Emily Bell, Ryan Cordell, Paul Fyfe, Isabel Galina Russell, Tessa Hauswedell, Clemens Neudecker, Julianne Nyhan, Mila Oiva, Sebastian Pado, Miriam Peña Pimentel, Lara Rose, Hannu Salmi, Melissa Terras, and Lorella Viola. 2020. "The Atlas of Digitised Newspapers: Reports from Oceanic Exchanges." https://doi.org/10.6084/m9.figshare.11560059.

Beelen, Kaspar. 2025. "Newswords Data (Contextualized Word Counts)." https://zenodo.org/records/14996278.

Beelen, Kaspar, Jon Lawrence, Daniel C. S. Wilson, and David Beavan. 2023. "Bias and Representativeness in Digitized Newspaper Collections: Introducing the Environmental Scan." *Digital Scholarship in the Humanities* 38, no. 1: 1–22.

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

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models be too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, New York, NY, United States, 610–23.

Bode, Katherine. 2020. "Why You Can't Model Away Bias." *Modern Language Quarterly* 81, no. 1: 95–124.

Bode, Katherine, and Lauren M. E. Goodlad. 2023. "Data Worlds: An Introduction." *Critical AI* 1, nos. 1–2. https://doi.org/10.1215/2834703X-10734026.

Brake, Laurel. 2015. "Nineteenth-Century Newspaper Press Directories: The National Gallery of the British Press." *Victorian Periodicals Review* 48, no. 4: 569–90.

Chasalow, Kyla, and Karen Levy. 2021. 'Representativeness in Statistics, Politics, and Machine Learning." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, New York, NY, United States, 77–89.

D'Ignazio, Catherine, and Lauren F. Klein. 2023. *Data Feminism*. MIT press, Cambridge, Massachusetts, United States.

Fyfe, Paul. 2016. "An Archaeology of Victorian Newspapers." *Victorian Periodicals Review* 49, no. 4: 546–77.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64, no. 12: 86–92. Publisher: ACM New York, NY, USA

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. 2019. "Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech." *Econometrica* 87 (4): 1307–40.

Gliserman, Susan. 1969 "Mitchell's "Newspaper Press Directory": 1846-1907." *Victorian Periodicals Newsletter* 4: 10–29.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2022. *Text as Data: a new Framework for Machine Learning and the Social Sciences*. Princeton University Press, Princeton, New Jersey, United States.

Hampton, Mark. 2004. *Visions of the Press in Britain, 1850-1950*. University of Illinois Press, Champaign, Illinois, United States.

Haraway, Donna. 2013. *Simians, Cyborgs, and Women: The Reinvention of Nature*. Routledge, New York, NY, United States.

Hauswedell, Tessa, Julianne Nyhan, Melodee H. Beals, Melissa Terras, and Emily Bell. 2020. "Of Global Reach Yet of Situated Contexts: An Examination of the Implicit and Explicit Selection Criteria that Shape Digital Archives of Historical Newspapers." *Archival Science* 20, no. 2: 139–65.

Hills, Thomas T., Eugenio Proto, Daniel Sgroi, and Chanuki Illushka Seresinhe. 2019. "Historical Analysis of National Subjective Wellbeing Using Millions of Digitized Books." *Nature Human Behaviour* 3, no. 12: 1271–5.

Hobbs, Andrew. 2013. "The Deleterious Dominance of The Times in Nineteenth-Century Scholarship [in en]." *Journal of Victorian Culture* 18, no. 4: 472–97. https://doi.org/10.1080/13555502.2013.854519.

Hobbs, Andrew. 2018. *A Fleet Street in Every Town: The Provincial Press in England, 1855-1900*. Open Book Publishers, Cambridge, United Kingdom.

Holland, Sarah, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2020. "The Dataset Nutrition Label." *Data Protection and Privacy* 12, no. 12: 1.

Hovy, Dirk. 2018. "The Social and the Neural Network: How to Make Natural Language Processing about People Again." In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, Association for Computational Linguistics, New Orleans, Louisiana, USA, 42–9.

Hovy, Dirk, and Shrimai Prabhumoye. 2021. "Five Sources of Bias in Natural Language Processing." *Language and Linguistics Compass* 15, no. 8: e12432.

Jo, Eun Seo, and Timnit Gebru. 2020. Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, New York, NY, United States, 306–16.

Jones, Aled. 2016. *Powers of the Press: Newspapers, Power and the Public in Nineteenth-Century England*. Routledge, New York, NY, United States.

Kelly, Bryan, Asaf Manela, and Alan Moreira. 2021. "Text Selection." *Journal of Business & Economic Statistics* 39, no. 4: 859–79.

King, Edmund. 2005. "Digitisation of Newspapers at the British Library [in en]." *The Serials Librarian* 49, nos. 1–2: 165–81. https://doi.org/10.1300/J123v49n0107.

Kitchin, Rob. 2014. "Big Data, New Epistemologies and Paradigm Shifts." *Big Data & Society* 1, no. 1: 2053951714528481.

Koss, Stephen E. 1981. *The Rise and Fall of the Political Press in Britain*. Vol. 2. London: Hamish Hamilton.

Lansdall-Welfare, Thomas, and Nello Cristianini. 2020. "History Playground: A Tool for Discovering Temporal Trends in Massive Textual Corpora." *Digital Scholarship in the Humanities* 35, no. 2: 328–41. https://doi.org/10.1093/llc/fqy077.

Lansdall-Welfare, Thomas, Saatviga Sudhahar, James Thompson, Justin Lewis, FindMyPast Newspaper Team, and Nello Cristianini. 2017. "Content Analysis of 150 Years of British Periodicals." *Proceedings of the National Academy of Sciences* 114 (4): E457–65.

Latour, Bruno. 2005. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press, Oxford, United Kingdom.

Lawrence, Jon. 2009. *Electing Our Masters: The Hustings in British Politics from Hogarth to Blair*. OUP Oxford, Oxford, United Kingdom.

Lawrence, Jon. 2017. "Class and Gender in the Making of Urban Toryism, 1880–1914." In *European Political History 1870–1913*, edited by T. Mergel and B. Ziemann. 195–218. Routledge, New York, NY, United States.

Lee, Alan J. 1976. *The Origins of the Popular Press in England, 1855-1914*, Rowman & Littlefield Pub Inc, London, United Kingdom.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331, no. 6014: 176–82.

Milligan, Ian. 2013. "Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010." *Canadian Historical Review* 94, no. 4: 540–69.

Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. 2008. "Fightin'words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis* 16, no. 4: 372–403.

Noble, Safiya Umoja. 2018. "Algorithms of Oppression: How Search Engines Reinforce Racism." In *Algorithms of Oppression*. New York University Press, New York, NY, United States.

O'Malley, Tom. 2015. "Mitchell's Newspaper Press Directory and the Late Victorian and Early Twentieth-Century Press." *Victorian Periodicals Review* 48, no. 4: 591–606. https://doi.org/10.1353/vpr.2015.0057.

Padilla, Thomas. 2017. *Always Already Computational: CollectionsData*. OSF. https://doi.org/10.17605/OSF.IO/MX6UK.

Putnam, Lara. 2016. "The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast." *The American Historical Review* 121, no. 2: 377–402.

Quinault, R. E. 1979. "Lord Randolph Churchill and Tory Democracy, 1880–1885." *The Historical Journal* 22, no. 1: 141–65.

Ryan, Yann, and Luke McKernan. 2021. "Converting the British Library's Catalogue of British and Irish Newspapers into a Public Domain Dataset: Processes and Applications. *Journal of Open Humanities Data* 7.

Sambasivan, Nithya, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M. Aroyo. 2021. "Everyone Wants to do the Model Work, not the Data Work": Data Cascades in High-Stakes AI." In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, New York, NY, United States, 1–15.

Shah, Deven Santosh, H. Andrew Schwartz, and Dirk Hovy. 2020. "Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview". In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5248–5264. Association for Computational Linguistics.

Shaw, Jane. 2005. *10 Billion Words: The British Library British Newspapers 1800-1900 Project Some Guidelines for Large-Scale Newspaper Digitisation [in en]*. Oslo, Norway.

Stoler, Ann Laura. 2002. "Colonial Archives and the ARTS of Governance." *Archival Science* 2, no. 1: 87–109.

Taddy, Matt. 2013. "Multinomial Inverse Regression for Text Analysis." *Journal of the American Statistical Association* 108, no. 503: 755–70.

Thylstrup, Nanna Bonde, Daniela Agostinho, Annie Ring, Catherine D'Ignazio, and Kristin Veel. 2021. *Uncertain Archives: Critical Keywords for Big Data*. MIT Press, Cambridge, Massachusetts, United States.

Tolfo, Giorgia, Olivia Vane, Kaspar Beelen, Kasra Hosseini, Jon Lawrence, David Beavan, Katherine McDonough, Estelle Bunout, Maud Ehrmann,

https://doi.org/10.1017/chr.2025.10007 Published online by Cambridge University Press

and Frédéric Clavert. 2021. *Hunting for Treasure: Living with Machines and the British Library Newspaper Collection*, Studies in Digital History and Hermeneutics, 25, De Gruyter, Berlin, Germany.

Westerling, Kalle, Kaspar Beelen, Tim Hobson, Katherine McDonough, Nilo Pedrazzini, Daniel CS Wilson, and Ruth Ahnert. 2025. "LwMDB: Open Metadata for Digitised Historical Newspapers from British Library Collections." *Journal of Open Humanities Data* 11, no. 1.

Wilson, Daniel C. S. 2023. "Working at Scale: What Do Computational Methods Mean for Research Using Cases, Models and Collections?" *Science Museum Group Journal* no. 18. https://doi.org/10.15180/221805.

Zaagsma, Gerben. 2023. "Digital History and the Politics of Digitization." *Digital Scholarship in the Humanities* 38, no. 2: 830–51.

Appendix A. Label remapping

Liberal: "whig; philosophical", "democratic; liberal", "progressive", "liberal; ultra", "liberal; presbyterian", "liberal; anglican", "anti-whig", "liberal; progressive", "liberal; anti-episcopal", "secularist", "independent; liberal", "liberal; moderate", "secularist; republican", "liberal; neutral", "whig; liberal", "liberal; methodist", "liberal; reform", "moderate; liberal; presbyterian", "advanced; liberal; non-comformist", "baptist; liberal", "ultra; liberal", "liberal", "liberal; free-trade", "whig; free-trade; free-church", "liberal; non-comformist", "liberal; liberal; progressive", "advanced; liberal", "free-trade", "liberal; progressive; methodist", "gladstonian", "liberal; roman-catholic", "independent; reformer", "non-comformist; liberal", "radical; advanced; liberal", "liberal; patriotic", "high-church; liberal", "liberal; whig", "advanced; liberal; progressive", "liberal; religious", "liberal; evangelical", "liberal; dissenting", "democratic; progressive", "liberal; independent", "whig", "democratic; anti-poor-law", "advanced", "independent; advanced; liberal", "liberal; democratic", "moderate; liberal", "independent; progressive", "national; liberal", "liberal; anti-poor-law", "whig; ministerial", "advanced; liberal; popular; progressive", "liberal; radical", "liberal; popular; progressive", "high-church; whig", "republican"

Independent: "independent; constitutional; progressive; whig", "liberal; conservative; evangelical", "independent", "liberal; conservative; agricultural", "liberal; independent; constitutional; progressive", "independent; constitutional; progressive", "independent; catholic", "conservative; liberal", "independent; unsectarian", "liberal; independent; constitutional", "high-church; independent", "independent; roman-catholic", "independent; protestant", "liberal; independent; constitutional; whig", "independent; neutral", "neutral; independent", "unionist; independent", "independent; constitutional; radical", "liberal; independent; constitutional; progressive; radical"

Neutral: "no-politics", "moderate", "non-political", "non-political"; "non-sectarian", "non-party", "impartial", "neutral", "non-political; "unsectarian"

Conservative: "unionist", "imperialist", "independent; unionist", "progressive; conservative", "constitutional", "conservative; protective", "protestant; conservative", "tory; old", "unionist; unionist; liberal", "catholic; conservative", "independent; constitutional", "conservative", "liberal; unionist", "neutral; conservative", "anti-radical", "unionist; liberal", "unionist; conservative", "constructive; tory", "constitutional; independent", "constitutional; tory", "patriotic; labour", "conservative; unionist", "high-church; conservative", "conservative; unsectarian", "high; tory", "tory", "patriotic", "church-of-england; constitutional", "independent; conservative", "liberal; conservative", "moderate; conservative", "conservative; protectionist", "conservative; high-church", "tariff-reform", "conservative; independent", "national; conservative", "constitutional; protestant"

National: "national", "sinn-fein", "nationalist; catholic", "repeal", "nationalist; liberal", "national; catholic", "nationalist", "independent; nationalist", "national; independent", "irish-national; catholic", "catholic; national", "home-rule"

Religious: "catholic", "salvation army", "roman-catholic; ultramontane", "church-of-england; neutral", "denominational", "congregational", "ultramontane", "protestant", "unsectarian", "unitarian", "undenominational", "anglo-catholic", "evangelical", "church-of-england", "anglican", "orange", "evangelical; unsectarian", "evangelical; protestant", "non-political; protestant", "presbyterian", "orange-protestant", "religious-equality", "roman-catholic", "non-sectarian", "church-of-ireland", "conservative; anglo-catholic", "high-church"

Radical: "radical", "democratic", "liberal; radical; labour", "radical; reform", "independent; radical", "labour", "social; democratic", "socialist; labour", "progressive; radical", "progressive; radical; labour", "democratic; republican", "advanced; radical", "radical; liberal", "democratic; popular; progressive", "chartist", "liberal; labour", "secularism", "socialism", "socialist", "radical; home-rule", "radical; labour"

Other:"industrial", "anti-corn-law", "coalition", "scientific", "temperance", "cosmopolitan"

Appendix B. Representation of minority categories

When including minority voices, we were especially interested in newspapers identified by oppositional categories, such as "radical," "labour" and "socialist." These are sometimes subsumed within the liberal press, but most purported to speak for a more plebeian readership and to challenge the political status quo. While a small category within the newspaper landscape, these radical titles purported to speak for the majority of the population - i.e., the working classes. In a sense these were majoritarian "minority" titles whose voice could be viewed as more important than their numbers would suggest.

In Figure B1, we measure the extent to which such *Radical* titles, and also those without any label (*nan*), contribute to bias. From a proportional perspective, there exists no overall anti-radical bias (partial KL scores close to zero for the whole period). The *Radical* titles are few in number and proportionality therefore tends to obscure the absence of such minority voices.

However, the other panels in Figure B1 – for example, the one based on the reweighted target distribution – clearly indicate an under-representation of radical titles. While the small numbers involved make measurements potentially error-prone, the historical trends are interesting. Negative bias diminishes between the 1850s and 1870s, but afterwards this trend reverses, suggesting that in times of the expansion/democratization of the press, the digitised sample increasingly underrepresents radical voices. There remains some uncertainty about this measurement, but it is potentially significant.

Lastly, we should note that unlabelled titles, which often appeared in the directories with listings consisting of little more than their title, are also less likely to be digitised. The significance of this bias is necessarily hard to evaluate because of the paucity of contextual information about these newspapers.

Appendix C. Fightin' words semantic clusters

1. *Partisan:* party-political words that signal distinct political identities (e.g., "liberal" or "conservative").

https://doi.org/10.1017/chr.2025.10007 Published online by Cambridge University Press

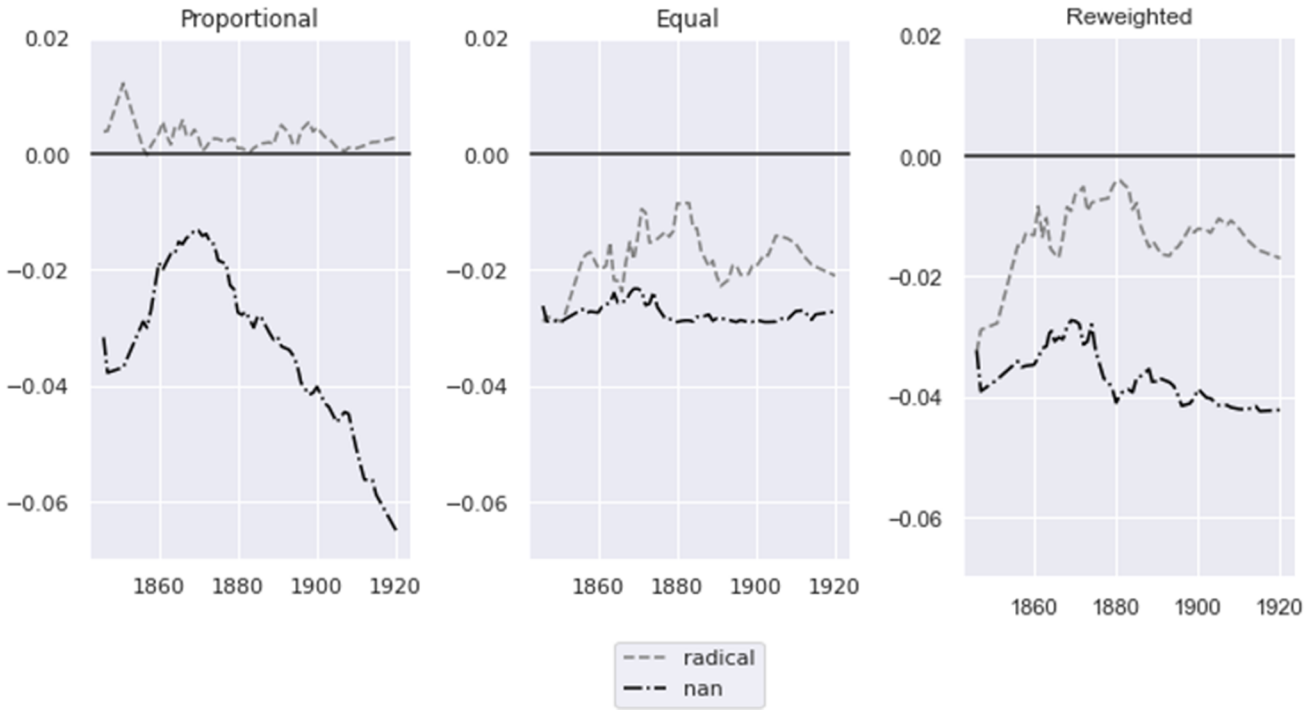

Figure B1. Contribution of secondary political labels to bias scores for each measure of representativeness.

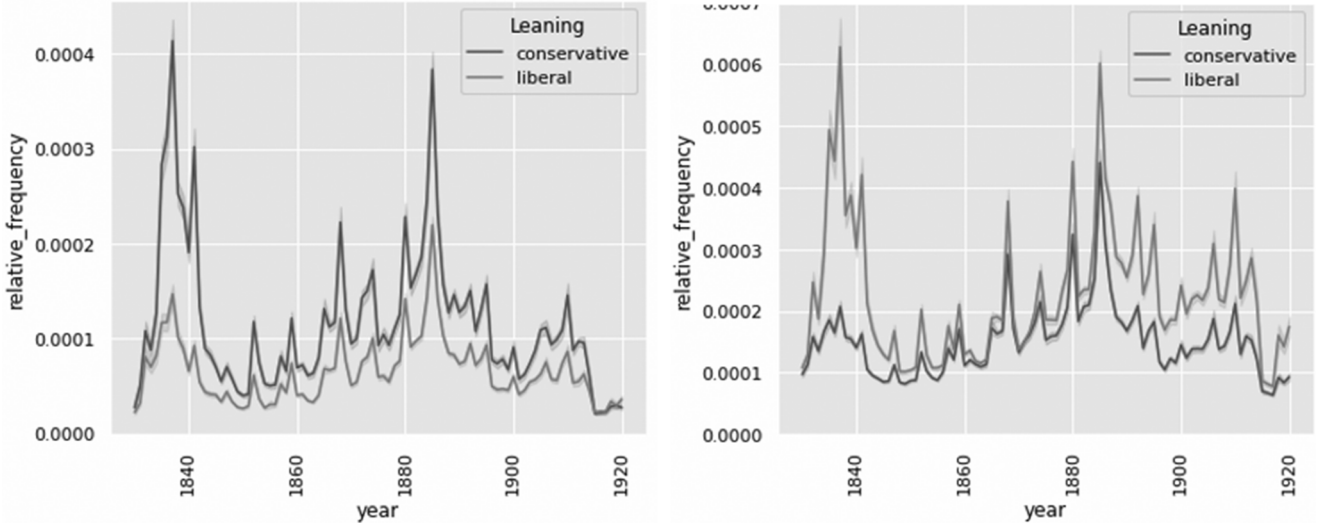

Figure D1. Relative frequency of *Liberal* and *Conservative* partisan words in party-aligned newspapers, by year (left distinctive Conservative words, right distinctive Liberal words). Note that while the temporal patterns are similar the range of values on the *Y*-axis are different.

2. *Politics:* political institutions, practices and processes (e.g., "parliament," "elections" and "legislation").
3. *Religion:* vocabulary used to signal religious practice, denominations, offices, identities, etc.
4. *Economy:* words related to commercial, financial and industrial practices and institutions including raw materials and transport.
5. *Legal and crime:* words related to court proceedings, criminal behaviour, legal practices, etc.
6. *Morality and social behaviour:* words expressing a moral evaluation (desirable or undesirable) or policies with a strong moral component (e.g., "temperance").
7. *Education and young people:* school and teaching, including young people as a distinct demographic category.
8. *Social groups:* words expressing social classification or referring to distinct classes within society including occupational groups (except those related to clusters: e.g., religion, law, education, agriculture, etc.)
9. *Place:* expressions of place and locality (excluding actual place names).
10. *Agriculture:* agricultural practices, occupations, tools, and words relating to crops, livestock and rural life (excluding sport).

https://doi.org/10.1017/chr.2025.10007 Published online by Cambridge University Press

11. *Military:* titles and ranks, military instruments, war related words.
12. *Royalty:* royal titles and practices related to the monarchy.
13. *Time and periodization:* temporal expressions, words anchoring discourse in time.
14. *Sport:* sport related activities, practices and objects.

Appendix D. Partisanship over time and OCR errors

Figure D1 offers some clues to better understand these trends by disaggregating the relative frequencies of these partisan (cluster 1) words by party. This shows a small increase in *Conservative* newspapers' use of Conservative partisan words between 1900 and 1910, but this was dwarfed by earlier peaks in the 1830s and mid-1880s, and also by the intensity of Liberal partisan language after 1900. Overall, words associated with Liberal partisanship appear more frequently (note the different scales), but both sides used their opponents' distinctive lexicon fairly freely from 1850 onwards. Only in the early period, before 1850, did party newspapers seemingly have radically diverging partisan vocabularies. This may be because party labels remained novel in this period; the tories had only recently reconstituted themselves as the Conservative Party (1834), and until 1857 the Liberals remained a broad coalition of reformers rather than a formal party. However, the peak is less pronounced in Figure 13 because we have less data for this period, which dampens the distinctiveness scores.

However, this focus on thematic clusters ignores another important finding of the *FW* method, namely the very uneven occurrence of OCR errors across the digital corpus. As we see in Table 2, a few non-words were sufficiently unambiguous to be coded into categories ("etreet," "hous," "yeeterday" and "dail" – all from the *Liberal* sub-corpus), but most were indecipherable. Overall, we found that 42.5% of the words in the Liberal *FW* list were OCR errors, as were 13% of the neutral words, compared with almost none (1.5%) among the distinctive words generated for the *Conservative* sub-corpus. One hypothesis is that *Liberal* and also *Neutral* newspapers were printed on poorer quality paper which may reproduce less accurately. Certainly, we found that OCR errors were also much more common in cheaper as opposed to more expensive newspapers, irrespective of their partisanship. As a simple test, we contrasted "cheaper" newspapers, those costing one penny or less per issue, with "expensive" newspapers of more than one penny. For the cheaper press, 38% of distinctive word tokens constituted likely OCR errors. The same was true for less than 1% of tokens characteristic of the more "expensive" newspapers. The causes of this pattern require further research, but we can already confidently say that researchers need to bear in mind the likely effects of differential OCR quality on their findings. Whether they are conducting simple interface searches or sophisticated linguistic analyses, their findings are likely to be heavily skewed towards the content, and hence the worldview, of dearer and more *Conservative* newspapers and their readers.

https://doi.org/10.1017/chr.2025.10007 Published online by Cambridge University Press