OXFORD

Reading the unreadable: creating a dataset of 19th century English newspapers using image-to-text language models

Jonathan Bourne[1],*

[1]Centre for Advanced Spatial Analysis, University College London, Gower Street, London, WC1E 6BT, United Kingdom

*Corresponding author. Centre for Advanced Spatial Analysis, University College London, Gower Street, London, WC1E 6BT, United Kingdom.
E-mail: ucabbou@ucl.ac.uk

Abstract

Oscar Wilde said, 'The difference between literature and journalism is that journalism is unreadable, and literature is not read'. Unfortunately, the digitally archived journalism of Oscar Wilde's 19th century often has no or poor quality Optical Character Recognition (OCR), reducing the accessibility of these archives and making them unreadable both figuratively and literally. This paper helps address the issue by performing OCR on 'The Nineteenth Century Serials Edition' (NCSE), an 84k-page collection of 19th-century English newspapers and periodicals, using Pixtral 12B, a pre-trained image-to-text language model. The OCR capability of Pixtral was compared to four other OCR approaches, achieving a median character error rate of 1%, 5x lower than the next best model. The resulting NCSE v2.0 dataset features improved article identification, high-quality OCR, and text classified into four types and seventeen topics. The dataset contains 1.4 million entries, and 321 million words. Example use cases demonstrate analysis of topic similarity, readability, and event tracking. NCSE v2.0 is freely available to encourage historical and sociological research. As a result, 21st-century readers can now share Oscar Wilde's disappointment with 19th-century journalistic standards, reading the unreadable from the comfort of their own computers.

Keywords: digital archives; historical document analysis; LLM; machine learning; NLP; OCR.

1. Introduction

Britain went through an enormous period of change in the 19th century. Inventions such as the steam train, telegraph and anaesthetic were life-changing for millions. It was also a time of immense societal change as Britain began the transition from what modern eyes would consider a sort of industrial feudalism into a liberal democracy. For example, slavery was banned (UK-Government 1833), The proportion of the population eligible to vote increased dramatically (Phillips and Wetherell 1995; Johnson 2013), married women were allowed to own property (UK-Government 1882), and children under 10 were banned from working in Mines (Ashley-Cooper 1842). Throughout this period, improvements in literacy (Stone 1969) and printing technology meant that the changes occurring throughout the country were reported on by journalists and discussed in the many local and national newspapers and periodicals. Indeed, so expansive was the change in literacy and the popularity of the press that

even journalism went through substantial changes. The traditional elite-centred journalism gave way to the 'New Journalism' (Hampton 2008; Twycross 2024) of the late 1800's. This 'New Journalism' focused more on the mass market audience, but was described as 'feather brained' (Arnold 1887) for its perceived low-brow style and possibly being the reason for the quip (Wilde 1891) quoted in the abstract of this paper.

Many 19th-century newspapers and periodicals have survived into the 21st century in original form or as microfilm images. Such archival newspapers are invaluable to modern scholars to understand what happened during the 19th century and how it was perceived at the time. Such understanding can help guide modern society when considering major changes now, dealing with polarizing issues, or morally ambiguous subjects. To make use of these archival resources requires searchable digital databases. However, whilst there has been a substantial effort to digitize historical

archives (Terras 2011), the process has proved challenging (Chiron *et al.*, 2017; Smith and Cordell 2018).

One such historical Archive is the Nineteenth-century Serial Edition (NCSE) (Brake *et al.*, 2008). A collection of six periodicals spanning almost the entire nineteenth century, including over 4,000 issues and approximately 84,000 pages of data. The original attempt at Optical Character Recognition (OCR), performed between 2005 and 2008, resulted in heavily corrupted text. Recently, there have been attempts to recover the original text using Transformer Language Model (LM) based post-OCR correction techniques (Bourne 2024a,b). However, these approaches were only partially successful.

Recently the release of easy-to-use image-to-text models (Agrawal *et al.*, 2024; Beyer *et al.*, 2024; Wei *et al.*, 2024) and deep learning approaches to page layout analysis (Auer *et al.*, 2024; Dell *et al.*, 2024; Almutairi 2025), suggest that the time is ripe for a more robust open source approach to historical archive digitization. These advances allow for the creation of a simple two-part process for extracting formatted text from historical newspapers, which is easily adaptable to new languages.

1.1 Related work

Due to the importance of the digitisation of historical archives and the challenges faced, there has been substantial research in the area of OCR and automated page layout analysis (Girdhar *et al.*, 2024). However, whilst there are many recent OCR and Layout models (Zhang *et al.*, 2021; Li *et al.*, 2022; Huang *et al.*, 2022; Da *et al.*, 2023; Wei *et al.*, 2024s), most of them are designed for extracting text from benchmarks such as scientific articles (Zhong *et al.*, 2019), receipts (Park *et al.*, 2019), or forms (Jaume *et al.*, 2019), not historical documents. This is an important distinction as historical documents have dense text-heavy layouts (Sven and Matteo 2022), resulting in poor layout parsing and text extraction.

Post-OCR correction using text-to-text autoregressive Language Models has recently gained traction as an approach to dealing with the poor OCR of historical documents (Boros *et al.*, 2024; Thomas *et al.*, 2024; Bourne 2024a, b; Kanerva *et al.*, 2025). This approach is based on the ability of the transformer architecture to maintain long-range dependencies across the text (Vaswani *et al.*, 2017), its ability to have a representation of language (Kallini *et al.*, 2024), and its skill at infilling (Devlin *et al.*, 2019) and crucially generative power. Whilst this has had some success, seeing reductions of over 50% in error rates (Thomas *et al.*, 2024; Bourne 2024b), this approach is inherently limited as it uses the corrupted output of the OCR process as opposed to the original images.

The increasing prevalence and ability of multi-modal language models (Agrawal *et al.*, 2024; Dubey *et al.*, 2024; OpenAI 2024; Team 2024; Qwen *et al.*, 2025), allows the input of images and the return of text. Many of these models have shown the ability to detect and read text; however, as yet, their ability to cope with the challenges posed by historical documents is under-explored. The value they provide is the ability to respond to prompts made of interleaved text and images and then draw on both the spatial context within the image and the broader knowledge context embedded in their parameters to provide relevant responses.

Such an ability means image-to-text models are not restricted to simply extracting text but can also be used to embed the entire image and query the embedding (Faysse *et al.*, 2024); in many ways, this allows issues related to layout and OCR to be entirely circumvented. However, OCR reduces the image to its core information, that is, the words on the page, which are easily verifiable, can be searched and read by a human and require very little storage space. This is not to say that there is no place for image embedding in the digitisation of historical documents, but the vast majority of the value they contain is not related to the image and layout but to the writing itself. As such, this paper focuses on OCR.

Figure 1 provides an example of how combining an image with a prompt which requires the knowledge stored in the model's parameters to infer culturally specific meaning. The figure shows an image provided to various language models (Agarwal *et al.*, 2024; Anthropic 2024; OpenAI 2024; Team 2024) models. The image alludes to 'The Importance of Being Earnest' by Oscar Wilde, a culturally significant play in Britain. Given the prompt 'Looking at the attached image to which 19th-century comedy play does it refer? and why?' the models generally returned the correct answer, a typical response and explanation from Anthropic (2024) is as follows

This image appears to reference Oscar Wilde's 'The Importance of Being Earnest' (1895). The emoji

Q: To which 19th century comedy does the below image refer?

A: The importance of Being Earnest

Figure 1. Multi-modal Language models can interleave text and images, allowing them to infer image based cultural references from textual cues. This is essential for effective OCR.

equation shows a baby + handbag = shocked face, which refers to the play's plot point about a baby being accidentally left in a handbag at Victoria Station. This is a crucial element of the play's story, where it's revealed that one of the main characters was found as a baby in a handbag at the railway station's cloakroom. This mix-up forms one of the play's most famous and absurd plot points, leading to the memorable line about being found in 'a handbag' delivered with shock and horror by Lady Bracknell.

This response demonstrates the importance of interleaving text and images. It is this ability to understand context and cultural references that makes multi-modal language models particularly promising for historical document OCR. It is these abilities that this paper attempts to leverage to accurately extract the text from the images stored in the NCSE dataset given the context of the prompt and text itself.

1.2 Objectives and contributions of the paper

Whilst the performance of the large frontier models will consistently outperform the small ones, they can be prohibitively expensive for a relatively small gain. This paper uses Pixtral 12b (Agrawal *et al.*, 2024), which balances cost with performance, to demonstrate the value of pre-trained image-to-text models as a cost-effective OCR tool for historical documents. It then applies the model to the NCSE (Brake *et al.*, 2008) collection in a pipeline to create a new, cleaned, classified and searchable dataset of 19th-century newspapers for the benefit of historians and social scientists. It demonstrates the value of the process by providing an initial analysis of the newly created dataset. Specifically, the paper seeks to answer the following core questions.

- How can images be pre-processed to improve the performance of Pixtral as a system of OCR?
- Is an image-to-text model effective at performing OCR on historical documents?

In addition, it demonstrates the value of the resultant dataset, the NCSE v2.0, by answering the following questions.

- What were the topic differences between the newspapers/periodicals?
- Were there any differences in the readability of the text?
- Over a comparable time-period how does the magnitude of coverage of a topic vary between periodicals?

Beyond the demonstration of image-to-text LMs as efficient OCR tools, the key output of this paper is the

NCSE v2.0 dataset. This dataset is an update of the original NCSE dataset (which can be considered V1.0), with significantly improved OCR quality and bounding box locations, in addition the entries are classified into types of text and topic.

Note throughout this work the words 'newspaper' and 'periodical' will be used interchangably.

2 Data

This paper has two different kinds of data. The data used to test the process and the data used to construct the NCSE v2.0 dataset.

2.1 Test data

This paper will use two datasets to test the quality of the OCR. The BLN600 (Booth *et al.*, 2024) and the NCSE test dataset. The BLN600 is a collection of 600 19th-century articles primarily related to crime. The NCSE test set is an improved version of the data used in Bourne (2024a,b). The NCSE test set comprises thirty-one randomly selected pages from the NCSE dataset. The bounding boxes were manually drawn and classed using the Labelbox platform (Labelbox 2025). The boxes were then transcribed with line breaks such that the text is line aligned. The images are cropped to make using the dataset easier so that each bounding box comes in an image-text pair. This resulted in 358 image-text pairs. A summary of the data used to evaluate the bounding box quality and the OCR is shown in Table 1.

2.2 Data for NCSE v2.0

The entire NCSE dataset is 4,263 pdf files containing a total of 84,509 page images[1]. As shown in Table 2, the dataset is quite diverse, with the periodicals covering a range of different years, significant differences in the number of pages, and the physical size of those pages between periodicals. The Northern Star is over twice as high and twice as wide as the smallest periodical, The English Woman's Journal. These size differences impact processing due to the relative size of the text compared to the page. However, the physical size of

Table 1. The number of pages, bounding boxes, and word counts by periodical BLN600 is treated as a single periodical.

Periodical	Pages	Bboxes	Words
Leader	3	71	6,671
English Woman's Journal	7	30	40,58
Monthly Repository	8	53	6,083
Northern Star	2	78	13,477
Publisher's Circular	6	63	4,257
Tomahawk	4	63	6,327
BLN600	NA	600	297,000

Table 2. The newspapers and periodicals of the NCSE and their key information.

Title	Years	Issues	Pages	Width (cm)	Height (cm)
Monthly Repository and Unitarian Chronicle	1806–37	486	17,734	33	48
Northern Star	1837–52	2226	17,965	51	78
Leader	1850–60	1005	24,391	36	49
English Woman's Journal	1858–64	91	5,663	24	34
Tomahawk	1867–70	188	1,996	35	38
Publishers' Circular	1880–90	285	16,760	34	51

the letters between periodicals is likely relatively similar as historically, the font size of print has fallen in a relatively narrow range (Legge and Bigelow 2011).

In terms of storage, the total size of the dataset is 60 GB. In the original NCSE project, the newspapers were scanned as TIFF files before being converted to PDF format for distribution in the publicly available dataset. The embedded images in these PDFs were stored at significantly higher resolution than necessary for effective OCR. To make these image files easier to process, this paper takes advantage of the fact that the images are all black and white and converts them into 2-bit binary PNG files at 120 dots per inch (except the Northern star, which is much physically larger, so 200 dpi is used). This resolution is sufficient for OCR purposes while substantially reducing computational and storage requirements. The PNG format is chosen as it is lossless, ensuring no additional quality degradation beyond the resolution downsampling, whilst the file format is accepted by the Pixtral API. Converting the images in this way reduces the total dataset size to approximately 27GB, a saving of 55%. In addition, separating each page into a single file allows independent processing.

3. Method

This section is broken into four parts: Layout detection and bounding boxes, OCR Extraction, Text classification, and Analysis of the NCSE v2.0. Figure 2 shows the overall processing pipeline. Each stage of the pipeline has clear inputs and outputs, making substitution and comparison easy. In addition, it allows the performance at each stage to be evaluated independently. Although the details of each stage and the evaluation metrics are introduced in the following subsections, an overview is provided in Table 3.

Without a very highly controlled test set where the position of all characters are known, measuring the interaction between evaluation methods and providing a final interpretable holistic value is challenging. As such this paper takes a modular approach and focuses on evaluating bounding-box and OCR separately with the limitation that error propagation is not evaluated.

3.1 Layout detection and bounding boxes

In the case of archival newspapers, layout detection refers to identifying articles, images, headlines, and so on found on the printed page; these regions of interest are then defined by bounding boxes.

Whilst the original NCSE dataset contained bounding boxes, these were found to be of very poor quality, often missing significant sections of the page, overlapping with other bounding boxes, or crossing multiple columns of text. As such, it was decided that the bounding boxes would be re-made using a pre-trained layout detection model.

Typically, layout detection uses some form of deep learning, whether this be language model-based (Huang *et al.*, 2022; Shehzadi *et al.*, 2023) or a vision model using a convolutional neural network (CNN). This paper considers only models using the CNN due to the combination of high performance and low cost. Specifically, the paper uses DocLayout-Yolo. DocLayout-Yolo is based on the popular Yolo architecture using Yolo-v10, a 15.4 million parameter model[2] trained on 300k synthetic images. In addition, DocLayout-Yolo achieves State-Of-The-Art across several benchmarks.

Two alternative models were qualitatively evaluated; these were the Yolov8 model used by Dell *et al.* (2024) and also the RT-DET model used in the Docling pipeline, which was trained on 2k and 81k images, respectively. However, a qualitative analysis of the performance showed that the Docling model produced a large number of erratic bounding boxes, whilst the model by Dell *et al.* (2024) appeared to consistently miss areas of the page and having sizing issues; this may be due to it's relatively small training set.

One of the disadvantages of the DocLayout-Yolo model is that it is trained to work with a wide range of modern document types, not archival newsprint; this means that it can struggle with dense text-heavy images and also can classify the bounding boxes with impossible classes, such as 'equation'. The model could be fine-tuned. However, sufficient training data is not available, and the data used by Dell *et al.* (2024) was not made publicly available. As such, the bounding
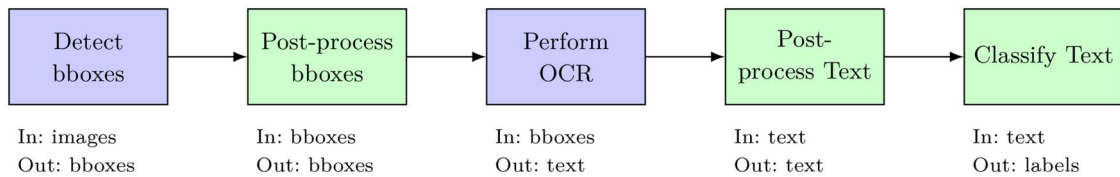
Figure 2. Overall project process showing inputs and outputs at each stage. Blue boxes use the image data, green boxes only require text, or structured data.

Table 3. Overview of pipeline phases with corresponding evaluation metrics and datasets used for testing.

	Phase	Test metrics	Datasets
1	Create Bounding Boxes	Coverage (%), Overlap (%)	NCSE
2	Perform OCR	Character Error Rate	NCSE, BLN600
4	Article Classification	Topic F1, Class F1	NCSE

boxes will be post-processed to increase the quality of archival newsprint.

3.1.1 Post-processing bounding-boxes

The post-processing of bounding boxes is designed to minimise errors related to erroneous class (such as marking 'title' as 'equation'), poor alignment with columns, overlaps between bounding boxes, and gaps within columns. The post-processing does not use the images themselves; instead, it takes advantage of the inherent structure of the newspaper layout. The process utilizes the fact that text and images are rectangular, bounding boxes are exclusive and should not overlap, columns are generally equally sized, and the printed material covers as much of the page as possible.

The process of post-processing the bounding boxes is shown in Algorithm 1 (for detail, see the code for the function 'postprocess_bbox' in the 'bbox_functions' module). Overall, the process is as follows. The DocLayout-Yolo model has a bounding box class 'abandoned' which typically is page, numbers and periodical titles at the top and bottom of the page; this conveniently provides a way of getting rid of unnecessary boxes and miss-classified boxes in that region of the page. Next, the number of text columns on the page are found, and text boxes are assigned to the correct column number, and widths and centroids regularised using the inferred column counts and widths. The system is flexible enough to hand changes on a page; invalid box classes in the rest of the page, such as 'equation' or 'table caption' are often titles and so are mass re-classed as titles. The reading order of the bounding boxes is then found and used to adjust the bounding box's lower edge to the upper edge of the subsequent box This adjustment has the advantage of removing many of the overlapping boxes. Box widths are then extended such that boxes that do not cross the width of the column are extended to column limits, which prevents errors

---

Algorithm 1 Bounding Box Processing Algorithm

Require: *bbox dataframe, minimum height threshold, fill columns*
 1: Reclassify boxes which will be abandoned
 2: Remove abandoned boxes
 3: Assign columns
 4: Re-classify invalid classes as 'title'
 5: Create reading order
 6: Adjust bbox lower limit to fill voids
 7: Extend bbox x-limits to fill column
 8: **if** *fill columns* **then**
 9:     Fill columns with bounding boxes of class 'text'
10: **end if**
11: Remove boxes with height $<$ *minimum height threshold*
12: Merge small boxes
13: Re-adjust bbox lower limit
14: Re-calculate reading order

---

where lines are only partially included. Then, optionally, voids before the first bounding box in the column and after the last bounding box are filled with text bounding boxes; this is to ensure that columns are correctly filled. Then, for efficiency reasons, boxes less than a minimum height are removed, and smaller boxes are merged for efficiency when sending to the LM. The lower limits of boxes are re-adjusted, and the final reading order is calculated. Note for detail on any of the sub-processes in Algorithm 1 please refer to the 'bbox_functions' module of the 'function_modules' library in the codebase (Bourne 2025a).Three different approaches will be taken. The first will be the output of DocLayout-Yolo, which is minimally processed to convert all the non-valid ('equation', etc.) classes to 'title'. The second approach will be the process shown in Algorithm 1, without column filling, whilst the third will be with column filling.
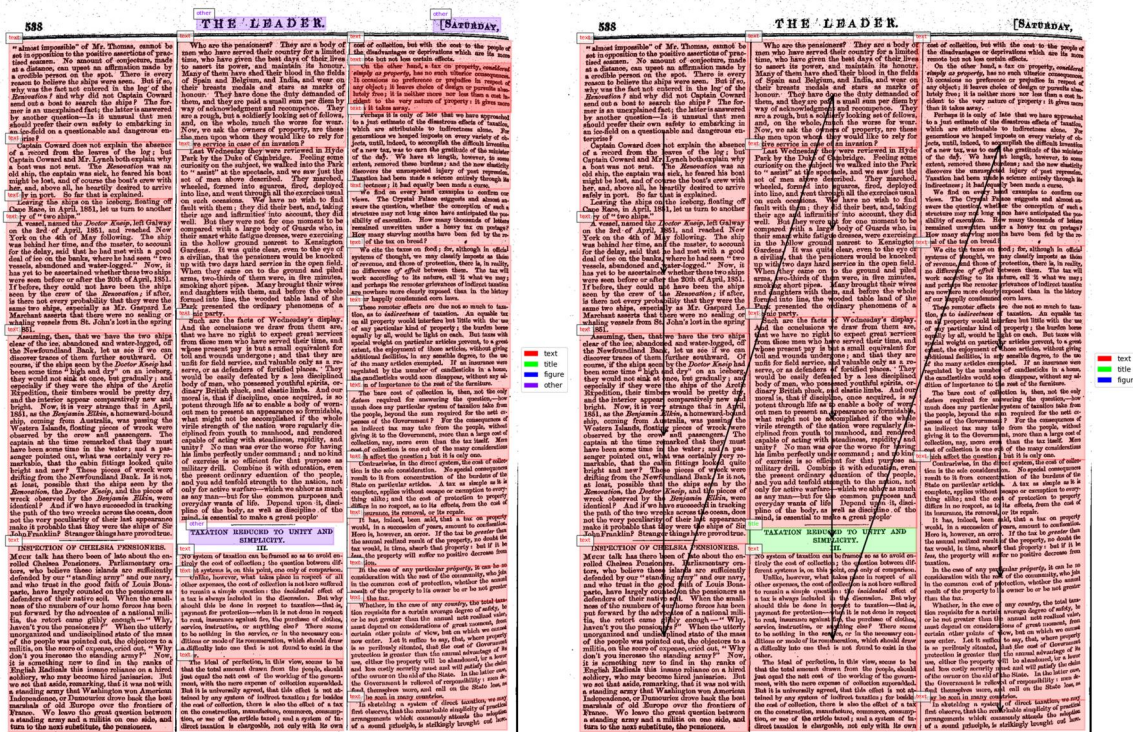
Figure 3. The affect of post-processing bounding boxes is shown for page 'CLD-1852-06-05_page_14'. Post-processing caused box re-classification and the removal of overlaps as well as the addition of the reading order shown as black arrows.

An example of the effect of post-processing bounding boxes is shown in Fig. 3 for page 'CLD-1852-06-05_page_14'. The title box was mis-classed when the original bounding boxes were created (blue in original, green post-processing). In addition, there was an overlap in the top right corner (seen as an area of darker red), which was fixed by the post-processing function. The blue boxes at the top of the page were correctly classed as abandoned by DocLayout-Yolo and were removed during post-processing. In addition, the reading order is shown as black arrows across the page; this is essential for correctly re-assembling the text post-OCR.

While the process can handle a diverse range of layouts, it also struggles to deal with issues such as asymmetrical columns, large overlaps in bounding boxes, and unusual combinations of bounding box class.

3.1.2 Evaluating post-processing of bounding-boxes

Evaluating the quality of bounding boxes and any post-processing without a custom test set is challenging. This is partly because the 'correctness' of bounding boxes is subjective relative to the philosophy of how page design is viewed and the training approach of the object detection algorithm. For example, using the popular Intersection Over Union (IOU) Tanimoto (1958) metric is problematic when granularity differs between ground truth and predictions. For example, for the case when the test set contains bounding boxes at paragraph level, but the predictive model creates the largest possible bounding box by merging paragraphs in the same column. In such a case, the IOU score will be very low despite the quality of the bounding box being high.

Due to these challenges, the quality of the bounding boxes will be evaluated using two complementary metrics: the percentage overlap of all bounding boxes and the fraction of the printed area covered by bounding boxes. These two metrics are granularity-agnostic and do not require aligned ground truth. However, the metrics are simple and won't work well in some instances, for example, empty or nearly empty pages, or when a single bounding box is erroneously very large and covers all other correct boxes. However, they provide a pragmatic solution that does not require a test set.

Each page's Coverage and Overlap scores are calculated using the expressions shown in equations (1 to 3). First, a mask is created the same size as the image

$$M(i,j) = \sum_{k=1}^{n} I[p(i,j) \in b_k] \qquad (1)$$

Where for an image of height h and width w pixels, $B = \{b_1, b_2, \ldots, b_n\}$ is the set of n bounding boxes on

the page, $p(i,j)$ is the pixel at position i,j and $I[\cdot]$ is the indicator function which is 1 if pixel i, j is in bounding box k. This results in an image mask where the value of each pixel is the total number of bounding boxes the pixel is covered by.

Using this mask the Coverage is calculated using

$$\text{Coverage} = \frac{\sum_{i=1}^{h} \sum_{j=1}^{w} (M(i,j) > 0)}{wh} \qquad (2)$$

such that the value is the fraction of all pixels who are covered by at least one bounding box. Similarly Overlap is calculated as

$$\text{Overlap} = \frac{\sum_{i=1}^{h} \sum_{j=1}^{w} (M(i,j) > 1)}{wh} \qquad (3)$$

which is the fraction of all pixels covered by at least two bounding boxes. As the values are all relative to the number of pixels in the original image, both Overlap and Coverage are bounded between 0 and 1.

It should be noted that these two metrics do not indicate that the bounding boxes are correct or appropriate. For example, a single bounding box that covered the entire print area of the page shown in Fig. 3, would appear perfect.

3.2 OCR Extraction

With the bounding box post-processing complete, it is time to consider the OCR approach itself. The process can be broken into three stages: image pre-processing, OCR evaluation and text post-processing.

As this paper is using an autoregressive LLM that has not been fine-tuned, an appropriate prompt needs to be given. Exploring prompt strategy is beyond the scope of this paper, and the prompt is simply designed to cover each of the three main categories 'text' (including 'title'), 'figure' and 'table'. The prompts used are shown in Table 4. Although three categories of

Table 4. Prompts used for converting the data from the different image types into text form.

Prompt type	Prompt text
Text	The text in the image is from a 19th century English newspaper, please transcribe the text including linebreaks. Do not use markdown use plain text only. Do not add any commentary.
Figure	Please describe the graphic taken from a 19th century English newspaper. Do not add additional commentary
Table	Please extract the table from the image taken from a 19th century English newspaper as a tab separated values (tsv) text file. Do not add any commentary

data are being extracted, this paper only focuses on the text class.

3.2.1 Pre-processing images for the LM

Pixtral uses a custom 16x16 vision encoder (Dosovitskiy et al., 2021) to represent images (Agrawal et al., 2024). The model is designed to handle rectangular images of any aspect ratio by flexibly dividing the image into sixteen equally sized patches. They also employ 2D relative rotary position encodings (Su et al., 2023) instead of absolute embeddings to handle changes in the aspect ratio and image size. This innovation is valuable for processing newspaper content because articles are rectangular but have no fixed aspect ratio. In addition, for fixed dots per inch, the size of the image may vary depending on the physical size of the paper. However, care needs to be taken because if the aspect ratio becomes too extreme, the patch embeddings may become quite long or wide relative to individual characters, affecting the ability of the model to understand the text. To check whether the aspect ratio impacts quality, OCR will be performed on the test sets by cropping the article image using aspect ratios of 1 (square crop), 1.5, 2, Inf (no crop). The piecewise images will have a 20% overlap to allow the full text to be reconstructed. If long bounding boxes negatively impact the OCR process, the cropped images will have a lower CER. Figure 4 shows how bounding box 'CLD-1853-07-30_page_2_B0C2R7' is split up when a square crop is used with 20% overlap. The algorithm increases the area of overlap in the last box to ensure it is square. As the third box is only a single line, boxes two and three overlap almost completely. The overlap can be seen in the left image as areas of darker blue.

Whilst cropping the image can handle potential deformation issues within the representation created by the Language Model itself, the proportions of the image are not the only potential source of transcription errors. Rotated pages and warping caused by the image not lying completely flat can also cause difficulties in OCR (Brown and Seales 2001; Girdhar et al., 2024). A second experiment will be performed to automatically correct for page rotation using 'deskewing'. As such, four pre-processing tests will be performed using all combinations of image cropping (True, False) and deskewing (True, False).

3.2.2 Evaluation of OCR quality

OCR quality will be compared using the BLN600 and NCSE test datasets. The NCSE test dataset will be cropped to article or paragraph level and paired with the ground truth text.

Four different OCR approaches will be compared: Pixtral ('pixtral-12b-2409'), Tesseract (Smith 2007) (pytesseract 0.3.13), Efficient OCR (Carlson et al., 2024) (3.2.2), EasyOCR (1.7.2), and GOT (0.1.0).
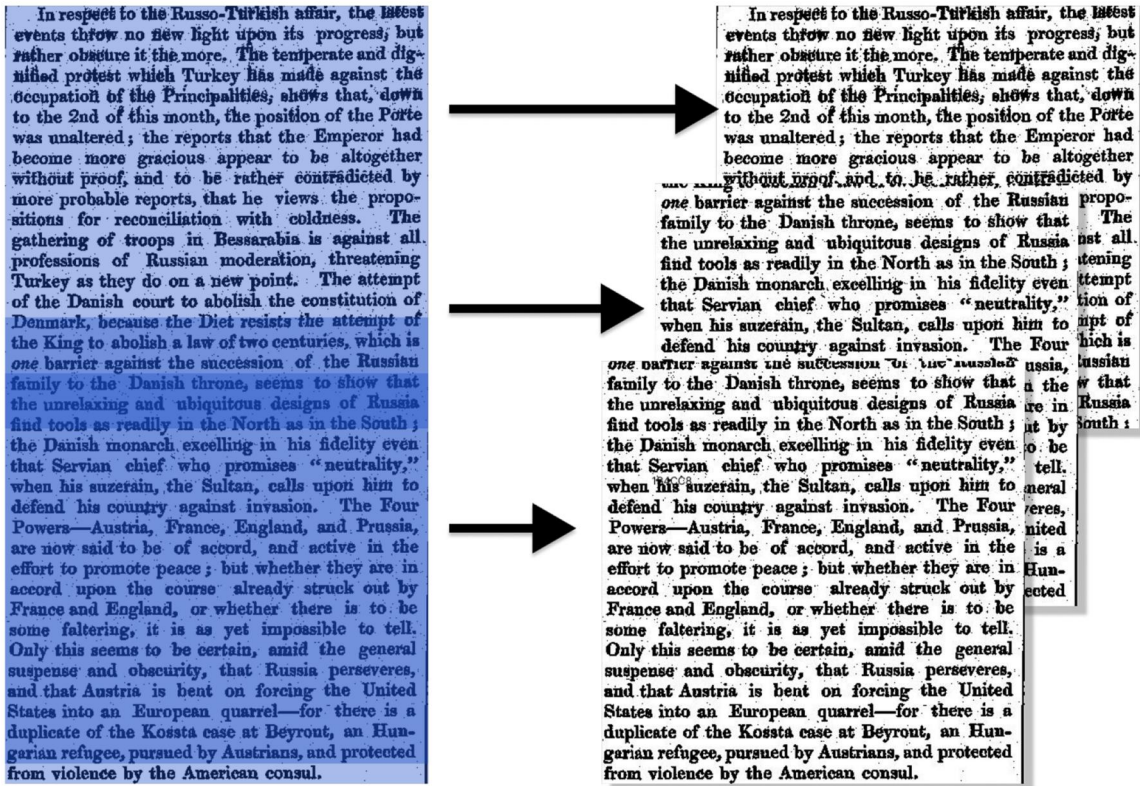
Splitting up a long bounding box

Figure 4. Bounding box 'CLD-1853-07-30_page_2_B0C2R7' is split into three boxes. Overlap is shown as darker shades of blue.

Tesseract, first made open-source in 2005, has become the defacto OCR open-source software that can be considered a baseline level of quality. EasyOCR is newer but has become very popular for its ease of use and ability to extract text from photographs of natural scenes. Efficient OCR is the model used in (Dell *et al.*, 2024) and is the only model designed specifically for archival texts. GOT is a small image-to-text language model developed specifically for OCR purposes and the only other image-to-text model being evaluated. To directly compare the impact of model size, Pixtral-Large ('pixtral-large-2411') will also be used. This model is conceptually the same as Pixtral 12B but 124B parameters, meaning it should perform substantially better. However, Pixtral-Large will not be used to perform OCR on the entire dataset as, at the time of writing, it was 13x more expensive for input tokens and 40x more expensive for output tokens.

The Pixtral models will be used by calling the Mistral API, while all other models will be run using an NVIDIA L4 with 24GB.

Each model will process all images, and the resultant text will be scored against the ground truth using the Character Error Rate (CER).

Character Error Rate (CER) is a commonly used metric for analysing the quality of text recovery in OCR tasks.

$$CER = \frac{S + D + I}{S + D + C} \quad (4)$$

CER is defined as the sum of the character substitutions (S), deletions (D), and insertions (I), divided by the substitutions, deletions, and correct characters (C). Although not particularly sophisticated, CER provides a simple, intuitive and broadly reliable measure of text quality. It should be noted that CER can be greater than 1. This can happen if there is a substantial number of insertions, which can be the case even when the original text is otherwise recovered perfectly. Such occurrences can occur when the LM hallucinates writing text not present in the original. Using the median

when measuring the CER across a range of documents is important, as a major hallucination event in a single document can significantly skew the final result when aggregating with the mean. The impact of using mean vs median can be seen in the contrasting results of (Boros *et al.*, 2024) and (Bourne 2024a) and is directly tested and discussed in (Bourne 2024a).

This paper is concerned with the accurate transmission of the meaning of the text; as such, line breaks, capitalisation, and extra spacing will be ignored when calculating the CER.

3.2.3 Post-processing of OCR results

Two types of post-processing were performed on the dataset. The first is light touch stripping any beginning and end symbols such as ' ' '; the text of class 'title' that is longer than 50 characters will be re-classed as text, and for consistency, line breaks will be removed such that a paragraph is a single line. After this, the dataset can be considered complete.

The second type of post-processing continues after the first, is more aggressive in its changes, and results in a dataset organised into article groups. This post-processing will take advantage of a quirk of 19th-century media, before the arrival of 'New Journalism', which was that there was minimal text formatting (Hampton 2008; Twycross 2024) and titles were predominantly all upper-case and paragraph separated. As such, text boxes will be split into paragraphs, and each paragraph will be tested as to whether it is a title or not using Algorithm 2. The Algorithm ensures that the entire 'Paragraph' in upper case is at least 5 letters and contains at least two vowels; in the case these three tests are True, the class will be changed to 'title';

Algorithm 2 Is Title

Require: text string s
Ensure: boolean indicating if string meets title criteria
 1: **function** IsTitle(s)
 2:    **if** $s \neq$ toUpper(s) **then**
 3:       **return** false
 4:    **end if**
 5:    $lettersOnly \leftarrow$ removeNonLetters(s) ▷
    Remove non-letters and spaces
 6:    **if** length($lettersOnly$) < 5 **then**
 7:       **return** false
 8:    **end if**
 9:    $vowelCount \leftarrow$ countMatches(s, [AEIOU])
10:    **if** $vowelCount < 2$ **then**
11:       **return** false
12:    **end if**
13:    **return** true
14: **end function**

---

otherwise, it will be kept as 'text'. This approach allows for cases where the text contains a title at the beginning, middle or end or for cases where there are multiple titles within a text string. The choice of letters and vowels was made after manually checking the titles of a random sample of texts. After separating these embedded titles and re-classing, consecutive titles will be merged into a single title block. Articles will then be constructed by merging title rows with all subsequent text, figure, and table rows until the next title. This will result in a dataset where each row is, in theory, a single article. The main limitation of this post-processing stage is that creating a ground truth is challenging as the amount of post-processing required is a result of the bounding-box method used and any bounding-box post-processing applied. As such, the existing ground truth cannot be used. Because of this, the title separation will only be evaluated qualitatively.

3.3 Classification

There will be two classification models created for the data. Classifying the type of text, classifying and the topic of the text using the Tier 1 IPTC subject codes (IPTC 2019). The classes of these groups will be as follows.

- Text Type: article; advert; poem/song/story; other
- Topic: arts, culture, entertainment and media; crime, law and justice; disaster accident and emergency incident; economy, business and finance; education; environment; health; human interest; labour; lifestyle and leisure; politics; religion; science and technology; society; sport; conflict, war and peace; weather

This creates, four text-type classes and 17 topic classes.

To create the text-type model, 2,000 samples of class 'text' will be selected from each newspaper to create a training set of 12,000 articles. This sampling approach ensures that each newspaper is equally represented and that there are sufficient data points to give examples of each class. Mistral Large (mistral-large-2411) will classify the texts into 'Text-Type'. A ModernBert Model (Warner *et al.*, 2024) will then be trained and used to classify the entire dataset. Next, another 12,000 examples of class text or title (minimum of 30 tokens) will be selected; this dataset will be classified into one of the IPTC topics using Mistral Large. Like the Text-type classifier, a ModernBert Model will be trained and used to predict across the dataset. The training data will be split 80/20 for train and validation, the details are shown in Table 5.

The trained Text type and Text topic models will be evaluated against the LLM classified dataset using the F1 score, shown in equations (5 to 7). The analysis

Table 5. Number of observations used for training and validating each classification model.

Classification model	Training	Validation
Text type	9,600	2,400
IPTC topic	9,600	2,400

Table 6. Training parameters for the ModernBERT classifiers.

Parameter	Value
Model parameters	149M
Learning rate	2e-5
Batch size	32
Maximum tokens	512
Epochs	5
Warmup ratio	0.1
Optimizer	AdamW
GPU	24GB NVIDIA L4

will use both macro and micro F1 to understand overall model performance, where micro F1 is the F1 for all observations in the data, whilst macro F1 is the mean F1 for individual classes. For the single class classifier used to determine the exclusive text type classes, micro F1 is the same as accuracy.

$$\text{precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (5)$$

$$\text{recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (6)$$

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (7)$$

The ModernBert model is used for several reasons; it is very cost-efficient, having only 149M parameters, architectural changes relative to the original BERT family of models (Devlin *et al.*, 2019; Liu *et al.*, 2019; He *et al.*, 2021) have substantially improved both speed and performance. It should be noted that research has shown that traditional machine-learning approaches outperform auto-regressive language models like Mistral Large (Bohacek and Bravansky, 2024). However, the labelling approach will be considered acceptable when creating a training set. The ModernBert model will provide a consistent and cost-effective method to classify the entire dataset.

The basic training details are given in Table 6. All training and inference was done on an NVIDIA L4 with 24GB of ram. To reduce training time texts were truncated to the first 512 tokens. During inference batch size was increased substantially to 256. For more details on training and inference see the python scripts text_type_trainer.py, IPTC_trainer.py and infer_folder.py available at (Bourne 2025a).

3.4 Analysis
The analysis of the NCSE v2.0 dataset will be simply a demonstration of potential use cases. A selection of 2,000 texts from the class 'article' will be randomly selected from each newspaper to create a subset of 12,000 observations. The distribution of topic labels will be compared, and the similarity between the newspapers in terms of the distribution will be measured using cosine similarity and hierarchical clustering (Müllner 2011). In addition, the Flesch-Kincaid readability (Flesch 1948) will be calculated using the bootstrapped

median of the sub-set (Efron and Tibshirani 1993). Finally, The Leader and Northern Star have a period of overlap covering the 'Great Exhibition' of 1851 (Luckhurst 1951). The Great Exhibition was a world fair of culture and technology making it news worthy in itself. However, it was also housed in the 'Crystal Palace' which was at time the largest glass building in the world (Addis 2006). As such there was likely considerable discussion of the exhibition and the Crystal Palace at the time. For this analysis the count of database entries, which include the phrases 'Crystal Palace' or 'Great Exhibition', will be plotted by month for both newspapers.

4. Results
The results section is divided into five subsections, covering different aspects of the OCR and dataset creation process.

4.1 Post-processing bounding-boxes
Using the 358 bounding boxes from the 31 pages of the ground truth testset and the raw outputs from DocLayout-Yolo, the relative performance of the bounding box post-processing can be evaluated with the coverage and overlap metrics. The overall results are shown in Table 7 and demonstrate that both post-processing methods outperform the raw DocLayout-Yolo bounding boxes, increasing coverage and reducing overlap. The post-processing methods have similar coverage to the ground truth bounding boxes but considerably higher overlap. At page level, both methods generally reduce overlap or leave it unchanged while increasing coverage. However, in the case of the Northern Star, the column-filling approach provides an increase of almost 10 percentage points over the more straightforward post-processing method and 24 percentage points compared to no post-processing. This increase takes the total median coverage to 100% compared to 92% and 79% for simple post-processing and no-post processing, respectively. However, this increased coverage comes at the cost of an increase in mean overlap from 17% to 24%. The Northern Star behaves

Table 7. Coverage and overlap statistics by post-processing method.

Source	Coverage mean	Coverage std	Overlap mean	Overlap std
Ground truth	0.97	0.02	0.01	0.01
Post process	0.97	0.10	0.04	0.19
Post process fill	0.99	0.05	0.05	0.19
No post processing	0.88	0.10	0.13	0.24
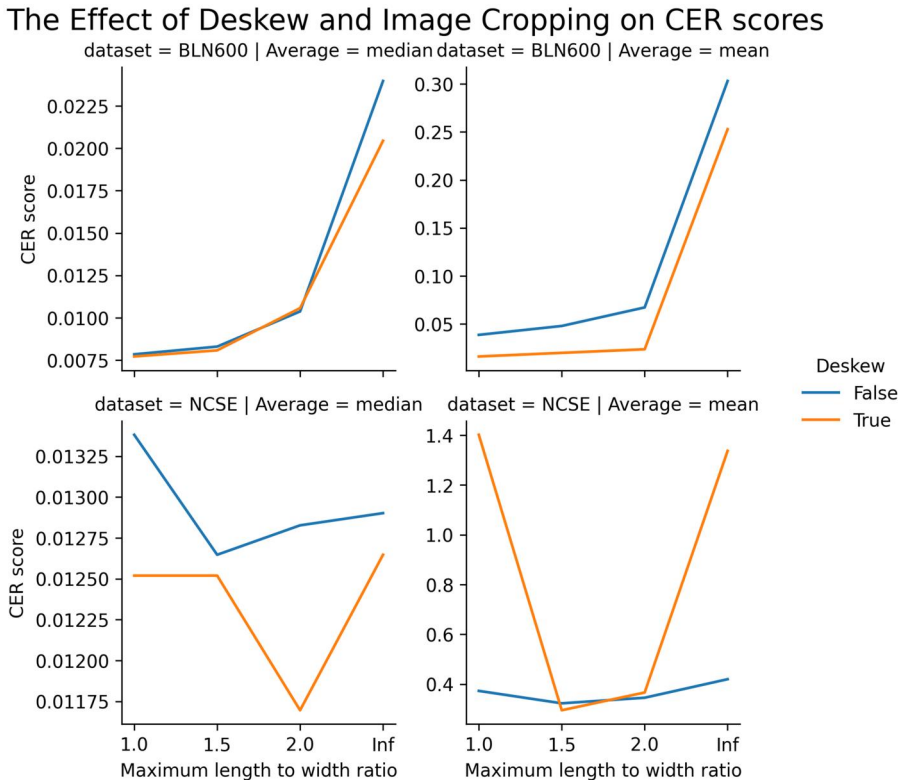
Figure 5. Pre-processing the text-images by enforcing a maximum length-to-width ratio can significantly decrease the CER of the returned text. There is a much larger decrease in the mean CER than the median, indicating that the probability of major errors in a text is substantially reduced by pre-processing. Note, x-axis is the same for all sub-plots.

differently from the other periodicals because it has much larger pages, which contain more columns.

4.2 Image pre-processing

Figure 5 shows the result of the experiments in pre-processing the region of interest before sending to the LM. The results show that without cropping, the mean CER is much larger than the median due to substantial OCR errors. These errors occur for both datasets and with and without deskewing. The errors tend to be re-peating a token or groups of tokens until the token limit is reached or simply commenting that the image looks like a newspaper text. Such responses can get CER values much higher than one, even into the

hundreds. The impact of deskew is more nuanced, as it causes a consistent and significant reduction in the mean CER for BLN600, but the effect on the mean NCSE CER is unclear. In addition, the median CER is only marginally affected. Interestingly, the impact that a larger image window has on the total number of tokens is very different between the two datasets. BLN600 has a mean total token of 2,740 for no crop and 8,950 for a square crop, an increase of 320%. In contrast, the NCSE test set has 2,210 for no crop and 2,635 for square crop, an increase of only 20%. This difference is likely due to the ratio of the bounding boxes themselves; BLN600 uses the entire image as a bounding box, whilst NCSE is more like paragraph level. As a result of this analysis,

Table 8. The results show Pixtral outperforms all other models, when evaluated using median CER.

Model	BLN	NCSE
GOT	0.41	0.06
EasyOCR	0.28	0.26
EffOCR	0.35	0.62
Pixtral	**0.01**	**0.01**
Tesseract	0.05	0.05

this paper will not use deskew; the image crop ratio will be 1.5, as will the merge-height in the bounding-box post-processing discussed in section 3.1.1.

4.3 Comparison of OCR quality

Table 8 compares Pixtral, Tesseract, EasyOCR, and Efficient OCR and GOT for both the NCSE test set and BLN600. The table shows that Pixtral comprehensively outperforms all other methods. In addition, the other methods failed to beat the Tesseract baseline.

Whilst Table 8 shows the median results, the mean results were also calculated and compared. All models had lower median scores than the mean, except for Efficient OCR, which scored better on the mean NCSE test set. The difference between Mean and Median performance was particularly noticeable for Pixtral and GOT on the NCSE dataset. Pixtral got a mean of 0.27, and GOT got a mean of 1.25. These errors were primarily driven by getting locked into repeating a set of characters until the token limit was reached. Pixtral Large was also used to extract the text from the datasets to test the impact of parameter size on performance. Both Pixtral models have the same median CER. However, the mean is quite different. Pixtral large has a mean CER score of only 0.01 and 0.06 on the BLN600 and NCSE test set, respectively, much lower than Pixtral 12B. This substantial difference is because unlike Pixtral 12B or GOT, Pixtral Large does not get stuck repeating phrases.

However, the overall difference between the two Pixtral models is slight; of the 958 observations across both datasets, 98% Pixtral-Large CER scores were less than 0.1; in contrast, Pixtral 12B achieved this 96% of the time, showing just how skewed the distribution is and the importance of using the median for evaluation.

As well as having the best overall performance, Pixtral was the model most likely to have zero errors, with 7% of the responses having a CER of 0. GOT had 1.5% with zero error, and Tesseract had 0.5%. All results from EasyOCR and EffOCR had some errors.

4.3.1 Performance on example text

Figure 6 shows an image taken from Northern Star 1st April 1843, [3]. The print is not particularly clear, with some smudging and streaking evident. To provide a

Figure 6. An excerpt from the Northern Star 1st April 1843, with densely packed text and poor quality printing. This produced broadly average results across the models.

Table 9. The first 200 characters of the Northern Star Example in ascending order of error.

Model	CER	Result
Pixtral	0.01	accrington. a public meeting was held last night, monday, march 27th, in the large weaving shed belonging to messrs. dixon and co. to adopt a petition to the house of commons against the bill now bef
Tesseract	0.11	accrington. a posuic mzerme was held last night, monday, march zish, in the large weaving shed belonging to! messra. dizon and c3, to adopt @ petition to the—honse of commons against the bill now b
EasyOCR	0.28	aggringtox, 4 prauc mxgroe pag beld jast pigbì mondaj, marcb 271b, in the feaving shed belonging i0 messra diron ad c5: adopt 8 petition lo the honse ol csmmors against the billnow before far hiamen
EffOCR	0.58	accblngtqn, 4 feb1ic meeiie8 wa held lat niphfmonday 4i4a8eu q *q, iii me 1afes weaving snfq petpdp1nr 10 442ihth4 9ne deter regulation of the factqf3f:. ! avi g@eql14& an t3nea1ou to the i1fagt 1abd
GOT	34.97	accrngton. a peblc mrtng wacheld last night, monday, march 27th, in the large weaving shed belonging to messrs. dixon and co. to adopt a petition to the honsc of commons agmist the bill now before par

more intuitive example of the errors, the first 200 characters of the returned texts are shown in Table 9, along with the CER for the entire text. The models generally perform close to the test set median, except GOT which became stuck in a loop repeating 'Honsc of Commons agmist the bill now before Par-' until the token limit was reached, producing a massive CER of 35. This kind of error is typical of, and unique to, the errors observed in LLM-based OCR, and both GOT and Pixtral suffer from it. If it had not been for this repeating error, GOT would have been a close second to

Pixtral. The example also shows how critical low errors are to create a usable reference text; even error rates of 11% produce a substantial amount of unreadable text.

4.4 Creating the NCSE v2.0

With the image pre-processing method decided and the clear result of the relative performance of the Pixtral model, the entire NCSE dataset was processed, creating the NCSE v2.0. Creating the dataset required 3.5 billion prompt tokens sent to the model and 599 million completion tokens generated, for a total token count of approximately 4 billion tokens. At the time of writing the cost of using the Pixtral api with batching (upto 24 hour delay but 50% cheaper) was $0.075 per million for both prompt and generation, this makes the cost of OCR for NCSE v2.0 about $4e^9 \times \$0.075e^{-6} = \$300$. The data set contains 82,690 pages of recovered data, which is 98% of the original image pages, 321 million words and 1.9 billion characters. As would be expected given the variance shown in Table 2, there are large differences in the volume and type of information between periodicals; the details of this can be seen in Table 10. The dataset has 1.1 million text entries, 198 thousand titles, descriptions of 17 thousand figures, and 16 thousand tables. After the more aggressive post-processing of the OCR, there were 841 thousand articles, which, given the total number of text boxes, seems high, suggesting the post-processing method is unreliable for improving title separation.

The ModernBert models performed relatively well. The text-type classifier had an accuracy/micro F1 of 0.85 across the four classes and a macro F1 of 0.78, reflecting the over-representation of the 'article' class.

The IPTC-topic classifier had a similar level of performance; across the seventeen classes, it had a micro F1 of 0.79 and a macro F1 of 0.72, reflecting the poor performance of some of the smaller classes. As a result, the models were used to classify the entire dataset by text type and topic.

4.5 Analysis

Table 11, shows the most frequently occurring topics by each periodical. There were substantial differences in the popularity of topics between the newspapers, which makes sense considering their focus. The English Woman's Journal's top topics of Human Interest, Society and Labour are unsurprising given that they were advocating for equal rights for women; likewise, Almost 50% of the topics of the Monthly repository were related to Religion. The Leader had substantial coverage of Conflict War and Peace as it was active during the Crimean War (Badem 2010; Goldfrank 2014) and the Second Opium War (Phillips 2012). Although there was a lot of variety in the most popular topic for each periodical, the least popular were broadly similar, with the Environment, Sport, and Weather seldom covered among any of the periodicals. The topic similarity of the periodicals was measured using cosine similarity visualised in Fig. 7 using hierarchical clustering. The Figure shows that the Northern Star and the Leader are very similar, followed closely by the Tomahawk; these three newspapers share a common interest in politics. In contrast, the English Woman's Journal, Publishers Circular, and Monthly Repository are quite distinct due to their focus on Woman's rights, Arts/Entertainment, and Religion, respectively.

Table 10. Key information on the NCSE v2.0 dataset.

Periodical	Issues	Pages	Figures	Tables	Text	Titles
English Woman's Journal	91	5,638	16	27	11,766	1,541
Leader	1005	23,691	1539	8,351	2,17,207	43,985
Monthly Repository and Unitarian Chronicle	329	17,425	353	157	71,571	7,525
Northern Star	2215	17,930	1907	3,476	7,78,923	1,20,210
Publishers' Circular	236	16,174	12524	3,180	60,873	22,097
Tomahawk	187	1,832	959	534	7,161	3,000

Table 11. Most common IPTC topic categories by periodical.

| Periodicals	Top categories
English Womans Journal	Human Interest, Society, Labour
Leader	Politics, Conflict War Peace, Business/Finance
Monthly Repository	Religion, Politics, Human Interest
Northern Star	Politics, Crime Law Justice, Labour
Publishers Circular	Arts/Entertainment, Education, Business/Finance
Tomahawk	Politics, Human Interest, Arts/Entertainment
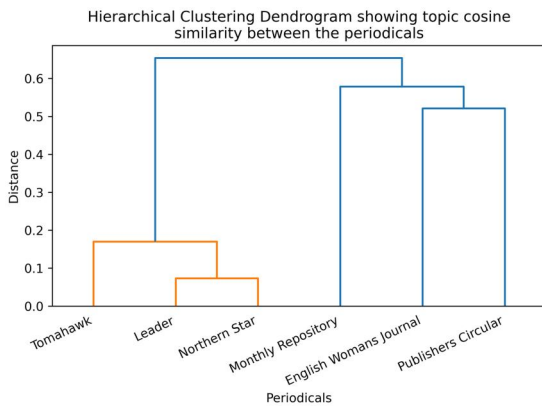
Figure 7. Periodical similarity based on topic distribution shows that the Leader, Northern Star, and Tomahawk are relatively similar due to the heavy focus on Politics; in contrast, the other three periodicals have quite distinct topic distributions. Distance is measure between 0 and 1.
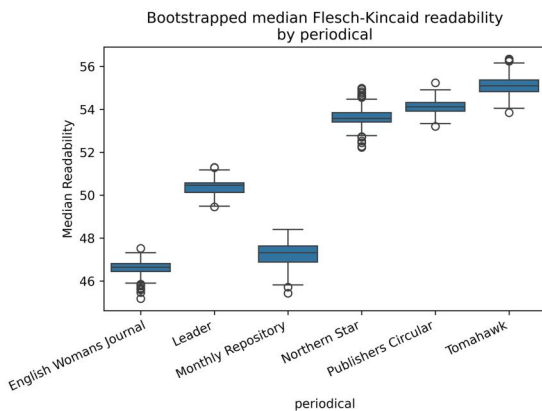

Figure 8. Although there are differences in the average readability, all the periodicals are broadly similar. By modern standards, they would be classed as relatively difficult to read.

The median bootstrapped Flesch-Kincaid readability of the sampled periodicals was analysed as is shown in Fig. 8. Despite their substantial topic differences and large temporal spread, they all had a similar reading level, roughly the range 45–55 out of 100, which by modern standards is considered fairly difficult to read. It should also be noted that the archaic nature of the newspapers may add some noise to the readability score.

Finally, Fig. 9 shows the Leader and Northern Star's coverage of the Great Exhibition in the Crystal Palace. The figure shows that coverage jumps in May when the exhibition started and drops in December as the exhibition closed in November. Although both Newspapers follow the same pattern, the Northern Star has more coverage overall, with 745 mentions to 604.
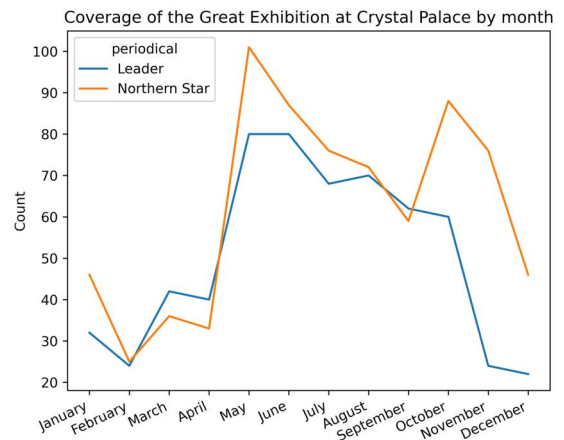

Figure 9. A comparison of the coverage of the 1851 Great Exhibition for the Leader and Northern star.

5. Discussion

With only light pre-processing of images, an off-the-shelf image-to-text model dramatically outperformed four purpose-built OCR engines on historical document digitisation tasks. Pixtral obtained a median CER of 0.01 on both datasets, an error rate that can be described as 'excellent' quality (Hodel *et al.*, 2021). In addition, no model, apart from Pixtral, beat the Tesseract baseline, indicating that the current deep learning training sets and benchmarks are not appropriate for the noisy text-dense images common with historical documents.

Despite being the only method explicitly designed for historical documents, Efficient OCR substantially underperformed its published results. This could be because, whilst Efficient OCR had a substantial silver training set, there was only a small hand-transcribed ground truth. In addition, the test set was only 275 lines of text. This may have led to over-fitting during training and/or an unrepresentative test set, leading to overconfidence in performance.

In many ways, this paper raises the same implicit critique of OCR that CoPali (Faysse *et al.*, 2024) does, which is that traditional OCR fundamentally misses the point when it comes to how text, particularly news media, is read and how it is designed to be read, which is as a whole. By combining the OCR abilities of the Pixtral LM with the page-parsing ability of DocLayout-Yolo, this paper creates a functional solution to the tension between character extraction and the greater semantic context. This combination results in very high-quality OCR within the bounding boxes. However, it lacks the page-level representation of CoPali, which can lead to missing areas of text and overlapping text boxes. However, this drawback is mitigated by raw text output, which is much more accessible and shareable.

The results in this paper show that the most serious issue relating to using language models with OCR is the risk of hallucination, particularly in getting locked into repeating phrases or characters. This phenomenon was experienced by both Pixtral and GOT but not by Pixtral Large. Cropping images substantially reduces the risk of major errors; however, fine-tuning would be the most effective approach without increasing model size. Fine-tuning has been shown to allow smaller models to compete with larger equivalents on the specific tasks they have been fine-tuned on. Therefore, it is reasonable to believe that fine-tuning Pixtral-12B could perform similarly to Pixtral-Large but at a much lower cost. However, fine-tuning does not entirely solve the issue but can be combined with other approaches, Poznanski et al. (2025) detect repeating n-grams in the final output and re-run the extraction, exploiting stochastic variance to avoid the repeating error on subsequent generation.

As mentioned earlier, an issue with fine-tuning is the lack of datasets. However, research has shown that LMs can be trained and fine-tuned on entirely synthetic data (Gunasekar et al., 2023; Li et al., 2023; Tan and Wang, 2024; Wei et al., 2024; Bourne 2024b). Synthetic historical documents could be generated by creating a system that converts text to a synthetic newspaper layout. Subsequently, corrupting the synthetic images with noise (Gupte et al., 2021) would allow the training of Language models robust to the issues observed in this paper. Such highly controlled data would solve many of the measuring issues this work has faced, such as estimating the quality of title identification, and promoting a straightforward way of measuring total image-to-text transcription quality, which was mentioned at the beginning of the method.

This paper also shows the limitation of using Language models for post-OCR correction (Soper et al., 2021; Boros et al., 2024; Thomas et al., 2024; Bourne, 2024a,b; Kanerva et al., 2025). In Bourne (2024a), the base OCR error was 0.17 and the best performance 0.07. However, it is now clear that the original NCSE dataset had very poor bounding boxes. In contrast, in this paper, with accurate bounding boxes, the base case Tesseract achieved a CER score of 0.05. As such, it can be seen that post-OCR correction is no substitute for a robust bounding-box process. Such a process would have high page coverage and low overlap, accurately bounding the elements of the page. Due to the lack of sufficient training data, the creation of synthetic datasets would also provide a substantial boost in training object detection models and possibly help avoid the errors observed when applying layout models trained on modern data (Auer et al., 2024; Zhao et al., 2024). However, consideration can also be given to the speed quality trade-off as CNN-type models such as YOLO are extremely fast but are not able to consider the page holistically in the same way as an image-to-text model such as (Faysse et al., 2024). The impact on process quality may be such that it is worth having more expensive bounding box detection to increase the page coverage and decrease the overlap of the bounding boxes, using some sort of image-to-text model.

Classifying the text and articles into text-type and IPTC-topic is valuable as a method of understanding the text. In this paper, the classification method was crude, with a silver set of data created using an LLM, then a much simpler classifier trained and used to predict across the entire dataset. The performance of the two classifiers was acceptable for the paper. However, the training dataset could be made more robust using data-centric cleaning methods (Northcutt et al., 2021). Similar to how the IPTC classification system was used for the texts, adverts could be classified by sector using already existing taxonomies (IAB 2025).

Although brief, the analysis of the NCSE v2.0 gives an idea of possible use cases when historical documents are stored as text. It is easy to see that the study could be extended using other techniques such as sentiment analysis (Medhat et al., 2014; Wankhade et al., 2022), vector embedding (Pilehvar and Camacho-Collados 2020; Smits et al., 2025) and Retrieval Augmented Generation (Lewis et al., 2020; Hogan et al., 2025), or Discourse Analysis (Atkinson et al., 2000) to name a few.

6. Conclusion

This paper has demonstrated that pre-trained image-to-text models like Pixtral 12B are highly effective for performing OCR on archival documents. Pixtral significantly outperformed the other approaches, having a CER of 0.01 on both datasets, 5x lower than the second-best approach, Tesseract. This performance was despite not being designed specifically for OCR. The results demonstrate the flexibility and general-purpose nature of modern LMs. In addition to the OCR quality, the process only cost about \$3.7 per 1,000 pages, making it appropriate for low-resource projects.

The most significant limitations in the current process are poor bounding box quality and the Language Model getting stuck and repeating sections of text. Both of these issues could be substantially improved in future work using synthetic data and fine-tuning.

The process developed in this paper resulted in the creation of the NCSE v2.0 dataset. The dataset contains 84,000 pages of data across 1.4 million entries and 321 million words. All entries are categorized by type and topic. Analysis of the NCSE v2.0 showed that with high-quality OCR and clear classification, quantitive comparative analysis of archival documents is made substantially easier.

It is hoped that the method developed in this paper and the resultant dataset 'NCSE v2.0' will reduce the difficulty of studying early news media because while the idea that 'The one duty we owe to history is to re-write it' (Wilde 1891) is certainly to be avoided in historical document OCR; everyone can agree that history should be both readable and read.

Author contributions

Jonathan Bourne (Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing—original draft, Writing—review & editing)

Data availability

The Data used in this paper is available at the UCL data repository 'NCSE v2.0: A Dataset of OCR-Processed 19th Century English Newspapers' (Bourne 2025b). The repository contains the NCSE test set, the NCSE v2.0 at the bounding box level, the bounding-box dataset, and the IPTC and text-type silver sets, all data is made available under a Creative Commons Attribution 4.0 International License (CC BY). All code for the project is available at the github repository for the project (Bourne 2025a).

The original NCSE page images can be found at the relevant King's College London Repositories (Turner 2024a,b,c,d,e,f), or linked to at the central project archive (Turner 2024g).

Notes

1. This is about 12k lower than the NCSE metadata; this difference is due to multiple editions of the same paper being present in the metadata but not the image data.

2. This is very small compared to most language models; for reference, it is almost 1000 times smaller than the Pixtral language model, used for the OCR
3. For reference to the data set, the bounding box ID is 'NS2_1843-04-01_page_4_B0C1R1'

References

Addis, B. (2006) 'The Crystal Palace and Its Place in Structural History', *International Journal of Space Structures*, **21**: 3–19. https://doi.org/10.1260/026635106777641199.

Agarwal, R. *et al.* (2024) Many-shot in-context learning. arXiv: 2404.11018 [cs], http://arxiv.org/abs/2404.11018, preprint: not peer reviewed.

Agrawal, P., *et al.* Pixtral 12B. arXiv: 2410.07073 [cs], http://arxiv.org/abs/2410.07073, October 2024, preprint: not peer reviewed.

Almutairi, A. (2025) 'Newspaper Elements Detection and Newspaper Pages Categorization Using CNNs and Transformers', *International Journal on Document Analysis and Recognition (IJDAR)*, **28**, 225–237. https://doi.org/10.1007/s10032-024-00503-9.

Anthropic (2024) The Claude 3 Model Family: Opus, Sonnet, Haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7c babf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

Arnold, M. (1887) 'Up to Easter', *The Nineteenth Century*, **21**, 629–643.

Ashley-Cooper, A. (1842.) An Act to prohibit the Employment of Women and Girls in Mines and Collieries, to regulate the Employment of Boys, and to make other Provisions relating to Persons working therein. King's Printer of Acts of Parliament. https://www.legislation.gov.uk/ukpga/Vict/5-6/99/contents/enacted.

Atkinson, P., Bauer, M. W., and Gaskell, G. (2000) *Qualitative Researching with Text, Image and Sound: A Practical Handbook for Social Research*. SAGE.

Auer, C., *et al.* Docling technical report. arXiv: 2408.09869 [cs], http://arxiv.org/abs/2408.09869, August 2024, preprint: not peer reviewed.

Badem, C. (2010) *The Ottoman Crimean War (1853-1856)*. Brill. https://library.oapen.org/handle/20.500.12657/38173.

Beyer, L., *et al.* PaliGemma: A versatile 3B VLM for transfer. arXiv: 2407.07726 [cs], http://arxiv.org/abs/2407.07726, October 2024, preprint: not peer reviewed.

Bohacek, M. and Bravansky, M. (2024) 'When XGBoost Outperforms GPT-4 on Text Classification: A Case Study', in An. Ovalle, K.-W. Chang, Y. T. Cao, N. Mehrabi, J. Zhao, A. Galstyan, J. Dhamala, A. Kumar, and R. Gupta (eds) *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pp. 51–60. Mexico City, Mexico: Association for Computational Linguistics.

Booth, C. W., Thomas, A., and Gaizauskas R. (2024) BLN600: A Parallel Corpus of Machine/Human Transcribed Nineteenth Century Newspaper Texts. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2440–2446,

Torino, Italia, May. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.219.

Boros, E., et al. (2024) Post-correction of Historical Text Transcripts with Large Language Models: An Exploratory Study. *LaTeCH-CLfL 2024*. The 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. The Association for Computational Linguistics.

Bourne, J. CLOCR-C: Context leveraging OCR correction with pre-trained language models. https://arxiv.org/abs/2408.17428v1, August 2024a, preprint: not peer reviewed.

Bourne, J. Scrambled text: training language models to correct OCR errors using synthetic data. arXiv: 2409.19735 [cs], http://arxiv.org/abs/2409.19735, September 2024b, preprint: not peer reviewed.

Bourne, J. (2025a) Codebase for creating the NCSE v2.0 dataset. https://github.com/JonnoB/reading_the_unreadable. original-date: 2024-06-01T06:22:56Z.

Bourne, J. (2025b) NCSE v2.0: A Dataset of OCR-Processed 19th Century English Newspapers, February.

Brake, L., Mussel, J., and Turner, M. (2008) Nineteenth-Century Serials Edition (NCSE) (2008; 2018). https://ncse.ac.uk/periodicals/.

Brown, M. S., and Seales, W. B. (2001) Document restoration using 3D shape: a general deskewing algorithm for arbitrarily warped documents. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pp. 367–374 vol. 2. doi: 10.1109/ICCV.2001.937649. https://ieeexplore.ieee.org/abstract/document/937649.

Carlson, J., Bryan, T., and Dell, M. (2024) 'Efficient OCR for Building a Diverse Digital History', in Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8105–8115, Bangkok, Thailand: Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.440. https://aclanthology.org/2024.acl-long.440/.

Chiron, G., et al. (2017) Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 1–4. doi: 10.1109/JCDL.2017.7991582. https://ieeexplore.ieee.org/document/7991582.

Da, C., et al. Vision grid transformer for document layout analysis. arXiv: 2308.14978, http://arxiv.org/abs/2308.14978, August 2023, preprint: not peer reviewed.

Dell, M., et al. (2024) 'American Stories: A Large-Scale Structured Text Dataset of Historical U.S. Newspapers', in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pp. 80744–80772. Red Hook, NY: Curran Associates Inc.

Devlin, J., et al. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 [cs], http://arxiv.org/abs/1810.04805, May 2019, preprint: not peer reviewed.

Dosovitskiy, A., et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv : 2010.11929 [cs], http://arxiv.org/abs/2010.11929, June 2021, preprint: not peer reviewed.

Dubey, A., et al. The Llama 3 Herd of models. arXiv: 2407.21783 [cs], http://arxiv.org/abs/2407.21783, July 2024, preprint: not peer reviewed.

Efron, B., and Tibshirani, R. (1993) *An Introduction to the Bootstrap*. Philadelphia, PA: Chapman & Hall/CRC.

Faysse, M., et al. ColPali: efficient document retrieval with vision language models. arXiv: 2407.01449 [cs], http://arxiv.org/abs/2407.01449, October 2024, preprint: not peer reviewed.

Flesch, R. (1948) 'A New Readability Yardstick', *Journal of Applied Psychology*, **32**, 221–233. https://doi.org/10.1037/h0057532

Girdhar, N., Coustaty, M., and Doucet, A. (2024) 'Digitizing History: Transitioning Historical Paper Documents to Digital Content for Information Retrieval and Mining—A Comprehensive Survey', *IEEE Transactions on Computational Social Systems*, **11**, 6151–6180. https://doi.org/10.1109/TCSS.2024.3378419

Goldfrank, D. M. (2014) *The Origins of the Crimean War*. London: Routledge. https://doi.org/10.4324/9781315836911

Gunasekar, S., et al. Textbooks are all you need. arXiv: 2306.11644 [cs], http://arxiv.org/abs/2306.11644, October 2023, preprint: not peer reviewed.

Gupte, A., et al. Lights, Camera, Action! A framework to improve NLP Accuracy over OCR documents. arXiv: 2108.02899 [cs], http://arxiv.org/abs/2108.02899, August 2021, preprint: not peer reviewed.

Hampton, M. (2008) 'New Journalism, Nineteenth-Century', in *The International Encyclopedia of Communication*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781405186407.wbiecn012.

He, P.. et al. DeBERTa: decoding-enhanced BERT with disentangled attention. arXiv: 2006.03654 [cs], http://arxiv.org/abs/2006.03654, October 2021, preprint: not peer reviewed.

Hodel, T. et al. (2021) 'General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example', *Journal of Open Humanities Data*, **7**. doi: 10.5334/johd.46.

Hogan, A., et al. Large language models, knowledge graphs and search engines: a crossroads for answering users' questions. arXiv: 2501.06699 [cs], http://arxiv.org/abs/2501.06699, January 2025, preprint: not peer reviewed.

Huang, Y., et al. LayoutLMv3: pre-training for document AI with unified text and image masking. arXiv: 2204.08387, http://arxiv.org/abs/2204.08387, July 2022.

IAB (2025) InteractiveAdvertisingBureau/Taxonomies. https://github.com/InteractiveAdvertisingBureau/Taxonomies.

IPTC (2019) News Categories Taxonomy for the Media. https://iptc.org/standards/subject-codes/.

Jaume, G., Ekenel, H. K., and Thiran, J.-P. FUNSD: a dataset for form understanding in noisy scanned documents. In *ICADAR-OST 2019*. arXiv: 1905.13538, http://arxiv.org/abs/1905.13538, October 2019, preprint: not peer reviewed.

Johnson, N. (2013) The History of the Parliamentary Franchise. https://researchbriefings.files.parliament.uk/documents/RP13-14/RP13-14.pdf.

Kallini, J., et al. Mission: impossible language models. arXiv: 2401.06416 [cs], http://arxiv.org/abs/2401.06416, August 2024, preprint: not peer reviewed.

Kanerva, J., et al. OCR error post-correction with LLMs in historical documents: no free lunches. arXiv: 2502.01205 [cs], http://arxiv.org/abs/2502.01205, February 2025. preprint: not peer reviewed.

Labelbox (2025) Labelbox. https://labelbox.com.

Legge, G. E., and Bigelow, C. A. (2011) 'Does Print Size Matter for Reading? A Review of Findings from Vision Science and Typography', *Journal of Vision*, **11**, 8. https://doi.org/10.1167/11.5.8.

Lewis, P., *et al.* (2020) 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks', in *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Li, J., *et al.* DiT: self-supervised pre-training for document image transformer. arXiv: 2203.02378, http://arxiv.org/abs/2203.02378, July 2022, preprint: not peer reviewed.

Li, Y., *et al.* Textbooks are all you need II: phi-1.5 technical report. arXiv: 2309.05463 [cs], http://arxiv.org/abs/2309.05463, September 2023, preprint: not peer reviewed.

Liu, Y., *et al.* RoBERTa: a robustly optimized BERT pretraining approach. arXiv : 1907.11692 [cs], http://arxiv.org/abs/1907.11692, July 2019. preprint: not peer reviewed.

Luckhurst, K. W. (1951) 'The Great Exhibition of 1851', *Journal of the Royal Society of Arts*, **99**, 413–456.

Medhat, W., Hassan, A., and Korashy, H. (2014) 'Sentiment Analysis Algorithms and Applications: A Survey', *Ain Shams Engineering Journal*, **5**, 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011.

Müllner, D. Modern hierarchical, agglomerative clustering algorithms. arXiv: 1109.2378 [cs, stat] version: 1, http://arxiv.org/abs/1109.2378, September 2011, preprint: not peer reviewed.

Northcutt, A., Athalye, C. G. and Mueller, J. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. arXiv: 2103.14749 [cs, stat]. arXiv: 2103.14749, http://arxiv.org/abs/2103.14749, November 2021, preprint: not peer reviewed.

OpenAI. GPT-4 technical report. http://arxiv.org/abs/2303.08774, March 2024, preprint: not peer reviewed.

Park, S., *et al.* (2019) CORD: A Consolidated Receipt Dataset for Post-OCR Parsing. https://openreview.net/forum?id=SJl3z659UH.

Phillips, A. (2012) 'Saving Civilization from Empire: Belligerency, Pacifism and the Two Faces of Civilization during the Second Opium War', *European Journal of International Relations*, **18**, 5–27. doi: 10.1177/1354066111416020.

Phillips, J. A., and Wetherell, C. (1995) 'The Great Reform Act of 1832 and the Political Modernization of England', *The American Historical Review*, **100**, 411–436. https://doi.org/10.2307/2169005.

Pilehvar, M. T., and Camacho-Collados, J. (2020) *Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning*. Morgan & Claypool Publishers.

Poznanski, J., *et al.* olmOCR: unlocking trillions of tokens in PDFs with vision language models. arXiv: 2502.18443 [cs], http://arxiv.org/abs/2502.18443, July 2025. preprint: not peer reviewed.

Qwen, A., *et al.* Qwen2.5 technical report. arXiv : 2412.15115 [cs], http://arxiv.org/abs/2412.15115, January 2025. preprint: not peer reviewed.

Shehzadi, T. *et al.*, Bridging the Performance Gap between DETR and R-CNN for Graphical Object Detection in Document Images. arXiv: 2306.13526 [cs], http://arxiv.org/abs/2306.13526, June 2023, preprint: not peer reviewed.

Smith, D. A., and Cordell, R. (2018) *A Research Agenda for Historical and Multilingual Optical Character Recognition*. North Eastern University. http://hdl.handle.net/2047/D20297452.

Smith, R. (2007) 'An Overview of the Tesseract OCR Engine', in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, vol. 2, pp. 629–633. doi: 10.1109/ICDAR.2007.4376991.

Smits, T. *et al.* (2025) 'A Fully-Searchable Multimodal Dataset of the—Illustrated London News, 1842—1890', *Journal of Open Humanities Data*, **11**. https://openhumanitiesdata.metajnl.com/articles/10.5334/johd.284.

Soper, E., Fujimoto, S., and Yu, Y.-Y. (2021) 'BART for Post-Correction of OCR Newspaper Text', in W. Xu, A, Ritter, T. Baldwin, and A. Rahimi (eds), *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pp. 284–290. Association for Computational Linguistics. doi: 10.18653/v1/2021.wnut-1.31.

Stone, L. (1969) 'Literacy and Education in England 1640-1900', *Past and Present*, **42**, 69–139. https://www.jstor.org/stable/650183.

Su, J.. *et al.* RoFormer: enhanced transformer with rotary position embedding. arXiv: 2104.09864 [cs], http://arxiv.org/abs/2104.09864, November 2023. preprint: not peer reviewed.

Sven, N.-M., and Matteo, R. Page layout analysis of text-heavy historical documents: a comparison of textual and visual approaches. arXiv: 2212.13924, http://arxiv.org/abs/2212.13924, December 2022. preprint: not peer reviewed.

Tan, C., and Wang, J. 1.5-Pints technical report: pretraining in days, not months – your language model thrives on quality data. arXiv: 2408.03506 [cs], http://arxiv.org/abs/2408.03506, August 2024, preprint: not peer reviewed.

Tanimoto, T. T. An Elementary Mathematical Theory of Classification and Prediction. Technical Report PB167360, International Business Machines Corp., New York., 1958. URL https://ntrl.ntis.gov/NTRL/dashboard/searchResults/titleDetail/PB167360.xhtml. Num Pages: 12.

Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. http://arxiv.org/abs/2403.05530, December 2024, preprint: not peer reviewed.

Terras, M. M. (2011) 'The Rise of Digitization', in R. Rikowski (ed.) *Digitisation Perspectives*, pp. 3–20. Rotterdam: SensePublishers. https://doi.org/10.1007/978-94-6091-299-3_1

Thomas, A., Gaizauskas, R., and Lu, H. (2024) 'Leveraging LLMs for Post-OCR Correction of Historical Newspapers', in *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pp. 116–121, Torino, Italia: ELRA and ICCL.

Turner, M. (2024a) ncse: The Monthly Repository, 1806-1837. URL https://kcl.figshare.com/articles/dataset/ncse_The_Monthly_Repository_1806-1837/25108519/1.

Turner, M. (2024b) ncse: The English Woman's Journal, 1858-1864. https://kcl.figshare.com/articles/dataset/ncse_The_English_Woman_s_Journal_1858-1864/25109275/1.

Turner, M. (2024c) ncse: The Northern Star, 1838-1852. https://kcl.figshare.com/articles/dataset/ncse_The_Northern_Star_1838-1852/25106895/1.

Turner, M. (2024d) ncse: The Tomahawk, 1867-1870. https://kcl.figshare.com/articles/dataset/ncse_The_Tomahawk_1867-1870/25109047/1.

Turner, M. (2024e) ncse: The Leader, 1850-1860. https://kcl.figshare.com/articles/dataset/ncse_The_Leader_1850-1860/25103786/1.

Turner, M. (2024f) ncse: The Publishers' Circular, 1880-1890. https://kcl.figshare.com/articles/dataset/ncse_The_Publishers_Circular_1880-1890/25101957/1.

Turner, M. (2024g) Nineteenth-Century Serials Edition: Investigating an Archival Future. https://kclpure.kcl.ac.uk/portal/en/projects/nineteenth-century-serials-edition-investigating-an-archival-futu/datasets/.

Twycross, A. (2024) 'The Rise of the New Journalism', in A. Twycross (ed.) *British Newspaper Strips: A Contextual History*, pp. 11–25. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-61213-8_2.

UK-Government. Slavery Abolition Act 1833 (repealed 19.11.1998) (1833) Statute Law Database. https://www.legislation.gov.uk/ukpga/Will4/3-4/73/section/12/1991-02-01.

UK-Government. Married Women's Property Act 1882 (1882) King's Printer of Acts of Parliament. https://www.legislation.gov.uk/ukpga/Vict/45-46/75/enacted.

Vaswani, A., *et al.* Attention is all you need. arXiv: 1706.03762 [cs], arXiv: 1706.03762, http://arxiv.org/abs/1706.03762, December 2017, preprint: not peer reviewed.

Wankhade, M., Rao, A. C. S., and Kulkarni, C. (2022) 'A Survey on Sentiment Analysis Methods, Applications, and Challenges', *Artificial Intelligence Review*, **55**, 5731–5780. https://doi.org/10.1007/s10462-022-10144-1.

Warner, B., *et al.* Smarter, better, faster, longer: a modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. arXiv: 2412.13663 [cs], http://arxiv.org/abs/2412.13663, December 2024, preprint: not peer reviewed.

Wei, H., *et al.* General OCR theory: towards OCR-2.0 via a unified end-to-end model. arXiv: 2409.01704, http://arxiv.org/abs/2409.01704, September 2024, preprint: not peer reviewed.

Wilde, O. (1891) 'The Critic as Artist', in *Intentions*, 3rd edn. Methuen and Co. https://www.gutenberg.org/cache/epub/887/pg887-images.html.

Zhang, P., *et al.* VSR: A unified framework for document layout analysis combining vision, semantics and relations. arXiv: 2105.06220, http://arxiv.org/abs/2105.06220, May 2021. preprint: not peer reviewed.

Zhao, Z., *et al.* DocLayout-YOLO: enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception. arXiv: 2410.12628, http://arxiv.org/abs/2410.12628, October 2024. preprint: not peer reviewed.

Zhong, X., Tang, J., and Yepes, A. J. PubLayNet: largest dataset ever for document layout analysis. arXiv: 1908.07836, http://arxiv.org/abs/1908.07836, August 2019. preprint: not peer reviewed.