

5-2019

Application of the Image Analysis for Archival Discovery Team's First- Generation Methods and Software to the Burney Collection of British Newspapers

Elizabeth Lorang

University of Nebraska-Lincoln

Leen-Kiat Soh

University of Nebraska-Lincoln

Chulwoo Pack

University of Nebraska-Lincoln

Yi Liu

University of Nebraska-Lincoln

Delaram Rahimighazikalayeh

University of Nebraska-Lincoln

Follow this and additional works at: <https://digitalcommons.unl.edu/cdrhgrants>

Part of the [Digital Humanities Commons](#), [Library and Information Science Commons](#), and the [Literature in English, British Isles Commons](#)

Lorang, Elizabeth; Soh, Leen-Kiat; Pack, Chulwoo; Liu, Yi; Rahimighazikalayeh, Delaram; and O'Brien, John, "Application of the Image Analysis for Archival Discovery Team's First- Generation Methods and Software to the Burney Collection of British Newspapers" (2019). *CDRH Grant Reports*. 7.
<https://digitalcommons.unl.edu/cdrhgrants/7>

This Article is brought to you for free and open access by the Center for Digital Research in the Humanities at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in CDRH Grant Reports by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

Authors

Elizabeth Lorang, Leen-Kiat Soh, Chulwoo Pack, Yi Liu, Delaram Rahimighazikalayeh, and John O'Brien

APPLICATION OF THE IMAGE ANALYSIS FOR ARCHIVAL DISCOVERY TEAM'S FIRST-GENERATION METHODS AND SOFTWARE TO THE BURNEY COLLECTION OF BRITISH NEWSPAPERS¹

ELIZABETH LORANG, YI LIU, CHULWOO PACK, LEEN-KIAT SOH, DELARAM
RAHIMIGHAZIKALAYEH, AND JOHN O'BRIEN

MAY 2019

1. BACKGROUND

The purpose of the study presented and analyzed here is to explore the generalizability of the Image Analysis for Archival Discovery (Aida) team's approaches across newspaper corpora. Up until this study, we have focused our training and testing data on U.S. newspapers of the nineteenth century. The current study, "Application of the Image Analysis for Archival Discovery Team's First-Generation Methods and Software to the Burney Collection of British Newspapers," is the first test of our approaches—methods and software—to a different newspaper corpus, specifically the *17th and 18 Century Burney Newspapers Collection*. This study stands as the first complete attempt at applying Aida's software and methods to non-*Chronicling America* newspapers, as a step toward understanding the potential of our approaches across digitized historic newspapers. In taking this step, our goals were (1) to test how well the software and a classifier model developed on *Chronicling America* newspapers performed on newspapers from a different corpus, a corpus that represents both a different geographical region and time period as well as newspapers digitized at an early stage in newspaper digitization history; (2) to explore whether classification results would be improved by training a new classifier model on Burney Collection images. Overall, we sought to explore how robust and extensible the first-generation Aida approach is and to better understand which parts of our methods might be brought over to new corpora "as is," and which may need to be calibrated for specific contexts.

We used 184 full-page images of historic newspapers from the Burney Collection for this study. The complete Burney collection features nearly one million pages. We began with this small number of pages because doing so allows us to investigate the performance of the classifier on the collection in greater depth and also affords us flexibility to iterate our testing more frequently. The 184 newspaper pages, all of which feature poetry or poetic content, were selected by team member and subject expert John O'Brien. They represent the many types of newspaper layouts present in the Burney Collection, as well as the time period covered in the collection. In addition, the page images

¹ This project was made possible in part by the Institute of Museum and Library Services (grant award LG-71-16-0152-16).

exemplify some “best case” and “worst case” scenarios for quality of the images. A full list of newspaper pages is presented in Appendix 1: Pages Processed and Analyzed from the Burney Collection.

We pursued 4 scenarios for this study, which allowed us to explore the efficacy of using a classifier trained on *Chronicling America* newspapers to identify content in the Burney Collection and to assess the effectiveness of improvements we made to our software across two versions (version 0.2.0 and version 0.3.0). The four scenarios are:

1. The Aida team’s initial first-generation approach (software version 0.2.0) deployed on the 184 Burney Collection pages, with a classifier model trained on *Chronicling America* newspapers.
2. The Aida team’s initial first-generation approach (software version 0.2.0) deployed on the 184 Burney Collection pages, with a classifier model trained on Burney Collection newspapers.
3. The Aida team’s improved first-generation approach (software version 0.3.0) deployed on the 184 Burney Collection pages, with a classifier model trained on *Chronicling America* newspapers.
4. The Aida team’s improved first-generation approach (software version 0.3.0) deployed on the 184 Burney Collection pages, with a classifier model trained on Burney Collection newspapers.

See projectaida.org for a range of presentations and publications that provide full detail on our first-generation methods, including page segmentation to derive the image snippets, as well as feature extraction and classification. Both versions (0.2.0 and 0.3.0) of the first-generation software use an artificial neural network (ANN) to classify image snippets from digitized newspaper pages as containing or not containing poetic content. The major changes from version 0.2.0 to 0.3.0 were improved binarization of the images, which led to better segmentation of page images (as demonstrated in section 3) and to more effective feature extraction, which improved classification.

The remainder of this report first presents our key findings (section 2), presents a detailed summary and analysis of our approach and results (section 3), and concludes with a discussion of the current challenges and next steps (section 4). Sections 2 and 4 will have most relevance for readers primarily interested in major take-aways. Those wishing to drill down into our methods and analyses should see section 3, as well as the appendices to this report, code for software versions 0.2.0 and 0.3.0, and corresponding output data.²

² Appendices are packaged with this report as a separate file. Code is available via GitHub at <https://github.com/ProjectAida/aida/releases/tag/v0.2.0> and <https://github.com/ProjectAida/aida/releases/tag/v0.3.0>. Data are available at <https://osf.io/xn7tv/> (see the dataset, “Data for ‘Application of the Image Analysis for Archival Discovery Team’s First-Generation

2. KEY FINDINGS

As currently conceived and implemented, the Aida team’s first-generation approach is not successful when applied to the Burney Collection of newspapers. Neither software version 0.2.0 nor version 0.3.0 retrieve poetic content at a rate sufficient for broader deployment at this time.

A baseline for comparison is a recent experiment we did to assess the effectiveness of improvements in the software from version 0.2.0 to version 0.3.0. In that experiment, we analyzed both training and testing results of each software version on 215 image snippets from *Chronicling America*, where we also had expert-determined ground truth for each image snippet. The baseline data were obtained by 10-fold cross validation, a common way to evaluate the strength of a classifier when training and testing data are from the same dataset. In this test on *Chronicling America* data, version 0.2.0 achieved an overall accuracy of 65.29% in testing. Version 0.3.0 achieved an overall accuracy of 67.47% in testing.

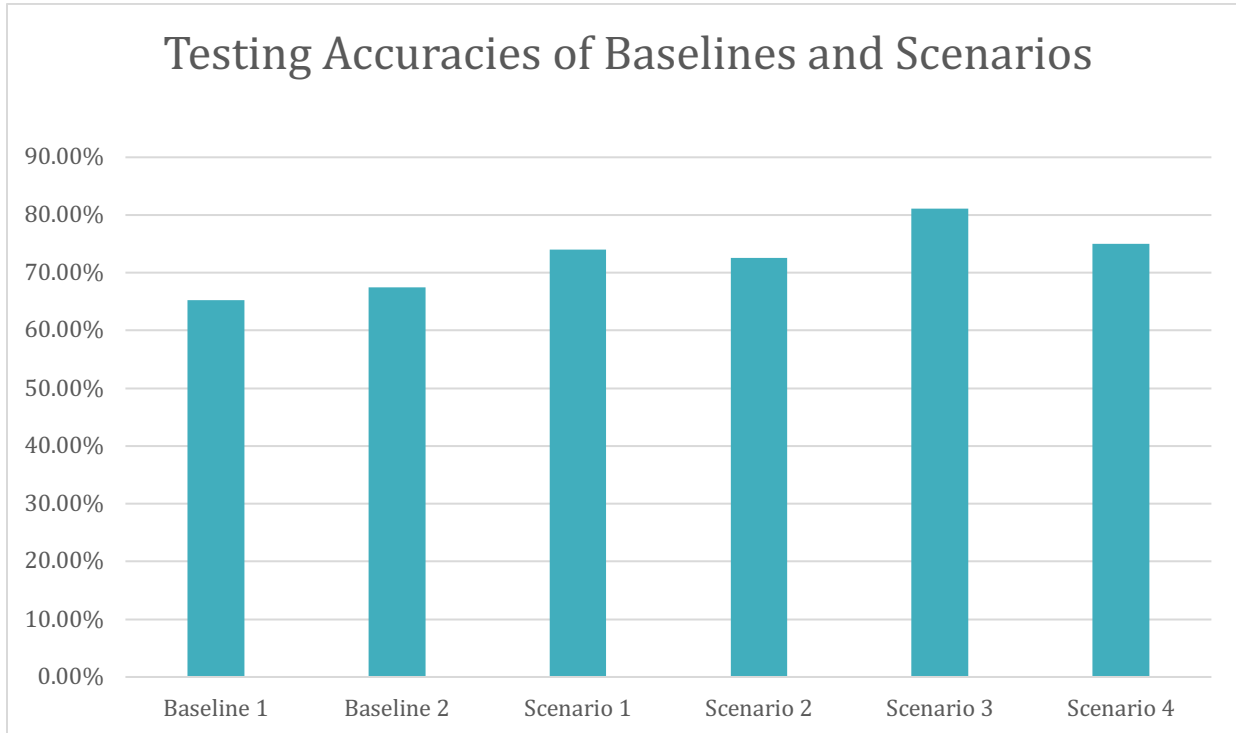
The testing accuracy in each of the 4 scenarios examined in this study initially appears to be a noticeable improvement, reaching accuracies of 74.05% (v.0.2.0, Chronicling America-based classifier); 81.14% (v.0.3.0, Chronicling America-based classifier); 72.6% (v.0.2.0, Burney-Based classifier); and 75.01% (v.0.3.0, Burney-based classifier). See Table 1 and Graph 1.

	Software Version	Classifier Model	Test Set	Test Accuracy
Baseline 1	0.2.0	Chron-Am	Chron-Am image snippets	65.29%
Baseline 2	0.3.0	Chron-Am	Chron-Am image snippets	67.47%
Scenario 1	0.2.0	Chron-Am	Burney image snippets	74.05%
Scenario 2	0.2.0	Burney	Burney image snippets	72.6%
Scenario 3	0.3.0	Chron-Am	Burney Collection image snippets	81.14%
Scenario 4	0.3.0	Burney	Burney Collection image snippets	75.01%

Table 1. Test accuracies of four scenarios from this study, compared with recent baseline (10-fold tests).

Methods and Software to the Burney Collection of British Newspapers”). Readers may also wish to see <https://projectaida.org>.

In every scenario, testing accuracy in the four Burney Collection scenarios outperformed the testing accuracy when we used software versions 0.2.0 and 0.3.0 trained and tested on Chronicling America image snippets (Baseline 1 and Baseline 2). In fact, the best testing accuracy resulted from using software version 0.3.0 with a classifier model trained on Chronicling America image snippets and then tested on Burney Collection image snippets—surpassing even the accuracy of the version 0.3.0 baseline (trained and tested on Chronicling America data). Overall accuracy, however, tells only part of the story.



Graph 1. Visualization of testing accuracies for both baselines and scenarios.

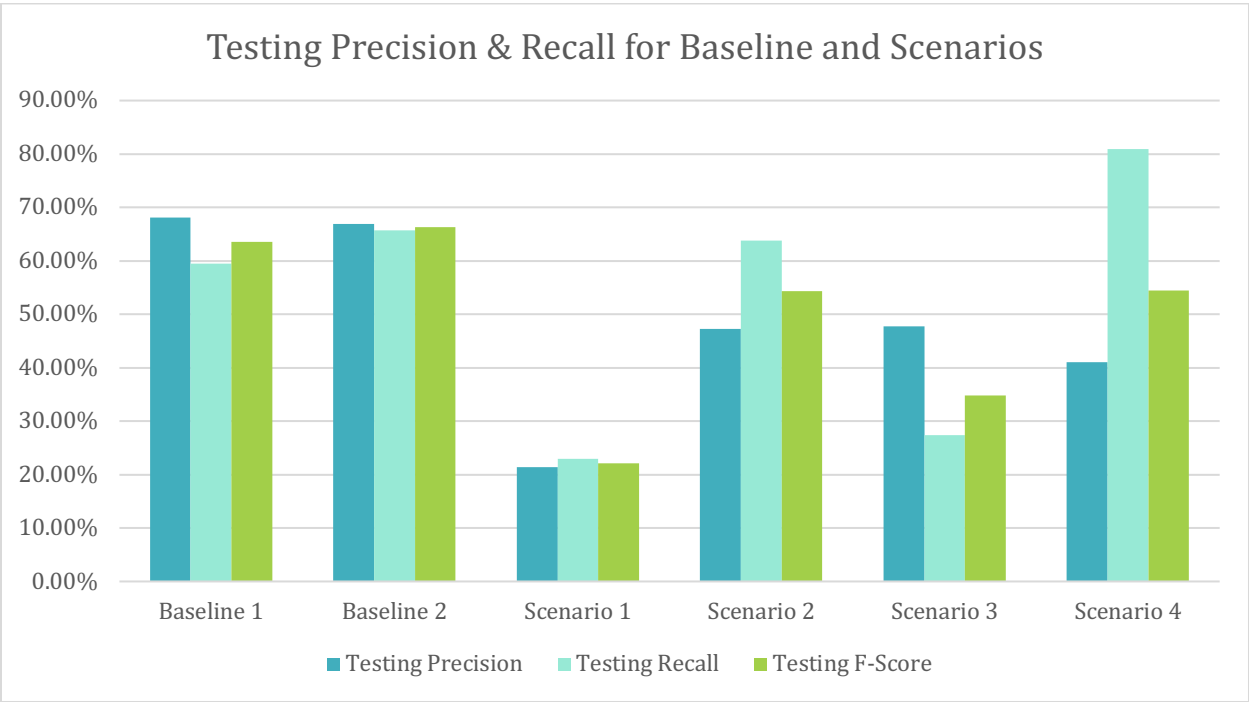
The generally high accuracy rates are legitimate, but they are also skewed. They are skewed because all of the scenarios did relatively well in accurately classifying page image snippets that did *not* have poetic content (true negatives). The overall accuracy measure includes true negatives as part of its calculation. But a more complete picture emerges when we look at precision and recall values, which assess whether we correctly identified the content we are interested in—those image snippets containing poetic content.

While we achieved high values for both precision and recall for our *training* data across all scenarios in this study, when we tested each scenario on new image snippets, all scenarios failed to identify poetic content at an acceptable rate (see Table 2 and Graph 2). Scenario 3, which looked most promising based on overall accuracy (over 81%), recalled less than 28% of the total snippets with poetic content—the very snippets we are most interested in.

	Software Version	Classifier Model	Training Precision	Training Recall	Training F-score	Testing Precision	Testing Recall	Testing F-score
Baseline 1	0.2.0	Chron-Am	85.21%	80.37%	82.72%	68.09%	59.49%	63.50%
Baseline 2	0.3.0	Chron-Am	91.27%	91.74%	91.50%	66.90%	65.69%	66.29%
Scenario 1	0.2.0	Chron-Am	90.00%	82.65%	86.17%	21.36%	22.92%	22.11%
Scenario 2	0.2.0	Burney	92.68%	77.55%	84.44%	47.29%	63.79%	54.31%
Scenario 3	0.3.0	Chron-Am	93.81%	92.86%	93.33%	47.75%	27.44%	34.85%
Scenario 4	0.3.0	Burney	89.32%	93.88%	91.54%	41.03%	80.91%	54.45%

Table 2. Precision and recall values for each scenario, at both training testing stages, compared with recent baseline (10-fold tests).

In our recent baseline tests, our improved first-generation approach (software version 0.3.0) achieved precision and recall values that approached 67% and 66%, respectively. These values were encouraging, especially since 10-fold validation can lead to lower accuracy values than if deployed on the entire dataset (that is, the values reported here are likely to be lower than if we tested them much more broadly). Nonetheless, they still leave ample room for improvement. Overall, precision and recall values for the four scenarios in this study are both more highly varied and significantly lower than this baseline, with some exceptions.



Graph 2. Visualization of testing precision, recall, and f-scores for both baselines and scenarios.

In the four scenarios, classifiers trained on *Chronicling America* data were least successful in retrieving poetic content from the Burney Collection test set. The best case, as far as retrieving the most poetic content, was Scenario 4 (software version 0.3.0, Burney Collection classifier deployed on Burney Collection image snippets), which successfully recalled nearly 81% of the snippets with poetic content; however, it did so at the expense of flooding the results with false positives, having an overall precision rate of just over 41%. Overall precision rates were low across all four scenarios, because of the high numbers of false positives in each scenario. In all but one scenario, the number of false positives exceeded the number of true positives. For example, in Scenario 4, we identified 407 true positives and 585 false positives. Section 3 details the results for each scenario.

Furthermore, our first-generation approach to newspaper page segmentation sees us discard anywhere from 25% to 50% of newspaper pages from the outset, because they are not suitable for our page segmentation methods. Between the high number of pages we must currently discard and the low or imbalanced precision and recall values, we do not recommend further implementation of our first-generation approach, as implemented here, to the Burney Collection. See Section 4 for further exploration of the challenges and next steps.

3. METHODS & DETAILED RESULTS

This section presents further detail on the methods used and results stemming from each of the four scenarios tested for this study. Here, we detail the influence of page segmentation, explain our strategy for determining ground truth and establish ground truth for the scenarios, and provide further detail and statistics for each scenario.

3.1 PAGE SEGMENTATION

The first step in preparing images for classification is to cut full page images into overlapping image snippets. As part of this process, we computationally evaluate full-page images according to several criteria, including whether we can find two or more columns in the page image, whether more than half of the page cannot be determined to have columns, and whether the standard deviation of the distance between columns is above 150. Full page images that pass these pre-tests go on to the segmentation process. Those that do not pass these segmentation pre-tests are discarded from further processing and are logged in an output text file for further human analysis. The purpose of these pre-tests is to minimize the number of bad output snippets for feature extraction and classification; bad output snippets lead to significantly skewed classification results.

Improvements from software version 0.2.0 to 0.3.0 increased our effectiveness in handling two-column newspapers. In addition, the enhanced approach to binarization enhanced features of the newspaper page images that we rely on for segmentation, and therefore more images passed these segmentation pre-tests. Overall, we saw an increase of 19% from version 0.2.0 to 0.3.0 in pages that passed segmentation and were subsequently processed into image snippets for feature extraction and classification.

Software Version 0.2.0

When we processed the 184 Burney Collection pages with software version 0.2.0, roughly 56.5% of the page images passed all segmentation pre-tests and proceeded to segmentation. The remaining page images failed a specific test or caused our software to throw an exception because some parameter of each image was outside the bounds configured in our software.

Category	# of Pages
Pages that passed segmentation pre-tests	104
Pages that failed segmentation pre-tests	31
Pages that caused software to throw an exception	49

Among the page images that failed segmentation pre-tests, 26 failed because their column widths were non-standard, above our allowed threshold (standard deviation was above 150); 5 images failed segmentation pre-tests because our software could find columns on \leq half of the newspaper page. Below are sample images that failed according to each of these criteria. See Appendix 2: Page Image Segmentation Results, Initial Approach for a complete list of page images, their segmentation result, and criteria for failing segmentation (when relevant).

The 104 pages that passed the segmentation pre-tests were processed into a total of 1,179 image snippets. These snippets were the basis for the feature extraction and classification processes.

Software Version 0.3.0

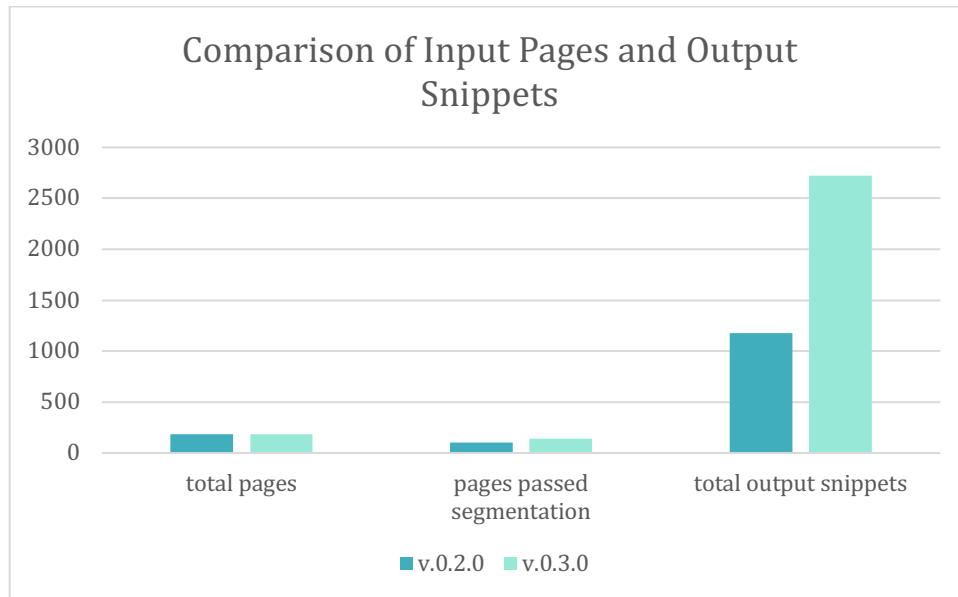
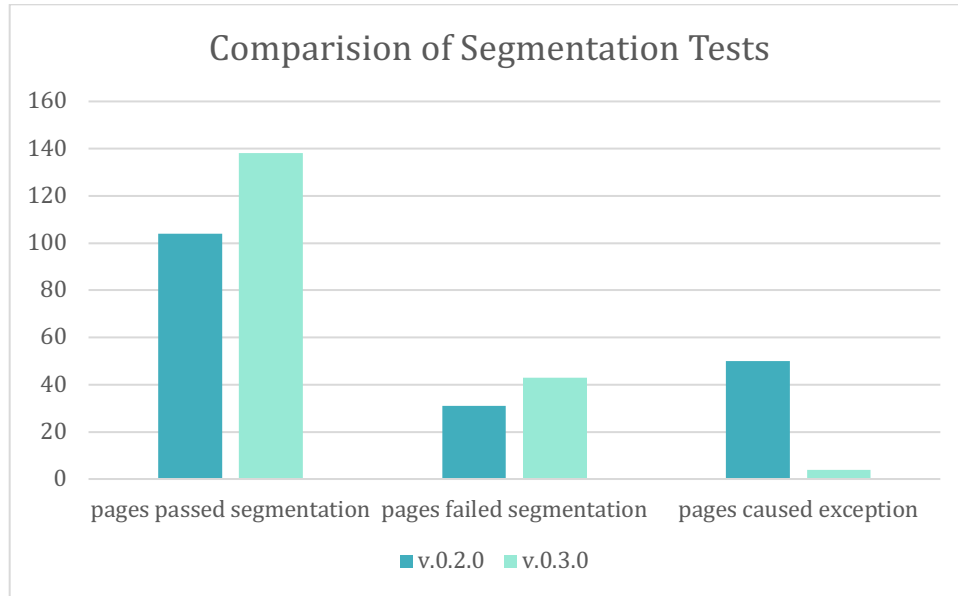
When we processed the 184 Burney Collection pages with software version 0.3.0, 75% of the page images passed all segmentation pre-tests and proceeded to segmentation. The remaining pages failed a specific test or caused our software to throw an exception because some parameter of each image was outside the bounds configured in our software.

Category	# of Pages
Pages that passed segmentation pre-tests	138
Pages that failed segmentation pre-tests	43
Pages that caused software to throw an exception	3

Among the page images that failed segmentation pre-tests, 39 failed because their column widths were non-standard, above our allowed threshold (standard deviation was above 150); 4 images failed segmentation pre-tests because our software could find columns on \leq half of the newspaper page. Below are sample images that failed according to each of these criteria. See Appendix 3: Page Image Segmentation Results, Improved Approach for a complete list of page images, their segmentation result, and criteria for failing segmentation (when relevant).

The 138 pages that passed the segmentation pre-tests were processed into a total of 2,725 image snippets. These snippets were the basis for the feature extraction and classification processes.

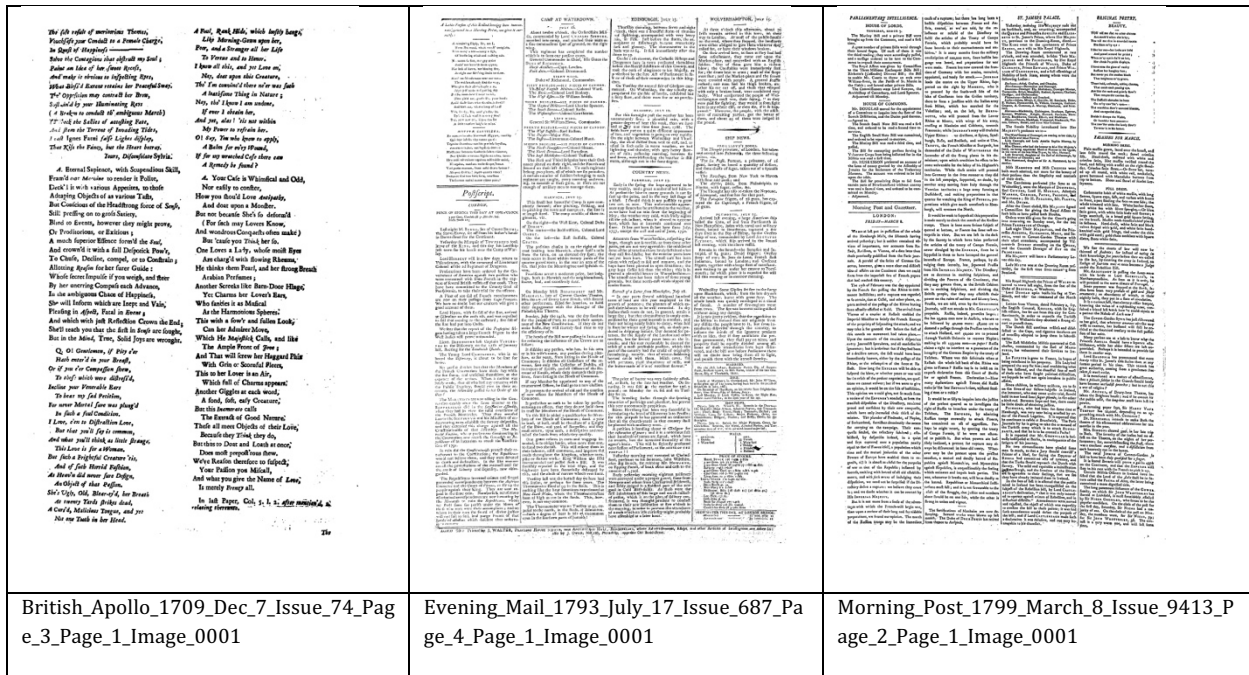
Comparison



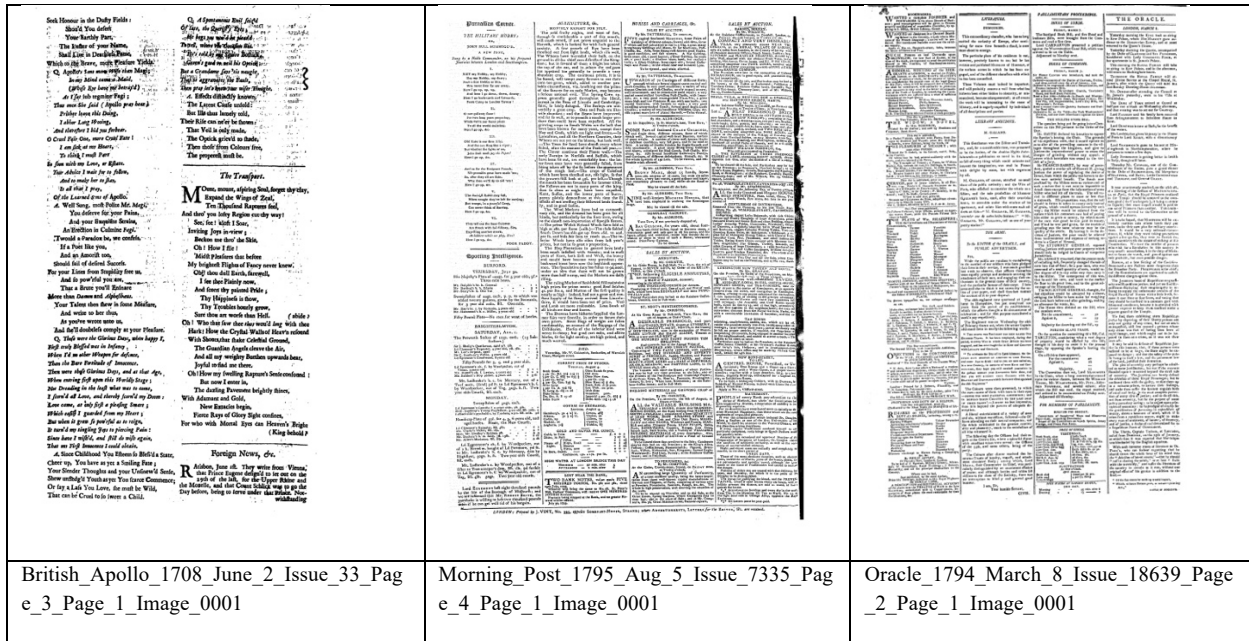
Problematic Page Images

Standard deviation above 150

In order to prevent generating bad snippets based on incorrectly determined column breaks, the algorithm examines whether the standard deviation of the distance between columns is above 150. Standard deviations above 150 are caused by features such as skew with a dense amount of characters or significant horizontal whitespace, and these images do not proceed to segmentation.



Software version 0.2.0, representative page images where the standard deviation of distance between columns is above 150.



Software version 0.3.0, representative page images where the standard deviation of distance between columns is above 150.

Columns detected on half of the page only

Observationally, this type of error usually occurs in cases where the document image has significant character density or skew. The following images are representative of those where our software

either located the first column more than halfway across the page or the final column less than halfway across the page.

<p>Evening Mail 1789 Sep 28-30 Issue 92 Page 2 Page 1 Image 0001</p>	<p>Oracle 1794 Dec 1 Issue 18865 Page 3 Page 1 Image 0001</p>	<p>WO2_B0210EVENMAIL_1796_01_08-0003 Page 1 Image 0001</p>

Software version 0.2.0, representative images where our software locates the first column more than halfway across the page or the final column less than halfway across the page.

<p>Observator 1704 June 14 Issue 25 Page 2 Page 1 Image 0001</p>	<p>Whitehal Evening Post 1758 21 Jan Issue 1850 Page 1 Page 1 Image 0001</p>	<p>WO2_B0143CALCCHRO_1788_02_21-0005 Page 1 Image 0001</p>




Software version 0.3.0, representative images where our software locates the first column more than halfway across the page or the final column less than halfway across the page.

Exceptions

When two column breaks—the leftmost and rightmost side of the page image—are obtained, then the whole page is passed to the snippet generation module. However, in some cases, the ratio of width to height of segmented column—the whole image in this scenario—is smaller than expected, and such images generate an out-of-bounds exception. They are not segmented into image snippets.

		
<p>Whitehall_Evening_Post_1759_Sep_8_Issue_2103_Page_4_Page_1_Image_0001</p>	<p>Weekly_Journal_or_British_Gazetteer_1727_Feb_4_Issue_91_Page_1_Page_1_Image_0001</p>	<p>WO2_B0143CALCCHRO_1788_03_20-0004_Page_1_Image_0001</p>

Software version 0.2.0, representative images where our software failed to fully evaluate a page, because a feature measurement fell outside of allowed parameters.

		
<p>Whitehall_Evening_Post_1800_Jan_4_Issue_8183_Page_1</p>	<p>Morning_Post_1778_April_11_Issue_1710_Page_4_Page_1</p>	<p>World_1791_Dec_5_Issue_1538_Page_3_Page_1_Image_0</p>

Software version 0.3.0, representative images where our software failed to fully evaluate a page, because a feature measurement fell outside of allowed parameters.

3.2 ESTABLISHING GROUND TRUTH

In order to fully assess the classification accuracy of our system, we manually classified all output snippets generated with both software versions. We classified each snippet as “true” or “false,” where “true” snippets contain poetic content (either complete poems or multiple lines of poems), and “false” snippets contain no poetic content. To achieve ground truth for each snippet, two members of the project team separately classified each snippet as “true” or “false.” We then diffed their classification lists and generated a list of snippets where they disagreed on their classification. The same team members then separately evaluated each disagreement snippet again, without consulting their earlier classification value, and recorded their new classification. We then diffed these disagreement classification files, and in all cases where there was still disagreement about the classification, team members discussed ground truth classification in person, to reach agreement on how a snippet should be classified.

Software Version 0.2.0

Following this process, we determined that of the 1,179 output snippets generated with software version 0.2.0, 301 contained a poem or poetic content (true snippets), and 878 snippets were absent of poetic content (false snippets).

True Snippets (Contain Poem)	301
False Snippets (Do Contain Poem)	878
Total Snippets	1,179

See Appendix 4: Ground Truth Determinations for All Snippets, Initial Approach for a complete list of snippets and their ground truth value.

Software Version 0.3.0

Following this process, we assessed that of the 2,725 output snippets generated with software version 0.3.0, 503 contained a poem or poetic content (true snippets), and 2,222 snippets were absent of poetic content (false snippets).

True Snippets (Contain Poem)	503
False Snippets (Do Contain Poem)	2,222
Total Snippets	2,725

See Appendix 5: Ground Truth Determinations for All Snippets, Improved Approach for a complete list of snippets and their ground truth value.

With this comprehensive ground-truth evaluation, we were then able to establish the accuracies of the classifiers trained on Chronicling America image snippets and the accuracies of the classifiers based on Burney Collection image snippets.

3.3 CLASSIFICATION

For both software versions 0.2.0 and 0.3.0, we tested two approaches to classification: 1) classifying Burney Collection image snippets based on a classifier trained on Chronicling America image snippets (ChronAm model); and 2) classifying Burney Collection image snippets based on a classifier trained on Burney Collection image snippets (Burney model). We tested both classifiers to explore how generalizable extracted feature measurements from one corpus may be to another, and to see if using a classifier trained on the same corpus to be analyzed would significantly increase classification accuracy.

3.3.1 ChronAm Classifier Model

We first used a classifier model trained on image snippets from Chronicling America. This classifier model was trained on 215 image snippets, which were manually created by members of the project team from full page images downloaded from Chronicling America. 98 image snippets contained poetic content, and 117 image snippets did not contain poetic content.

In this scenario, the code compared features of new Burney Collection image snippets against features of Chronicling America image snippets known to be true or false, in order to determine whether a new Burney Collection image snippet contains poetic content (true) or not (false).

Software Version 0.2.0

The accuracy of the classifier model itself *when tested and trained on the same data* was 90.00% (precision) and 82.65% (recall). When deployed, the overall precision and recall using the ChronAm classifier model on Burney Collection image snippets were 21.36% and 22.92%, respectively.

		Actual	
		Contains Poem	Doesn't Contain Poem
Predicted	Contains Poem	69 (True Positive)	254 (False Positive)
	Doesn't Contain Poem	232 (False Negative)	624 (True Negative)

Snippet-level classification of ChronAm classifier model tested on 1,179 Burney Collection image snippets.

See Appendix 6: Classifier Model Classification of Snippets, Initial Approach with Chronicling America Model for the classification value of each snippet processed.

We used the snippet-level classification of true positive snippets to determine how many pages we might return as containing poetic content at the page level. Based on the snippets correctly classified as true, we would successfully identify 50 out of 184 pages as containing poetic content (in reality, all 184 pages contain poetic content). This application of a ChronAm classifier model

with our initial software on Burney Collection images allows us to retrieve 27.17% of pages that contain poetic content, based on accurate snippet-level classification.

		Actual
		Page Contains Poem
Predicted	Page Contains Poem	50 (True Positive)
	Page Doesn't Contain Poem	134 (False Negative)

Snippet-level classification abstracted to page-level classification (does a given page contain poetic content?). ChronAm classifier model deployed on Burney Collection image snippets.

Software Version 0.3.0

The accuracy of the classifier model itself (when tested and trained on the same data) was 93.81% (precision) and 92.86% (recall). Overall precision and recall using the Chronicling America-based classifier on Burney Collection image snippets were 47.75% and 27.44%, respectively.

		Actual	
		Contains Poem	Doesn't Contain Poem
Predicted	Contains Poem	138 (True Positive)	151 (False Positive)
	Doesn't Contain Poem	365 (False Negative)	2071 (True Negative)

Snippet-level classification

See Appendix 7: Classifier Model Classification of Snippets, Improved Approach with Chronicling America Model for the classification value of each snippet processed.

We again used the snippet-level classification to determine how many pages we might return as containing poetic content at the page level.

		Actual
		Page Contains Poem
Predicted	Page Contains Poem	6 (True Positive)
	Page Doesn't Contain Poem	178 (False Negative)

Snippet-level classification abstracted to page-level classification (does a given page contain poetic content?)

Based on the snippets classified as true, we would successfully identify 6 out of 184 pages as containing poetic content (in reality, all 184 pages contain poetic content). This application of a Chronicling America-based classifier with our improved software on Burney Collection images allows us to retrieve 3.26% of pages that contain poetic content, based on accurate snippet-level classification.

3.3.2 Burney Classifier Model

We also used a classifier model trained on image snippets from the Burney Collection. We again used 215 image snippets for training, with the same ratio of true and false snippets as in the ChronAm classifier model (98 image snippets contained poetic content; 117 image snippets did not contain poetic content). In this case, however, the image snippets were automatically generated by our page segmentation software. There is some variation in the overall quality of the automatically generated snippets.

In this scenario, the code compared features of new Burney Collection image snippets against features of existing Burney Collection image snippets known to be true or false, in order to determine whether a new Burney Collection image snippet contains poetic content (true) or not (false).

Software Version 0.2.0

The accuracy of the classifier model itself when *tested and trained on the same data* was 92.68% (precision) and 77.55% (recall). When deployed, the overall precision and recall using the Burney classifier model on new Burney Collection image snippets were 47.29% and 63.79%, respectively.

		Actual	
		Contains Poem	Doesn't Contain Poem
Predicted	Contains Poem	192 (True Positive)	214 (False Positive)
	Doesn't Contain Poem	109 (False Negative)	664 (True Negative)

Snippet-level classification all snippets

See Appendix 8: Classifier Model Classification of Snippets, Initial Approach with Burney Collection Model

We used the snippet-level classification of true positive snippets to determine how many pages we might return as containing poetic content at the page level. Based on the snippets correctly classified as true, we would successfully identify 74 out of 184 pages as containing poetic content (in reality, all 184 pages contain poetic content). This application of a Chronicling America-based classifier with our initial software on Burney Collection images allows us to retrieve 40.22% of pages that contain poetic content, based on accurate snippet-level classification.

		Actual
		Page Contains Poem
Predicted	Page Contains Poem	74 (True Positive)
	Page Doesn't Contain Poem	110 (False Negative)

Snippet-level classification abstracted to page-level classification (does a given page contain poetic content?). Burney classifier model deployed on Burney Collection image snippets.

Software Version 0.3.0

The accuracy of the classifier model itself (when tested and trained on the same data) was 89.32% (precision) and 93.88% (recall). When tested on new image snippets, overall precision and recall using the Burney Collection-based classifier were 41.03% and 80.91%, respectively, with the improved first-generation approach.

		Actual	
		Contains Poem	Doesn't Contain Poem
Predicted	Contains Poem	407 (True Positive)	585 (False Positive)
	Doesn't Contain Poem	96 (False Negative)	1637 (True Negative)

Snippet-level classification, all snippets

See Appendix 9: Classifier Model Classification of Snippets, Improved Approach with Burney Collection Model for the classification value of each snippet processed.

Again, we used the "true positive" snippet-level classification to determine how many pages we might return as containing poetic content at the page level. Based on the snippets classified as true, we would successfully identify 114 out of 184 pages as containing poetic content. This application of a Burney Collection-based classifier with our improved software on Burney Collection images allows us to recall 61.96% of pages that contain poetic content, based on accurate snippet-level classification.

		Actual
		Page Contains Poem
Predicted	Page Contains Poem	114 (True Positive)
	Page Doesn't Contain Poem	70 (False Negative)

Snippet-level classification abstracted to page-level classification (does a given page contain poetic content?)

4. DISCUSSION

How do we begin to understand or make sense of these results? Do aspects of our methods and approach continue to hold promise, even with the statistics reported here? What might it take to develop an approach that is more successful for the purposes of classifying textual content based on visual features? This section addresses features of the corpus as well as assesses potential problem-areas in the current approach and methods. By way of conclusion, we address next steps that respond to the challenges uncovered in this analysis, as we seek to develop an approach that is adaptable and extensible.

The *Burney Collection of Seventeenth and Eighteenth-Century Newspapers* is in many ways a difficult corpus to work with. These difficulties, some of which are unique to the collection and others that are typical of similar newspaper digitization projects, make the Burney Collection a useful one for testing our methods and software. The challenges we have encountered derive from issues native to the original print collection itself, from the ways in which it was digitized, and from assumptions encoded into our methods and software, among other challenges.

The digitized Burney Collection was produced in 1992–96 by using microfilm copies created beginning in the 1960s from the original pages, predating standards for microfilming, which emerged first in the late 1970s. The Burney Collection was the first newspaper digitization project carried out by the British Library and was done as a test case to discover problems and create a workflow that could be extended to the Library’s other newspaper collections.

In terms of the represented materials themselves, the collection is not really composed of only “newspapers” in a modern sense of the term. Particularly in the seventeenth-century part of the collection, much of what is included are pamphlets, news sheets, and political ephemera. These texts make sense as part of a history of the emergence of the newspaper in its modern form at the start of the eighteenth century (the first daily newspaper, the *Daily Courant*, was published in London in 1702), but there are many other ways in which they do not fit modern conceptions of what a “newspaper” was like and what it consists of. Thus, even though the British Library and Gale/Cengage advertise that the Burney Collection has more than a million “newspaper” pages, there’s a percentage of that—perhaps 15% or 20%—that does not count for our purposes; the other pages are primordial ancestors who go further back on the newspaper family tree.

In addition, the Burney Collection presents challenges for our approach that emerge both in the original documents and from the digital versions (of microfilm reproductions). Eighteenth-century newspapers are far less *regular* than nineteenth century newspapers. In the Burney Collection, newspapers come in one-, two-, three-, and four-column formats, with some going as high as six columns. Column widths are not standard and are not completely uniform in most of the newspapers. In addition, like all texts from the letter-press era, text lines are frequently less even than those printed on the mechanical presses of the nineteenth century. And, as with other ephemeral texts printed in this era, newspapers often exhibit worn or broken type. Pages also have shrunk unevenly over time and have been torn and repaired. Further, there is a fair amount of bleed-through on the pages. Human readers might fairly effectively filter out the noise of bleed-

through without thinking about it, but it presents challenges for our current computational approach.

The features of the Burney Collection described above pose some challenges for our approach. For example, the varied and non-standard columns as well as bleed-through are difficult for our current approach to page segmentation. Our current approach is predicated on being able to find some consistent column breaks and then extrapolate from those as to where to break the page into columns. Irregularity in columns as well as bleed-through can undermine these efforts. Likewise, damage to the page, bleed-through, and extraneous non-textual information on the page (“noise”) can make it difficult for our current approaches to measure signal features present in a snippet. *That is, our current approaches to segmentation and to classification are not necessarily robust enough to handle common features of the images, whether those common features are inherent to the original newspapers, emerged during microfilming, or become present in digitization.*

The results described above also demonstrate that each of the classifiers tested are overfitted to the training data. The much higher rates of precision and recall for the training sets, as compared to significantly lower rates of precision and recall for the testing sets, are evidence of overfitting. Put another way, the high training accuracy indicates that our model is able to capture a set of representative features for distinguishing between two classes. During testing, however, those features are not fully representative of the larger set. The training set is not representative enough to provide rich information for our model. In addition, the complexity of the visual cues we are trying to teach the classifier to learn to recognize may be too subtle (due to noise) to adequately capture. This challenge in capture compounds the challenges of representativeness even further. *While overfitting will no doubt be present in some way across all of our efforts—unless we could develop a test set that manages to account for all possible variations—we can reduce the overfitting to a more acceptable level.*

Given our observations about the corpus and its history, as well as the results described above, our major next steps at this phase are three-fold. *First, we plan to analyze both the Chronicling America and Burney Collection corpora at the page-level.* We are using image processing-based measures, such as for contrast, range effect, skew orientation, noise effect, and layout compactness and complexity. Deriving these measurements and analyzing them will help us better understand similarities and differences at the page-level both within and across the corpora. We anticipate that doing so will help shift what are largely anecdotal and incomplete assessments of the corpora (of the images as well as of the newspapers themselves) to a more holistic understanding. In addition, this work will allow us to better understand and index the images for future use. As of May 2019, we have page-level data for 10,000 pages from Chronicling America and 10,000 pages from the Burney Collection. We are in the process of analyzing this data and will use the 10k analyses to inform next steps in page-level evaluation.

Another area of work emerging in part from this analysis, as well as from earlier work with Chronicling America and other newspaper corpora, is to *change our approach to zoning/segmentation for newspaper page images.* We have observed that our current approach to segmentation, including the pre-tests we have developed for passing only “good” images through to

segmentation, both excludes too many input images from the outset and does not produce consistently good-quality image snippets, even for those page images that pass the segmentation tests. With our current approach to feature extraction, even seemingly small problems with segmentation create problems for accurately measuring features. Inaccurate feature measurements then decrease the accuracy of classification. As a result, we are currently testing alternative approaches to segmentation that are more akin to the zoning that happens as part of pre-processing for optical character recognition processes but that also work well on dense, complex layouts such as those present in historic newspapers.

In addition, we are experimenting with *more powerful neural networks, including deep-learning-based designs such as convolutional neural networks, for learning features and completing classifications of the input images*. In these deep learning approaches, we continue to train the classifiers with true and false images, but we no longer define the features, nor tell the system how to measure the features, that we believe signal poetic content.

The improvements to zoning and to applying deep-learning approaches are currently in process. Since a major goal of our IMLS grant is to assess the efficacy of our approach and methods in order to consider what it might take to scale these approaches and methods, we will explore several scenarios going forward. These include:

- A. Improved (second-generation) zoning → first-generation classification
- B. First-generation zoning → deep learning (second-generation) classification
- C. Second-generation zoning → deep learning (second-generation) classification

We plan to explore these scenarios across both the Burney Collection and Chronicling America corpora. Both of the second-generation strategies—to zoning and to classification—are higher-resources strategies, which take more time and more computational resources than our first-generation strategies. In addition to understanding the implications of the individual components in our overall approach, we seek to develop the lightest-weight system possible for accomplishing the goals of classification. Therefore, a fourth scenario we may explore is no zoning (whole page) → deep learning (second-generation) classification. While such an approach eliminates zoning and the time and overhead of that approach, it may be too computationally expensive for classification and/or the features that signal poetic and other types of content within a page may become too muted when evaluated in the whole.

Ultimately, then, while the current study shows that our first-generation approaches to zoning, feature extraction, and classification do not yield adequate accuracy for finding poetic content in the Burney Collection, we remain optimistic about the viability of the larger conceptual framework or method. The current study and our other ongoing work have led us to understand the corpus and our methods at a greater level of detail and also more holistically.