Impact of OCR Quality on BERT Embeddings in the Domain Classification of Book Excerpts

Ming Jiang, Yuerong Hu, Glen Worthey, Ryan C. Dubnicek, Ted Underwood and J Stephen Downie

University of Illinois, Urbana-Champaign, USA

Abstract

Digital humanities (DH) scholars have been increasingly interested in using BERT for document representation in computational text analysis. However, most word embeddings, including BERT embeddings, have been developed using "clean" corpora, while DH research is usually based on digitized texts with optical character recognition (OCR) errors. Will these errors introduced by the digitization process reduce BERT's performance and distort the research findings? To shed light on the impact of OCR quality on BERT models, we conducted an empirical study on the resilience of BERT embeddings (pre-trained and fine-tuned) to OCR errors by measuring BERT's ability to enable classification of book excerpts by subject domain. We developed specialized parallel corpora for this task consisting of matching pairs of OCR'd text (19,049 volumes) and "clean" re-keyed text (4,660 volumes) from English-language books in six domains published from 1780 to 1993. This study is the first to systematically quantify OCR impact on contextualized word embedding techniques with a use case of OCR'd book datasets curated by digital libraries (DL). Experimental results show that pre-trained BERT is less robust when used on OCR'd texts; however, fine-tuning pre-trained BERT on OCR'd texts significantly improves its resilience to OCR noise in classification tasks according to the changes of classifier performance. These findings should assist DH scholars who are interested in using BERT for scholarly purposes.

Keywords

Optical Character Recognition, BERT Resilience, Word Embeddings, Text Analysis, Parallel Corpora, HathiTrust, Digital Humanities, Digital Libraries, Data Curation

1. Introduction

The accessibility of ever-growing digitized textual curations in digital libraries (DL) and the rapid development of natural language processing (NLP) techniques have opened up a variety of new research opportunities to humanities scholars for computational text analysis [19, 12, 13]. In recent years, BERT (Bidirectional Encoder Representations from Transformers) has been widely used as a fundamental text representation tool in text-based computing, for it focuses on encoding the contextual meaning of words into a vector space [7, 24]. There are two main reasons for its popularity. First, in encoding word tokens rather than word types (i.e., distinct words), BERT is helpful in identifying the correct meaning of a homonym within

CHR 2021: Computational Humanities Research Conference, November 17–19, 2021, Amsterdam, The Netherlands

mjiang17@illinois.edu (M. Jiang); yuerong2@illinois.edu (Y. Hu); gworthey@illinois.edu (G. Worthey); rdubnic2@illinois.edu (R.C. Dubnicek); tunder@illinois.edu (T. Underwood); jdownie@illinois.edu (J.S. Downie)

0000-0002-3604-166X (M. Jiang); 0000-0001-8375-9108 (Y. Hu); 0000-0003-2785-0040 (G. Worthey); 0000-0001-7153-7030 (R.C. Dubnicek); 0000-0001-8960-1846 (T. Underwood); 0000-0001-9784-5090 (J.S. Downie)

its context (e.g., *bank* in "river bank" and "savings bank"). Second, BERT can leverage the general linguistic knowledge it has learned from a massive, high-resource corpus such as Wikipedia to serve specialized and lower-resource downstream tasks, such as movie review sentiment classification [1]. So far, BERT has produced promising improvements in both (1) fundamental text analysis, e.g., text segmentation [1], named entity recognition [28, 16], and post-OCR correction [28, 20]; and (2) specific research topics, e.g., historical analysis of semantic change in lexical/grammatical constructions [24, 18, 9], literary genre analysis [30, 4], literary event detection [25], and computational narrative intelligence[23].

Digital humanities (DH) scholars working with computational analysis have been increasingly interested in using this technique for their research on digitized texts. However, a majority of large DL text curations and other historical text collections are machine-transcribed and include varying degrees of optical character recognition (OCR) noise. Such noise might decrease the generally impressive performance of BERT because it was originally developed on born-digital texts without OCR errors [7]. Even though existing OCR systems have significantly improved through advances in AI techniques (e.g., image recognition) and persistent efforts of digital curators (e.g., the Library of Congress, HathiTrust Digital Library), OCR noise can hardly ever be completely eliminated given its ubiquity, its uneven distribution, and the heterogeneous nature of its source texts. Meanwhile, advanced NLP techniques like BERT are generally limited in their transparency and interpretability, which is even worse when processing OCR'd texts. [17]. Such uncertainty might reduce the credibility of digital humanities research when applying BERT-based computations to OCR'd texts for further analysis.

Therefore, we believe BERT's performance on OCR'd texts is an important problem to look into. This study aims to empirically investigate this problem with three research questions: (1) Would the original BERT model [7] (pre-trained on Wikipedia and free Web books) work as well with OCR'd texts containing noise? (2) If we fine-tune the pre-trained BERT using a corpus with a certain amount of OCR noise, would this result in any improvements for processing OCR'd texts in downstream tasks? and (3) What are the quantifiable impacts of OCR quality on both pre-trained and fine-tuned BERT models?

To shed light on the interaction between OCR'd texts and BERT, we focused on measuring the ability of BERT to encode digitized texts' semantics and comparing the performance of BERT encoding on clean (i.e., re-keyed) versus OCR'd texts. The texts we used were book excerpts generated from ∼4,000 pairs of book volumes selected from a parallel corpus of digital English-language books, with 4,660 human-proofread "lean" volumes from Project Gutenberg (Gutenberg) and their matching pairs of 19,049 OCR'd volumes from HathiTrust Digital Library (HathiTrust) [12]. Books in this corpus cover six subject domains published from 1780 to 1993. We chose subject domain classification as the application downstream from BERT in order to quantify its encoding performance, because document classification in general is a popular application for digital humanists studying subject, genre, authorship, and many other features of their texts. [34, 27]. Specifically, we investigated both the generic embedding obtained from the pre-trained BERT model and the domain-adapted embedding by fine-tuning the pre-trained BERT on the downstream training corpus (i.e., either clean or noisy).

The remainder of this paper is organized as follows. In section 2, we review related work on BERT and OCR'd texts. In section 3, we provide detailed information about the parallel book dataset that we created and leveraged, and how we built the book excerpt corpora needed for our experiments. In section 4, we describe our research design and workflow. We also give explanations for the specific decisions made and methods adopted. In section 5, we present our experimental results and findings. Finally in section 6, we discuss our conclusions and future

Table 1
Statistics of three parallel training corpora

	Fiction	Social_Science	Agriculture	World_War_History	Medicine	Business	**Total**
Small_Balanced(SB)	167	167	167	167	166	166	**1000**
Small_Unbalanced(SU)	355	152	148	130	122	93	**1000**
Large_Unbalanced(LU)	1164	423	409	341	359	304	**3000**

work.

2. Related Work

BERT used in existing work for digital history and literary studies generally plays a text preprocessing role by encoding text information into vectors for further computation. Popular research topics in this field mainly focus on the diachronic analysis of literary texts [24, 18, 9, 30] and narrative understanding [25, 23]. Regarding data sources, commonly used corpora typically come from Project Gutenberg [25], the Corpus of Historical American English [9], and OCR'd text collections organized in DL [24, 18]. Although BERT has shown its power in representing clean texts, some empirical studies [24, 14, 6] have witnessed a drop of its performance on processing digitized texts containing OCR errors. Inspired by that, we are interested in advancing the understanding of BERT's applicability on OCR'd noisy texts.

Based on a literature review on OCR noise analysis, common error types include character misidentification (e.g., "inserted"→"insorted"), broken words (e.g., un-rejoined hyphenated words "talking"→"talk- ing"), incorrectly joined words (e.g., "the belief"→"th-ebelief"), and meaningless symbols (e.g., OCR attempts to recognize hand-written marginalia) [3, 8]. Given the various patterns and random distribution of OCR noise, even the state-of-the-art techniques for OCR correction cannot completely filter the OCR noise out.

Prior work on the impact of uncorrected OCR'd texts on other NLP tasks can be divided into two groups: (1) those quantifying impact by measuring the performance differences of a set of popular NLP techniques applied on a parallel corpus consisting of OCR'd and clean texts [11, 26, 5]; and (2) those analyzing OCR impact by interviewing scholar-users for their feedback on the use of digital archives and NLP techniques for computational textual analysis [29]. Popular NLP tasks adopted in existing studies include tokenization, sentence segmentation, named entity recognition, dependency parsing, topic modeling, information retrieval, text classification, collocation, and authorial attribution [11, 26, 5]. Most studies show that OCR errors lead to a consistent negative influence on NLP tasks, even for some tasks that have been considered "solved" (e.g., sentence segmentation)[26]. In this research, we extend prior work by studying the impact of OCR quality on BERT-based text representations, where we particularly explore BERT's ability to encode the intrinsic semantic features of OCR-impacted texts in comparison with its encoding of parallel clean texts.

3. Data and Corpora Preparation

The source data for this study is a parallel corpus of English monographs [12] collected from two real-world digital libraries: (1) Gutenberg for a human-proofread "clean" corpus; and, (2) HathiTrust for an OCR'd "noisy" corpus. This corpus has a total of 4,660 Gutenberg volumes

in 6 domains (i.e., fiction, social science, agriculture, medicine, business, world war history), each of which is matched with several different copies (4 on average) of the same work held in HathiTrust.

Since classification is a supervised learning task, we started by preparing three parallel data splits from the raw corpus for training, validation, and testing, respectively. Considering the many-to-one matching relationship between Hathitrust and Gutenberg volumes, in order to make the clean and OCR'd version of each data split, aligned by volume, and to avoid volume duplication in splits with clean data, we first split Gutenberg data by randomly selecting 10% of 4,660 Gutenberg volumes for validation (465 volumes), 10% for testing (467 volumes), and the rest for training. Then we randomly picked one paired HathiTrust copy of each Gutenberg volume to build corresponding training, validation and testing splits of OCR'd texts.

Following [2, 21], data distribution and downstream corpus size also influence the embeddings' encoding ability, in addition to text quality, especially for the fine-tuned BERT embedding. Taking these two variables into consideration, we modified the original parallel training split by resampling the data into three types of parallel training corpus: (1) a small balanced corpus (SB) containing 1000 books with an equal number of books per genre; (2) a small unbalanced corpus (SU) containing 1000 books with a different number of books per genre; and (3) a large unbalanced corpus (LU) containing 3000 books with a different number of books per genre. Table 1 shows the details of each type of training corpus. Given the highly skewed data distribution in the original parallel corpus (e.g., fiction volumes comprise 88%) [12], our unbalanced corpora were generated by a slight smoothing based on the exponentially smoothed weighting method [10], where we empirically set the smoothing factor as 0.3.

There are two main challenges in the encoding of book content by BERT. First, book-length texts and the computational cost of BERT models make it expensive to encode each volume's full text. Moreover, BERT models are restricted to processing at most 512 tokens at a time, which limits their encoding abilities on long sentences. To address these issues, we followed prior work [31, 32] by parsing the full content per volume into a set of word sequences with at most n tokens and randomly sampled k continuous word sequences as a text chunk to feed into BERT. Referring to prior studies' parameter settings and our own hardware computing constraints, we set $n = 128$ and $k = 15$ ($\sim$1920 tokens per chunk). Recent studies on subject domain and genre classification [31, 32] show that book chunks should be sufficient for predicting an entire book's subject, and with this premise, we decided to focus on parallel book excerpts for our study. Although this method could not process complete volumes, the random sampling strategy is helpful in augmenting the book content to be trained or tested as much as possible, which compensates for the limits on text length.

To make each classifier's predictions on clean versus OCR'd test set comparable, the sampled text chunks from each pair of test volumes were aligned by an existing text alignment algorithm [33]. We manually examined a random sample of chunk pairs to ensure alignment accuracy. Furthermore, for a statistical significance test of the classification results, we grouped all the sampled chunk pairs into a set of parallel testing folds. In the end, our parallel testing corpus consists of 20 parallel testing folds, where each parallel fold contains one unique pair of text chunks extracted from a pair of Gutenberg and HathiTrust volumes($20 \times 467 = 9340$ parallel examples in total).

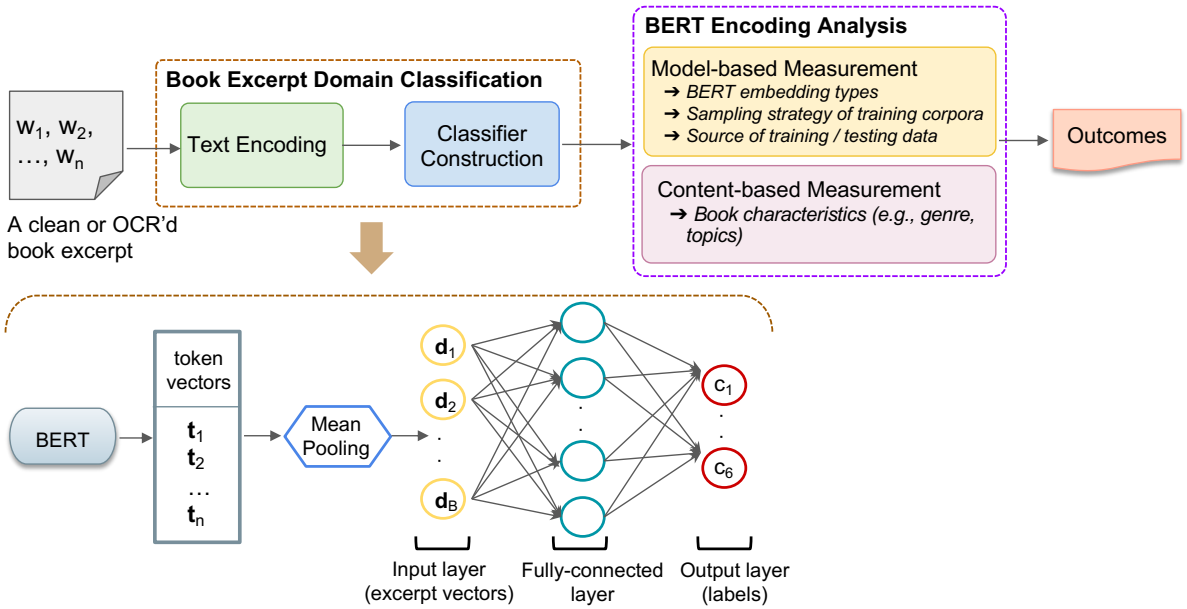

Figure 1: Overview of study workflow

4. Research Design and Workflow

The primary goal of this study is to analyze the performance of BERT embeddings in encoding book excerpts into n D-dimensional (D=768) token vectors for book domain classification based on the parallel clean and OCR'd texts. We measured and compared BERT embeddings' encoding ability in different classifiers using macro-averaged precision (P), recall (R), and F1 score (F1). Considering the potential influence of experimental settings on BERT embeddings' performance, we analyzed the classification outcomes based on the model settings and data characteristics respectively. Figure 1 visualizes the overall workflow of this study, which includes two stages: (1) building classifiers based on text representations offered by BERT embeddings on book excerpts; and (2) quantifying BERT embeddings' performance in different classification settings to analyze BERT embeddings' resilience to OCR noise.

4.1. Domain Classifier Construction

With the encoded BERT token representations per excerpt, we first generate a single chunk-level feature vector by averaging token vectors, one of standard practices popularly used in prior work [22], for further excerpt classification. With 2 types of BERT embedding, 3 types of training data sampling, and 2 aligned training corpora, in total, this study built 12 classifiers. Considering that our primary goal is to explore BERT embeddings' resilience against OCR errors rather than improving classification performance, we employed a fundamental multi-perception neural network model with three layers for building classifiers. With respect to the training process, by feeding the set of training examples, the model was expected to learn a weighting matrix for predicting the mapping probability per example into each domain class, where each training example was assigned to the domain with the highest probability. Following the standard practice of applying deep learning techniques for classification [1, 30],

our model was optimized by a cross-entropy loss function during training to maximize the model predictability (i.e., F1 score). To compare the consistency of predictions with and without OCR errors, we proposed two types of classifications: (1) both training and testing corpora are either clean or noisy (i.e., containing OCR errors); and (2) one is clean and the other is noisy.

The detailed implementation of model training is as follows. We used the Adam optimizer [15] to train all classification models with 20 epochs[1]. As to the learning rate, for pre-trained BERT-based classifiers, we set this parameter as 2.0e-3 for for the Gutenberg corpus and 2.5e-3 for the HathiTrust corpus respectively, while for fine-tuned classifiers, we set both of them 2.5e-5. Our empirical setting for this parameter was based on the resultant classifier's performance on the validation set in order to find the optimal one. The batch size was set as 40 (book excerpts) for all the models.

4.2. Analysis of BERT Encoding on Clean Versus OCR'd Texts

4.2.1. Model-based measurement

Based on the classification results of 12 generated classifiers on our parallel testing corpus, we analyzed the relations among BERT embedding types (i.e., pre-trained or fine-tuned BERT), the source of training and testing data, and the sampling strategy of training corpora by pairwise comparison of any two of three variables. Our goals were: (1) finding the optimal BERT embedding with the highest resilience against OCR errors; and (2) identifying the optimal sampling strategy for building the training corpus that most significantly improves the BERT embedding performance.

Given that the above analysis primarily focused on the comparison of BERT-based classifiers' overall performance, we further proposed a fine-grained investigation of BERT embeddings' resilience to OCR errors regarding the amount of noise. To conduct this investigation, we first prepared three subsets of OCR'd testing data containing different amounts of OCR errors. The level of OCR noise was measured by the character-level error rate (CER) based on the comparison of each OCR'd book excerpt with its paired clean text. After sorting the OCR'd excerpts by their CER in an ascending order, from this ranked excerpt list, we separately sampled 1500 excerpts at the top, middle, and the bottom position as the low-, medium-, and high-noisy testing subsets. Figure2 displays the distribution of CER in each testing subset, where the average CER per subset is around 0.40, 0.54, and 0.65, respectively. We then evaluated each classifier's predictability on each subset. Note that, in this analysis, we only considered those classifiers trained on the corpus with the identified optimal sampling strategy. To further look into the resilience of BERT embeddings with respect to the change of the downstream classification's training corpus source, rather than exploring each individual classifier's results, we measured the divergence of classification results between the classifier trained on the clean versus the OCR'd texts for each type of BERT embedding.

4.2.2. Content-based measurement

Although each book in the raw parallel corpus was assigned to a single subject domain tag, given the diversity of content-based characteristics (e.g., topics, genres, narrative styles) inherent in a book-sized text and its randomly sampled excerpts, it is possible that the input

[1]The number of epochs was optimized empirically by trying a set of values (i.e., 15, 20, 30, 50).

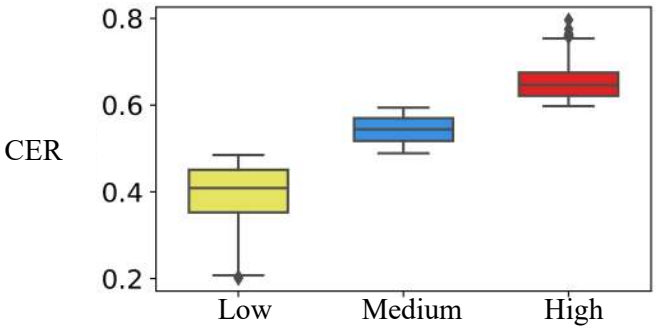

Figure 2: Distribution of the amount of OCR noise measured by CER in three sampled testing subsets. Each set contains 1500 examples.

Table 2
Classification results on three training corpora (%). P, R, and F1 denotes precision, recall, and F1 score, respectively. All evaluation indicators are at macro-level and represent the average value of results over 20 folds of testing samples. The highest F1 score per classification strategy in each training setting is highlighted in bold.

		G→G			G→H			H→G			H→H		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
SB	Pre-trained	49.88	71.67	53.24	49.97	70.44	52.64	47.14	74.50	53.33	46.97	73.25	53.05
	Fine-tuned	69.06	79.75	72.65	70.00	79.17	72.99	68.07	79.06	71.54	68.93	78.07	71.70
SU	Pre-trained	70.31	67.05	66.28	70.67	64.68	65.48	60.24	66.38	60.89	62.32	65.01	60.98
	Fine-tuned	75.23	77.71	74.39	75.25	77.50	74.71	75.79	78.83	**76.27**	74.94	79.45	**76.20**
LU	Pre-trained	64.30	74.16	67.71	65.79	72.69	67.88	59.59	73.44	64.38	60.17	72.82	64.86
	Fine-tuned	76.02	79.51	**76.60**	75.71	79.78	**76.71**	74.60	80.33	76.10	73.86	80.01	75.72

data itself might bring challenges for a BERT-based classifier to identify its annotated domain tag. Moreover, whether and how such challenges occur with OCR'd texts vary from those occurring with clean texts is uncertain. For instance, if all BERT-based classifiers fail to classify either clean or OCR'd excerpts of the same book correctly, one potential reason for this result could be that the original book includes more than one subject. In contrast, if all classification models work well on the clean texts only, it is likely that OCR noise is resulting in different predictions. To address these concerns, we started by exploring semantic associations among misclassified domains by visualizing the confusion matrix of each classifier. To further capture book excerpts' individual features for understanding their influence on classification, we then grouped the predictions made per classifier on individual excerpts by book, to measure the consistency of classifiers' prediction accuracy at the book level. This measurement is based on calculating the number of testing excerpts of the same book that were assigned to the same correct domain across different classifiers on average. Given the quantitative outcomes, we sampled some cases with poor prediction accuracy, and explored potential reasons for misclassification by close reading of the book content.

5. Outcomes and Findings

5.1. Resilience of BERT embeddings

Table 2 provides an overview of the classification results grouped by (1) source of training and testing data (Gutenberg or HathiTrust); (2) sampling strategy of parallel training corpus (small-balanced, small-unbalanced and large-unbalanced); and (3) type of BERT embedding (pre-trained or fine-tuned). Overall, we observe that classifiers built with fine-tuned BERT outperformed those built with pre-trained BERT by 20% (F1 score) based on the balanced training corpora and 10% (F1 score) based on the unbalanced training corpora. This result indicates that the fine-tuning process, intended to adapt the generic pre-trained BERT embedding space to fit into a specific text corpus (either clean or OCR'd), will substantially improve the encoding ability of BERT for digitized literary texts even with the distortion of OCR noise.

Regarding the influence of training sampling strategies to BERT encoding, in general, unbalanced corpora were more helpful in training classifiers than balanced corpora, which suggests that excessive artifact intervention of training data distribution indeed could hurt BERT's encoding ability. Table 3 further shows the paired t-test scores of the statistical difference of performance between any two comparable classifiers that differ only in either size or data distribution of training corpus. It is to be noted that differences between any two compared classifiers' performances over 20 testing folds follow an approximately normal distribution based on the Shapiro-Wilks Test. According to the results, pre-trained BERT-based classifiers are all sensitive to both size and data distribution in the training corpus (p-value < 0.05 at least). However, the increase in size of the OCR'd training corpus has no significant impact on fine-tuned BERT embedding. This observation may be understood as a positive signal to humanities scholars that a small training corpus is enough to achieve optimal performance of fine-tuned BERT when working with OCR'd texts. Comparatively, training corpus size (t-test score from -0.71 to 3.32 where p-value < 0.01 at most) is less influential on BERT embeddings' performance than is training data distribution (t-test score from 2.05 to 15.54 where the majority of p-values < 0.001).

Similar to the analysis of training sampling strategies, we compared classifiers' performance with respect to the source of training data. Table 4 shows the paired t-test results. Pre-trained BERT-based classifiers were significantly more sensitive to their training data source when these classifiers were built on unbalanced training corpora (p-value tends to be < 0.001). In particular, the growth of training corpus size increased such sensitivity (t-test score increased from 4.09*** to 5.85 when testing on the clean corpus, and from 3.49** to 4.31*** when testing on the OCR'd corpus). Meanwhile, for fine-tuned BERT, classifiers only showed their sensitivity to the source of training data in small unbalanced training corpora (t-test score was -2.86** when testing on the clean corpus, and -2.10* when testing on the OCR'd corpus). According to the F1 score of these classifiers' prediction results shown in Table 2, we found that, compared with fine-tuning on clean texts, fine-tuning on OCR'd texts improved BERT-based classifiers' performance by $\sim$2%, which suggests that potential OCR noise in the small-unbalanced corpus for BERT fine-tuning can boost the resulting embedding's encoding performance.

5.2. Impact of the amount of OCR noise on BERT encoding.

Given three testing sample sets with different levels of OCR noise (see details of data preparation in section 4.2.1), Table 5 shows the divergence of F1 score between classifiers built with

Table 3
Paired t-test shows the differences of classification results varied by training strategies. The statistical significance is represented by p-value (one-tailed), where *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$.

	G→G		G→H		H→G		H→H	
	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned
SU vs. SB	15.54***	2.33*	11.06***	2.05*	7.76***	6.44***	7.87***	5.07***
LU vs. SU	1.85*	2.65**	2.42*	1.99*	3.32**	-0.22	2.94**	-0.71

Table 4
Paired t-test shows the differences of classification results varied by training data source. The statistical significance is shown by p-value (one-tailed), where *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$.

	SB		SU		LU	
	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned
G→G vs. H→G	-0.13	1.41	4.09***	-2.86**	5.85***	0.73
G→H vs. H→H	-0.68	1.37	3.49**	-2.10*	4.31***	1.17

either pre-trained or fine-tuned BERT embeddings on each sample set. This divergence was calculated by the subtraction of classification results using OCR'd texts for training from the one using clean texts for training.

Overall, we found that classifiers obtained greater benefit from clean training data compared with OCR'd data *except* in the case of fine-tuned BERT-based classifiers making predictions on the low-noise testing data. Regarding the classification divergence across the three testing sample sets, we observed a gradual decrease in difference on testing samples with low (4.88%), medium (3.96%), and high (0.70%) level of OCR noise when classifiers employed pre-trained BERT for text encoding, while the pattern was the opposite in classifiers built with fine-tuned BERT (i.e., -1.96% for low noisy group, 1.43% for medium noisy group, and 3.79% for high noisy group). We further compared the absolute differences of classification results between two classifiers per embedding type, and found that testing samples with lower-level OCR noise were more sensitive to the training data source than those with higher-level noise in pre-trained BERT-based classifiers. On the contrary, for the classifiers built with the fine-tuned BERT, the largest performance difference was found in the testing set with a high amount of dirty OCR size.

Here are three major conclusions. First, the consistency of text quality in an embedding's pre-training corpus, downstream training, and downstream testing corpus is helpful in improving pre-trained BERT's applicability for literary text classification. Second, the heterogeneous nature of OCR noise can improve the generalization ability of fine-tuned embeddings to process texts with comparatively low levels of OCR noise. Finally, fine-tuned BERT-based classifiers are more stable with regard to changes in the source of training corpus than pre-trained BERT-based classifiers, which further confirms that fine-tuned BERT outperforms pre-trained BERT in its resilience to OCR errors.

5.3. Error analysis by content-based measurement.

Figure 3 shows eight confusion matrix heatmaps for the eight classifiers trained on the large unbalanced corpora. In each matrix, the diagonal values in comparatively darker blue cells

Table 5
Divergence of classification results (F1 score) by changing the training corpus source from clean to OCR'd texts on three testing sample sets with different levels of OCR error.

	Low Noisy	Medium Noisy	High Noisy
LU Pre-trained	4.88%	3.96%	0.70%
LU Fine-tuned	-1.96%	1.43%	3.79%

represent the ratio of correct predictions, while the other values indicate the ratio of misclassifications (actual VS predicted). The higher the value is, the darker its corresponding cell color. For example, in the first matrix (fine-tuned, G→G), the value "0.45%" in the cell at the upper left corner indicates that 0.45% of "world war history" excerpts were misclassified as "agriculture" by the fine-tuned BERT-based classifier, which was trained and tested on Gutenberg texts. For both pre-trained and fine-tuned BERT-based classifications, we found that book excerpts in the business domain were more likely to be misclassified as fiction (25.4% on average) and social science (19.8% on average), while book excerpts in the medicine domain were more likely to be mistakenly classified as social science, especially with fine-tuned BERT-based classifiers trained on the OCR'd texts (32.86% misclassifications in H→G classification and 27.86% misclassifications in H→H). By looking more closely at social-science instances, we observed that the pattern of misclassifications was different in the classifier built with pre-trained BERT compared with that built with fine-tuned BERT. Specifically, in the classifications using pre-trained BERT for text encoding, prediction errors mainly concentrated in the domains of business (10% on average), medicine (8.5% on average), and fiction (7.5% on average). Meanwhile, for fine-tuned BERT-based classification, fiction (17% on average) and medicine (11% on average) were the top two misclassifications for social-science excerpts.

Comparing prediction errors with respect to the source of data for training and testing, we found that the pattern of misclassification in fine-tuned BERT-based classifications tended to be similar among all four types of classification. However, the ratio of errors per domain in pre-trained BERT-based classifications was likely to be different depending on the classifiers' training corpus source. For example, business instances tended to be misclassified as fiction (25%-28%) when the training corpus is clean, but as social science (23%-27%) when using OCR'd texts for training. Similarly, medicine instances have an markedly higher ratio of misclassification as social science (27.89%-32.86%) in the OCR'd training corpus compared with the clean one (11.43%-16.43%). These observations reaffirm that fine-tunbed BERT is more robust for processing OCR'd texts compared with pre-trained BERT.

We further looked into the prediction consistency of all BERT-based classifiers on each book in both clean and OCR'd versions. Given two aligned lists (i.e., clean and OCR'd) of book-level average prediction accuracy across different classifiers, we found that there was a large overlap of books with comparatively low accuracy in clean versus OCR'd corpus, which suggests that content-based characteristics of these particular books may be the main cause of recurring prediction mistakes. We verified this hypothesis by manually checking the books with the lowest prediction scores, and confirmed that these books had heterogeneous genre-related features which were confusing even for human readers. For instance, the book *The Story of My Life* by Helen Keller is generally considered a classic "social science" work because of its main subject and its many non-fiction features. However, this is a classic autobiography composed of touching stories of a great woman struggling with severe disability, first published in 1903.

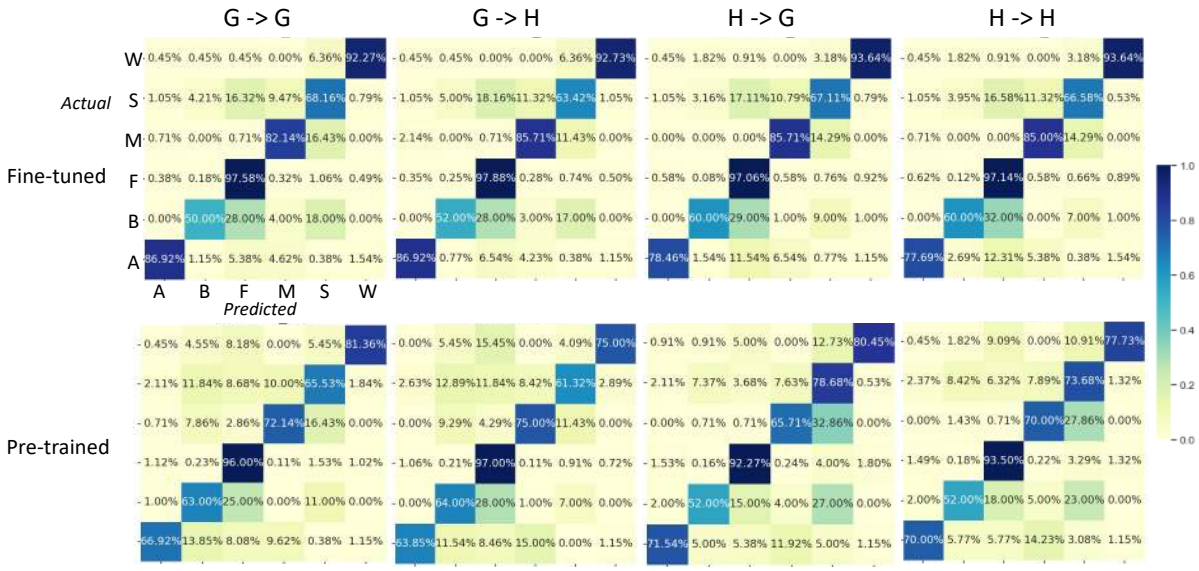

Figure 3: Confusion matrices of classification models built on the large-unbalanced training corpora. Labels "A", "B", "F", "M", "W", "S" represent "agriculture", "business", "fiction", "medicine", "world war history" and "social science".

Therefore, it is less surprising and even understandable for the models to label its instances as "medicine" or "fiction" based on their learning of the training data.

6. Conclusions and Future Work

We have investigated the resilience of pre-trained and fine-tuned BERT embeddings for encoding OCR'd texts through a case study of classifying book excerpts into subject domains. To the best of our knowledge, this is the first empirical study to systematically quantify the influence of OCR quality on BERT. By changing BERT embedding types and classification model settings, we built 12 BERT-based classifiers using book excerpt corpora extracted from a large parallel book corpus of aligned clean and OCR'd volumes sourced from two well-known digital libraries. Our analysis shows that the original BERT embedding pre-trained on born-digital texts is not resilient to OCR noise, at least according to its classification accuracy. However, fine-tuning the pre-trained BERT on OCR'd texts will significantly improve BERT's resilience to OCR noise, and hence will benefit downstream applications. Besides, fine-tuned BERT outperforms the pre-trained one in its encoding stability with regards to changes in training corpus size and training data source. For both types of BERT embedding, unbalanced training corpora benefit embeddings' resilience to OCR noise in downstream classifications. Our findings suggest that DH scholars should consider employing fine-tuned BERT for digitized-text-based scholarly research, particularly when their research involves document classification.

While our experiments yield significantly positive evidence for fine-tuned BERT embeddings' resilience to OCR noise in the use-case of document classification, the impact of OCR noise on BERT for other downstream tasks remain under-investigated. For example, it is possible that BERT could react to OCR noise differently at more fine-grained levels, such as sentence-level tasks (e.g., next sentence prediction, sentence-based sentiment analysis, etc.) and word-level (e.g., part-of-speech tagging, etc.). Therefore, future work focusing on BERT's performance on

OCR'd texts both at different text granularities and for different downstream NLP tasks would be useful to deepen our understanding of how OCR impacts this contextualized embedding technology. Furthermore, since our corpora consist exclusively of English-language books from the 18th and 19th centuries, expanding this study to curated datasets from other historical periods, languages, and publication types would be a very worthwhile future exercise.

References

[1] A. Adhikari, A. Ram, R. Tang, and J. Lin. "Docbert: BERT for document classification". In: *arXiv preprint arXiv:1904.08398* (2019).

[2] M. Antoniak and D. Mimno. "Evaluating the stability of embedding-based word similarities". In: *Transactions of the Association for Computational Linguistics* 6 (2018), pp. 107–119.

[3] G. T. Bazzo, G. A. Lorentz, D. S. Vargas, and V. P. Moreira. "Assessing the Impact of OCR Errors in Information Retrieval". In: *Proceedings of European Conference on Information Retrieval*. Springer. 2020, pp. 102–109.

[4] Ben. *Language Models & Literary Clichés: Analyzing North Korean Poetry with BERT*. 2020. URL: https://digitalnk.com/blog/2020/10/01/language-models-literary-cliches-analyzing-north-korean-poetry-with-bert/.

[5] G. Chiron, A. Doucet, M. Coustaty, M. Visani, and J.-P. Moreux. "Impact of OCR errors on the use of digital libraries: Towards a better access to information". In: *Proceedings of 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. Ieee. 2017, pp. 1–4.

[6] A. Cuba Gyllensten, E. Gogoulou, A. Ekgren, and M. Sahlgren. "SenseCluster at SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection". In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, 2020, pp. 112–118. URL: https://aclanthology.org/2020.semeval-1.12.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 2019, pp. 4171–4186.

[8] J. Esakov, D. P. Lopresti, and J. S. Sandberg. "Classification and distribution of optical character recognition errors". In: *Document Recognition*. Vol. 2181. International Society for Optics and Photonics. 1994, pp. 204–216.

[9] L. Fonteyn. "What about Grammar? Using BERT Embeddings to Explore Functional-Semantic Shifts of Semi-Lexical and Grammatical Constructions". In: *Proceedings of the Workshop on Computational Humanities Research*. 2020, pp. 257–268.

[10] E. S. Gardner Jr. "Exponential smoothing: The state of the art". In: *Journal of Forecasting* 4.1 (1985), pp. 1–28.

[11] M. J. Hill and S. Hengchen. "Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study". In: *Digital Scholarship in the Humanities* 34.4 (2019), pp. 825–843.

[12] M. Jiang, Y. Hu, G. Worthey, R. C. Dubnicek, B. Capitanu, D. Kudeki, and J. S. Downie. "The Gutenberg-HathiTrust parallel corpus: A Real-World Dataset for Noise Investigation in Uncorrected OCR Texts". In: *iConference 2021 (Poster)* (2021).

[13] M. L. Jockers. *Macroanalysis: Digital methods and literary history*. University of Illinois Press, 2013.

[14] V. Kanjirangat, S. Mitrovic, A. Antonucci, and F. Rinaldi. "SST-BERT at SemEval-2020 Task 1: Semantic Shift Tracing by Clustering in BERT-based Embedding Spaces". In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, 2020, pp. 214–221. URL: https://aclanthology.org/2020.semeval-1.26.

[15] D. P. Kingma and J. Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[16] K. Labusch, P. Kulturbesitz, C. Neudecker, and D. Zellhöfer. "BERT for Named Entity Recognition in Contemporary and Historical German". In: *Proceedings of the 15th Conference on Natural Language Processing*. 2019, pp. 8–11.

[17] T. Linzen, G. Chrupała, Y. Belinkov, and D. Hupkes, eds. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, 2019. URL: https://aclanthology.org/W19-4800.

[18] M. Martinc, P. K. Novak, and S. Pollak. "Leveraging contextual embeddings for detecting diachronic semantic shift". In: *arXiv preprint arXiv:1912.01072* (2019).

[19] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, et al. "Quantitative analysis of culture using millions of digitized books". In: *Science* 331.6014 (2011), pp. 176–182.

[20] T. T. H. Nguyen, A. Jatowt, N.-V. Nguyen, M. Coustaty, and A. Doucet. "Neural machine translation with BERT for post-OCR error detection and correction". In: *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*. 2020, pp. 333–336.

[21] S. Padma, S. S. Kumar, and R. Manavalan. "Performance analysis for classification in balanced and unbalanced data set". In: *Proceedings of the 6th International Conference on Industrial and Information Systems*. Ieee. 2011, pp. 300–304.

[22] S. Palachy. *Document Embedding Techniques*. 2019. URL: https://towardsdatascience.com/document-embedding-techniques-fed3e7a6a25d%5C#ecd3.

[23] L. Qin, A. Bosselut, A. Holtzman, C. Bhagavatula, E. Clark, and Y. Choi. "Counterfactual Story Reasoning and Generation". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 5043–5053.

[24] D. Schlechtweg, B. McGillivray, S. Hengchen, H. Dubossarsky, and N. Tahmasebi. "SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection". In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, 2020, pp. 1–23.

[25] M. Sims, J. H. Park, and D. Bamman. “Literary Event Detection”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 3623–3634.

[26] D. van Strien, K. Beelen, M. C. Ardanuy, K. Hosseini, B. McGillivray, and G. Colavizza. “Assessing the Impact of OCR Quality on Downstream NLP Tasks.” In: *Proceedings of the 12th International Conference on Agents and Artificial Intelligence, 1*. 2020, pp. 484–496.

[27] O. Suissa, A. Elmalech, and M. Zhitomirsky-Geffet. “Text analysis using deep neural networks in digital humanities and information science”. In: *Journal of the Association for Information Science and Technology* (2021).

[28] K. Todorova and G. Colavizzaa. “Transfer Learning for Historical Corpora: An Assessment on Post-OCR Correction and Named Entity Recognition”. In: *Proceedings of the Workshop on Computational Humanities Research*. 2020, pp. 310–339.

[29] M. C. Traub, J. Van Ossenbruggen, and L. Hardman. “Impact analysis of OCR quality on research tasks in digital archives”. In: *Proceedings of International Conference on Theory and Practice of Digital Libraries*. Springer. 2015, pp. 252–263.

[30] T. Underwood. *Do humanists need BERT?* 2019. URL: https://tedunderwood.com/category/methodology/genre-comparison/.

[31] J. Worsham and J. Kalita. “Genre Identification and the Compositional Effect of Genre in Literature”. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA, 2018, pp. 1963–1973.

[32] J. M. Worsham. *Towards literary genre identification: Applied neural networks for large text classification*. University of Colorado Colorado Springs, 2018.

[33] I. Z. Yalniz and R. Manmatha. “A fast alignment scheme for automatic OCR evaluation of books”. In: *Proceedings of 2011 International Conference on Document Analysis and Recognition*. Ieee. 2011, pp. 754–758.

[34] B. Yu. “An evaluation of text classification methods for literary study”. In: *Literary and Linguistic Computing* 23.3 (2008), pp. 327–343.