On Evaluation of Document Classification with RVL-CDIP

Stefan Larson[1]* **Gordon Lim**[2] **Kevin Leach**[1]

[1]Vanderbilt University
Nashville, TN, USA

[2]University of Michigan
Ann Arbor, MI, USA

Abstract

The RVL-CDIP benchmark is widely used for measuring performance on the task of document classification. Despite its widespread use, we reveal several undesirable characteristics of the RVL-CDIP benchmark. These include (1) substantial amounts of label noise, which we estimate to be 8.1% (ranging between 1.6% to 16.9% per document category); (2) presence of many ambiguous or multi-label documents; (3) a large overlap between test and train splits, which can inflate model performance metrics; and (4) presence of sensitive personally-identifiable information like US Social Security numbers (SSNs). We argue that there is a risk in using RVL-CDIP for benchmarking document classifiers, as its limited scope, presence of errors (state-of-the-art models now achieve accuracy error rates that are within our estimated label error rate), and lack of diversity make it less than ideal for benchmarking. We further advocate for the creation of a new document classification benchmark, and provide recommendations for what characteristics such a resource should include.

1 Introduction

Within the document understanding research area, the RVL-CDIP dataset (Harley et al., 2015) has emerged as the primary benchmark for evaluating and comparing document classifiers. RVL-CDIP is composed of 16 document type categories, including `resume`, `letter`, `invoice`, etc. Its large volume of training data—320,000 samples—facilitates benchmarking state-of-the-art deep learning and transformer-based architectures. While initially released as a computer vision benchmark in 2015, more recent state-of-the-art models now incorporate image, text, and page layout modalities. For instance, recent tri-modal models like DocFormer (Appalaraju et al., 2021), ERNIE-Layout (Peng et al., 2022), LayoutLMv3 (Huang et al., 2022), and Bi-VLDoc (Luo et al., 2022) now achieve classification accuracies ranging in the mid- to high-90s, with Bi-VLDoc reporting a state-of-the-art of 97.12% on the RVL-CDIP test set. This is a large improvement over earlier image-centric work, and we chart this improvement in Figure 1.

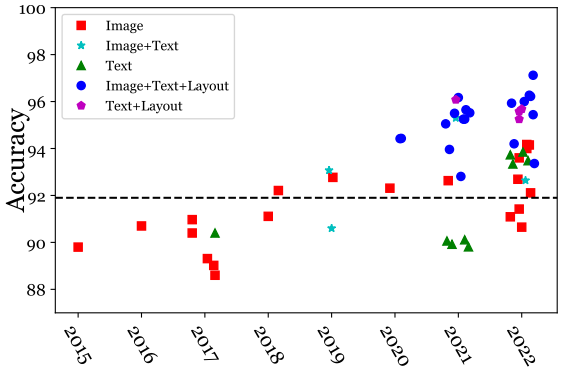


Figure 1: Model accuracy on RVL-CDIP by year and modality. The horizontal dashed line represents our estimated label error rate for RVL-CDIP's test set.

As model performance on RVL-CDIP improves, it becomes increasingly important to ensure that further gains are meaningful with respect to the classification task. This concern has been raised by prior work that has found that benchmark evaluation datasets often contain substantial amounts of label errors or noise (e.g., Northcutt et al., 2021a), substantial overlap between test and train data (e.g., Elangovan et al., 2021; Søgaard et al., 2021), and data collection artifacts that cause models to overfit to spurious cues (e.g., Gururangan et al., 2018; McCoy et al., 2019). Therefore, we cast a critical eye to the RVL-CDIP benchmark to answer: *Is RVL-CDIP still suitable for effectively measuring the performance of document classifiers?*

In doing so, we first observe a lack of clear label or annotation guidelines provided with the original introduction of RVL-CDIP. Therefore, we create verifiable label guidelines for the 16 RVL-CDIP categories. With these guidelines, we are then able to conduct a review of the data, and we find that

* Corresponding email: stefan.dataset@gmail.com

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2665–2678

label errors account for an estimated 8.1% of data RVL-CDIP's test split, a rate greater than the current state-of-the-art model accuracy error rate, indicating that contemporary high-performing models are overfitting to noise. We also observe relatively high rates of documents that have ambiguous or multiple valid labels, which is problematic given RVL-CDIP is a single-label classification benchmark. Additionally, we also observe a large overlap between test and train data splits, where there are (near-) duplicate documents seen in both train and test splits, as well as documents that share common templates. Lastly, our review of RVL-CDIP data uncovered a surprisingly large amount of sensitive personally-identifiable information, particularly in the `resume` category, where we found 7.7% of documents contained US Social Security numbers.

We argue that the characteristics that we observe make RVL-CDIP an unattractive benchmark for training and evaluating document classifiers. We end with recommendations for what qualities a new document classification benchmark should have.

2 Related Work

This section discusses related work in two areas: (1) prior work in document classification on RVL-CDIP, and (2) prior work on analyzing datasets.

2.1 RVL-CDIP and Document Classification

The RVL-CDIP corpus has been used as a benchmark for document classification since its introduction by Harley et al. (2015), who used it to evaluate convolutional neural network (CNN) image classifiers on the dataset's document images. Most immediate follow-up work followed Harley et al. (2015) and explored different image-based CNN models, as done in Csurka et al. (2016); Afzal et al. (2017); Tensmeyer and Martinez (2017); Das et al. (2018); Ferrando et al. (2020). Just relying on image features is limited, as much of a document's "essence" is informed by its textual content. Therefore, more recent work has incorporated the textual modality, including Audebert et al. (2019) and Dauphinee et al. (2019).

Even more recent work has capitalized on the transformer model architecture, often combining vision transformers with large transformer-based language models (and often combining these with a third modality based on page layout derived from detected optical character recognition (OCR) regions) as in LayoutLMv1 (Xu et al., 2020), LayoutLMv2 (Xu et al., 2021), LayoutLMv3 (Huang et al., 2022), DocFormer (Appalaraju et al., 2021), TILT (Powalski et al., 2021), and ERNIE-layout (Peng et al., 2022). These more recent transformer-based models have achieved state-of-the-art accuracy scores on RVL-CDIP, the most recent being Luo et al. (2022)'s Bi-VLDoc, which achieves a reported accuracy of 97.12% on RVL-CDIP. (For a listing of models benchmarked on RVL-CDIP since Harley et al. (2015), see Table 1.)

Model (Reported by)	Modality	Accuracy
Bi-VLDoc (Luo et al., 2022)	I, T, L	97.12
ERNIE-Layout-large (Peng et al., 2022)	I, T, L	96.27
UDOP-Dual (Tang et al., 2022)	I, T, L	96.22
DocFormer-base (Appalaraju et al., 2021)	I, T, L	96.17
StructuralLM-large (Li et al., 2021a)	T, L	96.08
UDOP (Tang et al., 2022)	I, T, L	96.00
LayoutLMv3-large (Huang et al., 2022)	I, T, L	95.93
LiLT-base (Wang et al., 2022a)	T, L	95.68
LayoutLMv2-large (Xu et al., 2021)	I, T, L	95.65
BROS-base (Wang et al., 2022a)	T, L	95.58
TILT-large (Powalski et al., 2021)	I, T, L	95.52
DocFormer-large (Appalaraju et al., 2021)	I, T, L	95.50
LayoutLMv3-base (Huang et al., 2022)	I, T, L	95.44
Donut (Kim et al., 2021)	I, T	95.30
Wukong-Reader-large (Bai et al., 2022)	I, T, L	95.26
Pham et al. (2022)	T, L	95.25
TILT-base (Powalski et al., 2021)	I, T, L	95.25
LayoutLMv2-base (Xu et al., 2021)	I, T, L	95.25
UDoc-star (Gu et al., 2021)	I, T, L	95.05
Wukong-Reader-base (Bai et al., 2022)	I, T, L	94.91
LayoutLMv1-base (Xu et al., 2020)	I, T, L	94.43
LayoutLMv1-large (Xu et al., 2020)	I, T, L	94.42
MATrIX (Delteil et al., 2022)	I, T, L	94.20
DocXClassifier-xl (Saifullah et al., 2022a)	I	94.17
DocXClassifier-large (Saifullah et al., 2022a)	I	94.15
DocXClassifier-base (Saifullah et al., 2022a)	I	94.00
UDoc (Gu et al., 2021)	I, T, L	93.96
Longformer-base (Pham et al., 2022)	T	93.85
Longformer-large (Pham et al., 2022)	T	93.73
MGDoc (Wang et al., 2022b)	I, T, L	93.64
Dessurt (Davis et al., 2022)	I	93.60
Bigbird-base (Pham et al., 2022)	T	93.48
Pramanik et al. (2022)	I, T, L	93.36
Bigbird-large (Pham et al., 2022)	T	93.34
Multimodal Ensemble (Dauphinee et al., 2019)	I, T	93.07
SelfDoc (Li et al., 2021b)	I, T, L	92.81
LadderNet (Sarkhel and Nandi, 2019)	I	92.77
Zingaro et al. (2021)	I, T	92.70
DiT-large (Li et al., 2022)	I	92.69
VLCDoC (Bakkali et al., 2022)	I, T	92.64
InceptionResNetV2 (Xu et al., 2021)	I	92.63
EfficientNet (Ferrando et al., 2020)	I	92.31
Region Ensemble (Das et al., 2018)	I	92.21
DiT-base (Li et al., 2022)	I	92.11
MAE-base (Li et al., 2022)	I	91.42
Stacked CNN Single (Das et al., 2018)	I	91.11
BEiT-base (Li et al., 2022)	I	91.09
VGG-16 (Afzal et al., 2017)	I	90.97
Csurka et al. (2016)	I	90.70
ResNext-101 (Li et al., 2022)	I	90.65
Audebert et al. (2019)	I, T	90.60
ResNet-50 (Afzal et al., 2017)	I	90.40
RoBERTa-large (Li et al., 2021a)	T	90.11
RoBERTa-base (Li et al., 2021a)	T	90.06
BERT-large (Li et al., 2021a)	T	89.92
BERT-base (Li et al., 2021a)	T	89.81
Tensmeyer and Martinez (2017)	I	89.31
GoogLeNet (Afzal et al., 2017)	I	89.02
AlexNet (Afzal et al., 2017)	I	88.60

Table 1: Model accuracy on RVL-CDIP for various image (I), text (T), and layout-based (L) document classification models, ordered by reported score. Models incorporating multiple modalities typically outperform uni-modal models.

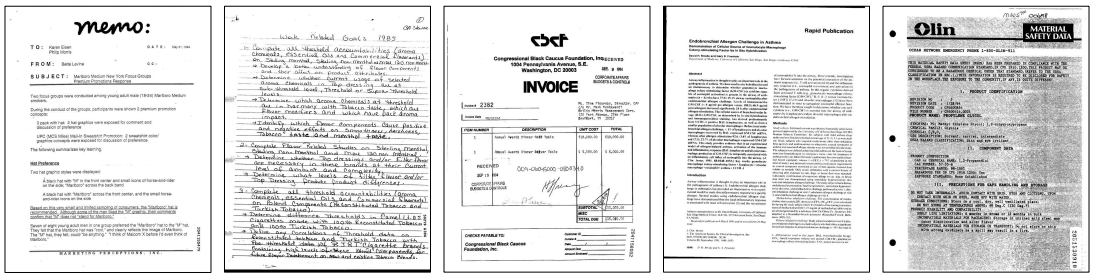


Figure 2: Samples from the RVL-CDIP dataset.

advertisement	memo
budget	news_article
email	presentation
file_folder	questionnaire
form	resume
handwritten	scientific_publication
invoice	scientific_report
letter	specification

Table 2: RVL-CDIP document type categories.

Despite these high scores, recent work has exposed gaps in models trained on RVL-CDIP. In particular, recent work has found that models trained on RVL-CDIP perform poorly on out-of-distribution data (Larson et al., 2022) and perturbed in-distribution data (Saifullah et al., 2022b). We also mention that RVL-CDIP is often used as a pre-training dataset, where models are first pre-trained on RVL-CDIP (and perhaps others) and then evaluated on other downstream tasks or datasets (e.g., Nguyen et al. (2021); Kanchi et al. (2022)). Other datasets like FUNSD are subsets of RVL-CDIP (Jaume et al., 2019).

2.2 Analysis of Datasets

Prior work has investigated the presence of label and annotation errors and corpus quality in NLP and image datasets. This work includes Abedjan et al. (2016); Radenović et al. (2018); Müller and Markert (2019); Pleiss et al. (2020); Northcutt et al. (2021a); Kreutzer et al. (2022); Ying and Thomas (2022); Chong et al. (2022). One common conclusion is that the utility of a benchmark evaluation dataset is lessened if the label error and/or ambiguity rate is close to- or exceeds model prediction error rate. This has been observed for various datasets, such as ATIS (Béchet and Raymond, 2018; Niu and Penn, 2019), and the CNN/Daily Mail reading comprehension task (Chen et al., 2016).

Orthogonal to label errors, prior work has also observed non-trivial overlap between test and train splits in datasets on which natural language processing and computer vision models are evaluated (e.g., Finegan-Dollak et al., 2018; Allamanis, 2019; Barz and Denzler, 2020; Lewis et al., 2021; Wen et al., 2022; Croft et al., 2023). Such work often argues that non-trivial amounts of overlap between test and train data can lead to "inflated" performance scores, as overlapping data can reward a model's ability to memorize training data (Elangovan et al., 2021), and to under-estimate out-of-sample error (Søgaard et al., 2021). Evidence of this can also be found in the multitude of studies that report lower model performance scores on newly-collected evaluation sets versus reported scores on benchmarks (e.g., Augenstein et al., 2017; Recht et al., 2019; Harrigian et al., 2020; Kim and Kang, 2022; Larson et al., 2022). In this paper, we investigate the presence of errors, ambiguous data, and overlapping test-train data for the RVL-CDIP benchmark dataset.

3 The RVL-CDIP Dataset

The RVL-CDIP dataset was introduced in Harley et al. (2015) as a benchmark for evaluating image-based classification and retrieval tasks.[1] Since then, RVL-CDIP has primarily been used as a document type classification benchmark. RVL-CDIP consists of 400,000 document images distributed across 16 document type categories, listed in Table 2. Example documents from RVL-CDIP are shown in Figure 2. Documents in RVL-CDIP were sampled from the larger IIT-CDIP Test Collection, which itself is a snapshot of the voluminous Legacy Tobacco Documents Library (LTDL) collection — at that time, LDTL contained approximately 7 million documents (Lewis et al., 2006).[2] These documents were made publicly available as part of legal pro-

[1] Harley et al. (2015) referred to RVL-CDIP as *BigTobacco*.

[2] The LTDL is now called the Truth Tobacco Industry Documents collection, and is included in the broader Industry Documents Library (IDL) hosted by the UCSF Library: https://www.industrydocuments.ucsf.edu/. For more background on the LTDL, see Schmidt et al. (2002) and Tasker et al. (2022).

Category	Description
advertisement	Advertisements from print-form media like newspapers and magazines. Also a small amount of scripts for television or radio advertisements. A small amount of "ad order instructions" and "ad insertion" documents.
budget	Includes various budget documents such as expense, spending, sales, cash, and accounting reports and forecasts; budgets; quotes and estimates; and income and bank statements. Also includes receipt-like documents such as political campaign contribution requests and other receipts, as well as checks and check stubs.
email	Scanned images of printed emails.
file_folder	Scanned images of folders and binders. Folder scans are often characterized by vertically oriented text (indicating a folder label). A moderate amount of file folders in RVL-CDIP contain handwritten text or notes. Some scanned folders may be indistinguishable from blank pages.
form	Form documents with form-like elements (e.g., lines or spaces for user-provided data entry). The form-like elements can appear empty or filled.
handwritten	Includes handwritten documents like handwritten letters and scientific notes.
invoice	Includes invoices, bills, and account statements.
letter	Letters, often with letterhead and commonly with "Dear..." salutations. The distinction between letters and memos is often unclear in RVL-CDIP.
memo	Memoranda or inter-office correspondence documents, often with clear "TO", "FROM", "SUBJECT" headings.
news_article	Includes news articles in the form of clippings from newspapers and other print-form news media, as well as a small amount of news articles from the web.
presentation	Includes scanned images of presentation and overhead slides, transcripts of speeches and statements. Also includes a large amount of press releases.
questionnaire	Includes customer surveys and questionnaires, as well as survey and questionnaire prompts for surveyors. Also includes questionnaires appearing to be part of legal proceedings and investigations. In RVL-CIDP, many questionnaires have a substantial amount of form-like elements.
resume	Includes resumes, curricula vitae (CVs), biographical sketches, executive biographies (e.g., those written in third-person), a small amount of business cards.
scientific_pub.	Mainly papers and articles from scientific journals and book chapters, but also includes book title pages. Also includes news articles from science newsletters. News articles from science newsletters are very similar to the news_article category.
scientific_rep.	Includes bioassay, pathology, and test reports; charts, graphs, and tables; research reports (including progress reports), research proposals, abstracts, paper drafts. Many reports and abstracts bear similarities to scientific publications. Many test result documents are similar to documents in the specification category.
specification	Data sheets (including safety data sheets); product, material, and test specifications. Also includes specification change reports.

Table 3: RVL-CDIP categories alongside our descriptions and notes.

ceedings and settlements against several American tobacco and cigarette companies and organizations, and as such, the documents in RVL-CDIP are almost exclusively related to the tobacco industry.[3]

Most document images in RVL-CDIP capture the initial page of a document; some common exceptions appear to be charts and tables (these are typically labeled as scientific_report) as well as presentation slides (labeled as presentation). Additionally, almost all of the documents (that contain readable text) are in English, although we did find small amounts of documents in other languages (including German, Dutch, French, Spanish, Portuguese, Italian, Japanese, Chinese, Arabic, and Hebrew) as part of our review. Examples of non-English RVL-CDIP samples are displayed in Figure 11 in the Appendix.

There are 320,000 training, 40,000 validation, and 40,000 test samples, but Harley et al. (2015) provides no information on how the data was partitioned into these splits, so we assume it was done randomly for each of the 16 document categories. Harley et al. (2015) report that the 16 categories were chosen, in part, because these categories had ample representation (i.e., at least 25,000 samples) in IIT-CDIP. Unfortunately, we are unaware of any published guidelines, criteria, rules, or documentation defining or describing each of the 16 RVL-CDIP categories, nor is it clear who or what provided the initial category labels in IIT-CDIP (nor in LTDL).[4] Thus, we describe how we developed label guidelines for each RVL-CDIP document type category in Section 3.1 below.

[3]For more background on the history of the litigation and documents, see Glantz et al. (1996); Ciresi et al. (1999); Tasker et al. (2022).

[4]Schmidt et al. (2002) and Tasker et al. (2022) indicate that type labels may have been ascribed to the documents by human workers employed at UCSF's LTDL.

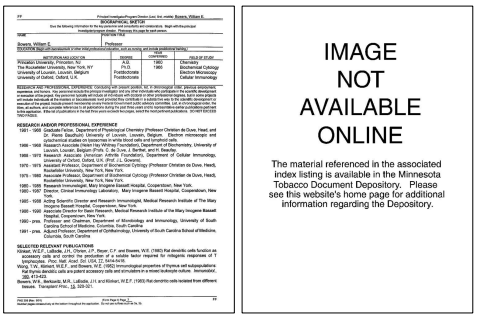
3.1 Establishing Label Guidelines

The RVL-CDIP dataset does not have a published list of descriptions, rules, or guidelines describing each of the 16 document type categories. We discuss an extensive analysis from which we develop such guidelines.

We established our list of guidelines by first sampling 1,000 documents from each of the 16 categories in the training set (for a total of 16,000 documents). We then reviewed these samples category by category. This review process helped us identify commonalities within each category, and helped us discover that many of the categories seem to have distinct groups of sub-types within them. For instance, we found that the `resume` category is largely composed of (1) resumes and curricula vitae, (2) "Biographical Sketch" documents (i.e., those required for grant applications for the National Institutes of Health (example shown in Figure 3a), (3) executive biographies, and (4) scanned business cards. Such cases reveal opportunities for refining and diversifying appropriate categories.

In another category, `advertisement`, we found samples mostly consisted of advertisements from print-form media like newspapers and magazines, as well as smaller amounts of scripts for television or radio advertisements. The `advertisement` category also included a small amount of document images identical to the one shown in Figure 3b. We found that this "IMAGE NOT AVAILABLE" document appears mostly in the `advertisement` category, yet it is an example of a document that we do not include in our label guidelines for this category, as it is not at all faithful to the semantic nature of the `advertisement` category.

Our annotation guidelines are listed in Table 3, along with our notes and observations. It was occasionally necessary to review multiple document categories prior to establishing rules. This was the case with the `budget` and `invoice` categories, each of which included non-trivial amounts of scanned check images and contribution requests. (Examples of cases like these are displayed in Figures 21-23 in the Appendix.) For cases like these, we annotated these sub-types in the relevant categories in order to estimate their relative frequencies. We then would append our annotation guidelines accordingly; for instance, 8.8% of `budget` documents and 3.8% of `invoice` documents that we reviewed were check images, so our guidelines specify that the `budget` category consists of check images, while `invoice` does not. Ultimately, our goal with establishing such guidelines is to provide repeatable, verifiable criteria that faithfully reflect the semantic nature of each category.



(a) (b)

Figure 3: Example "Biographical Sketch" `resume` document (a) and "IMAGE NOT AVAILABLE" document found mostly in `advertisement`.

4 Label Errors and Ambiguities in RVL-CDIP

Armed with better knowledge of what constitutes each of the 16 RVL-CDIP categories, we analyze the contents of the RVL-CDIP test set to estimate the amount of label errors and ambiguities found in this set.

We manually checked for errors in the RVL-CDIP test set by sampling 1,000 documents from each of the 16 categories (for a total of 16,000 documents). We used our label guidelines established in Section 3.1 to help us determine the validity of each of these 16,000 samples. We tracked several types of errors and ambiguities: (1) documents found in a category that clearly are mis-labeled and instead belong in a different RVL-CDIP category — we refer to this error type as *mis-labeled*; (2) documents that do not appear to have a single clear RVL-CDIP label — we refer to this label type as *unknown*; (3) documents that have mixed or multiple features that belong to at least two RVL-CDIP categories — we refer to this type as *mixed*. Examples of documents exhibiting these error types can be seen in Figure 4. We point out a particularly interesting *mixed* case: the first two *mixed* examples are nearly identical, but the original label is `news_article` in one case but `letter` in the second. More examples are shown in the Appendix in Figures 12–14.

Findings. Our estimated error rates in the RVL-CDIP test set are shown in Table 4. We estimate that error rates (i.e., combined rates for *mis-labeled* and *unknown*) range between 1.6% (in the case

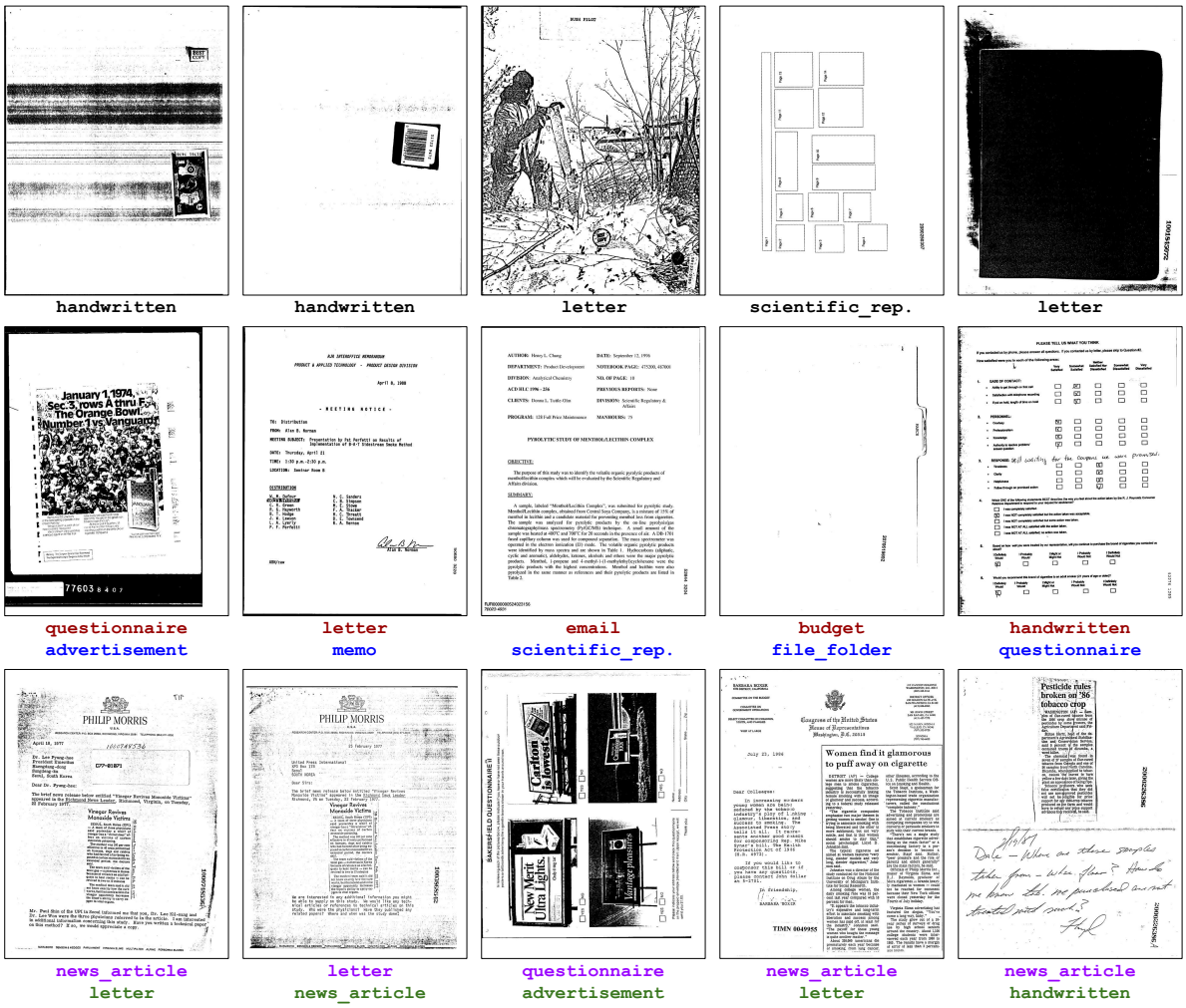


Figure 4: Example errors and ambiguities. Top row: *unknown*, middle row: *mis-label*, bottom row: *mixed*.

Category	*mis-labeled*	*unknown*	*Tot. Error*	*mixed*
advertisement	1.9%	4.5%	6.4%	3.5%
budget	9.7%	4.1%	13.8%	1.5%
email	1.7%	8.3%	10.0%	0.4%
form	4.4%	6.4%	10.8%	0.5%
file_folder	0.4%	3.1%	3.5%	1.9%
handwritten	2.5%	5.2%	7.7%	2.4%
invoice	9.7%	1.3%	11.0%	0.2%
letter	13.5%	3.4%	16.9%	0.5%
memo	2.0%	2.4%	4.4%	2.1%
news_article	4.6%	2.5%	7.1%	0.4%
presentation	1.8%	4.9%	6.7%	1.0%
questionnaire	5.6%	7.3%	12.9%	6.9%
resume	0.2%	1.4%	1.6%	0.4%
scientific_pub.	2.5%	1.8%	4.3%	0.0%
scientific_rep.	4.6%	3.9%	8.5%	5.6%
specification	1.5%	1.9%	3.4%	0.4%
Average	4.2%	3.9%	8.1%	1.7%

Table 4: Estimated label error and multi-label rates in the RVL-CDIP test set.

of resume) and 16.9% (in the case of letter). The average of each category's error rates is 8.1%, which is higher than the classification accuracy error rates reported by many state-of-the-art models listed in Table 1. In some cases, the majority of a category's errors were mis-labels of a particular type. For instance, about 59% of the erroneous letter documents we reviewed were actually memo documents. Similarly, 74% of the erroneous invoice documents were actually budget documents. Lastly, roughly 1.7% of RVL-CDIP's test set is data that have multiple valid labels.

5 Overlap Between Test and Train Splits

Our analysis also reveals a substantial degree of undesirable overlap between train and test samples within RVL-CDIP. To measure this overlap, we use an approach similar to Larson et al. (2019) and Elangovan et al. (2021), which, for each test sample in each document type category, finds the maximally similar sample in the same document

Figure 5: Example test-train pairs with corresponding maximum cosine similarity scores. These three example pairs show instances of near-duplicates (left and center) and documents that have highly similar structure (right).

Category	mean	median
advertisement	0.893	0.903
budget	0.963	0.968
email	0.976	0.982
form	0.948	0.956
file_folder	0.967	0.974
handwritten	0.945	0.952
invoice	0.962	0.966
letter	0.953	0.960
memo	0.957	0.961
news_article	0.919	0.936
presentation	0.929	0.945
questionnaire	0.961	0.968
resume	0.965	0.967
scientific_pub.	0.936	0.955
scientific_rep.	0.950	0.961
specification	0.972	0.978
Average	0.950	0.958

Table 5: Mean and median of maximum cosine similarity scores between train and test sets for each RVL-CDIP category.

type category's training split. We then average these maximum similarity scores together for each document category. That is, for each document category C in RVL-CDIP, we compute

$$\frac{1}{|test_C|} \sum_{b \in test_C} \max_{a \in train_C} sim(a, b)$$

where a and b are samples from category C's train and test splits, respectively. We use CLIP (Radford et al., 2021) to extract a 512-dimension feature embedding from each sample, and use cosine similarity for $sim(\cdot, \cdot)$. We note that this vector-based similarity technique is common practice in the image- and information retrieval (e.g., Babenko et al. (2014)).

Findings. Average and median of the maximum similarity scores for test-train pairs are shown in Table 5 for each RVL-CDIP category. Overall, we see a high degree of similarity across test and train data: mean scores range between 0.893 (`advertisement`) and 0.976 (`email`), with an average of 0.950. Ten of the 16 document categories

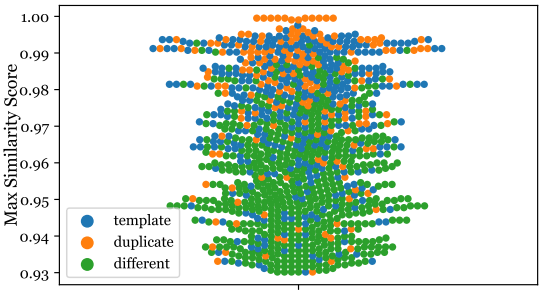


Figure 6: Sampled subset of maximal similarity scores for test-train pairs with scores between 0.93 and 1.0.

have average scores at- or above 0.95. The median score for each category is larger than the mean in all cases, indicating a long tail in the distribution of scores. Indeed, we see this in Figure 10 (in Appendix), which charts the distribution of similarity scores for all test data in RVL-CDIP. Figure 5 shows three examples of test-train pairs with similarity scores ranging between 0.937 and 0.981. Two of the three pairs in Figure 5 seem to be near-duplicates, where there appear to be minor differences in scanning or noise artifacts between each document. In the third (`invoice`) example, we see that the two samples are distinct, yet both share a large degree of similarity because both use the same document template (e.g., invoices from the same company that are structurally and visually similar but that contain different "data"). We show more example pairs in Figures 15–18 in the Appendix.

To help better understand the similarity scores, we conduct an experiment where we categorize each similarity pair into one of the following: *duplicate*, if the test-train pair represents the same document; *template*, if both documents in a pair use the same document template; and *different*, for all other pairs. We annotated a sample of 1,086 similarity pairs with maximum similarity scores ranging between 0.93 and 1.0. A visualization of the relationship between maximal similarity score and match type is shown in Figure 6, where we

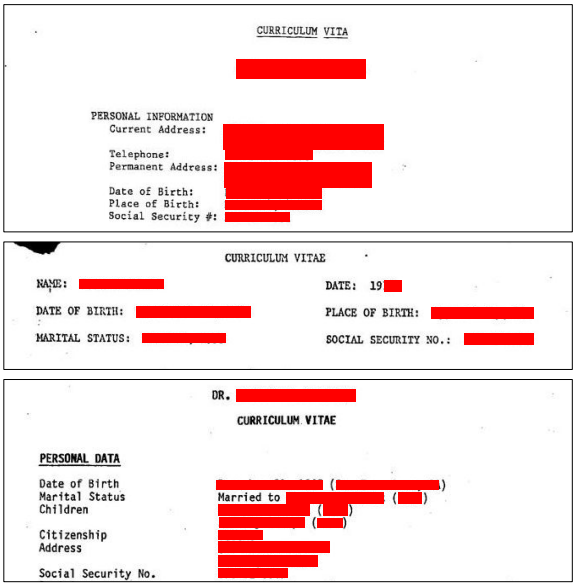


Figure 7: Example documents from RVL-CDIP showing sensitive personally identifiable information (PII; redacted by us).

observe that the likelihood of a pair being either a *duplicate* or *template* match increases with similarity score.

Considering the overall median maximal similarity score is 0.958, we can estimate a lower-bound for the rate of *duplicate* and *template* match pairs by scaling the proportion of documents above the median maximal score (i.e., half, or 0.5) by the fraction of *duplicate* and *template* matches above the median (0.958). This gives us 0.5×0.641, and therefore we estimate that at least 32% of samples from the RVL-CDIP test set have either a duplicate counterpart or a sample that shares a template layout in the training set. While there is generally no established acceptable number or percentage for test-train overlaps, prior work (e.g., Søgaard et al. (2021); Elangovan et al. (2021)) has argued that overlaps are undesirable, and that building generalizable, robust models entails evaluation against novel, unseen data points (e.g., Koh et al. (2021); Malinin et al. (2021); Larson et al. (2022)).

6 Presence of Sensitive Information

While reviewing samples from RVL-CDIP, we noticed that the `resume` category had a non-trivial quantity of documents that contain sensitive and personally-identifiable entities. Naturally, resumes typically contain a person's name and basic contact information (e.g., phone numbers or email addresses). However, we found a plethora of sensitive entities like citizenship and marital statuses, places and dates of birth, names of children and spouses, and national ID numbers like US Social Security and Canadian National ID numbers.

Out of a sample of 1,000 documents from the `resume` test set, we found that 7.7% contained a US Social Security Number. While we recognize that US Social Security numbers were not considered sensitive several decades ago (when many of the `resume` documents in RVL-CDIP were created), their presence in so many documents in a publicly accessible dataset[5] is still striking, especially considering the coexistence of this entity type with others like person names, dates and places of birth, etc. In particular, malicious Social Security numbers are often connected with fraud and identity theft crimes in the USA. Moreover, the sensitive entities discussed in this section are considered highly sensitive under many state and national laws.[6] Additionally, we found that 43.6% of the test resumes contain birth dates, 19.9% contain places of birth, 11.4% contain marital (or spousal or parental) statuses, and 8.9% contain citizenship statuses. Example documents containing sensitive PII can be seen in Figure 7.

Given the presence of sensitive PII in RVL-CDIP, it is reasonable to wonder if sensitive PII also appears in datasets derived from RVL-CDIP, like FUNSD (Jaume et al., 2019). Similarly, we also wonder if sensitive PII appears in datasets that were derived from the larger IIT-CDIP or UCSF Industry Documents Library corpora, such as Tobacco-800 (Zhu et al., 2007; Zhu and Doermann, 2007), Tobacco-3482 (Kumar et al., 2014), DocVQA (Mathew et al., 2021), and OCR-IDL (Biten et al., 2022). We will investigate this in future work.

7 Discussion and Recommendations

Given our findings concerning labeling errors, test/train overlap, and presence of sensitive information in the RVL-CDIP document classification benchmark, we discuss several concrete recommendations to raise awareness among researchers engaged in benchmarking classifiers using this dataset:

(0) Sub-Types in RVL-CDIP. Our investigation into RVL-CDIP revealed that many of the RVL-CDIP categories are in fact composed of several

[5] On 9 Feb. 2023, RVL-CDIP tallied "1,765 downloads last month" on the Hugging Face Datasets platform.

[6] For example, the State of Michigan's Social Security Number Privacy Act (2004).

sub-types. We encourage researchers and practitioners to be aware of this fact. For instance, curricula vitae, bigraphical sketches, executive biographies, and business cards are the four sub-types of the `resume` category. This finding has implications for modeling tasks where prior knowledge of the label set is assumed, like in zero-shot settings where each category may be specified to the model as a string, as done in Siddiqui et al. (2021). Additionally, unsupervised clustering analyses like Finegan-Dollak and Verma (2020) may exhibit low performance scores on RVL-CDIP due to many of the categories having distinct and disparate subtypes (e.g., radio scripts versus print advertisements in the `advertisement` category, or business cards versus biographical sketches in the `resume` category).

(1) Errors. Users of RVL-CDIP should be aware that there are many label errors and noisy samples with unknown labels in RVL-CDIP. Recall from Section 4 that an estimated 8.1% of test samples from RVL-CDIP contain label errors, with an additional 1.7% being ambiguous mixed or multi-label cases. This is problematic for benchmarking new models, since the estimated label error rate is now greater than state-of-the-art model accuracy error rates. Here, the implication is that high-capacity models like CNNs and transformers are now overfitting to noise. This is indeed the case for models like DiT (Li et al., 2022), which predict the "IMAGE NOT AVAILABLE" document to be an `advertisement` document due to its relative abundance in that category's training set.

(2) Ambiguities. Users of RVL-CDIP should be aware that there are many samples in RVL-CDIP that could have multiple valid document type labels. We estimate this number to be 1.7% of the RVL-CDIP test set. Like label errors, such mixed or multi-label cases make it challenging to evaluate a model effectively, as there are samples for which a model may make a wrong prediction according to the RVL-CDIP test label annotations, but in reality many of these wrong predictions could actually be reasonable.

(3) Test-Train Overlap. Practitioners and researchers should be aware that there is a high degree of overlap between the RVL-CDIP test set and the train set. Recall from Section 5 that almost a third of RVL-CDIP test samples have a near-duplicate in the training set for the same document type category, or a training sample that uses the same document template. This is undesirable, as testing models on data that is very similar to the training data can lead to "inflated" accuracy scores (Elangovan et al., 2021; Søgaard et al., 2021). Moreover, highly similar train and test splits do not facilitate the evaluation of a model's ability to generalize well to new in-domain data.

(4) Sensitive Information. There is an unsettling amount of sensitive information in the RVL-CDIP dataset, which naturally leads to information and data privacy concerns. We estimate that 7.7% of `resume` test samples contain Social Security numbers. While RVL-CDIP is already publicly available, researchers and practitioners should take care when disseminating samples or copies of RVL-CDIP. Moreover, we highlight that prior work (e.g., Carlini et al. (2021)) showed that it is possible to extract training data from machine learning models, making production deployments of models trained on RVL-CDIP an information privacy and security risk.

Suggestions for a future dataset. We suggest the development and adoption of a new benchmark for evaluating document classifiers. Several qualities of a such a benchmark would include (1) minimal label errors; (2) multi-label annotations, to allow for modeling more natural occurrences of documents; (3) minimal test-train overlap; (4) absence of sensitive information. Going beyond the points made in this paper, a new benchmark would do well to be (5) large-scale, consisting of 100+ or even 250+ document categories, to test a model's ability to handle breadth, and (6) multi-lingual, to benchmark language transfer approaches.

8 Conclusion

RVL-CDIP has been used as the *de facto* benchmark for evaluating state-of-the-art document classification models, but this paper provides an in-depth analysis of the RVL-CDIP dataset and shows that there are several undesirable characteristics of this dataset. We first provide a set of label guidelines for each RVL-CDIP category, and we use this to help us quantify the presence of errors in RVL-CDIP, finding that the RVL-CDIP test set contains roughly 8.1% label errors. We then observe that roughly a third of the test data is highly similar to the training set. Lastly we observe an unsettling amount of personally sensitive information in RVL-CDIP. Given these findings, we offer suggestions for a new document classification benchmark.

Limitations

The RVL-CDIP dataset has no official set of label guidelines, making error analyses challenging since we could not rely on pre-defined rules. For this reason we followed best practices to create annotation rules to help us in our error analysis. Detecting duplicates in RVL-CDIP is also challenging, as two documents may appear to be the same, but may have minor differences due to scanning artifacts or even different indexing labels (it appears that many of the documents have been scanned and included in IIT-CDIP more than once). Therefore we again have to rely on best judgement when labeling pairs as duplicates (or near-duplicates). Additionally, due to limitations in human resources, we were unable to exhaustively inspect all 400,000 RVL-CDIP samples for the presence of errors, ambiguities, sensitive information, etc., and thus had to rely on sampling the dataset in order to draw conclusions.

Acknowledgements

We thank Nicole Cornehl Lima, Ramla Alakraa, Zongyi Liu, Junjie Shen, Temi Okotore for help with data review, as well as the University of Michigan's Undergraduate Research Opportunity Program (UROP) for their support of these student researchers as well as support for Gordon. We also thank the anonymous EACL reviewers for their feedback.

References

Ziawasch Abedjan, Xu Chu, Dong Deng, Raul Castro Fernandez, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, Michael Stonebraker, and Nan Tang. 2016. Detecting data errors: Where are we and what needs to be done? *Proc. VLDB Endow.*, 9(12):993–1004.

Muhammad Zeshan Afzal, Andreas Kölsch, Sheraz Ahmed, and Marcus Liwicki. 2017. Cutting the error by half: Investigation of very deep cnn and advanced training strategies for document image classification. In *Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*.

Miltiadis Allamanis. 2019. The adverse effects of code duplication in machine learning models of code. In *Proceedings of the 2019 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*.

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. 2021. DocFormer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Nicolas Audebert, Catherine Herold, Kuider Slimani, and Cédric Vidal. 2019. Multimodal deep networks for text and image-based document classification. In *Proceedings of the Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle (APIA)*.

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83.

Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. 2014. Neural codes for image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Haoli Bai, Zhiguang Liu, Ziaojun Meng, Wentao Li, Shuang Liu, Nian Xie, Rongfu Zheng, Liangwei Wang, Lu Hou, Jiansheng Wei, Xin Jiang, and Qun Liu. 2022. Wukong-Reader: Multi-modal pre-training for fine-grained visual document understanding. *arXiv preprint arXiv:2212.09621*.

Souhail Bakkali, Zuheng Ming, Mickael Coustaty, Marçal Rusiñol, and Oriol Ramos Terrades. 2022. VLCDoC: Vision-language contrastive pre-training model for cross-modal document classification. *arXiv preprint arXiv:2205.12029*.

Björn Barz and Joachim Denzler. 2020. Do we train on test data? Purging CIFAR of near-duplicates. *Journal of Imaging*, 6(6).

Frédéric Béchet and Christian Raymond. 2018. Is ATIS too shallow to go deeper for benchmarking spoken language understanding models? In *Proceedings of InterSpeech*.

Ali Furkan Biten, Rubèn Tito, Lluis Gomez, Ernest Valveny, and Dimosthenis Karatzas. 2022. OCR-IDL: OCR annotations for industry document library dataset. *arXiv preprint arXiv:2202.12985*.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *Proceedings of the 30th USENIX Security Symposium*.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the CNN/Daily Mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Derek Chong, Jenny Hong, and Christopher Manning. 2022. Detecting label errors by using pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Michael V. Ciresi, Roberta B. Walburn, and Tara D. Sutton. 1999. Decades of deceit: Document discovery in the Minnesota tobacco litigation. *William Mitchell Law Review*, 25(2).

Roland Croft, M. Ali Babar, and M. Mehdi Kholoosi. 2023. Data quality for software vulnerability datasets. *arXiv preprint arXiv:2301.05456*.

Gabriela Csurka, Diane Larlus, Albert Gordo, and Jon Almazán. 2016. What is the right way to represent document images? *arXiv preprint arXiv:1603.01076*.

Arindam Das, Saikat Roy, Ujjwal Bhattacharya, and Swapan Kumar Parui. 2018. Document image classification with intra-domain transfer learning and stacked generalization of deep convolutional neural networks. In *Proceedings of the 24th International Conference on Pattern Recognition (ICPR)*.

Tyler Dauphinee, Nikunj Patel, and Mohammad Rashidi. 2019. Modular multimodal architecture for document classification. *arXiv preprint arXiv:1912.04376*.

Brian Davis, Bryan Morse, Bryan Price, Chris Tensmeyer, Curtis Wigington, and Vlad Morariu. 2022. End-to-end document recognition and understanding with Dessurt. *arXiv preprint arXiv:2203.16618v3*.

Thomas Delteil, Edouard Velval, Lei Chen, and Luis Goncalves. 2022. MATrIX - modality-aware transformer for information extraction. *arXiv preprint arXiv:2205.08094v1*.

Aparna Elangovan, Jiayuan He, and Karin Verspoor. 2021. Memorization vs. generalization : Quantifying data leakage in NLP performance evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Javier Ferrando, Juan Luis Domínguez, Jordi Torres, Raúl García, David García, Daniel Garrido, Jordi Cortada, and Mateo Valero. 2020. Improving accuracy and speeding up document image classification through parallel systems. In *Proceedings of the International Conference on Computational Science (ICCS)*.

Catherine Finegan-Dollak, Jonathan K. Kummerfeld, Li Zhang, Karthik Ramanathan, Sesh Sadasivam, Rui Zhang, and Dragomir Radev. 2018. Improving text-to-SQL evaluation methodology. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Catherine Finegan-Dollak and Ashish Verma. 2020. Layout-aware text representations harm clustering documents by type. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*.

Stanton A. Glantz, John Slade, Lisa A. Bero, Peter Hanauer, and Deborah E. Barnes. 1996. *The Cigarette Papers*. University of California Press.

Jiuxiang Gu, Jason Kuen, Vlad I. Morariu, Hangdong Zhao, Nikolaos Barmpalios, Rajiv Jain, Ani Nenkova, and Tong Sun. 2021. Unified pretraining framework for document understanding. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. Do models of mental health based on social media data generalize? In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. LayoutLMv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*.

Shrinidhi Kanchi, Alain Pagani, Hamam Mokayed, Marcus Liwicki, Didier Stricker, and Muhammad Zeshan Afzal. 2022. EmmDocClassifier: Efficient multimodal document image classifier for scarce data. *Applied Sciences*, 12(3).

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2021. OCR-free document understanding transformer. *arXiv preprint arXiv:2111.15664*.

Hyunjae Kim and Jaewoo Kang. 2022. How do your biomedical named entity recognition models generalize to novel entities? *IEEE Access*.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne

David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 2021 International Conference on Machine Learning (ICML)*.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*.

Jayant Kumar, Peng Ye, and David Doermann. 2014. Structural similarity for document image classification and retrieval. *Pattern Recognition Letters*, 43.

Stefan Larson, Gordon Lim, Yutong Ai, David Kuang, and Kevin Leach. 2022. Evaluating out-of-distribution performance on document image classifiers. In *Proceedings of the Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Stefan Larson, Anish Mahendran, Andrew Lee, Jonathan K. Kummerfeld, Parker Hill, Michael A. Laurenzano, Johann Hauswald, Lingjia Tang, and Jason Mars. 2019. Outlier detection for improved data quality and diversity in dialog systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.

D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard. 2006. Building a test collection for complex document information processing. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 2021a. StructuralLM: Structural pre-training for form understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*.

Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. 2022. DiT: Self-supervised pre-training for document image transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*.

Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I. Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. 2021b. SelfDoc: Self-supervised document representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chuwei Luo, Guozhi Tang, Qi Zheng, Cong Yao, Lianwen Jin, Chenliang Li, Yang Xue, and Luo Si. 2022. Bi-VLDoc: Bidirectional vision-language modeling for visually-rich document understanding. *arXiv preprint arXiv:2206.13155*.

Andrey Malinin, Neil Band, Alexander Ganshin, German Chesnokov, Yarin Gal, Mark J. F. Gales, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, Vyas Raina, Denis Roginskiy, Mariya Shmatova, Panos Tigar, and Boris Yangel. 2021. Shifts: A dataset of real distributional shift across multiple large-scale tasks. *arXiv preprint arXiv:2107.07455*.

Minesh Mathew, Dimosthenis Karatzas, and C.W. Jawahar. 2021. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Nicolas M. Müller and Karla Markert. 2019. Identifying mislabeled instances in classification datasets. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*.

Laura Nguyen, Thomas Scialom, Jacopo Staiano, and Benjamin Piwowarski. 2021. Skim-attention: Learning to focus via document layout. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Jingcheng Niu and Gerald Penn. 2019. Rationally reappraising ATIS-based dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Curtis Northcutt, Anish Athalye, and Jonas Mueller. 2021a. Pervasive label errors in test sets destabilize machine learning benchmarks. In *35th Conference on Neural Information Processing Systems Track on Datasets and Benchmarks*.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. 2021b. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research (JAIR)*, 70:1373–1411.

Qiming Peng, Yinxu Pan, Wenjin Wang, Bin Luo, Zhenyu Zhang, Zhengjie Huang, Yuhui Cao, Weichong Yin, Yongfeng Chen, Yin Zhang, Shikun Feng, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2022. ERNIE-layout: Layout knowledge enhanced pre-training for visually-rich document understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2022*.

Hai Pham, Guoxin Wang, Yijuan Lu, Dinei Florencio, and Cha Zhang. 2022. Understanding long documents with different position-aware attentions. *arXiv preprint arXiv:2208.08201*.

Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q. Weinberger. 2020. Identifying mislabeled data using the area under the margin ranking. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*.

Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Pałka. 2021. Going full-TILT boogie on document understanding with text-image-layout transformer. In *Proceedings of the 16th International Conference on Document Analysis and Recognition (ICDAR)*.

Subhojeet Pramanik, Shashank Mujumdar, and Hima Patel. 2022. Towards a multi-modal, multi-task learning based pre-training framework for document representation learning. *arXiv preprint arXiv:2009.14457v2*.

Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. 2018. Revisiting Oxford and Paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning (ICML)*.

Saifullah, Stefan Agne, Andreas Dengel, and Sheraz Ahmed. 2022a. DocXClassifier: High performance explainable deep network for document image classification. *techarxiv preprint*.

Saifullah, Shoaib Ahmed Siddiqui, Stefan Agne, Andreas Dengel, and Sheraz Ahmed. 2022b. Are deep models robust against real distortions? A case study on document image classification. In *Proceedings of the 26th International Conference on Pattern Recognition (ICPR)*.

Ritesh Sarkhel and Arnab Nandi. 2019. Deterministic routing between layout abstractions for multi-scale classification of visually rich documents. In *Proceedings of the Twenty-Eigth International Joint Conference on Artificial Intelligence (IJCAI)*.

Heidi Schmidt, Karen Butter, and Cynthia Rider. 2002. Building digital tobacco industry document libraries at the University of California, San Francisco Library/Center for Knowledge Management. *D-Lib Magazine*, 8(9).

Shoaib Ahmed Siddiqui, Andreas Dengel, and Sheraz Ahmed. 2021. Analyzing the potential of zero-shot recognition for document image classification. In *Proceedings of the 16th International Conference on Document Analysis and Recognition (ICDAR)*.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. 2022. Unifying vision, text, and layout for universal document processing. *arXiv preprint arXiv:2212.02623*.

Kate Tasker, Taketa Rachel, Charles Macquarie, and Ariel Deardorff. 2022. Digital archives and data science: Building programs and partnerships for health sciences research. In *Handbook of Research on Academic Libraries as Partners in Data Science*.

Chris Tensmeyer and Tony Martinez. 2017. Analysis of convolutional neural networks for document image classification. In *Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*.

Jiapeng Wang, Lianwen Jin, and Kai Ding. 2022a. LiLT: A simple yet effective language-independent layout transformer for structured document understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Zilong Wang, Jiuxiang Gu, Chris Tensmeyer, Nikolaos Barmpalios, Ani Nenkova, Tong Sun, Jingbo Shang, and Vlad Morariu. 2022b. MGDoc: Pre-training with multi-granular hierarchy for document image understanding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yuqiao Wen, Guoqing Luo, and Lili Mou. 2022. An empirical study on the overlapping problem of open-domain dialogue datasets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*.

Cecilia Ying and Stephen Thomas. 2022. Label errors in BANKING77. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*.

Guangyu Zhu and David Doermann. 2007. Automatic document logo detection. In *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR)*.

Guangyu Zhu, Yefeng Zheng, David Doermann, and Stefan Jaeger. 2007. Multi-scale structural saliency for signature detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Stefano Pio Zingaro, Giuseppe Lisanti, and Maurizio Gabbrielli. 2021. Multimodal side-tuning for document classification. In *Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR)*.