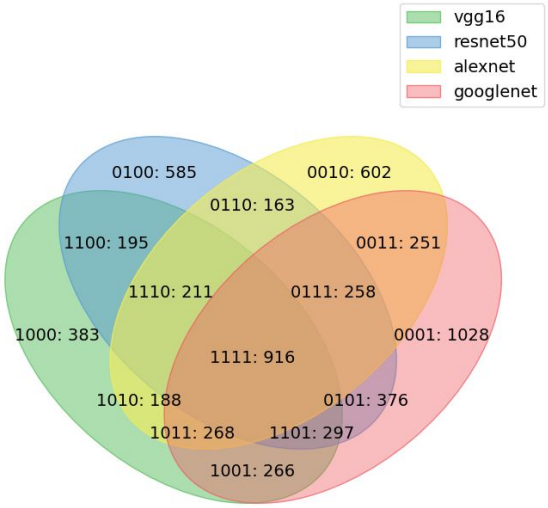


Figure 8: Venn diagram showing overlaps among predicted label errors in RVL-CDIP's test set using Cleanlab with four different classifier models. The labels are one-hot encoded in this figure (e.g., "1010" indicates the intersection between the Cleanlab predictions for VGG-16 and AlexNet).

A Appendix

This appendix is used to provide supplementary material. Appendix A.1 discusses using an automated label error dection tool called Cleanlab, and why we ultimately did not use it to aid us in our review of RVL-CDIP. Appendix A.2 provides supplementary visualizations in support of the main paper. Finally, Appendix A.3 details where to find the label annotations developed in this paper.

A.1 Cleanlab Discussion

Automated tools exist for detecting label errors in classification datsets. One such exemplary tool is Cleanlab, which uses confident learning algorithms to predict label errors in datasets (Northcutt et al., 2021b,a). We used the off-the-shelf version of Cleanlab,[7] which aims to identify label errors in a dataset given a model's predictions on that dataset. That is, Cleanlab uses the original data labels, the model's predicted labels, and the model's confidence scores to make a prediction of *error* or *not-error* for each sample.

We used Cleanlab on the RVL-CDIP test set (using models trained on the full RVL-CDIP training set from Larson et al. (2022)) with four different classifier models: GoogLeNet, AlexNet, ResNet, and VGG-16 (each uses the architecture from Szegedy et al. (2015), Krizhevsky et al. (2012), He et al. (2016), and Simonyan and Zisserman (2015),

[7] https://github.com/cleanlab/cleanlab

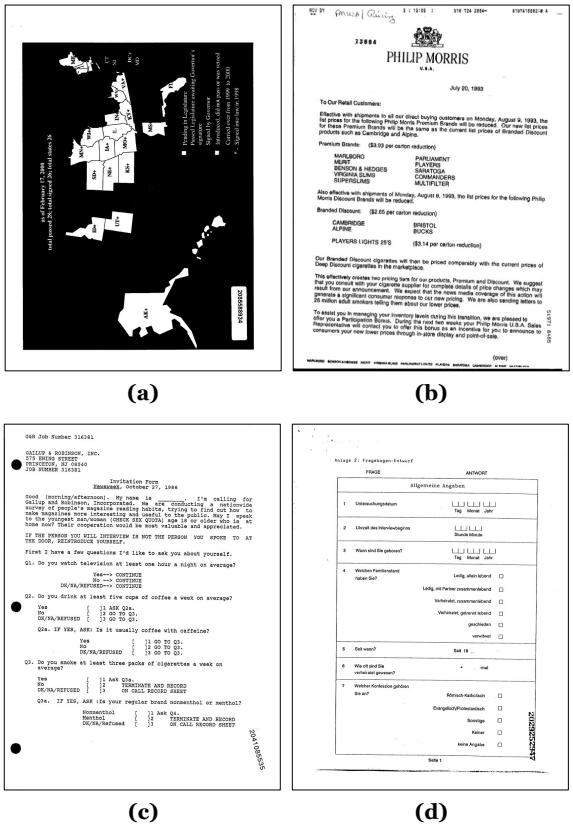


Figure 9: Problematic label error predictions from Cleanlab with VGG-16.

respectively). One initial observation was that there was not a "tight" amount of agreement across all four runs of the Cleanlab tool. This is visualized in Figure 8, where we see that only 916 out of 5,987 RVL-CDIP test samples were predicted as errors by all four runs of Cleanlab.

We also observed that many of the label error predictions made by Cleanlab were themselves problematic. For instance, Figures 9a and 9b show cases where Cleanlab incorrectly predicted a label error: Figure 9a is a valid `presentation` document, and Figure 9b is a valid `letter` document. Figures 9c and 9d show *ambiguous* cases where Cleanlab predicted a label error: Figure 9c shows a `form` document with questionnaire-like elements, while Figure 9d shows a `questionnaire` document with form-like elements. However, we argue that these false-positives are not entirely due to the Cleanlab tool, but instead due to the noisy nature of the RVL-CDIP training set: since Cleanlab uses model predictions, and since those models were trained on noisy data, the Cleanlab predictions are therefore bound to be imperfect. Similarly, we posit that the large amount of test-train overlap leads to brittle models, which also leads to imperfect predictions by Cleanlab. Indeed, Cleanlab's docu-

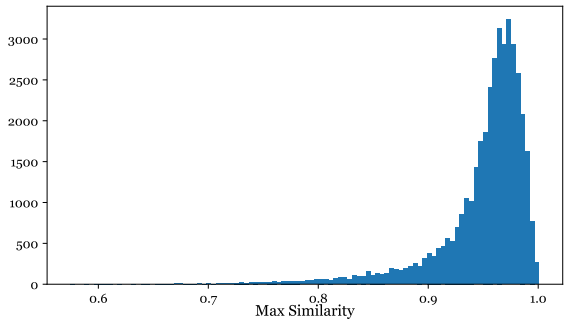


Figure 10: Maximum similarities between test and train samples for RVL-CDIP test data.

mentation warns that "Cleanlab performs better if the [model confidence scores] from your model are out-of-sample"[8] and we have argued in the main paper above that high amounts of test-train overlap lead to fewer test cases that are out-of-sample.

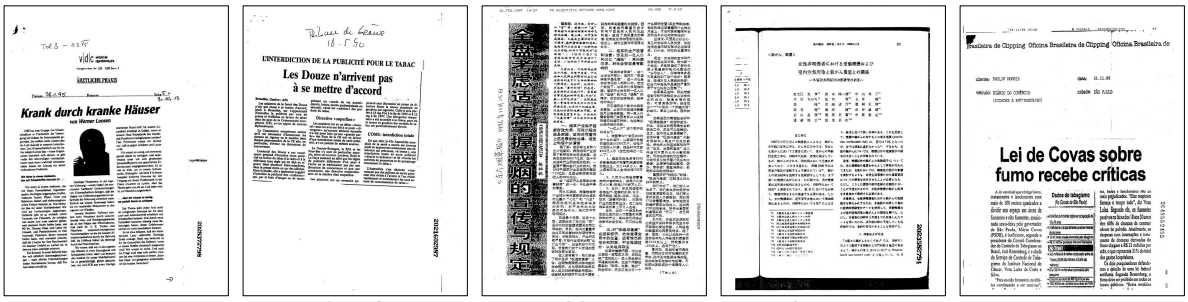
A.2 Supplementary Visualizations

Figure 10 charts maximum similarity scores between test and train samples for the RVL-CDIP test data. Figure 11 lists several non-English samples from RVL-CDIP. Figures 12–14 show example errors and ambiguous documents. Figures 15–18 display test-train pairs with corresponding similarity scores. Figure 19 show examples of "Biographical Sketch" documents from the `resume` category, illustrating the high level of similarity of this particular sub-type; Figure 20 shows a similar case for another category. Figures 21-23 show cases where two categories have the same sub-types of documents.

A.3 Data Availability

The data and metadata that we annotated as part of our error analysis (excluding data with sensitive information) is available at: `github.com/gxlarson/rvlcdip-errors`.

[8] `https://docs.cleanlab.ai/stable/index.html`



german news_article | french news_article | chinese news_article | japanese scientific_pub. | portuguese news_article

Figure 11: Example non-English samples from RVL-CDIP.

invoice | budget | email | resume | scientific_pub.

form | questionnaire | budget | advertisement | email

scientific_pub. | questionnaire | scientific_rep. | news_article | memo

specification | budget | presentation | form | scientific_rep.

Figure 12: Examples of *unknown* label errors with corresponding original RVL-CDIP labels.

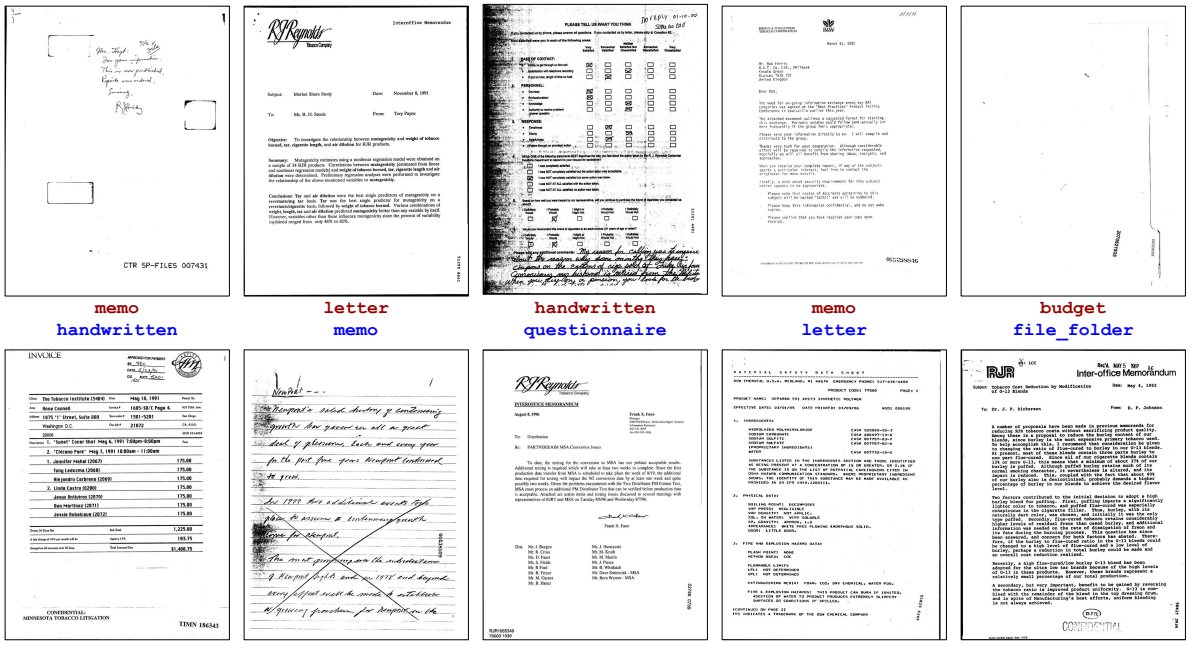


Figure 13: Examples of *mis-label* label errors with corresponding original (top) and corrected (bottom) RVL-CDIP labels.

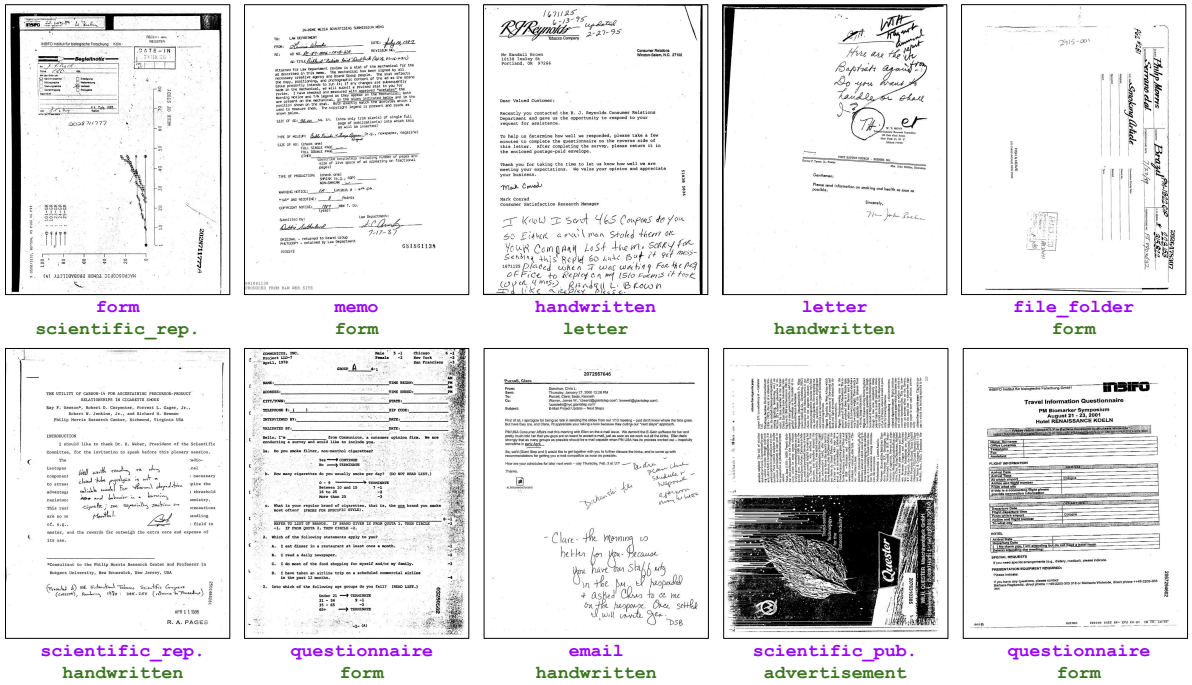


Figure 14: Examples of mixed or multi-label documents from RVL-CDIP. The original RVL-CDIP label is shown first (top) and the additional valid label second (bottom).

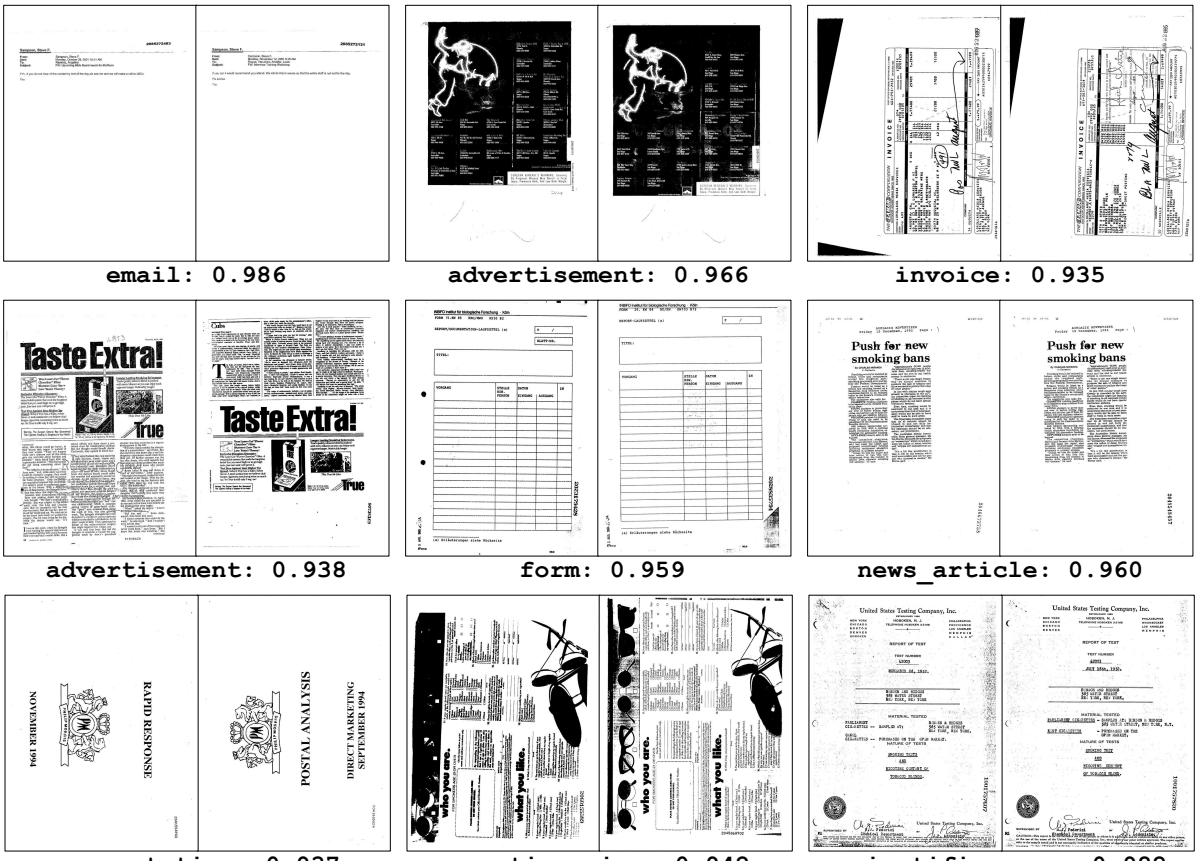
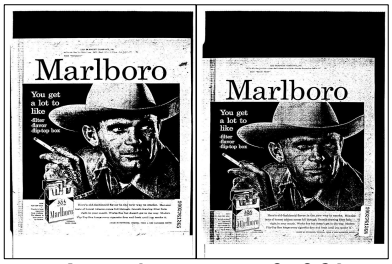
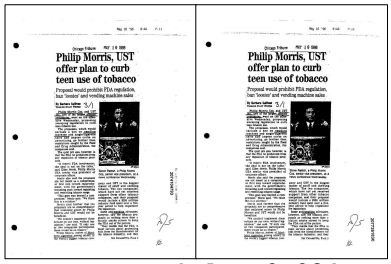


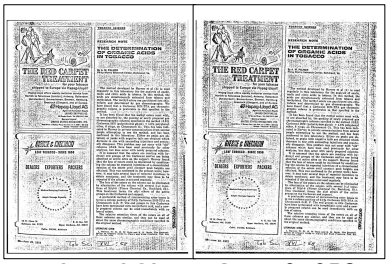
Figure 15: Examples of test-train pairs with corresponding cosine similarity scores.



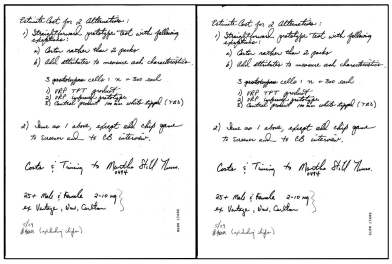
advertisement: 0.964



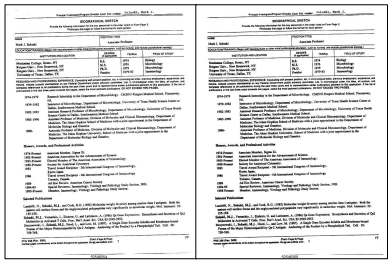
news_article: 0.984



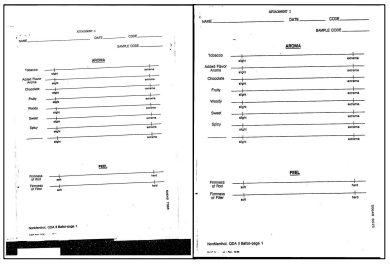
scientific_pub.: 0.953



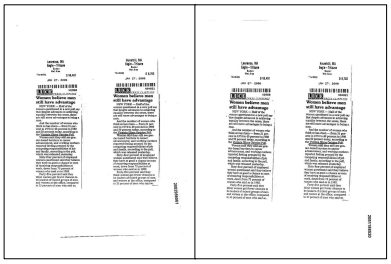
handwritten: 0.990



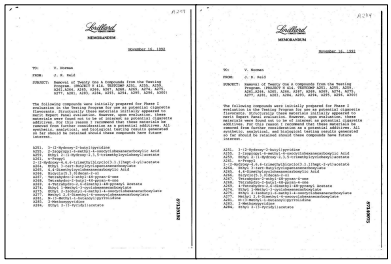
resume: 0.994



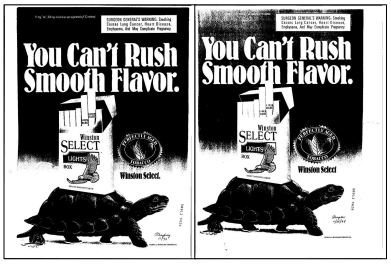
questionnaire: 0.950



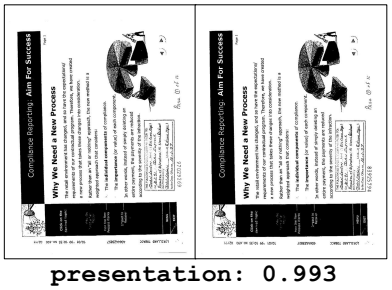
news_article: 0.956



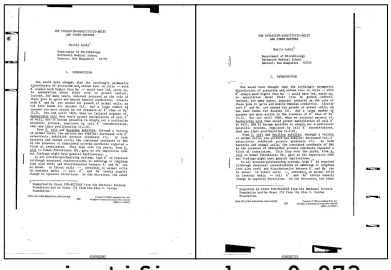
memo: 0.991



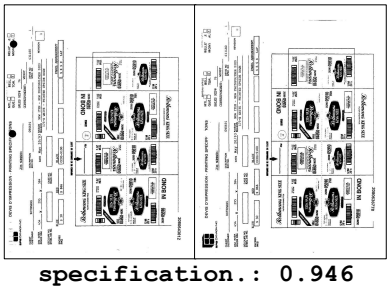
advertisement.: 0.939



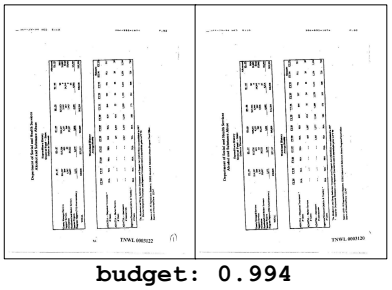
presentation: 0.993



scientific_pub.: 0.973

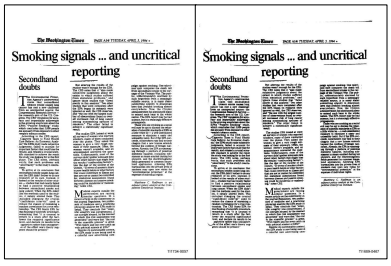


specification.: 0.946



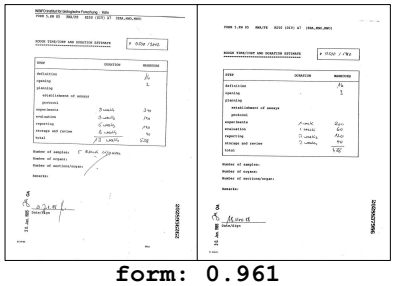
budget: 0.994

advertisement.: 0.944

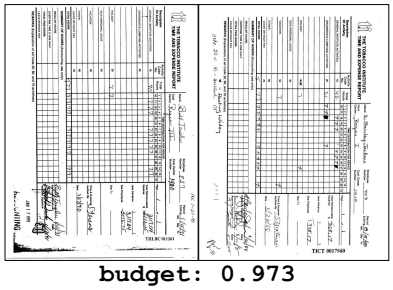


news_article.: 0.981

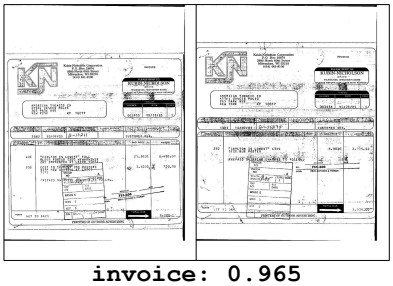
Figure 16: Example test-train duplicate pairs.



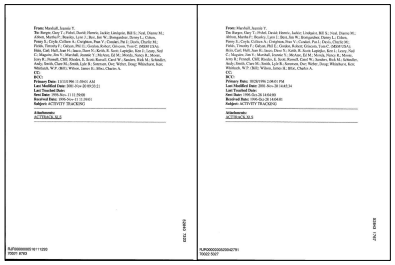
form: 0.961



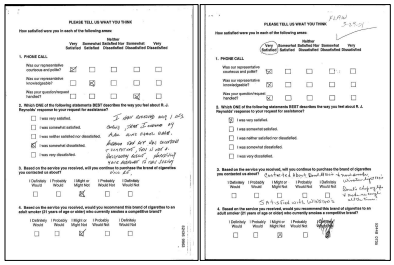
budget: 0.973



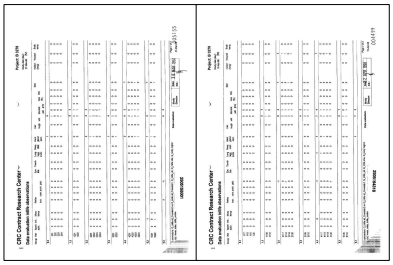
invoice: 0.965



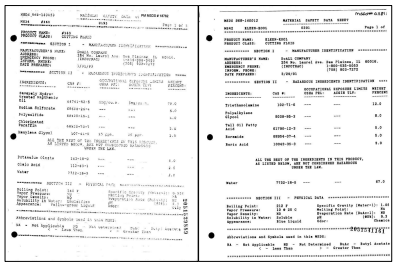
email: 0.992



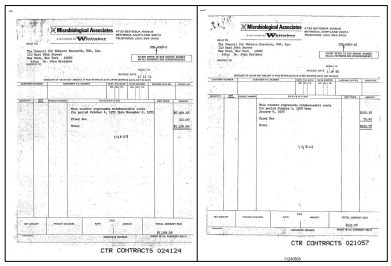
questionnaire: 0.964



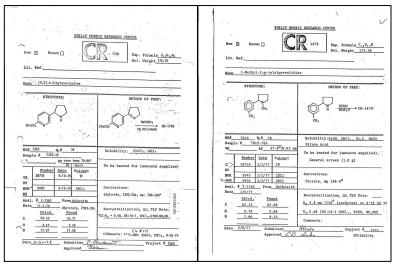
scientific_rep.: 0.994



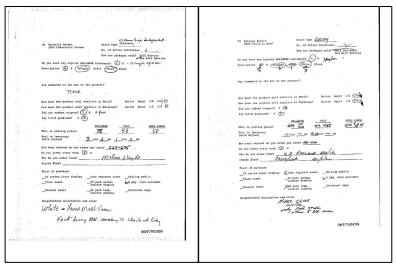
specification: 0.959



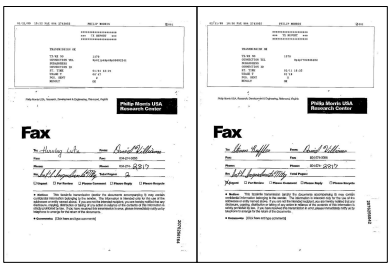
invoice: 0.986



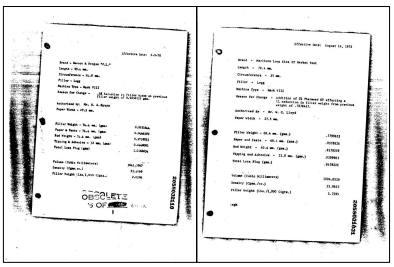
form.: 0.967



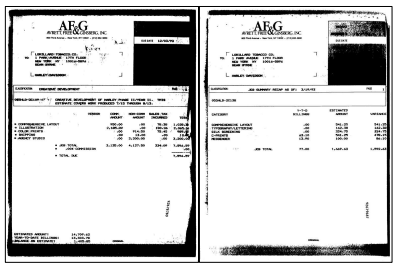
questionnaire: 0.959



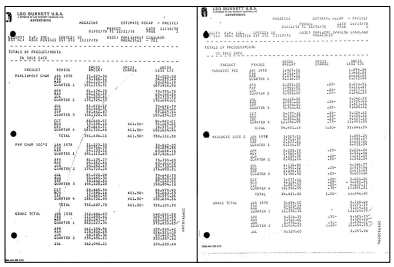
form: 0.986



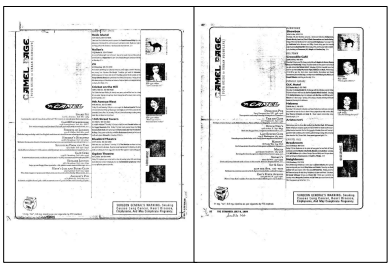
specification.: 0.963



invoice: 0.972



form: 0.967



advertisement.: 0.948

Figure 17: Example test-train pairs with a high level of similarity due to overlap in document templates.

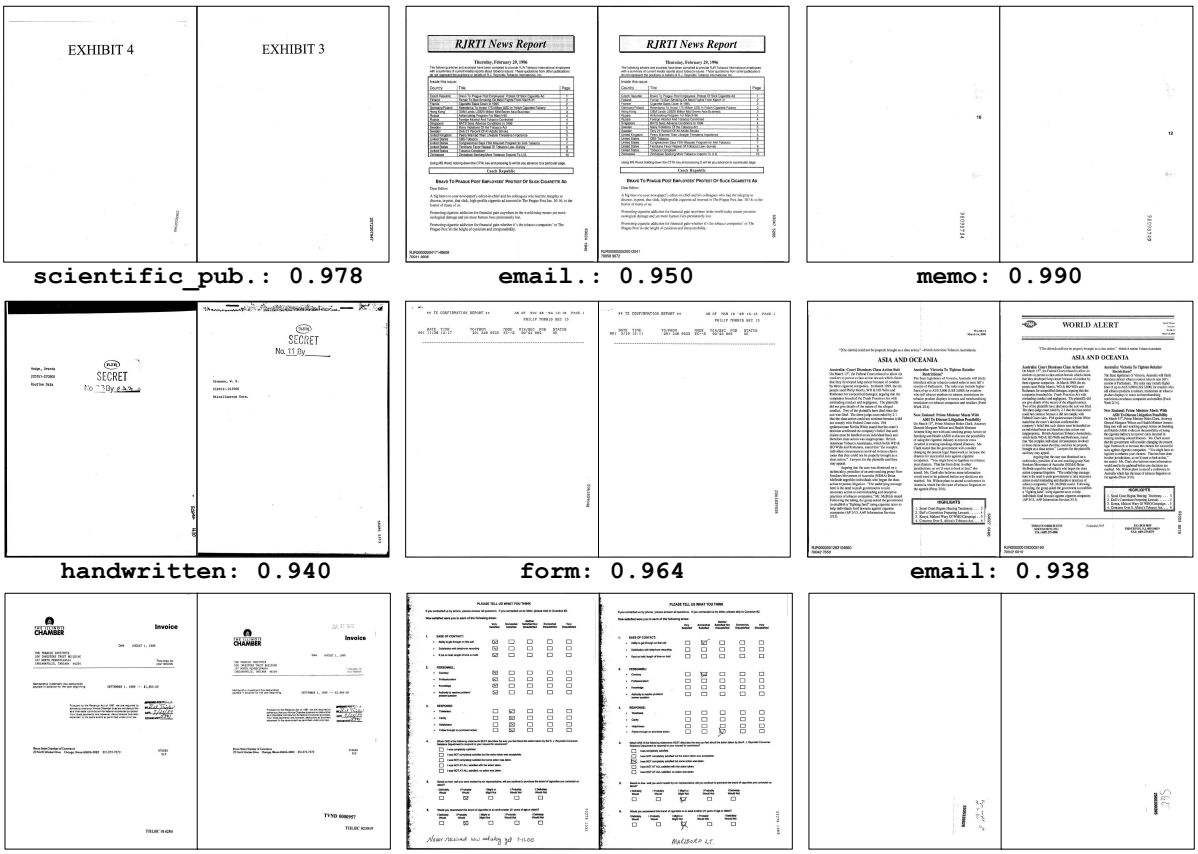


Figure 18: Example test-train pairs that have erroneous labels.

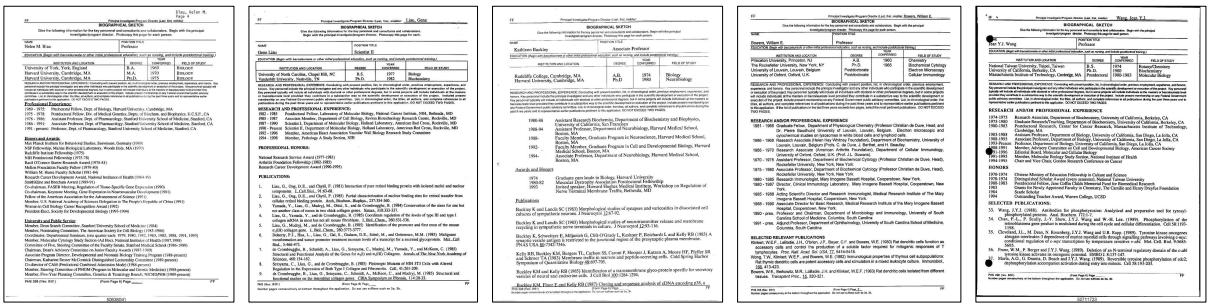


Figure 19: Examples of "Biographical Sketch" documents, which are abundant in the `resume` category.

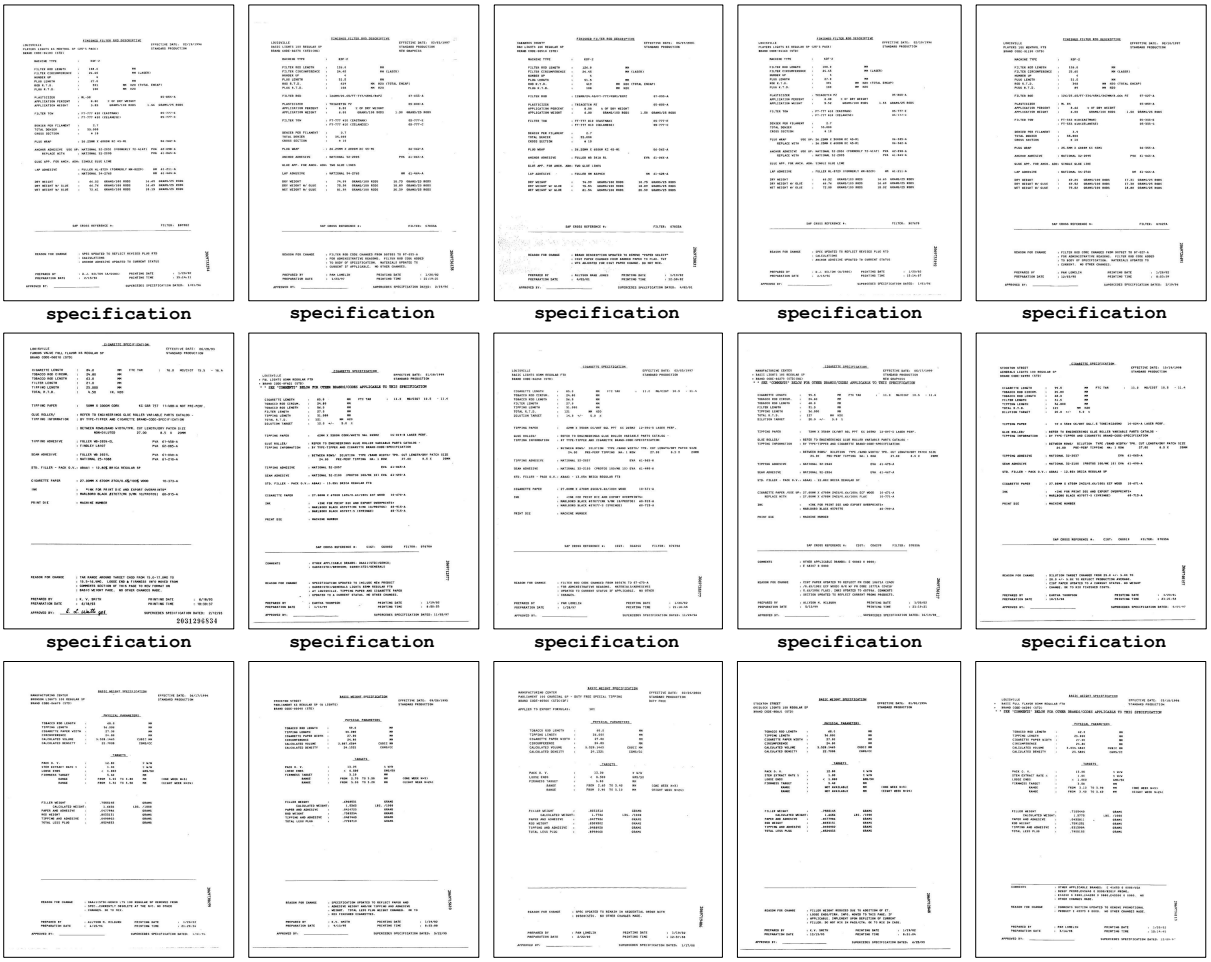


Figure 20: Examples of three common types of documents within the specification document categor: "Finished Filter Rod Descriptive" documents (top row), "Cigarette Specification" (middle row), "Basic Weight Specification" (bottom row).

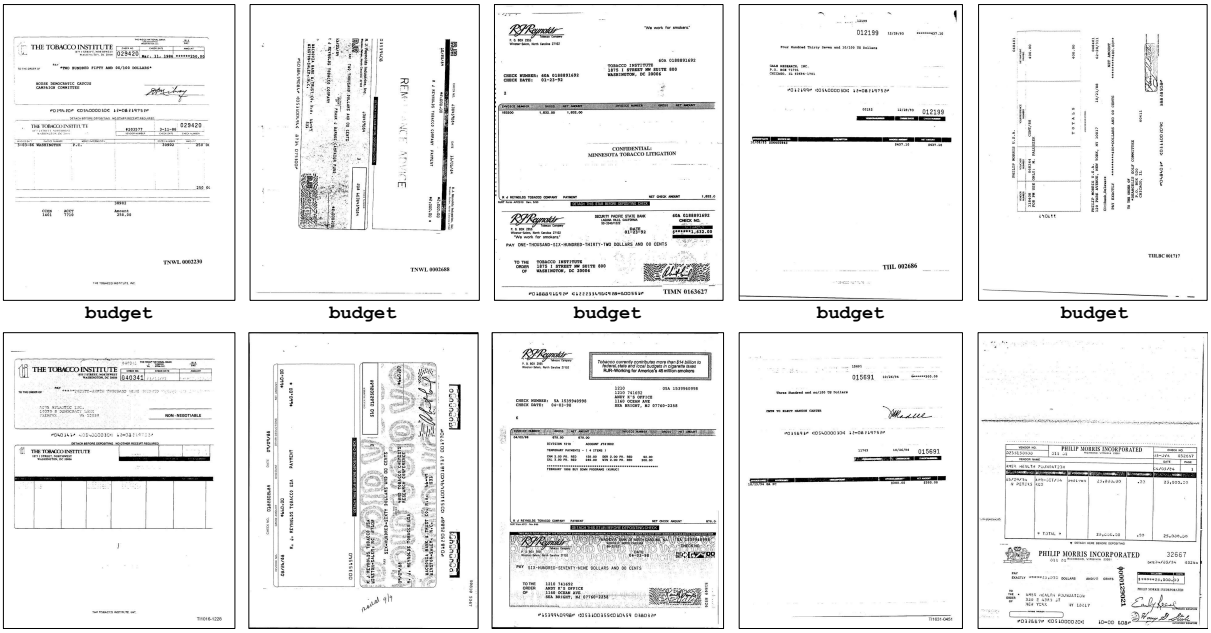


Figure 21: Examples of check images from the budget (top) and invoice document categories.

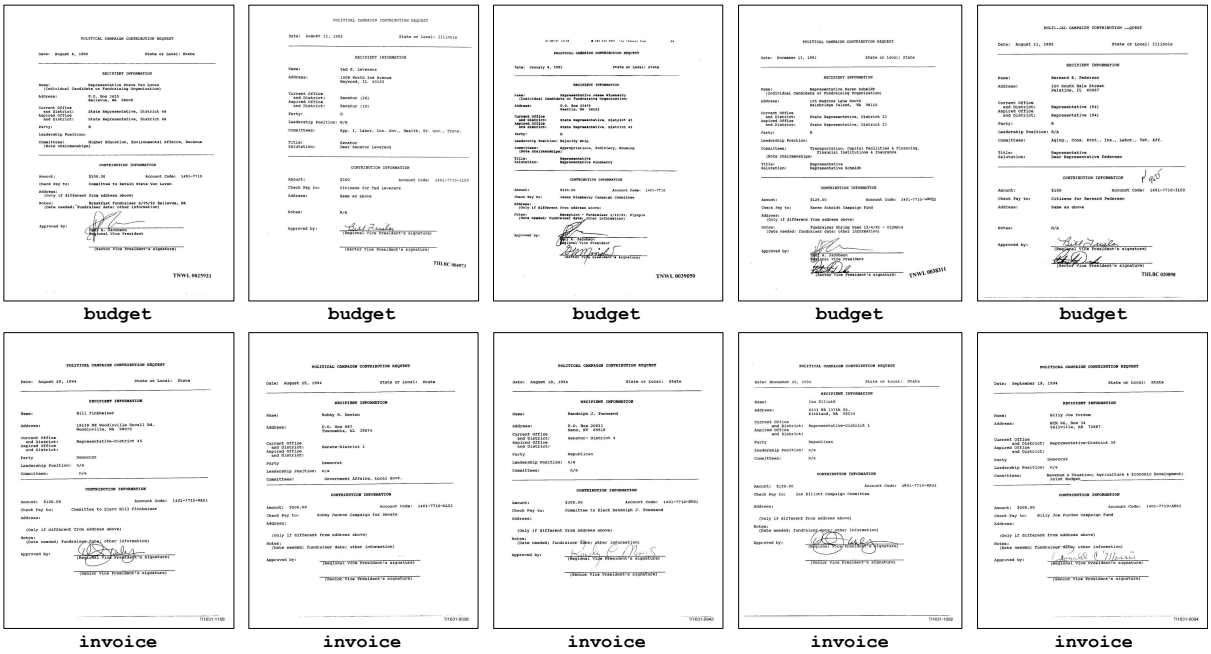


Figure 22: Examples of "Political Campaign Contribuiton Request" documents from budget (top row) and invoice (bottom row) categories.

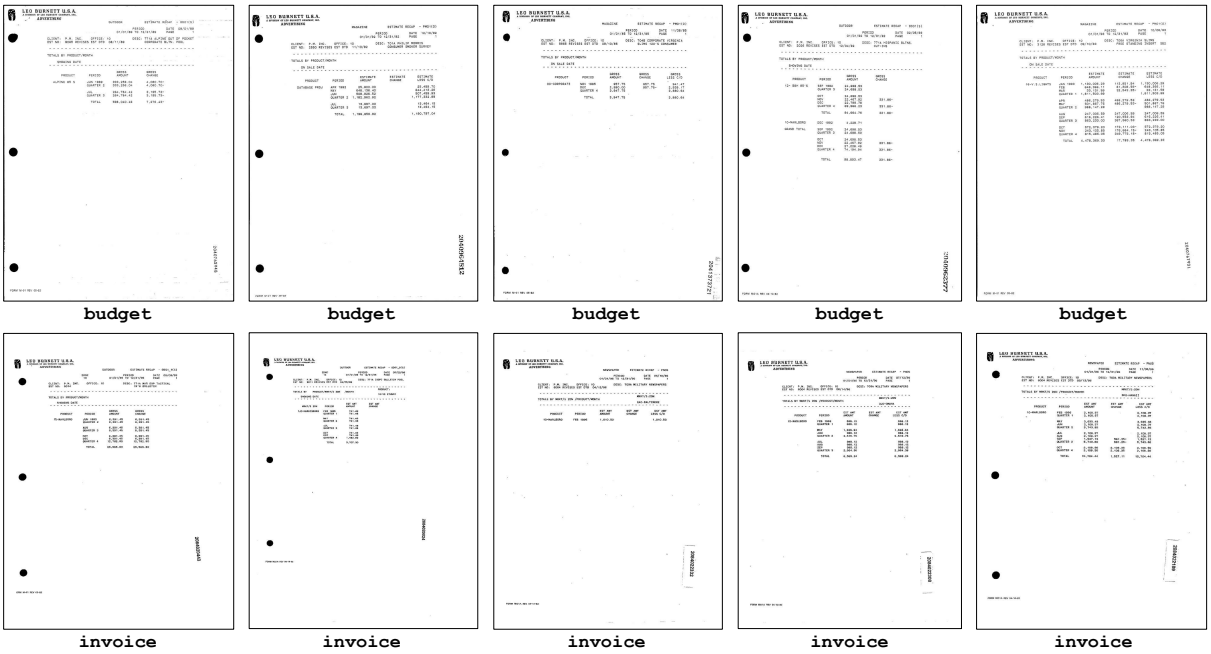


Figure 23: Examples of advertisement placement report images from the budget (top) and invoice document categories.