Table 5: Per-class accuracy scores on RVL-CDIP-*N* for each document classification model.

Model	budget	email	form	handwritten	invoice	letter	memo	news_article	questionnaire	resume	scientific_pub.	specification
VGG-16	0.793	0.848	0.743	0.403	0.737	0.901	0.553	0.686	0.718	0.696	0.974	0.230
ResNet-50	0.810	0.758	0.714	0.358	0.421	0.803	0.511	0.570	0.641	0.674	0.949	0.148
GoogLeNet	0.776	0.818	0.700	0.449	0.439	0.816	0.553	0.616	0.513	0.609	0.923	0.115
AlexNet	0.690	0.727	0.443	0.352	0.526	0.836	0.404	0.558	0.487	0.663	0.949	0.148
LayoutLMv2	0.897	0.848	0.529	0.261	0.333	0.836	0.511	0.512	0.769	0.565	0.923	0.164
DiT	0.862	0.970	0.914	0.624	0.860	0.954	0.723	0.849	0.821	0.734	0.923	0.410

A Additional Results

A.1 RVL-CDIP-*N* Confusion Matrices

Table 5 displays accuracy scores for each document category in RVL-CDIP-*N*. There are several patterns: all models perform well on `scientific_publication` but typically poorly on `handwritten` and `specification`. (See Figures 30 and 31 for a comparison between `specification` documents from RVL-CDIP-*N* and RVL-CDIP, and Figures 24 and 25 for a comparison between `handwritten` documents.) Tables 8 – 13 display model confusion matrices on RVL-CDIP-*N*.

A.2 Model Prediction Similarity

Tables 6 and 7 show pairwise similarity scores for each model on RVL-CDIP-*N* and RVL-CDIP-*O*, respectively. Here, we compute similarity between a pair of models by computing the Hamming similarity between predicted labels, or

$$sim(\mathbf{y}_a, \mathbf{y}_b) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(\mathbf{y}_a^i = \mathbf{y}_b^i)$$

where $\mathbf{y}_a^i$ is model a's predicted label on the i^{th} test document. We find that there is a higher degree of similarity in model predictions on RVL-CDIP-*N* than on RVL-CDIP-*O*.

A.3 In- versus Out-of-Domain Performance

Tables 14 and 15 chart AUC scores for MSP on RVL-CDIP versus RVL-CDIP-*O* (Table 14) and on RVL-CDIP-*N* versus RVL-CDIP-*O* (Table 15). Tables 16 and 17 chart AUC scores for Energy on RVL-CDIP versus RVL-CDIP-*O* (Table 17) and on RVL-CDIP-*N* versus RVL-CDIP-*O* (Table 17). We find that the Augraphy augmentations typically have little impact. The main finding in Tables 14 – 17 is that out-of-domain detection suffers on the more realistic RVL-CDIP-*N* versus RVL-CDIP-*O* setting, where both sets of test documents are out-of-distribution. This is in contrast with the *T*-*O* setting where we use the in-distribution RVL-CDIP test set as the in-domain data. We report similar findings using the FPR95 metric in Tables 18 – 21.

A.4 Confidence Scores

Figures 8 – 19 show relationships between confidence scores and performance on RVL-CDIP and RVL-CDIP-*N* accuracy as well as RVL-CDIP-*O* detection rate. We see that as confidence score threshold increases, the detection rate on RVL-CDIP-*O* increases while accuracy on RVL-CDIP and RVL-CDIP-*N* naturally decreases due to the decision rule defined in 4.2. Figures 8 – 19 also display distributions of model confidence scores for RVL-CDIP-*N* and RVL-CDIP-*O*. We see that there is a large amount of overlap between the two distributions for all models and for both MSP and Energy confidence score methods, which helps explain the drop in AUC scores on the *N*-*O* (RVL-CDIP-*N* versus RVL-CDIP-*O*) setting.

	VGG-16	*ResNet-50*	*GoogLeNet*	*AlexNet*	*LayoutLMv2*	*DiT*
VGG-16	1.00	0.63	0.61	0.62	0.57	0.65
ResNet-50		1.00	0.58	0.57	0.57	0.58
GoogLeNet			1.00	0.56	0.54	0.57
AlexNet				1.00	0.55	0.56
LayoutLMv2					1.00	0.55
DiT						1.00

Table 6: Pairwise similarity between predictions made by each model on RVL-CDIP-*N*.

	VGG-16	*ResNet-50*	*GoogLeNet*	*AlexNet*	*LayoutLMv2*	*DiT*
VGG-16	1.00	0.45	0.43	0.44	0.42	0.27
ResNet-50		1.00	0.46	0.42	0.43	0.33
GoogLeNet			1.00	0.41	0.40	0.30
AlexNet				1.00	0.39	0.28
LayoutLMv2					1.00	0.30
DiT						1.00

Table 7: Pairwise similarity between predictions made by each model on RVL-CDIP-*O*.

Table 8: Confusion matrix for VGG-16 on the RVL-CDIP-*N* data. True labels are rows, and predicted labels are columns.

	advertisement	budget	email	file_folder	form	handwritten	invoice	letter	memo	news_article	presentation	questionnaire	resume	scientific_pub.	scientific_rep.	specification
advertisement	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
budget	0.000	0.793	0.017	0.000	0.034	0.000	0.017	0.000	0.000	0.000	0.000	0.052	0.000	0.017	0.052	0.000
email	0.000	0.061	0.848	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.030	0.061	0.000	0.000	0.000
file_folder	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
form	0.000	0.043	0.000	0.000	0.743	0.000	0.014	0.014	0.000	0.000	0.043	0.029	0.029	0.000	0.000	0.086
handwritten	0.023	0.051	0.000	0.051	0.040	0.403	0.011	0.091	0.080	0.000	0.006	0.028	0.011	0.011	0.193	0.000
invoice	0.000	0.105	0.000	0.000	0.035	0.000	0.737	0.070	0.018	0.000	0.000	0.018	0.000	0.000	0.018	0.000
letter	0.000	0.007	0.020	0.007	0.000	0.000	0.000	0.901	0.000	0.013	0.033	0.000	0.020	0.000	0.000	0.000
memo	0.000	0.021	0.043	0.000	0.021	0.000	0.000	0.128	0.553	0.000	0.043	0.043	0.085	0.021	0.043	0.000
news_article	0.128	0.012	0.000	0.023	0.000	0.000	0.000	0.000	0.000	0.686	0.023	0.000	0.012	0.093	0.023	0.000
presentation	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
questionnaire	0.000	0.051	0.000	0.000	0.077	0.000	0.000	0.000	0.000	0.000	0.000	0.718	0.051	0.043	0.077	0.026
resume	0.022	0.011	0.022	0.000	0.033	0.000	0.016	0.016	0.000	0.000	0.016	0.049	0.696	0.974	0.033	0.033
scientific_pub.	0.000	0.000	0.000	0.000	0.026	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.974	0.000	0.000
scientific_rep.	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
specification	0.197	0.022	0.049	0.000	0.131	0.000	0.098	0.000	0.000	0.016	0.049	0.033	0.033	0.098	0.033	0.230

Table 9: Confusion matrix for ResNet-50 on the RVL-CDIP-*N* data. True labels are rows, and predicted labels are columns.

	advertisement	budget	email	file_folder	form	handwritten	invoice	letter	memo	news_article	presentation	questionnaire	resume	scientific_pub.	scientific_rep.	specification
advertisement	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
budget	0.000	0.810	0.000	0.000	0.052	0.000	0.017	0.000	0.000	0.000	0.000	0.052	0.000	0.017	0.069	0.000
email	0.000	0.061	0.758	0.000	0.030	0.000	0.000	0.030	0.000	0.000	0.000	0.030	0.061	0.000	0.000	0.030
file_folder	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
form	0.028	0.029	0.000	0.000	0.714	0.000	0.000	0.014	0.000	0.000	0.000	0.086	0.014	0.000	0.029	0.100
handwritten	0.028	0.051	0.006	0.045	0.034	0.358	0.000	0.063	0.063	0.006	0.074	0.034	0.068	0.006	0.148	0.017
invoice	0.018	0.123	0.018	0.000	0.123	0.000	0.421	0.070	0.018	0.000	0.053	0.018	0.000	0.000	0.035	0.018
letter	0.007	0.000	0.053	0.000	0.007	0.000	0.000	0.803	0.000	0.000	0.033	0.000	0.046	0.000	0.033	0.000
memo	0.043	0.000	0.064	0.000	0.021	0.000	0.000	0.000	0.511	0.021	0.170	0.000	0.043	0.000	0.106	0.021
news_article	0.221	0.023	0.023	0.035	0.012	0.000	0.000	0.000	0.000	0.570	0.047	0.000	0.000	0.058	0.000	0.012
presentation	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
questionnaire	0.000	0.051	0.000	0.000	0.103	0.000	0.000	0.000	0.026	0.000	0.000	0.641	0.026	0.000	0.103	0.026
resume	0.027	0.027	0.033	0.005	0.023	0.000	0.005	0.011	0.000	0.011	0.049	0.054	0.674	0.033	0.027	0.016
scientific_pub.	0.000	0.000	0.000	0.000	0.051	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.949	0.000	0.000
scientific_rep.	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
specification	0.180	0.033	0.016	0.049	0.148	0.000	0.049	0.000	0.000	0.000	0.098	0.082	0.033	0.082	0.082	0.148

Table 10: Confusion matrix for GoogLeNet on the RVL-CDIP-*N* data. True labels are rows, and predicted labels are columns.

	advertisement	budget	email	file_folder	form	handwritten	invoice	letter	memo	news_article	presentation	questionnaire	resume	scientific_pub.	scientific_rep.	specification
advertisement	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
budget	0.000	0.776	0.052	0.000	0.052	0.000	0.017	0.000	0.000	0.000	0.000	0.069	0.000	0.000	0.052	0.000
email	0.000	0.000	0.818	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.091	0.000	0.061	0.030
file_folder	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
form	0.014	0.029	0.000	0.000	0.700	0.000	0.000	0.000	0.000	0.000	0.000	0.0129	0.029	0.000	0.000	0.086
handwritten	0.017	0.067	0.017	0.051	0.023	0.449	0.006	0.051	0.057	0.006	0.034	0.017	0.074	0.011	0.125	0.000
invoice	0.000	0.070	0.000	0.000	0.211	0.000	0.439	0.088	0.018	0.000	0.035	0.018	0.053	0.000	0.070	0.000
letter	0.000	0.000	0.013	0.007	0.000	0.000	0.000	0.816	0.020	0.007	0.033	0.000	0.072	0.000	0.020	0.007
memo	0.000	0.021	0.021	0.000	0.021	0.000	0.000	0.085	0.553	0.021	0.064	0.000	0.213	0.000	0.000	0.000
news_article	0.279	0.000	0.000	0.012	0.000	0.012	0.000	0.000	0.000	0.616	0.024	0.000	0.000	0.035	0.000	0.023
presentation	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
questionnaire	0.000	0.103	0.026	0.000	0.026	0.000	0.000	0.026	0.000	0.000	0.051	0.513	0.128	0.000	0.0103	0.026
resume	0.011	0.011	0.027	0.000	0.038	0.000	0.000	0.000	0.005	0.038	0.082	0.098	0.609	0.109	0.049	0.023
scientific_pub.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.026	0.026	0.000	0.000	0.923	0.026	0.000
scientific_rep.	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
specification	0.115	0.115	0.000	0.016	0.148	0.000	0.033	0.016	0.000	0.033	0.082	0.016	0.082	0.164	0.066	0.115

Table 11: Confusion matrix for AlexNet on the RVL-CDIP-*N* data. True labels are rows, and predicted labels are columns.

	advertisement	budget	email	file_folder	form	handwritten	invoice	letter	memo	news_article	presentation	questionnaire	resume	scientific_pub.	scientific_rep.	specification
advertisement	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
budget	0.000	0.690	0.017	0.034	0.052	0.000	0.034	0.000	0.000	0.000	0.000	0.052	0.017	0.017	0.052	0.034
email	0.091	0.061	0.727	0.000	0.000	0.000	0.000	0.000	0.000	0.030	0.000	0.000	0.060	0.000	0.000	0.030
file_folder	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
form	0.029	0.071	0.000	0.000	0.443	0.000	0.071	0.029	0.029	0.000	0.014	0.071	0.071	0.000	0.000	0.171
handwritten	0.034	0.051	0.000	0.051	0.023	0.352	0.023	0.159	0.091	0.006	0.040	0.011	0.017	0.000	0.142	0.000
invoice	0.000	0.105	0.000	0.000	0.175	0.000	0.526	0.123	0.000	0.000	0.000	0.035	0.035	0.000	0.000	0.000
letter	0.013	0.000	0.039	0.000	0.000	0.000	0.000	0.836	0.007	0.000	0.007	0.000	0.079	0.007	0.013	0.000
memo	0.021	0.021	0.085	0.000	0.085	0.000	0.085	0.000	0.404	0.000	0.043	0.043	0.213	0.021	0.043	0.000
news_article	0.302	0.116	0.000	0.023	0.000	0.023	0.000	0.000	0.000	0.558	0.000	0.000	0.000	0.058	0.023	0.000
presentation	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
questionnaire	0.051	0.000	0.026	0.000	0.026	0.000	0.000	0.026	0.000	0.000	0.026	0.487	0.128	0.000	0.154	0.026
resume	0.033	0.022	0.038	0.000	0.038	0.000	0.005	0.005	0.005	0.005	0.033	0.016	0.663	0.033	0.065	0.043
scientific_pub.	0.000	0.000	0.000	0.000	0.026	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.949	0.026	0.000
scientific_rep.	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
specification	0.115	0.115	0.016	0.016	0.016	0.016	0.131	0.016	0.000	0.033	0.033	0.016	0.016	0.131	0.033	0.148

Table 12: Confusion matrix for LayoutLMv2 on the RVL-CDIP-*N* data. True labels are rows, and predicted labels are columns.

	advertisement	budget	email	file_folder	form	handwritten	invoice	letter	memo	news_article	presentation	questionnaire	resume	scientific_pub.	scientific_rep.	specification
advertisement	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
budget	0.000	0.897	0.017	0.017	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.052	0.017	0.000	0.000	0.000
email	0.000	0.000	0.848	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.030	0.030	0.061	0.000	0.000	0.000
file_folder	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
form	0.014	0.271	0.000	0.014	0.529	0.000	0.000	0.000	0.000	0.000	0.000	0.086	0.043	0.000	0.000	0.043
handwritten	0.125	0.097	0.011	0.045	0.028	0.261	0.006	0.216	0.000	0.000	0.068	0.017	0.034	0.006	0.080	0.006
invoice	0.000	0.509	0.000	0.000	0.070	0.000	0.333	0.000	0.000	0.000	0.035	0.018	0.035	0.000	0.000	0.000
letter	0.007	0.007	0.026	0.000	0.000	0.007	0.000	0.836	0.007	0.000	0.046	0.000	0.039	0.000	0.026	0.000
memo	0.021	0.043	0.043	0.000	0.021	0.000	0.000	0.106	0.511	0.021	0.128	0.043	0.043	0.000	0.021	0.000
news_article	0.290	0.023	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.512	0.012	0.000	0.012	0.140	0.012	0.000
presentation	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
questionnaire	0.000	0.128	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.051	0.769	0.051	0.000	0.000	0.000
resume	0.065	0.082	0.023	0.000	0.023	0.000	0.000	0.000	0.000	0.000	0.087	0.076	0.565	0.016	0.033	0.033
scientific_pub.	0.026	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.026	0.923	0.000	0.026
scientific_rep.	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
specification	0.279	0.230	0.016	0.000	0.115	0.000	0.016	0.000	0.000	0.000	0.034	0.016	0.016	0.066	0.049	0.164

Table 13: Confusion matrix for DiT on the RVL-CDIP-*N* data. True labels are rows, and predicted labels are columns.

	advertisement	budget	email	file_folder	form	handwritten	invoice	letter	memo	news_article	presentation	questionnaire	resume	scientific_pub.	scientific_rep.	specification
advertisement	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
budget	0.017	0.862	0.017	0.000	0.034	0.000	0.000	0.000	0.000	0.017	0.000	0.000	0.000	0.000	0.052	0.000
email	0.000	0.000	0.970	0.000	0.000	0.000	0.000	0.000	0.000	0.030	0.000	0.000	0.000	0.000	0.000	0.000
file_folder	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
form	0.000	0.029	0.029	0.000	0.914	0.000	0.000	0.000	0.000	0.000	0.00	0.029	0.000	0.000	0.000	0.000
handwritten	0.000	0.017	0.006	0.011	0.034	0.642	0.011	0.085	0.057	0.006	0.085	0.006	0.000	0.000	0.040	0.000
invoice	0.018	0.035	0.035	0.000	0.035	0.000	0.860	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.018	0.000
letter	0.013	0.000	0.007	0.000	0.000	0.000	0.000	0.954	0.000	0.000	0.026	0.000	0.000	0.000	0.000	0.000
memo	0.000	0.000	0.064	0.000	0.021	0.000	0.000	0.149	0.723	0.000	0.021	0.000	0.000	0.000	0.021	0.000
news_article	0.105	0.000	0.012	0.000	0.000	0.000	0.000	0.012	0.000	0.849	0.012	0.000	0.000	0.012	0.000	0.000
presentation	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
questionnaire	0.000	0.000	0.000	0.000	0.103	0.000	0.000	0.000	0.000	0.000	0.000	0.821	0.000	0.000	0.077	0.000
resume	0.022	0.000	0.049	0.000	0.005	0.000	0.000	0.000	0.000	0.016	0.005	0.000	0.734	0.038	0.076	0.054
scientific_pub.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.026	0.000	0.000	0.923	0.051	0.000
scientific_rep.	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
specification	0.115	0.066	0.000	0.000	0.230	0.000	0.000	0.000	0.000	0.016	0.016	0.000	0.000	0.000	0.148	0.410

Table 14: AUC scores using MSP on RVL-CDIP versus RVL-CDIP-*O* (*T-O*) .

Model	*T-O* micro	*T-O* micro w/Aug.	*T-O* macro	*T-O* macro w/Aug.
VGG-16	0.881	0.867	0.895	0.883
ResNet-50	0.849	0.852	0.871	0.879
GoogLeNet	0.838	0.839	0.859	0.862
AlexNet	0.871	0.882	0.889	0.895
LayoutLMv2	0.842	0.832	0.875	0.878
DiT-base	0.894	0.923	0.901	0.929

Table 15: AUC scores using MSP on RVL-CDIP-*N* versus RVL-CDIP-*O* (*N-O*) .

Model	*N-O* micro	*N-O* micro w/Aug.	*N-O* macro	*N-O* macro w/Aug.
VGG-16	0.649	0.631	0.720	0.712
ResNet-50	0.581	0.584	0.654	0.658
GoogLeNet	0.592	0.592	0.679	0.647
AlexNet	0.620	0.646	0.684	0.635
LayoutLMv2	0.620	0.620	0.717	0.706
DiT-base	0.728	0.777	0.780	0.754

Table 16: AUC scores using Energy on RVL-CDIP versus RVL-CDIP-*O* (*T-O*) .

Model	*T-O* micro	*T-O* micro w/Aug.	*T-O* macro	*T-O* macro w/Aug.
VGG-16	0.923	0.906	0.930	0.916
ResNet-50	0.844	0.868	0.880	0.899
GoogLeNet	0.847	0.854	0.869	0.878
AlexNet	0.909	0.920	0.922	0.932
LayoutLMv2	0.849	0.849	0.891	0.891
DiT-base	0.888	0.933	0.902	0.936

Table 17: AUC scores using Energy on RVL-CDIP-*N* versus RVL-CDIP-*O* (*N-O*) .

Model	*N-O* micro	*N-O* micro w/Aug.	*N-O* macro	*N-O* macro w/Aug.
VGG-16	0.645	0.646	0.720	0.707
ResNet-50	0.554	0.583	0.661	0.671
GoogLeNet	0.587	0.561	0.689	0.642
AlexNet	0.646	0.628	0.706	0.655
LayoutLMv2	0.643	0.643	0.699	0.684
DiT-base	0.753	0.731	0.792	0.764

Table 18: FPR95 scores using MSP on RVL-CDIP versus RVL-CDIP-*O* (*T-O*) .

Model	*T-O* micro	*T-O* micro w/Aug.	*T-O* macro	*T-O* macro w/Aug.
VGG-16	0.649	0.657	0.530	0.543
ResNet-50	0.731	0.733	0.649	0.599
GoogLeNet	0.748	0.745	0.619	0.641
AlexNet	0.702	0.664	0.592	0.582
LayoutLMv2	0.717	0.717	0.590	0.590
DiT-base	0.587	0.497	0.461	0.401

Table 19: FPR95 scores using MSP on RVL-CDIP-*N* versus RVL-CDIP-*O* (*N-O*) .

Model	*T-O* micro	*T-O* micro w/Aug.	*T-O* macro	*T-O* macro w/Aug.
VGG-16	0.916	0.923	0.777	0.830
ResNet-50	0.935	0.933	0.834	0.832
GoogLeNet	0.946	0.933	0.842	0.862
AlexNet	0.919	0.916	0.785	0.850
LayoutLMv2	0.932	0.932	0.795	0.806
DiT-base	0.847	0.886	0.650	0.707

Table 20: FPR95 scores using Energy on RVL-CDIP versus RVL-CDIP-*O* (*T-O*) .

Model	*T-O* micro	*T-O* micro w/Aug.	*T-O* macro	*T-O* macro w/Aug.
VGG-16	0.461	0.529	0.379	0.444
ResNet-50	0.650	0.603	0.527	0.482
GoogLeNet	0.665	0.654	0.559	0.566
AlexNet	0.525	0.423	0.433	0.378
LayoutLMv2	0.753	0.753	0.573	0.573
DiT-base	0.688	0.375	0.415	0.336

Table 21: FPR95 scores using Energy on RVL-CDIP-*N* versus RVL-CDIP-*O* (*N-O*) .

Model	*T-O* micro	*T-O* micro w/Aug.	*T-O* macro	*T-O* macro w/Aug.
VGG-16	0.924	0.895	0.796	0.812
ResNet-50	0.942	0.935	0.784	0.850
GoogLeNet	0.943	0.945	0.799	0.869
AlexNet	0.937	0.931	0.779	0.847
LayoutLMv2	0.939	0.939	0.801	0.806
DiT-base	0.843	0.852	0.614	0.663

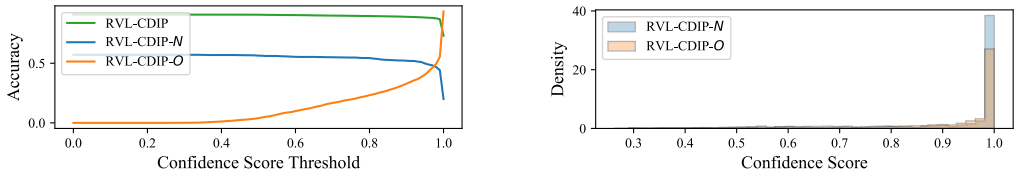

Figure 8: Relationship between confidence scores and accuracy (left) and distribution of confidence scores (right) for VGG-16 using MSP.

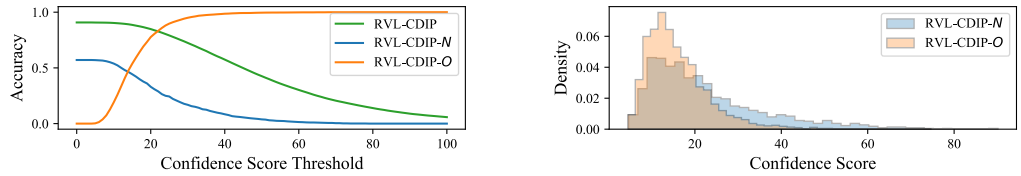

Figure 9: Relationship between confidence scores and accuracy (left) and distribution of confidence scores (right) for VGG-16 using Energy.

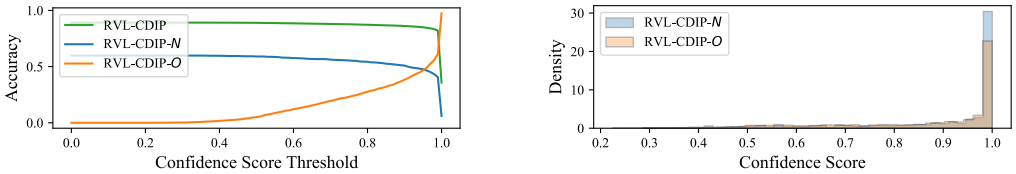

Figure 10: Relationship between confidence scores and accuracy (left) and distribution of confidence scores (right) for ResNet-50 using MSP.

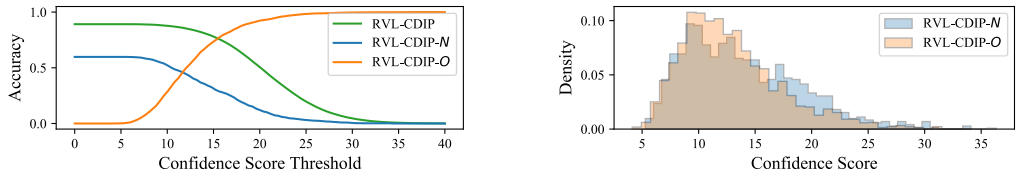

Figure 11: Relationship between confidence scores and accuracy (left) and distribution of confidence scores (right) for ResNet-50 using Energy.

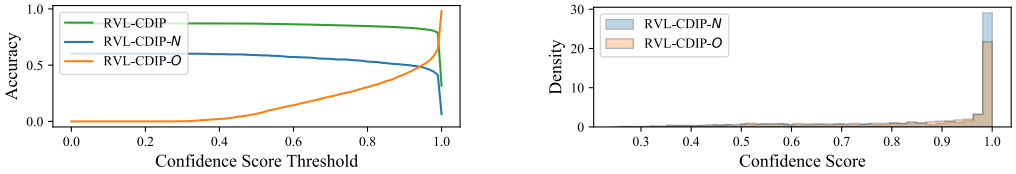

Figure 12: 2Relationship between confidence scores and accuracy (left) and distribution of confidence scores (right) for GoogLeNet using MSP.

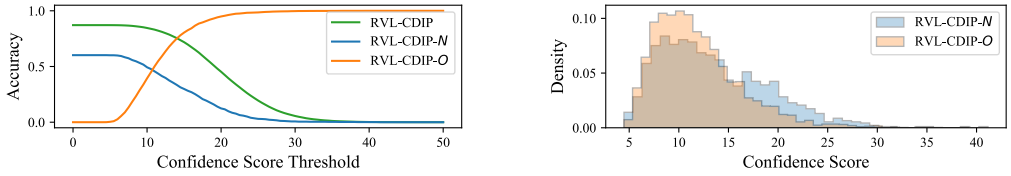

Figure 13: Relationship between confidence scores and accuracy (left) and distribution of confidence scores (right) for GoogLeNet using Energy.

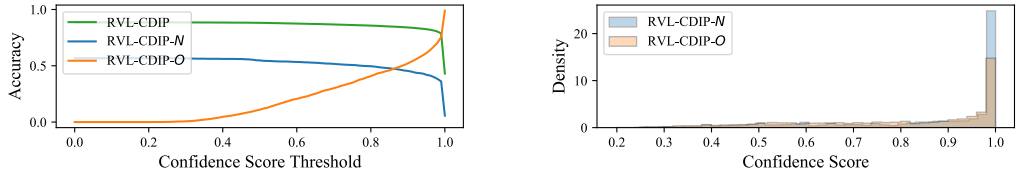

Figure 14: Relationship between confidence scores and accuracy (left) and distribution of confidence scores (right) for AlexNet using MSP.

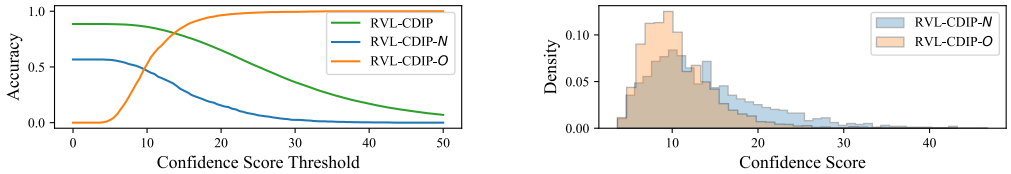

Figure 15: Relationship between confidence scores and accuracy (left) and distribution of confidence scores (right) for AlexNet using Energy.

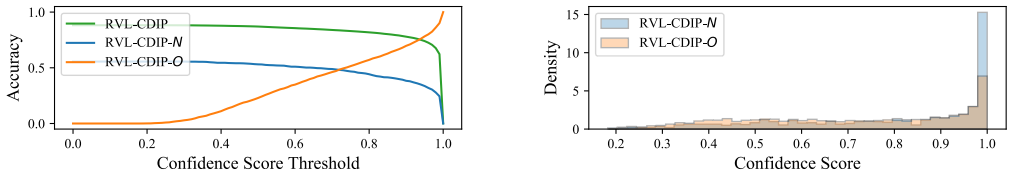

Figure 16: Relationship between confidence scores and accuracy (a) and distribution of confidence scores (b) for LayoutLMv2 using MSP.

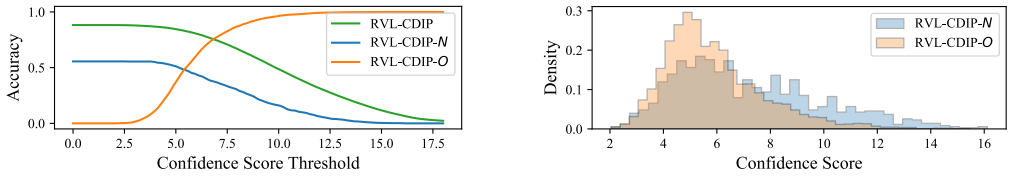

Figure 17: Relationship between confidence scores and accuracy (left) and distribution of confidence scores (right) for LayoutLMv2 using Energy.

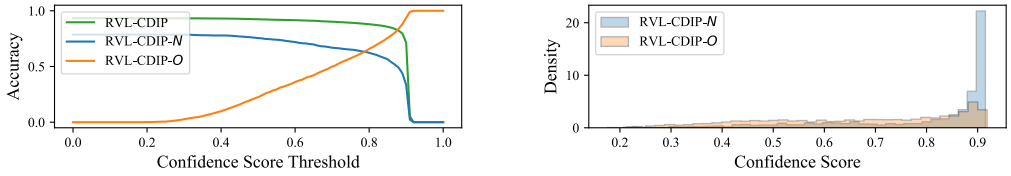

Figure 18: Relationship between confidence scores and accuracy (left) and distribution of confidence scores (right) for DiT using MSP.

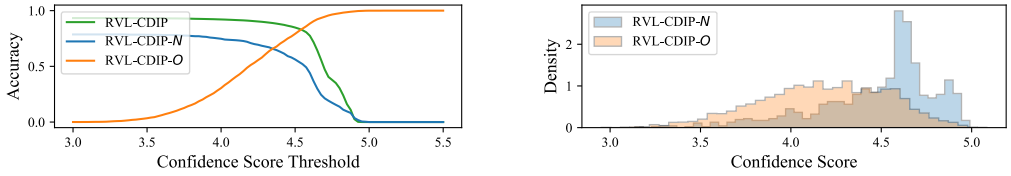

Figure 19: Relationship between confidence scores and accuracy (left) and distribution of confidence scores (right) for DiT using Energy.

B Comparison of RVL-CDIP and RVL-CDIP-*N*

Figures 20 – 43 compare samples from RVL-CDIP with those from RVL-CDIP-*N*.

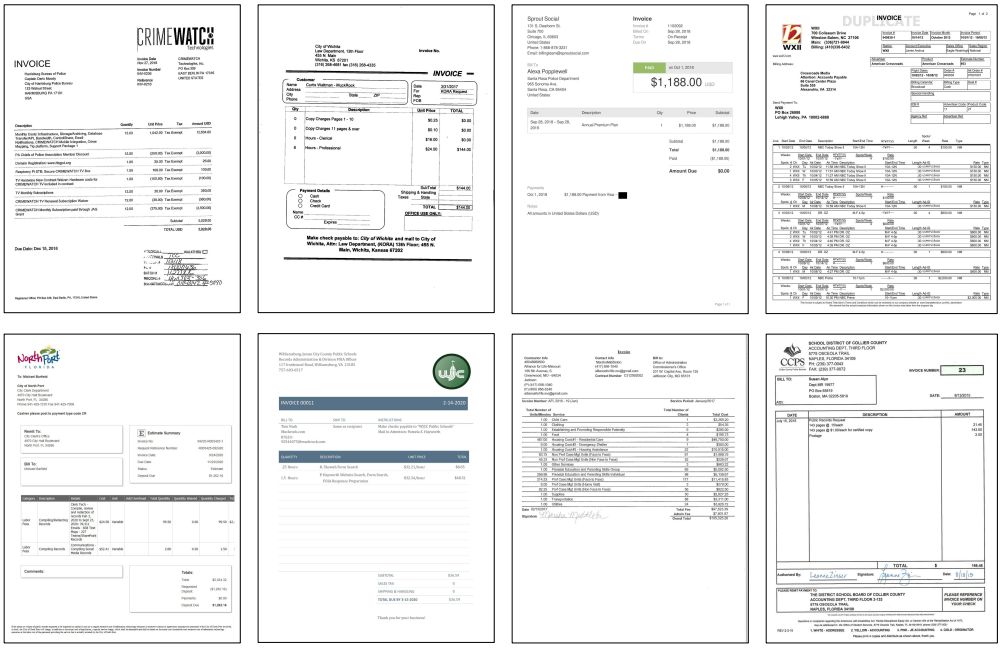

Figure 20: Samples of `invoice` documents from RVL-CDIP-*N*.

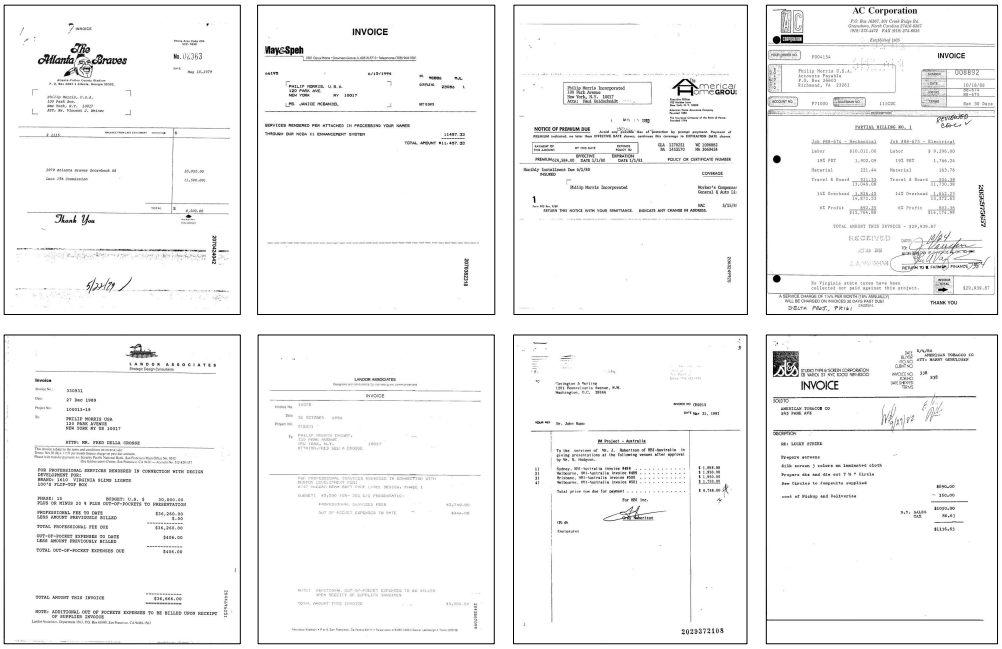

Figure 21: Samples of `invoice` documents from RVL-CDIP.

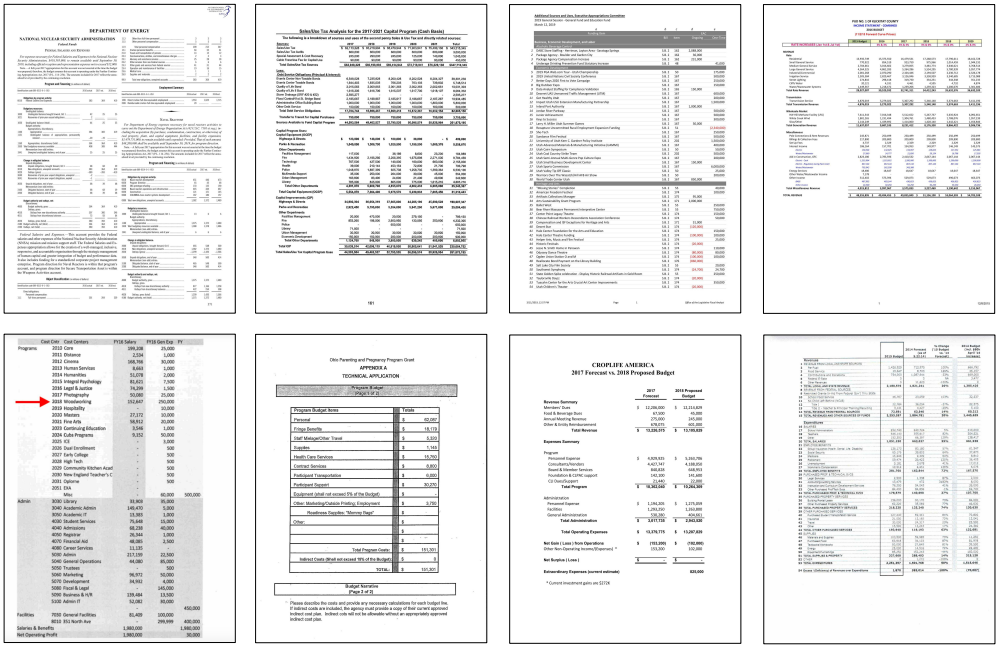

Figure 22: Samples of budget documents from RVL-CDIP-*N*.

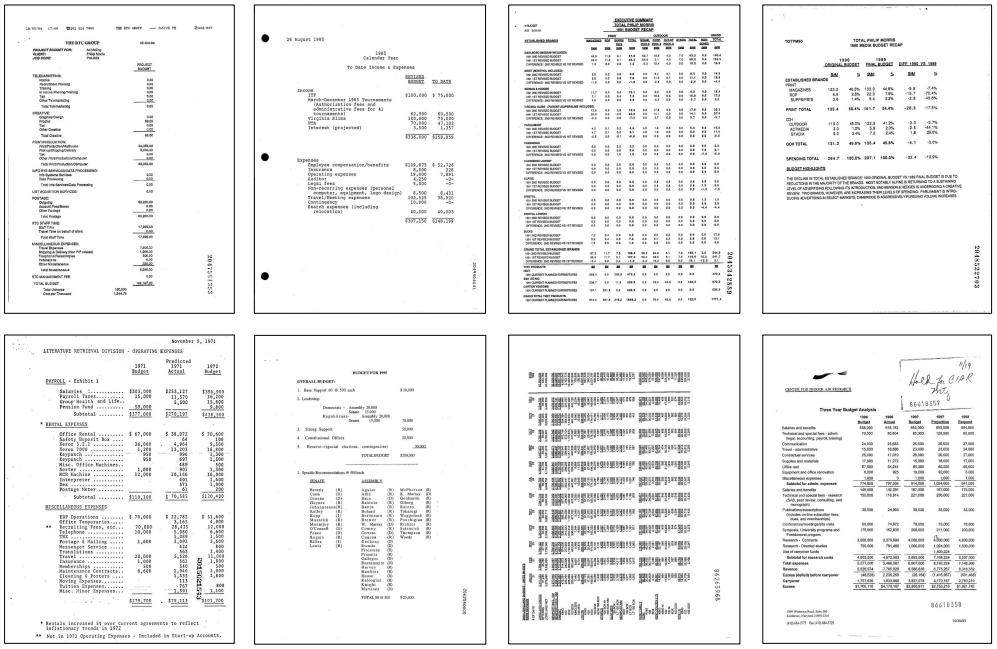

Figure 23: Samples of budget documents from RVL-CDIP.

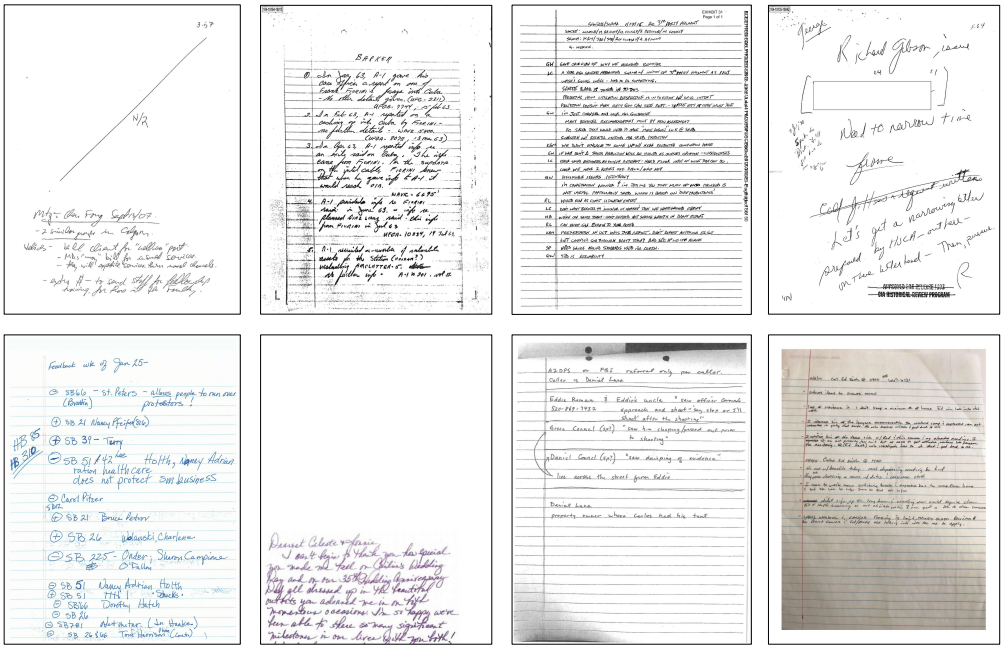

Figure 24: Samples of `handwritten` documents from RVL-CDIP-*N*.

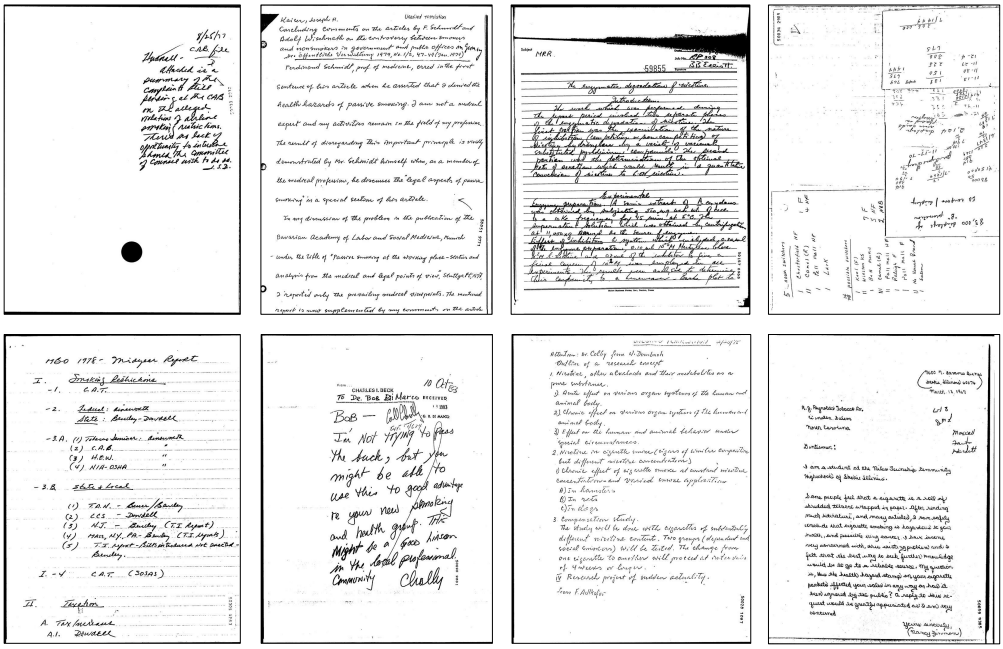

Figure 25: Samples of `handwritten` documents from RVL-CDIP.

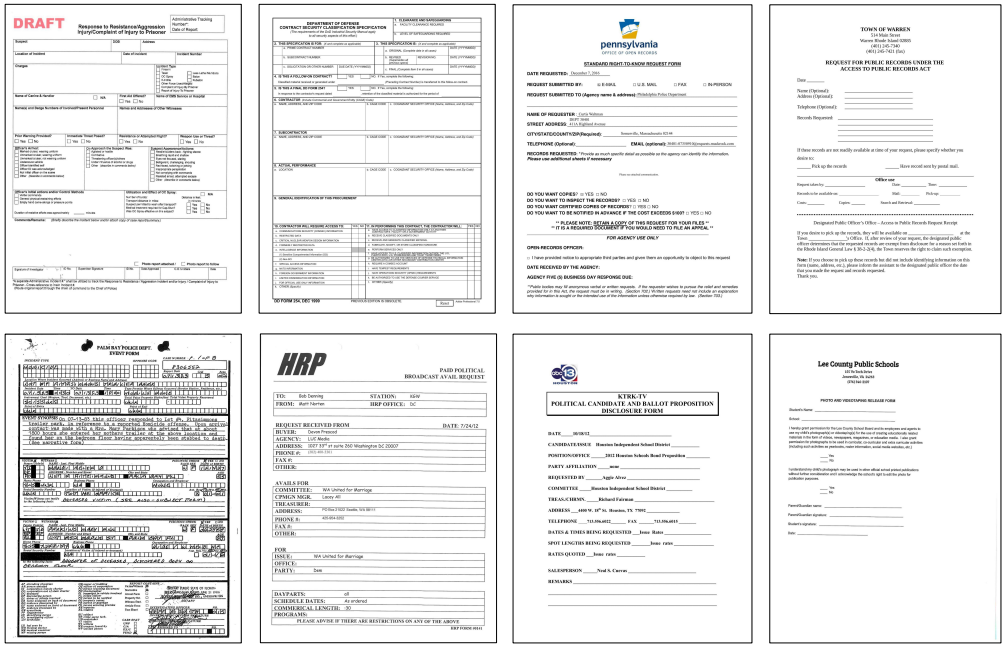

Figure 26: Samples of `form` documents from RVL-CDIP-*N*.

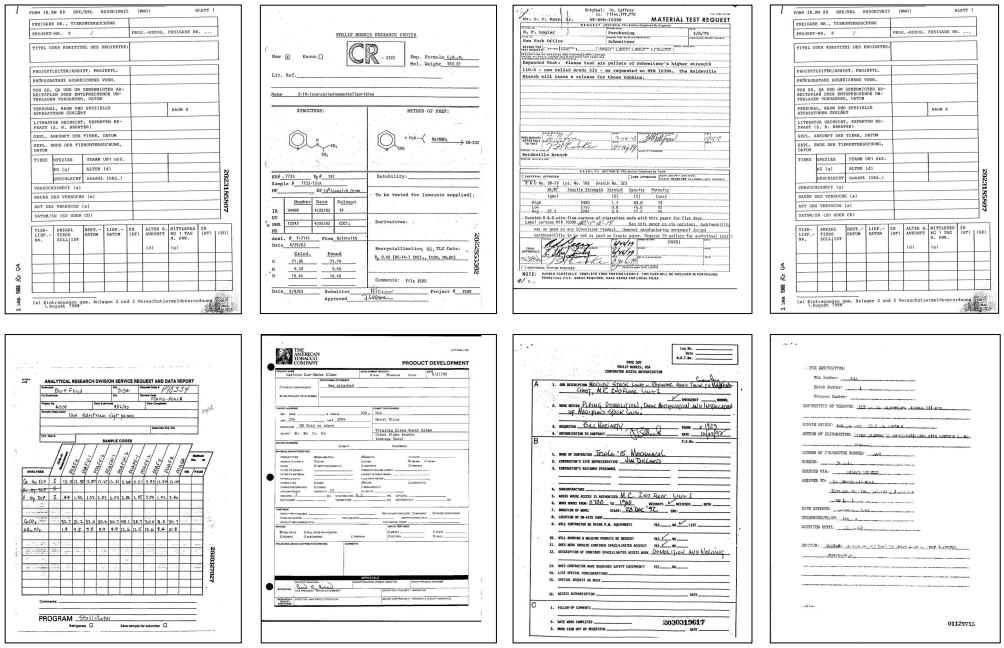

Figure 27: Samples of `form` documents from RVL-CDIP.

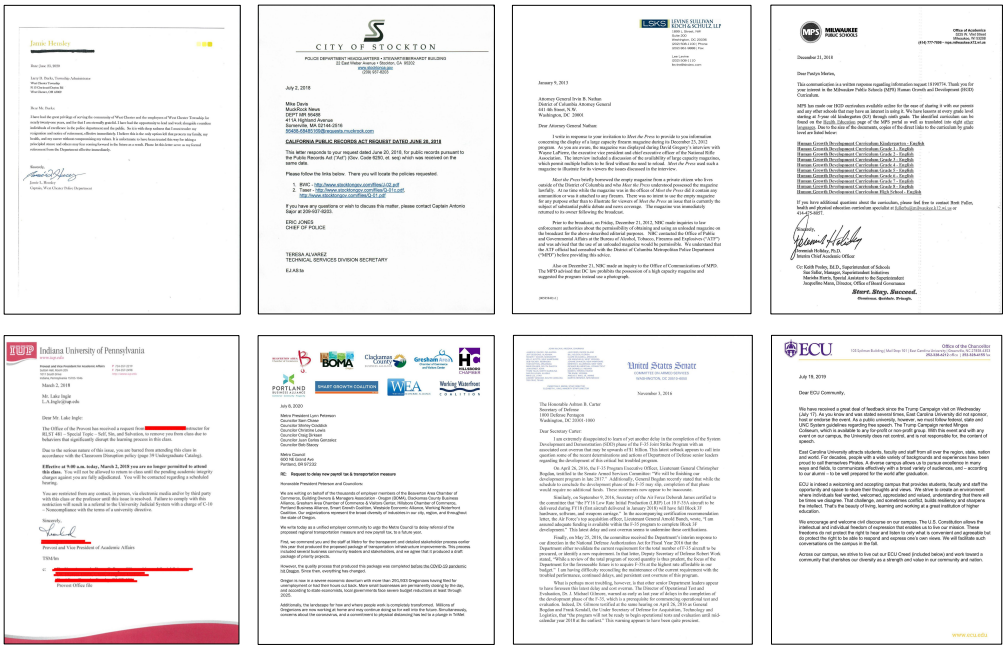

Figure 28: Samples of `letter` documents from RVL-CDIP-*N*.

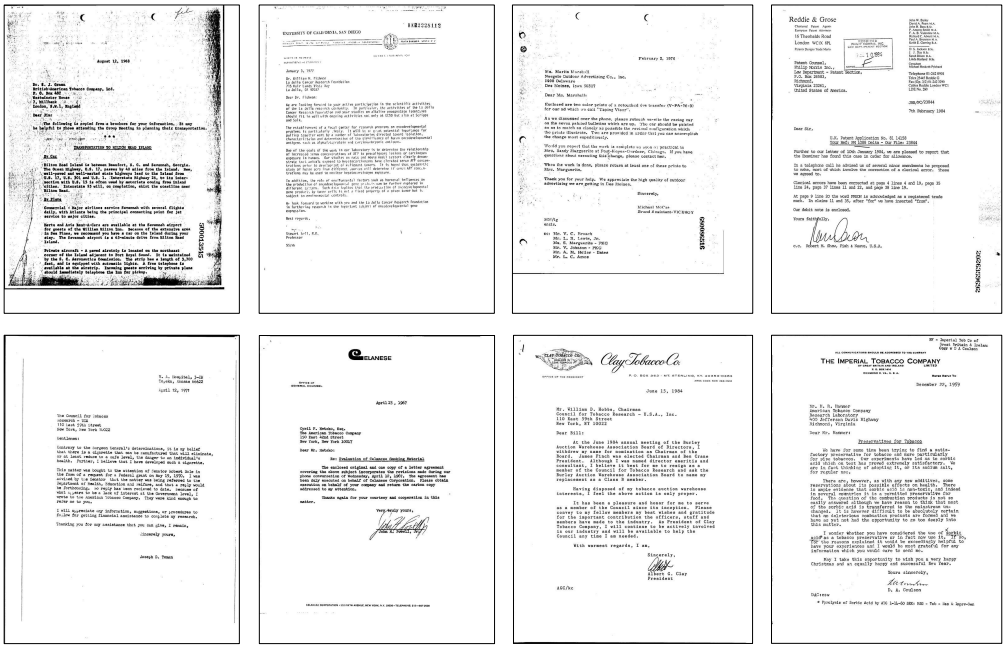

Figure 29: Samples of `letter` documents from RVL-CDIP.

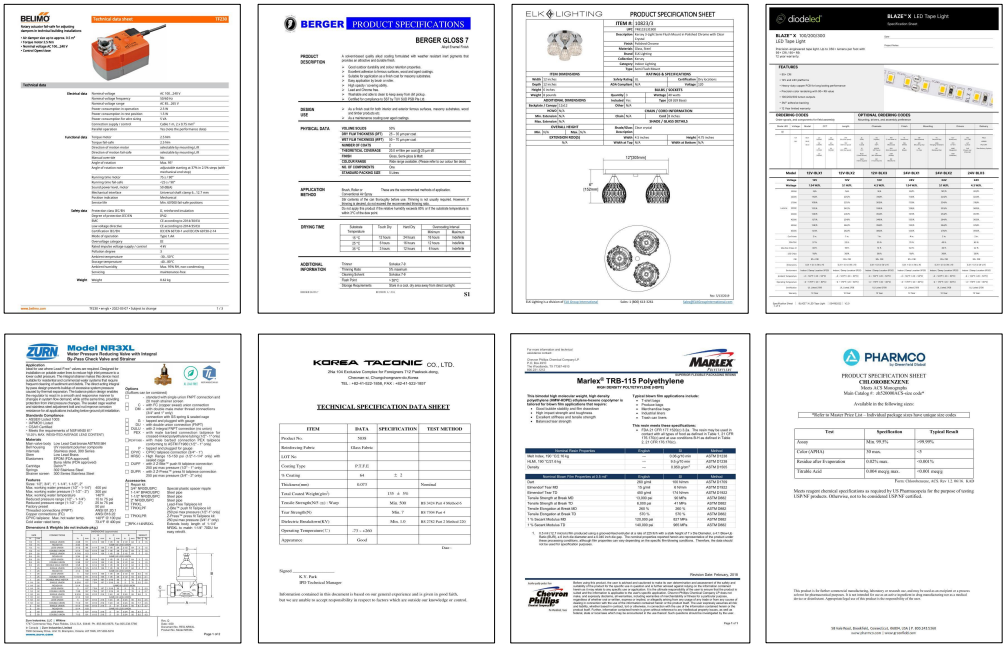

Figure 30: Samples of `specification` documents from RVL-CDIP-*N*.

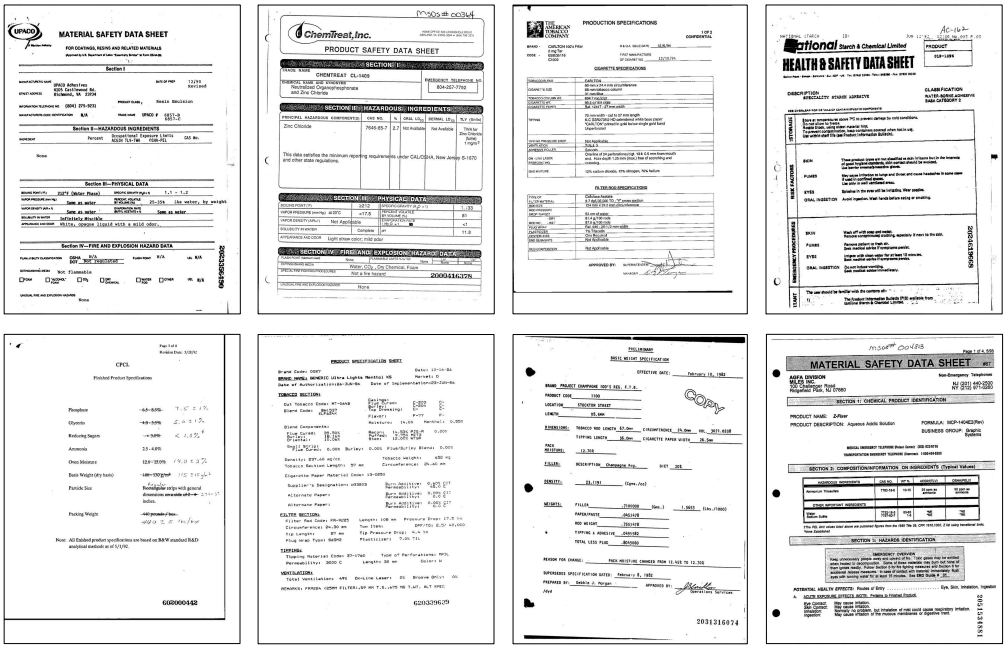

Figure 31: Samples of `specification` documents from RVL-CDIP.

Figure 32: Samples of `news_article` documents from RVL-CDIP-*N*.

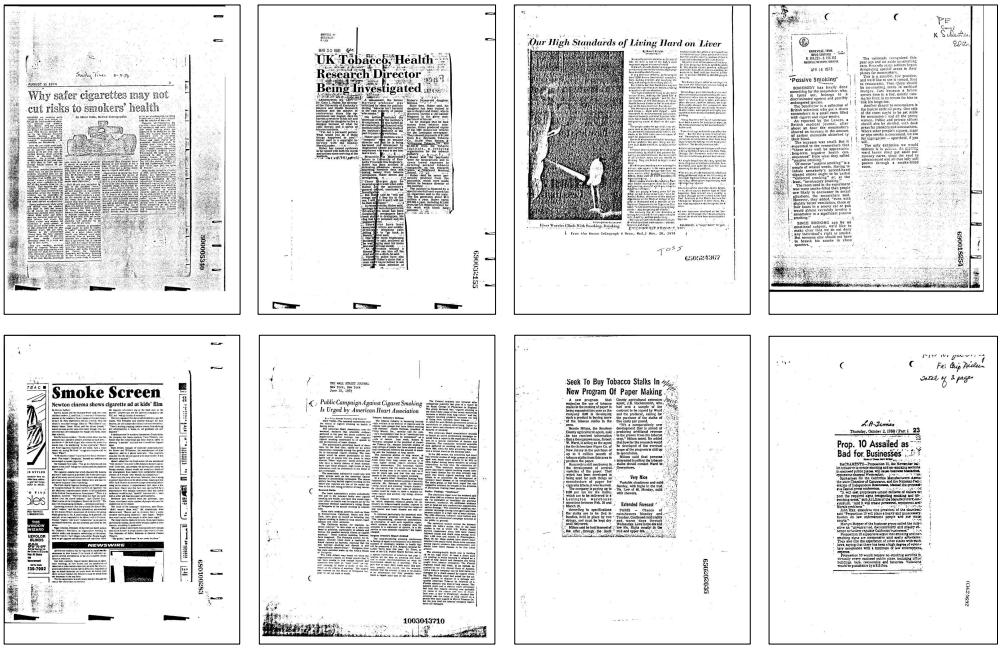

Figure 33: Samples of `news_article` documents from RVL-CDIP.

Figure 34: Samples of `scientific_publication` documents from RVL-CDIP-*N*.

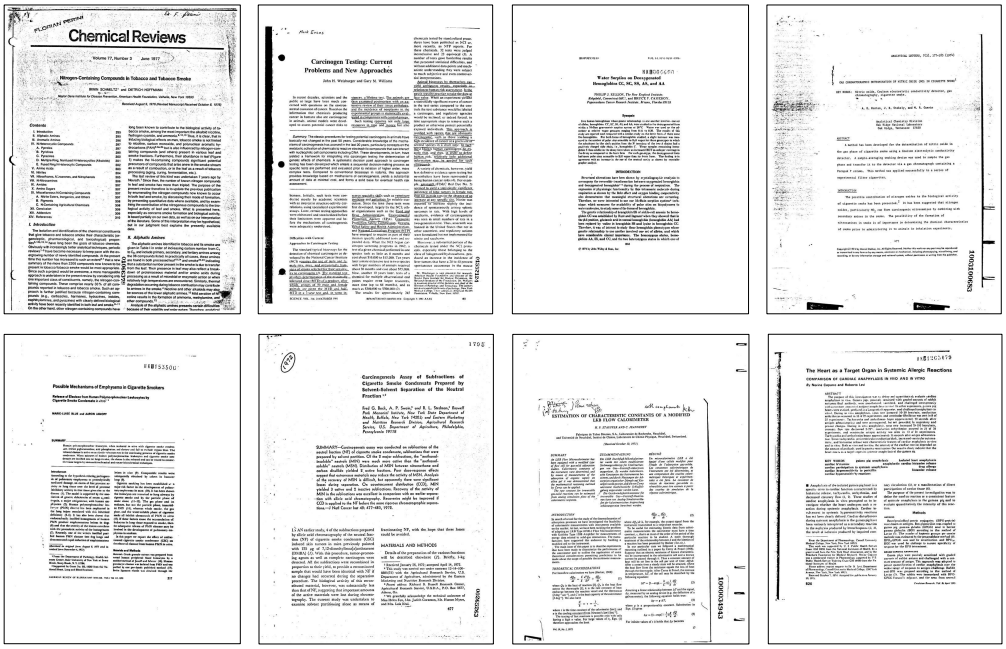

Figure 35: Samples of `scientific_publication` documents from RVL-CDIP.

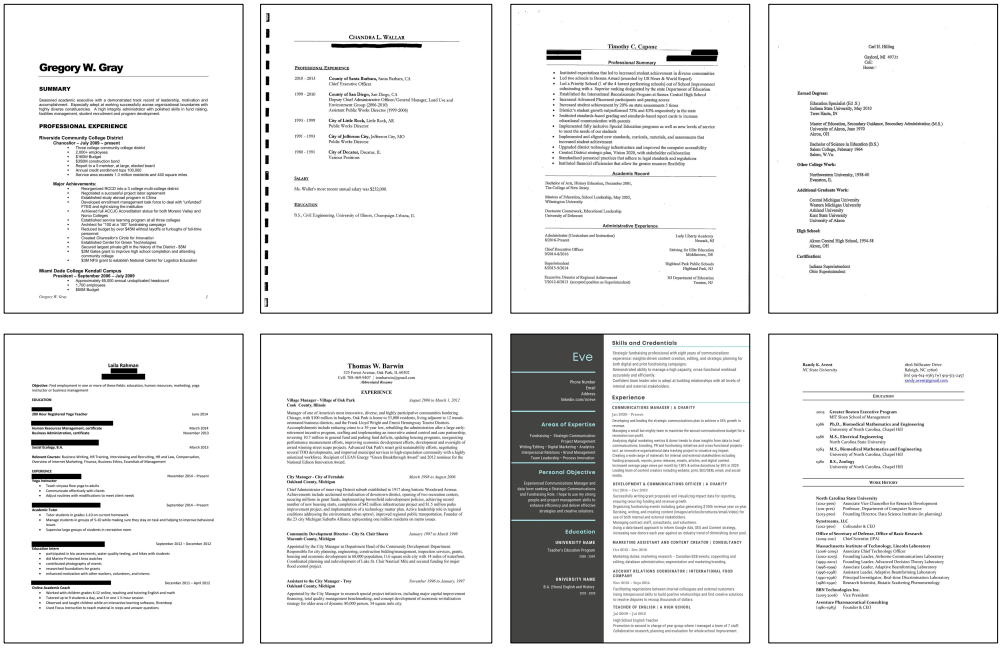

Figure 36: Samples of `resume` documents from RVL-CDIP-*N*.

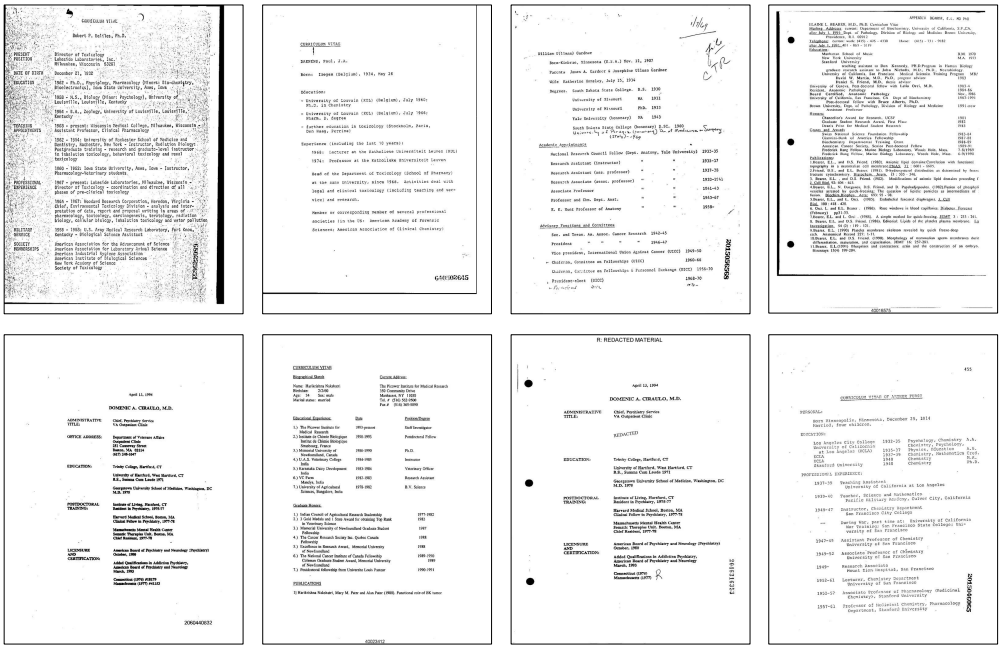

Figure 37: Samples of `resume` documents from RVL-CDIP.

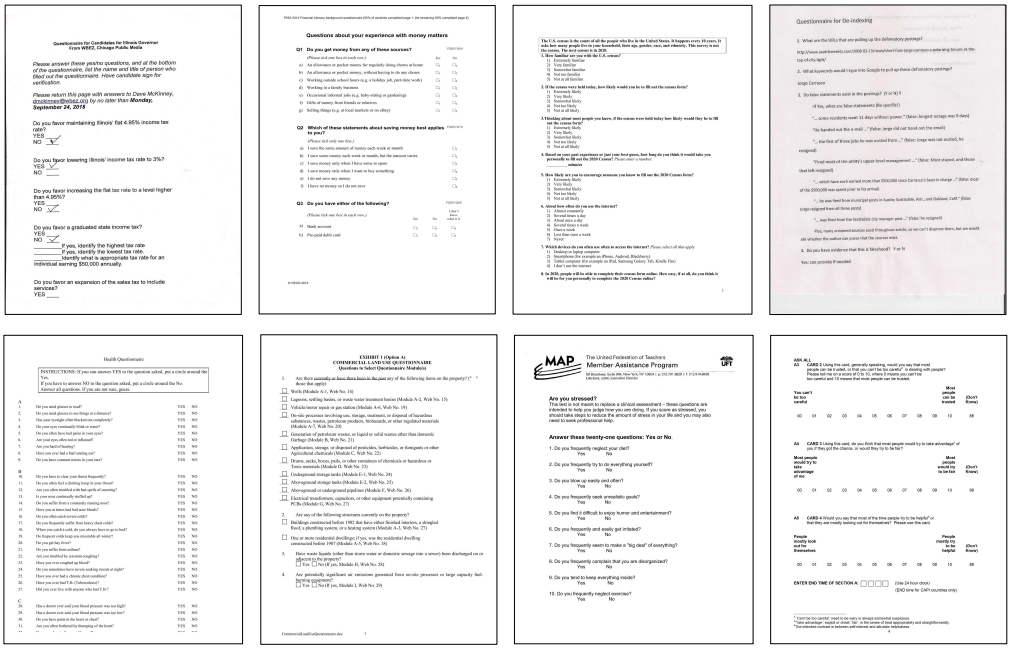

Figure 38: Samples of `questionnaire` documents from RVL-CDIP-*N*.

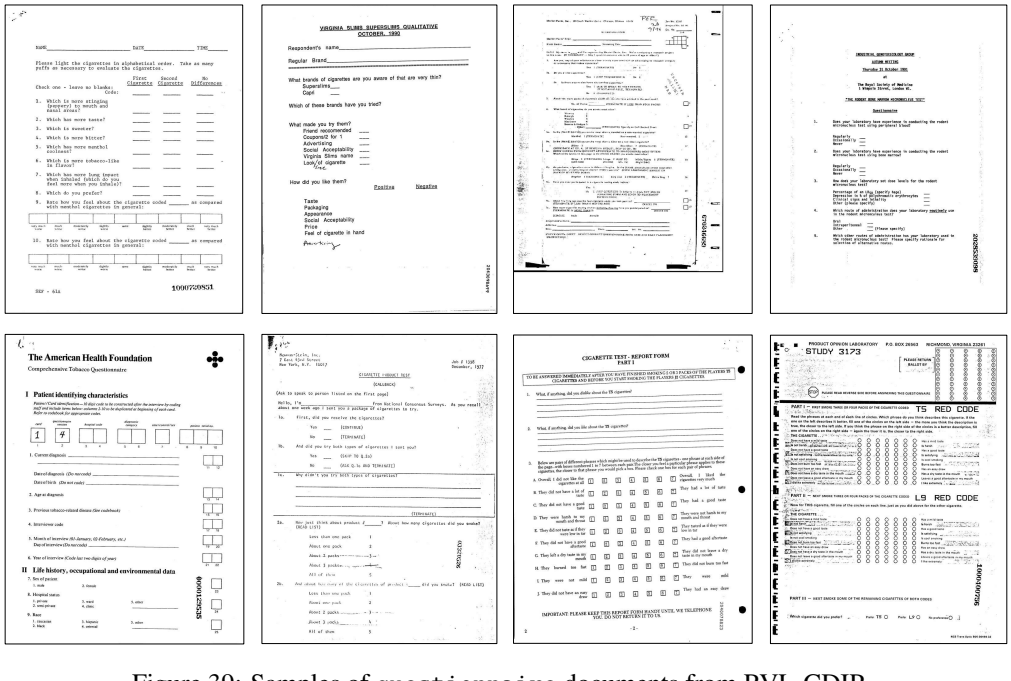

Figure 39: Samples of `questionnaire` documents from RVL-CDIP.

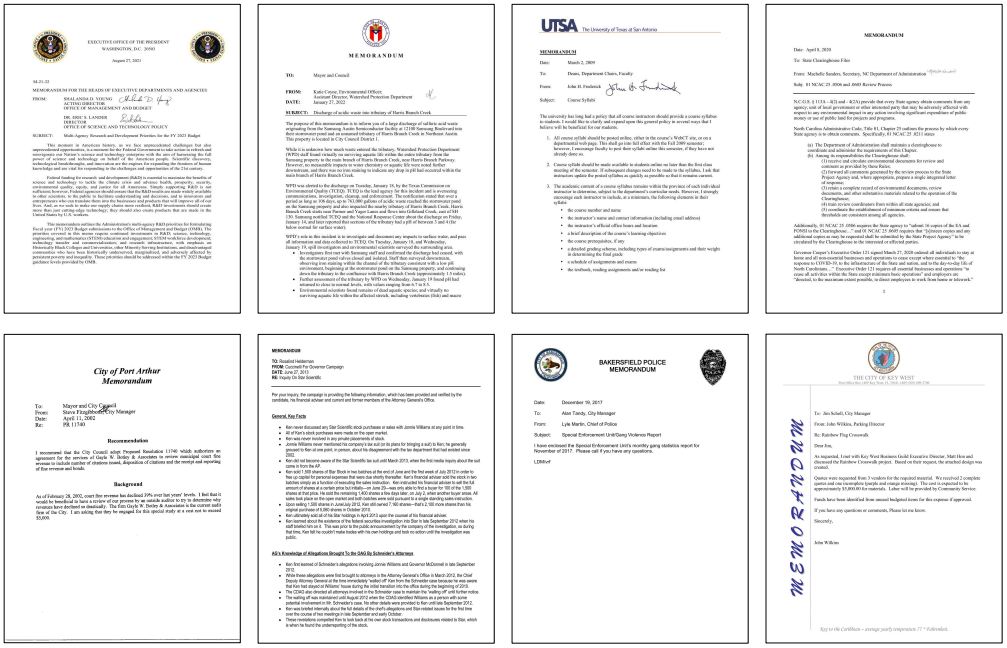

Figure 40: Samples of `memo` documents from RVL-CDIP-*N*.

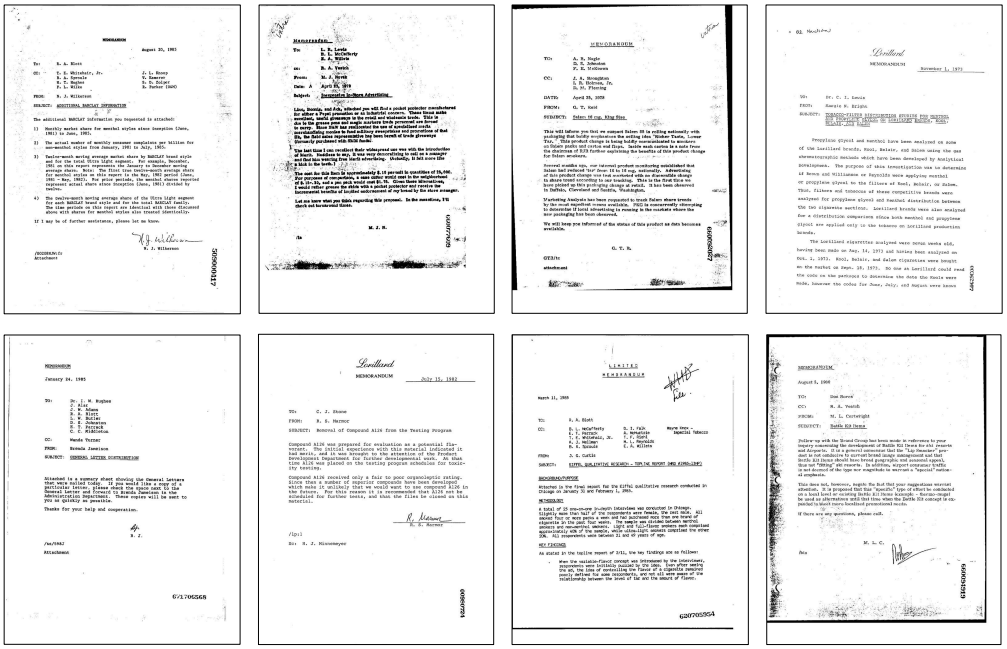

Figure 41: Samples of `memo` documents from RVL-CDIP.

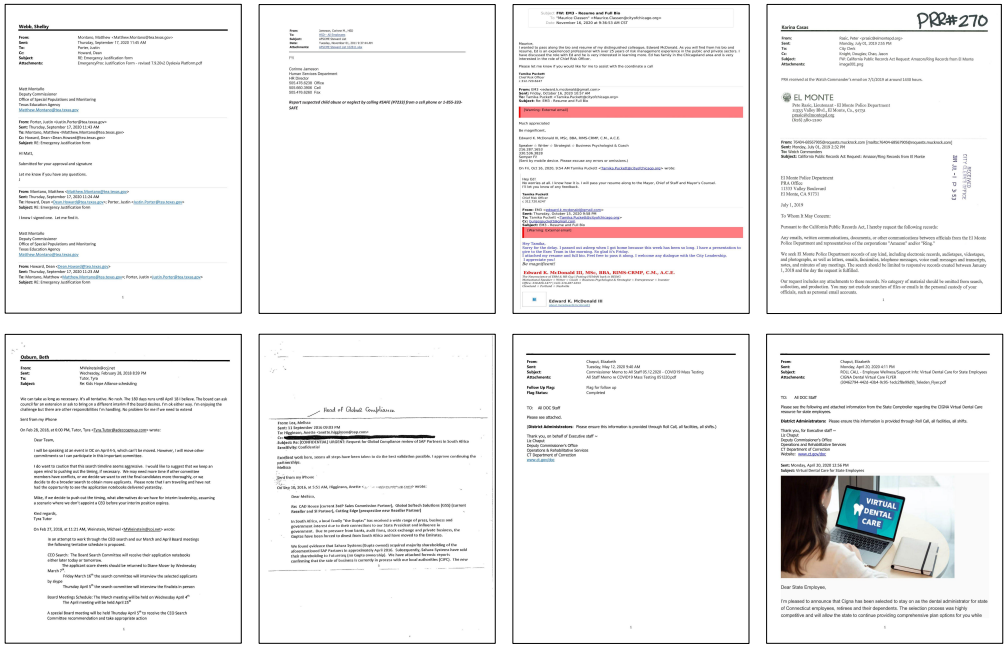

Figure 42: Samples of `email` documents from RVL-CDIP-*N*.

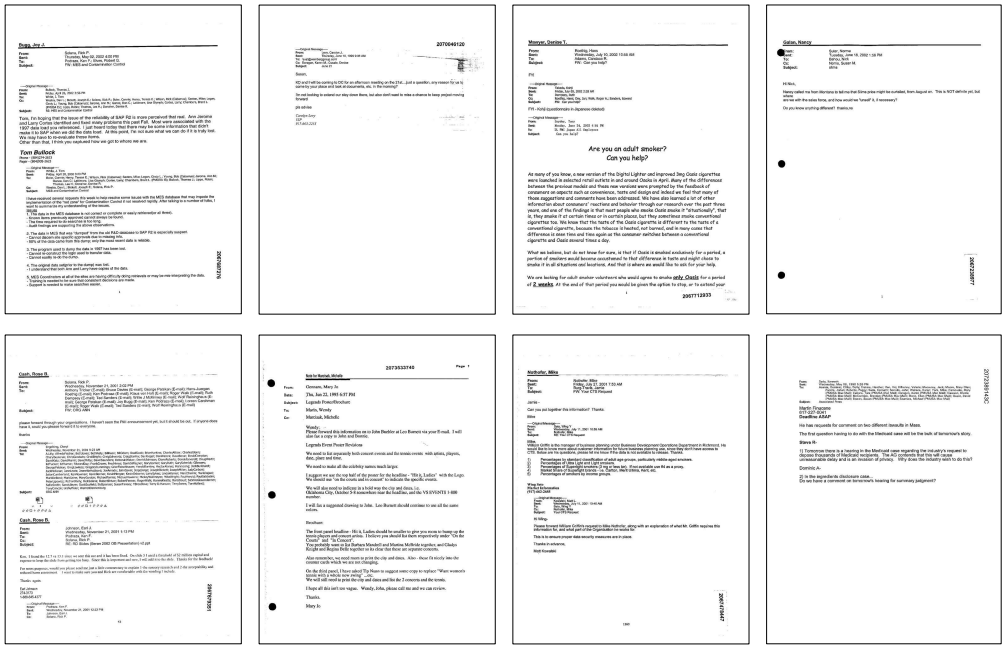

Figure 43: Samples of `email` documents from RVL-CDIP.