Behind the Smoke: Misleading mistakes on the Tobacco3482 Document Benchmark

Gordon Lim[1], Stefan Larson[2], Kevin Leach[2]
gbtc@umich.edu, [1]University of Michigan, [2]Vanderbilt University

1. Abstract

Tobacco3482 [1] is a widely-used document classification benchmark dataset. However, our manual inspection of the entire dataset uncovers widespread ontological issues. We analyze the mistakes of a top-performing model made and found that 35% of the supposed mistakes are not actually mistakes. Rather, the model predicted alternative valid labels or was presented with images that do not belong to any of the dataset's categories. Additionally, we report the incorrect label rate and instances of personally identifiable information (PII) discovered during our manual inspection. Supplementary material, including dataset annotations and code, is available at https://github.com/gordon-lim/tobacco3482-mistakes.

2. Tobacco3482

Tobacco3482 is a dataset of 3,482 document images, each labeled as one of 10 categories: *Advertisement* (ADVE)*, Email, Form, Letter, Memo, News, Note, Report, Resume,* or *Scientific*. These images were collected from the IIT Complex Document Information Processing (IIT-CDIP) Test Collection, which in turn sources its documents from the Truth Tobacco Industry Documents (TTID). None of Tobacco3482, IIT-CDIP, or TTID report formal labeling guidelines which raises concerns about the consistency and reliability of the labels within the Tobacco3482 dataset. During an initial review of the Tobacco3482 dataset, we observed several instances of unusual categorization within the dataset. Here, we provide some examples in the *Reports* category.

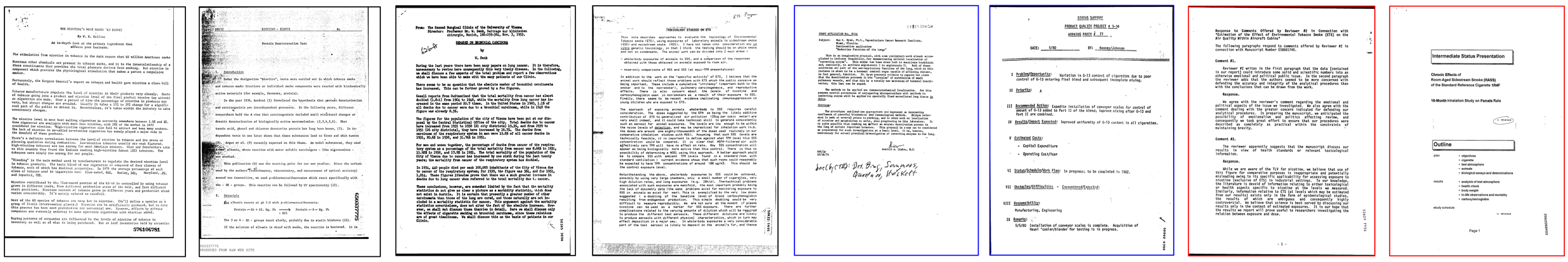


Figure 1. Example documents in the *Report* category Traditional report examples are boxed in black. These documents have a structured format with sections such as introduction, methodology, and results. Ambiguous report examples are boxed in blue (left: Grant application detailing a research proposal; right: Status summary for a product quality project). Mis-labeled report examples are boxed in red (left: response to reviewers; right: presentation slides).

3. Methods

Based on our initial review of the dataset, we exercised discretion in assigning new labels when category boundaries were unclear. We documented the process and provide a brief version of our guidelines in our paper. Informally, we were lenient towards the given Tobacco3482 label when the documents somewhat fit the definition of the category (blue examples in Figure 1). However, we strictly removed the label when the document image was far from the definition (red examples in Figure 1).

Given our corrected label annotations, we aimed to evaluate the impact on a model's evaluation. Specifically, we wanted to analyze the model's incorrect predictions and quantify how many of them were alternative valid labels or predictions made on images that did not belong to any of Tobacco3482's label categories. We used a DiT model [2], from HuggingFace. We selected this model because it is a top-performing model on RVL-CDIP [3], has not yet been benchmarked on Tobacco3482, and has accessible source code. To collect DiT's mistakes on the entire dataset, we performed our evaluation using 4-fold cross-validation, achieving an 84.1% top-1 accuracy.

4. Results

Overlapping categories. We identified 583 samples (16.7% of the dataset) that should have multiple Tobacco3482 category labels. Figure 2 compares the overlap between different Tobacco3482 categories in our multi-label annotations. The *Memo* and *Report* categories exhibit the most overlap, as reports can often be written in memo format for communication purposes. The *Report* and *Scientific* categories also have a lot of overlap, since reports can document results from scientific studies. Figures 3 and 4 provide examples of these overlapping categories.

Label errors. We also report 409 label errors, which corresponds to an 11.7% label error rate. This includes 258 labels that matched none of our newly-assigned labels, i.e., mis-labeled. Additionally, there were 151 document images that do not fall within any of Tobacco3482's label categories. We present examples in Figure 5.

DiT Mistakes. Following the above, we found that of the 554 so-called mistakes made by the DiT model on Tobacco3482, 196 of them were not actually mistakes. Specifically, in 147 instances, the model predicted an alternative valid label from our multi-label annotations. Additionally, in 49 instances, the model faced the unfair and impossible challenge of predicting an unknown label. If we consider these mistakes as correct, the DiT model's accuracy would increase to 89.7%. This marks a 5.7% improvement from the original 84.0% accuracy and brings it closer to a more recent method that achieved 90.7% accuracy using four times more parameters.

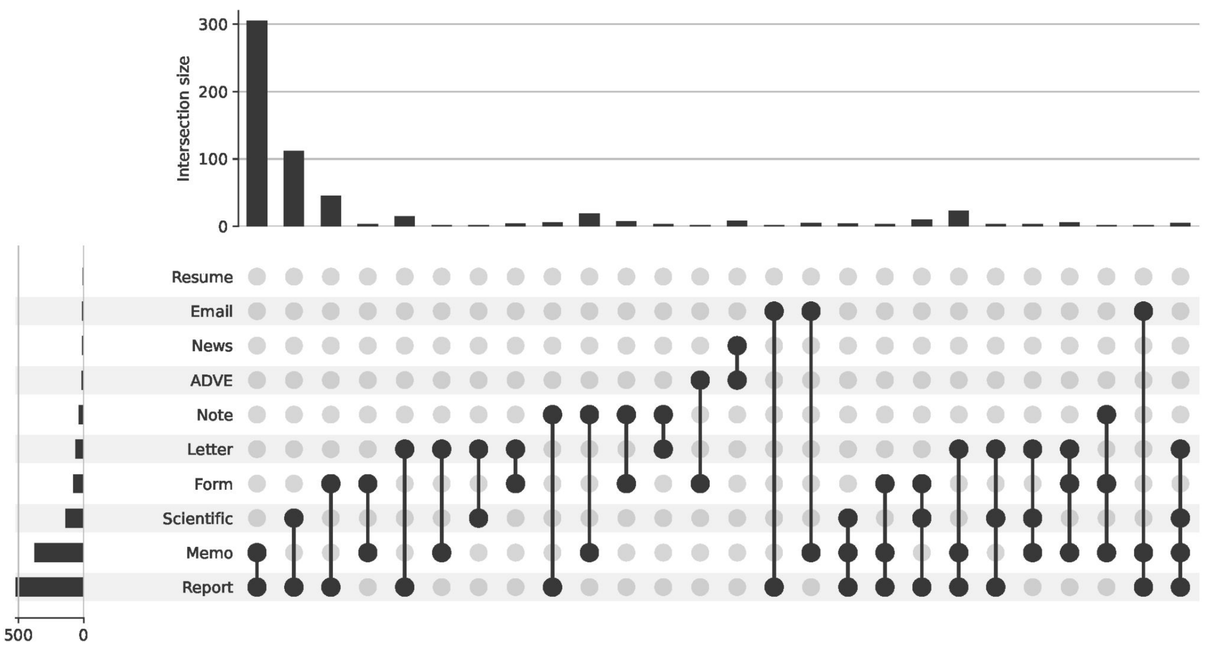
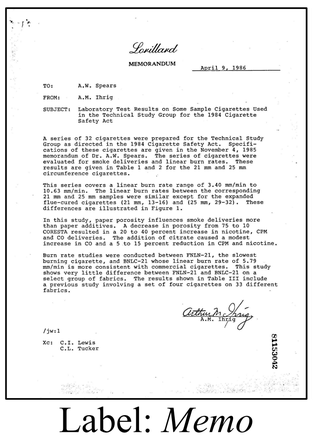
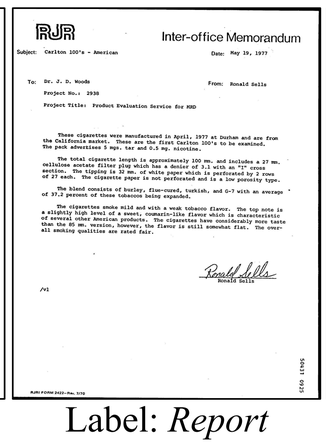


Figure 2. UpSet plot [4] of multi-label Tobacco3482 category annotations. A majority of the documents with multiple labels are documents that are both *Reports* and *Memos*.

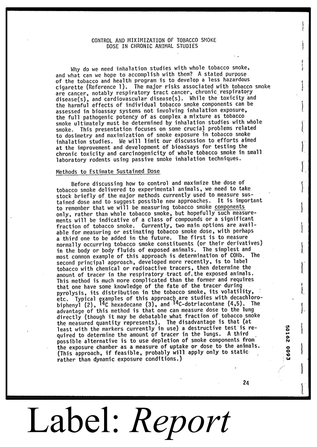


Label: *Memo*
Also: *Report*

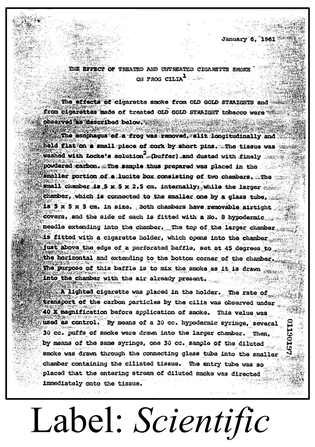


Label: *Report*
Also: *Memo*

Figure 3. Overlap of Memo and Report categories. Both are reports written in memos.



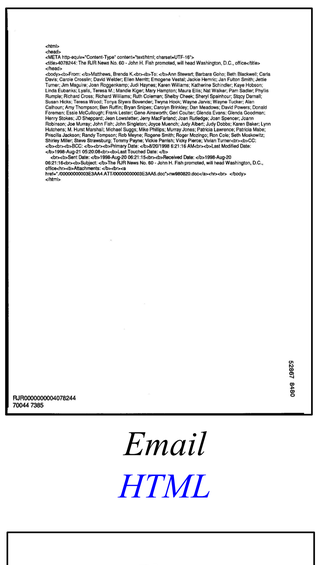
Label: *Report*
Also: *Scientific*



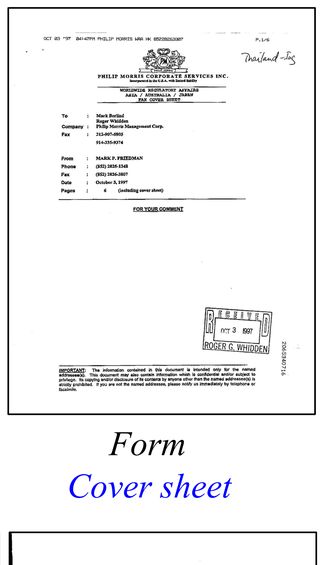
Label: *Scientific*
Also: *Report*

Figure 4. Overlap of Report and Scientific categories. Both are reports on scientific studies.

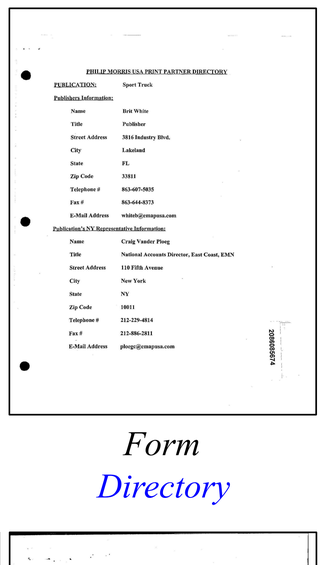
ADVE
Broken



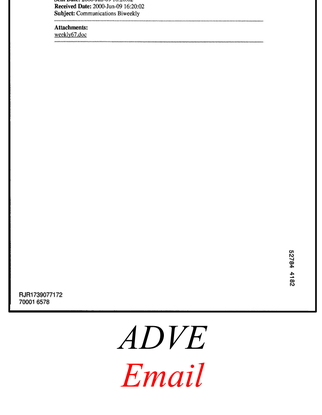
Email
HTML



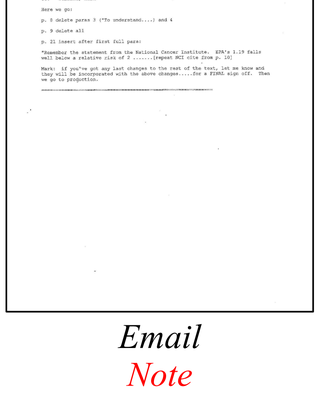
Form
Cover sheet



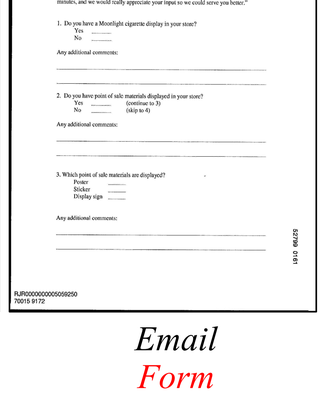
Form
Directory



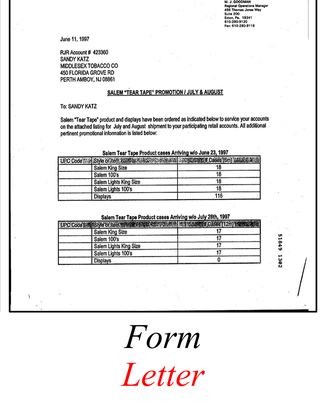
ADVE
Email



Email
Note



Email
Form



Form
Letter

Figure 5. Label errors in Tobacco3482. Tobacco3482 labels are shown in black. The top row shows examples with unknown labels (not one of Tobacco3482 labels) and suggested labels in blue, while the bottom row shows examples with mislabels and corrected Tobacco3482 labels in red.

5. Discussion, Future Work, and Conclusion

This short paper represents a first effort to measure the extent of label errors in the Tobacco3482 dataset, adding to the relevant literature alongside papers that do the same for other datasets, such as RVL-CDIP. We also present early evidence showing how such errors can make model accuracy scores on these benchmarks misleading, as many mistakes were not actually mistakes. As model performance approaches near-perfect levels (>95%), it is crucial that these accuracy metrics reflect a model's ability to learn meaningful features seen in the real world, rather than spurious features in the dataset.

Future work will build on our efforts by benchmarking more models on Tobacco3482 after addressing the issues we have documented, such as by using multi-label accuracy and removing label errors. Additionally, we plan to apply the same analysis to RVL-CDIP, as the paper highlighting errors in RVL-CDIP has not yet conducted benchmarking that takes these issues into account.

1. Jayant Kumar, Peng Ye, and David S. Doermann. 2014. Structural similarity for document image classification and retrieval. Pattern Recognition Letters, 43:119–126.
2. Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, Furu Wei. 2022. Dit: Self-supervised pre-training for document image transformer. Proceedings of the 30th ACM International Conference on Multimedia
3. Adam Harley, Alex Ufkes, Konstantinos Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. Proceedings of the International Conference on Document Analysis and Recognition.
4. Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. 2014. Upset: Visualization of intersecting sets. IEEE Transactions on Visualization and Computer Graphics 20(12), 1983–1992