Label Errors in the Tobacco3482 Dataset

Gordon Lim
University of Michigan
Ann Arbor, MI, USA
gbtc@umich.edu

Stefan Larson
Vanderbilt University
Nashville, TN, USA
stefan.larson@vanderbilt.edu

Kevin Leach
Vanderbilt University
Nashville, TN, USA
kevin.leach@vanderbilt.edu

Abstract

Tobacco3482 is a widely used document classification benchmark dataset. However, our manual inspection of the entire dataset uncovers widespread ontological issues, especially large amounts of annotation label problems in the dataset. We establish data label guidelines and find that 11.7% of the dataset is improperly annotated and should either have an unknown label or a corrected label, and 16.7% of samples in the dataset have multiple valid labels. We then analyze the mistakes of a top-performing model and find that 35% of the model's mistakes can be directly attributed to these label issues, highlighting the inherent problems with using a noisily labeled dataset as a benchmark. Supplementary material, including dataset annotations and code, is available at github.com/gordon-lim/tobacco3482-mistakes/.

1. Introduction

Document classification is a fundamental component of document processing pipelines used by organizations for efficient search and retrieval [4]. Recent advances in document classification have demonstrated continuous improvement in performance on the Tobacco3482 [19] document image dataset (e.g., [1,5,13,16]). However, a growing body of research on dataset quality casts serious doubt on the usefulness of many benchmark datasets for evaluating model performance [2,6,7,18]. Notably, widely-used datasets like ImageNet [3] has been found to have labeling issues, including incorrect labels and images that should be assigned multiple labels [12,14,15]. Such issues destabilize benchmarks and make low capacity models seem less performant than they are [15]. To our knowledge, no studies have yet quantified the extent of label issues in the Tobacco3482 dataset.

This short paper presents the first examination of the Tobacco3482 dataset for data label issues. We conduct a thorough review of the Tobacco3482 dataset, and find that 11.7% of samples from the dataset are either mislabeled or do not belong to any of Tobacco3482's labels,
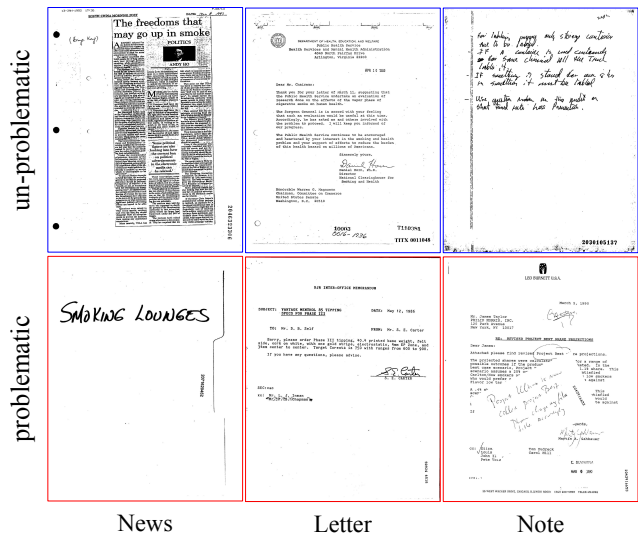

Figure 1. Top row: un-problematic document images from Tobacco3482. Bottom row: samples from Tobacco3482 that are erroneously labeled (left and center) or could have multiple valid labels (right). Bottom left: the document is not a news article. Bottom center: the document should be labeled as Memo and not Letter. Bottom right: the image contains both a letter (background) and a note (foreground), and thus could have two valid labels (Note and Letter).

and 16.7% of samples from the dataset have multiple valid labels. Examples of data labeling issues are shown in Figure 1. To contextualize the impact of these data labeling issues on model evaluation, we analyze the mistakes of a top-performing transformer-based document image classification model and find that 35% of the model's mistakes are actually valid alternative labels or images that were not represented by any of the dataset's labels. Overall, our findings highlight flaws in the Tobacco3482 dataset concerning data quality.

Category	Label Guidelines
ADVE	Media targeted towards a general audience to promote a product, service, or agenda. Can be found in newspapers. Excludes magazine covers and content drafts.
Email	Electronic mail, often including "Sent via electronic mail" or timestamps. Common features are "To/From", "Subject", "cc" fields, and digital attachments. Excludes HTML code.
Form	Documents with spaces (boxes, lines, etc.) to be filled in. Examples include registration forms, report templates, and data sheets. Excludes documents that print "Form" but do have fillable spaces.
Letter	Messages delivered on physical paper with physical addresses. Features include a letterhead, recipient's name, address, salutation (e.g., "Dear xxx"), and sign-off (e.g., "Sincerely,"). Excludes *memos* using "letter" in its subject.
Memo	Documents addressed to an organization or person, often denoted by "To/From". They might state "Memo", "Memorandum", or "Correspondence". Typical content include policy changes, meeting schedules, event invitations, recommendations, or actions. Excludes documents that print "File note" not addressed to anyone.
News	Articles resembling those in newspapers, typically with multiple columns and select font enlarged. They may also include newspaper clippings. Excludes scientific articles published in scientific news platforms.
Note	Handwritten or typed brief messages. Typically smaller than letter-sized paper and have minimal layout features. Excludes handwriting on other documents and *Reports* with "Note" in title.
Report	Formal documents presenting information, often with section headers such as introduction, methodology, and results. Topics can be scientific or non-scientific and might include data reports, findings, or series of events. May also presented using bullet lists or tables. Excludes documents that are primarily applications, requests, or proposals.
Resume	Formal documents containing personal career and/or education information, often titled "Curriculum Vitae", "Resume" or "Biographical Sketch". They include personal information like name, title, birth date, education, and professional experience.
Scientific	Documents containing scientific knowledge, typically written by scientists, often featuring scientific terms, mathematical formulas, and section headers like Abstract. Examples include research papers, grant applications, and project reports. Excludes business reports and product development reports.

Table 1. Guidelines and examples illustrating how we assigned new labels for the Tobacco3482 dataset.

2. A Review of Tobacco3482

In this section we first discuss the background and content of Tobacco3482, then we conduct a review of the dataset in order to establish label guidelines.

2.1. Background on Tobacco3482

Tobacco3482 [19] consists 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 sampled from the IIT Complex Document Information Processing (IIT-CDIP) Test Collection, which in turn sources its documents from the Truth Tobacco Industry Documents (TTID).[1] TTID is a document collection containing internal documents from major US tobacco companies, released as part of legal settlements [9]. Documents in TTID have one or multiple manual document category annotations[2] that include the 10 Tobacco3482 categories. Using TTID's API, we searched for Tobacco3482 documents by searching for the Bates numbers in their filenames, resulting in the identification of 1,707 documents. Of these, 134 documents (3.8% of the entire Tobacco3482 dataset) have multiple Tobacco3482 category annotations. This indicates that at least some of the documents from Tobacco3482 may have valid alternate labels. Despite this, the Tobacco3482 dataset still assigns only a single label to all of its documents. Furthermore, there are no formal labeling guidelines reported by Tobacco3482, IIT-CDIP, or TTID, which raises concerns about the consistency and reliability of the labels within the Tobacco3482 dataset.

2.2. Data Label Guidelines and Review

We seek to quantify the amount of label errors in the Tobacco3482 dataset. Since the Tobacco3482 dataset (and its parent TTID corpus) contains no labeling guidelines, we conducted an initial review of the Tobacco3482 dataset in order to establish annotation guidelines. We documented the process and provide a brief version of our guidelines, along with descriptive examples, in Table 1. This procedure was also conducted in prior work by Larson, et al. (2023) [7], who established label guidelines in order to analyze label error rates in the RVL-CDIP dataset.

Like Larson, et al. (2023) [7], we then reviewed the Tobacco3482 dataset using our guidelines, and re-annotated any document images that were problematically labeled. We tracked three types of problematic label types: (1) *unknown*, where the document image does not fit within the guidelines of any of the 10 Tobacco3482 categories; (2) *mis-labeled*, where the document image should be re-annotated with a different, more correct Tobacco3482 category; and (3) *multiple labels*, where the document image contains multiple valid category labels. Examples of each problematic label type are displayed in Figure 2. We quantify the amount of each problematic label type below.

Unknown and Wrong Labels. We found 151 documents that did not belong to any of the 10 Tobacco3482 categories according to our guidelines in Table 1. Additionally, we found 258 documents that were incorrectly labeled
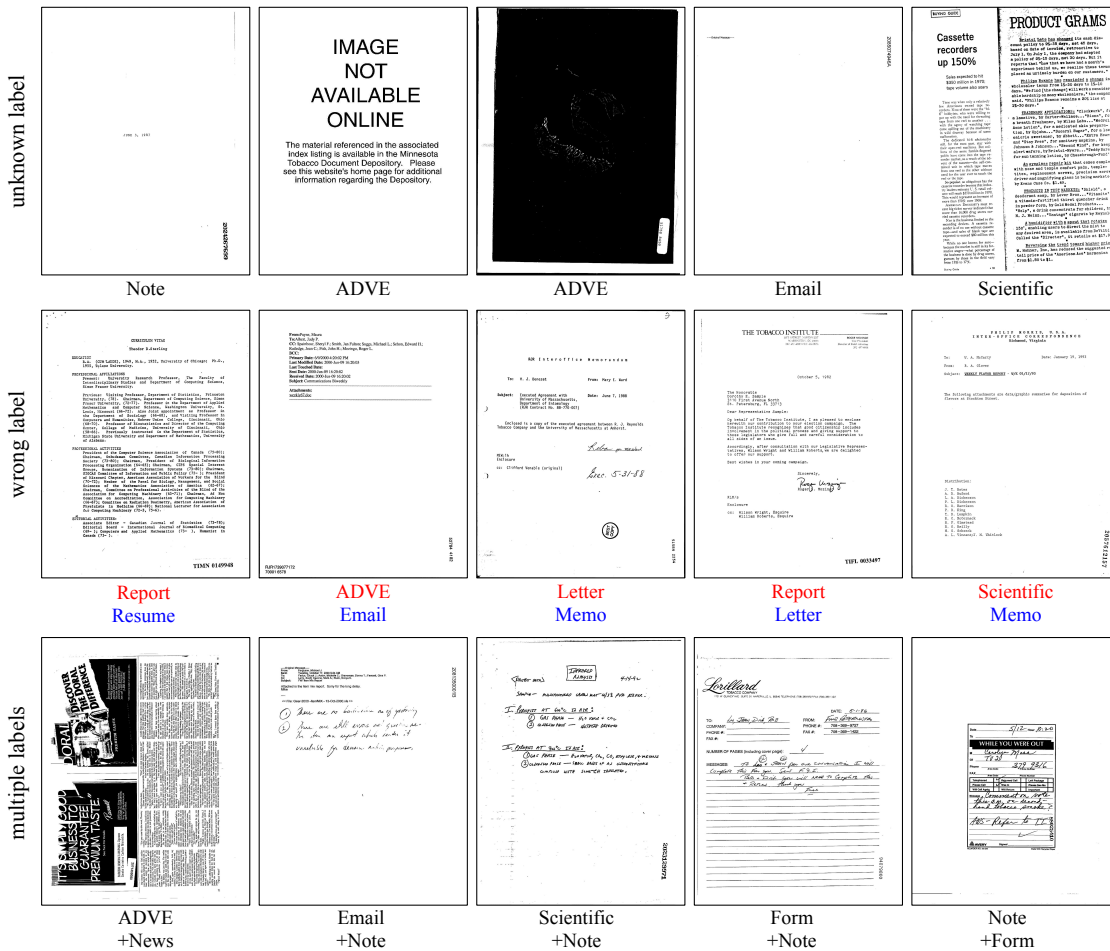

Figure 2. Examples of problematic samples from Tobacco3482. Top row: documents where the valid label is unknown. Middle row: documents that have the wrong original label (shown in red (top label) with corrected label shown in blue (bottom label)). Bottom row: documents that have multiple valid labels (original label shown on top, with additional valid label shown on bottom).

and have a valid alternative. Together, these two label issues account for 409 samples, or 11.7% of the entire Tobacco3482 dataset. Table 2 displays the rates of each label error type for each Tobacco3482 category. The Scientific category contains many errors; roughly 14.6% of its documents are unknown, and 23.8% documents are mis-labeled. Roughly 24% of the Letter category is mis-labeled, where most of the corrections should be to the Memo category (see the companion repository for a confusion matrix[3]).

Multiple Lables. We identified 583 samples (16.7% of the dataset) that should have multiple valid Tobacco3482 category labels. Figure 3 compares the overlap between different Tobacco3482 categories on 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 contain much overlap, since report-style documents can encapsulate the results of scientific studies, and vice-versa. (See the companion repository for examples.)

Category	Num. Samples	% Unknown	% Mis-Label
ADVE	230	7.39	2.17
Email	598	2.51	2.68
Form	431	9.05	4.18
Letter	567	0.35	24.0
Memo	620	0.16	0.48
News	188	3.19	0.00
Note	201	6.97	4.48
Report	265	7.17	3.40
Resume	119	0.00	0.00
Scientific	261	14.6	23.8

Table 2. Number of samples for each category, and label error rates for unknown and mis-labeled samples in Tobacco3482.

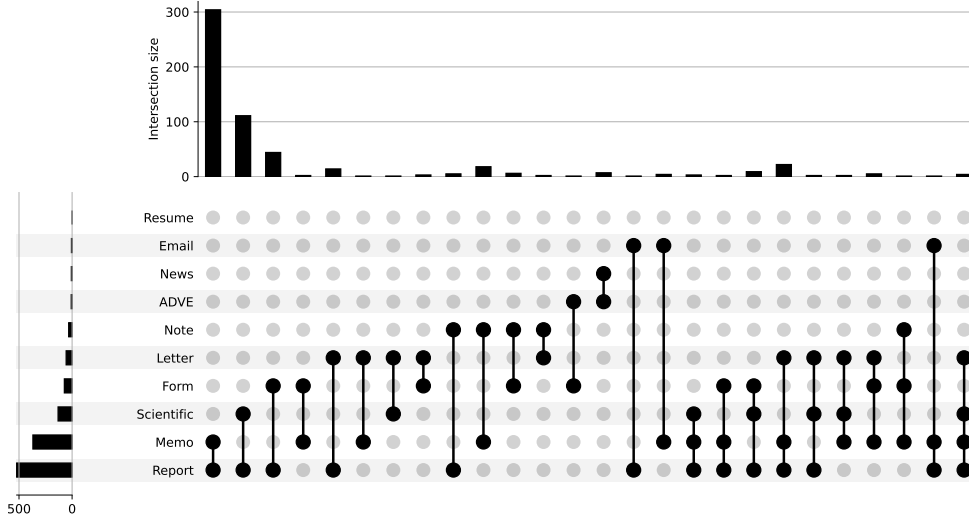[3] github.com/gordon-lim/tobacco3482-mistakes/

Figure 3. UpSet plot [10] of multi-label Tobacco3482 category annotations. A majority of the documents with multiple labels are documents that are both Reports and Memos.

3. Impact of Label Issues on the Classification Task

To contextualize the impact of these data labeling issues on model evaluation, we analyze the mistakes of a DiT model [11]. We used a DiT model from Hugging Face[4], using RVL-CDIP [4] pre-trained weights. We selected this model because it is a top-performing model on RVL-CDIP, has not yet been benchmarked on Tobacco3482, and has accessible source code. To collect DiT's mistakes on the entire original (i.e., un-corrected) Tobacco3482 dataset, we performed our evaluation using 4-fold cross-validation, achieving a 84.1% top-1 accuracy.

Following the above, we found that of the 554 mistakes made by the DiT model on the un-corrected Tobacco3482 dataset, 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 [16].

4. Discussion

We inspected the Tobacco3482 dataset for problematic annotation labels and found there to be a substantial amount of various types of label issues. Within the document understanding realm, this finding is unfortunately shared in

common with the larger RVL-CDIP document classification corpus, in which Larson, et al. (2023) [7] observed similar label problems. Thus, published model performance on the original versions of Tobacco3482 and RVL-CDIP may not be indicative of a model's capabilities. In a recent orthogonal finding to ours, Saifullah, et al. (2024) [17] observed that both Tobacco3482 and RVL-CDIP contain significant feature biases like numeric codes in the document images that impact model performance. Like us, Saifullah, et al. raise concern over whether published model performance on the original Tobacco3482 dataset is a valid measure of a model's true capabilities. Moreover, recent work by Larson, et al. (2024) [8] found large amounts of sensitive personal information in both Tobacco3482 and RVL-CDIP; we thus caution researchers in the document understanding community on using Tobacco3482 and other datasets derived from TTID due to the presence of label errors, data bias, and sensitive material.

5. 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 [7]. We also present early evidence showing how such errors can make model accuracy scores on these benchmarks misleading, as many classification mistakes committed by the DiT classifier 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.

References

[1] Muhammad Nabeel Asim, Muhammad Usman Ghani Khan, Muhammad Imran Malik, Khizar Razzaque, Andreas Dengel, and Sheraz Ahmed. Two stream deep network for document image classification. In *Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019. 1

[2] Björn Barz and Joachim Denzler. Do we train on test data? purging CIFAR of near-duplicates. *Journal of Imaging*, 6(6), 2020. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 1

[4] Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, 2015. 1, 4

[5] Mohammad Hassanpour and Hamed Malek. Document image classification using SqueezeNet convolutional neural network. In *Proceedings of the 5th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, 2019. 1

[6] Rayson Laroca, Valter Estevam, Alceu S. Britto, Rodrigo Minetto, and David Menotti. Do we train on test data? the impact of near-duplicates on license plate recognition. In *Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN)*, 2023. 1

[7] Stefan Larson, Gordon Lim, and Kevin Leach. On evaluation of document classification with RVL-CDIP. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2023. 1, 2, 4

[8] Stefan Larson, Nicole Cornehl Lima, Santiago Pedroza Diaz, Amogh Manoj Joshi, Siddharth Betala, Jamiu Tunde Suleiman, Yash Mathur, Kaushal Kumar Prajapati, Ramla Alakraa, Junjie Shen, Temi Okotore, and Kevin Leach. De-identification of sensitive personal data in datasets derived from IIT-CDIP. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. 4

[9] David Lewis, Gady Agam, Shlomo Argamon, Ophir Frieder, D. Grossman, and Jefferson Heard. 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*, 2006. 2

[10] Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. UpSet: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992, 2014. 4

[11] Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. DiT: Self-supervised pre-training for document image transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 4

[12] Sasha Luccioni and Kate Crawford. The nine lives of ImageNet: A sociotechnical retrospective of a foundation dataset and the limits of automated essentialism. *Journal of Data-centric Machine Learning Research*, 2024. 1

[13] Lucia Noce, Ignazio Gallo, Alessandro Zamberletti, and Alessandro Calefati. Embedded textual content for document image classification with convolutional neural networks. In *Proceedings of the 2016 ACM Symposium on Document Engineering*, 2016. 1

[14] Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *J. Artif. Int. Res.*, 70:1373–1411, May 2021. 1

[15] Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the 35th Conference on Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. 1

[16] Saifullah, Stefan Agne, Andreas Dengel, and Sheraz Ahmed. DocXclassifier: towards a robust and interpretable deep neural network for document image classification. *International Journal on Document Analysis and Recognition (IJDAR)*, 27(3), 2024. 1, 4

[17] Saifullah, Stefan Agne, Andreas Dengel, and Sheraz Ahmed. The reality of high performing deep learning models: A case study on document image classification. *IEEE Access*, 12:103537–103564, 2024. 4

[18] Cecilia Ying and Stephen Thomas. Label errors in BANKING77. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, 2022. 1

[19] Guangyu Zhu and David Doermann. Automatic document logo detection. In *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR)*, 2007. 1, 2