The Eastasouth Journal of Information System and Computer Science
Vol. 1, No. 01, August, pp. 196 - 206

Context-Aware Document Intelligence for Automated Compliance Evidence Extraction from Heterogeneous Financial Documents

Varsha Shah
Independent Researcher, Seattle, WA, USA

Article Info

Article history:

Received, Aug, 2023
Revised, Aug, 2023
Accepted, Aug, 2023

Keywords:

Benchmark Datasets;
Compliance Evidence Extraction;
Document Intelligence;
Financial Natural Language Processing;
Layout-Aware Transformers;
Multimodal Document Understanding;
Named Entity Recognition;
Regulatory Technology

ABSTRACT

Regulatory compliance activities across the financial sector produce a vast and growing amount of unstructured evidentiary material, such as loan agreements, non-disclosure agreements, annual reports, disclosure forms, scanned correspondence, and more — all of which auditors and regulators must review to assure compliance with disclosure, anti-money-laundering and data-protection requirements. Text-only Natural Language Processing pipelines are unable to handle this type of content because the facts relevant to compliance are often embedded in the format of a page, such as signature blocks, tabular disclosures and heading to clauses. This paper summarizes the results of two and a half years of work, including twenty-five papers from the field of transformer language modeling, the field of layout-aware document representation learning, the field of financial natural language processing and the field of applied regulatory-technology systems, all of which were published between 2015 and 2022. It explores how the performance of text-only encoders jumped from an F1 score of around sixty percent on benchmarks for form understanding to the layout-aware multimodal architectures that exceeded the ninety percent mark on the same benchmark, and reviews applied systems that lowered the manual anti-money-laundering review workload by more than fifty percent while maintaining detection rates above eighty percent. The review also outlines five benchmark datasets that test extraction tasks that are related to compliance and suggests a generalized six-stage extraction pipeline extracted from the reviewed architectures. The results show that layout-conditioned pretraining yields the highest improvements, not just a parameter scale, and that the human verification is a constant characteristic of all the deployed systems studied.

Corresponding Author:

Name: Varsha Shah
Institution: Independent Researcher, Seattle, WA, USA
Email: varshachaudhari17@gmail.com

1. INTRODUCTION

The growing volume of documentation required for compliance with disclosure, anti-money-laundering and data protection duties relies on the financial institutions and their regulators. The evidence base is highly varied and consists of born-digital contracts, scanned document forms, tabular financial statements and free-text management commentary, all using spatial layout features like headers, signature blocks and tabular grids that are also complemented by natural language. The idea of adopting

Regulatory technology (RegTech) has been interpreted as a direct solution to the compliance burden exerted on banks as a result of subsequent prudent and conduct-based regulation [1]. Now there are proposals for supervisory technology (SupTech) initiatives to enable regulators to manage the resulting flow of filings, but there are also governance and operational risks associated with SupTech initiatives [2].

In conventional NLP pipelines, often relying on sequential text encoders, a fact whose compliance relevance is often necessary is not always fully captured by the text alone, but also by the location on the page. Without the benefit of surrounding layout, a given number of tokens in a form will not likely make a reliable determination of whether a clause containing "Governing Law," or a total-assets amount on a balance sheet, or a checkbox response in a disclosure form is present, especially if OCR adds to the noise. The breakthroughs brought by transformer language representation learning [3] has led to a first significant improvement in the comprehension of text in context, and further architectures have integrated textual, positional and visual signals into a unified pretraining scheme [4]. This is the motivation behind the concept of context-aware document intelligence, which is adopted in this review to refer to document context-aware systems, in which the extraction of entities, clauses or fields is conditioned on the document's combined text layout image context instead of text context.

This review aims to provide an overview of methodological and empirical approaches to the document-intelligence and financial natural-language-processing literature up to 2022, which can be summarized as a set of techniques for extracting compliance evidence from documents. The review includes four questions: (i) how did representation-learning architectures develop from text-only to layout-aware to multimodal encoders; (ii) what are benchmark datasets and metrics that have been used to test compliance-relevant extraction tasks; (iii) how have these architectures been adapted to applied compliance systems like fraud detection, anti-money-laundering triage, and privacy-policy auditing; and (iv) what architectural pattern has reoccurred across these systems? The literature is synthesized in Section 2, the process of reviewing literature is described in Section 3, comparison results, benchmark statistics, and a generalized extraction pipeline are provided in Section 4, and implications are given in Section 5 based on the insight available by the present year of 2022.

2. LITERATURE REVIEW

2.1 Foundation Language Representation Models

Vectors corresponding to the vocabulary tokens have gradually been replaced by vectors that depend on the context in which the token appears. In [5] it was shown that features from a bidirectional language model (rather than a fixed lookup table) consistently improved language understanding across a variety of tasks. Bidirectional Encoder Representations from Transformers (BERT) extends this idea by using a masked-language-modeling task with the Transformer encoder architecture that achieves excellent performance on eleven natural-language-understanding tasks, including question answering [3]. An optimised variant of this framework dropped the objective of predicting the next sentence, increased the pretraining data and length, improved downstream accuracy, and made clear that the scale of pretraining and optimization schedule, apart from novelty of architecture, have a significant impact on the quality of the representations [6]. These successive encoders were evaluated using a common evaluation target: the Stanford Question Answering Dataset (SQuAD), which consists of over one hundred thousand pairs of questions and answers taken from Wikipedia [7].

2.2 Self-Attention and the Transformer Architecture

The common denominator among the above encoders is the

Transformer, which substitutes recurrent computation with a self-attention mechanism, where all positions of an input sequence are connected to all other positions in a single action [8]. The scaled dot-product attention function for query Q, key K and value V is defined as:

$$Attention(Q, K, V) = softmax(QK^T / \sqrt{d_k}) V \quad (1)$$

where d_k is the dimensionality of the key vectors and the scaling term $1/\sqrt{d_k}$ helps to avoid the vanishing gradients characteristic of the softmax function during weight updates [8]. With multiple attention heads computed in parallel, the model can simultaneously attend to a variety of information from different representation subspaces, as used in document-understanding architectures for relating distant and layout-adjacent text segments, such as a form label to its corresponding value placed in a separate column [9].

2.3 *Layout-Aware and Multimodal Document Understanding*

Text-only encoders ignore the two-dimensional structure that conveys a significant portion of the compliance relevant information for scanned and semi-structured documents. To tackle this shortcoming, LayoutLM integrated the two-dimensional position information of the text into the pretraining process of eleven million document pages, yielding an entity-level F1 score of 79.27 percent, which is approximately nineteen and thirteen points higher than that of the equivalently-sized BERT and RoBERTa baselines, respectively [4]. The method was further extended in LayoutLMv2, where visual features were directly integrated into the pretraining objectives along with the text-image alignment and matching objectives, which achieved FUNSD F1 score of 82.76 percent and 84.20 percent, respectively, at base and large configurations [10]. The masking objectives for text and image patches were combined into a single pre-training framework, and the convolutional backbone was discarded from the previous versions, bringing the FUNSD F1 score to 92.08 percent, and still being relatively parameter-efficient [11]. As shown in Figure 1, the accuracy improvements for each generation were measurable when compared to the previous generation's accuracy, suggesting that joint text-layout-image conditioning — not just parameter scale — has been the main driver of accuracy gains for form-understanding tasks.

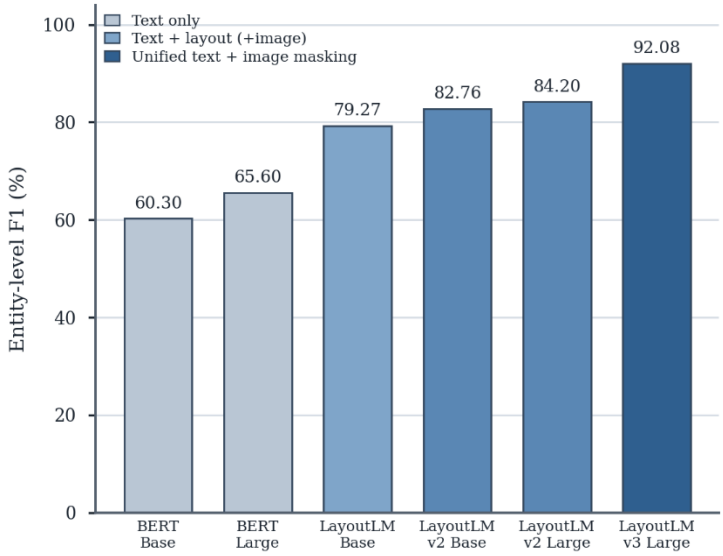

Figure 1. Progression of Text-Only and Layout-Aware Models on the FUNSD Benchmark
Note: Values denote entity-level F1 scores as reported in the corresponding primary sources [3], [4], [10], [11].

2.4 *Sequence Labeling Foundations for Entity and Clause Extraction*

Compliance evidence extraction is often seen as a sequence-labeling or span-extraction task, requiring the identification and classification of evidence in a running text of a compliance document. A bidirectional long short-term memory (LSTM) encoder followed by a conditional random field (CRF) decoder proved that character- and word-level neural features could outperform or be comparable to feature-engineered NER systems using an external gazetteer [12]. Later, a survey classified deep-learning methods for NER into three types, namely, distributed-representation, context-encoder, and tag-decoder, to offer a taxonomy for positioning document-specific systems [13]. Transfer learning has also been shown to work well in fine-tuning general-domain encoders on specialized data sets — for example, when domain vocabulary is different from general domains, finetuning on financial and biomedical corpora led to improvements in recognition [14], a pattern that holds true for the financial-domain models in Section 2.5.

2.5 *Domain-Adapted Financial Language Models*

The representation of financial disclosure language is underrepresented in general-purpose encoders, as is the polarity conventions of financial disclosure language. FinBERT was further pretrained and fine-tuned on the Financial PhraseBank corpus to classify financial sentiment, with an accuracy of 0.86 and an F1 score of 0.84 on the held-out test split, surpassing baselines of long-short-term-memory networks trained on the same data [15]. A variant of FinBERT, trained on 4.9 billion tokens of financial corpora (which included corporate financial filings, analyst reports, and earnings call transcripts), was used to learn the task of general financial information extraction more than sentiment polarity, showing that a common base architecture could be trained to different tasks using different corpora in a specific compliance domain [16].

2.6 *Benchmark Datasets for Compliance-Relevant Extraction*

A few document-intelligence specific benchmarks have been used to assess progress in document-intelligence architectures. FUNSD contains 199 training and testing forms divided into 149/50 and is annotated with 9,707 semantic entities, of which 1,825 are in headers, 2,176 are in answers, and 6,706 are in questions. The Kleister datasets also allow evaluation of long documents with a formal structure, such as NDA's in Kleister NDA (2,160 entities in 540 NDA's) and Charity (21,612 entities in 2,788 annual financial reports), with the highest reported baseline accuracy of 83.57 percent F1 on the Charity subset [17]. CUAD assigns the task to legal clause identification, consisting of 510 commercial contracts labeled with over 13,000 labels at the level of clauses for each of 41 categories [18]. The DocVQA project aims to redefine extraction as question answering over document images, consisting of 50,000 questions, formulated on over 12,000 images, and answered with human accuracy of 94.36% [19]. On the classification level, the classification corpus RVL-CDIP, comprising 400,000 scanned documents from 16 different classification categories, has been used as a standard benchmark for document routing to downstream pipelines and has been accompanied by an initial evaluation of the use of convolutional features for document-image classification which demonstrated that the convolutional features beat handcrafted descriptors [20]. The size of the annotations contained in these benchmarks has been summarized in Table 2 and plotted in Figure 2: they span more than an order of magnitude and are due to the different granularity of the forms, contracts, and computing question-answering tasks.

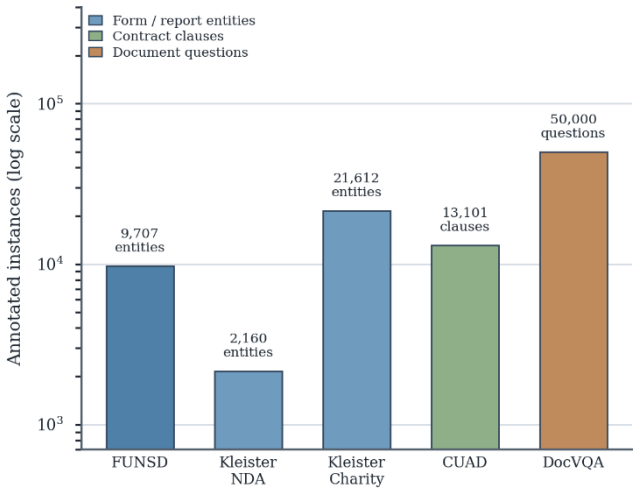

Figure 2. Scale of Compliance-Relevant Document Understanding Benchmarks
Note: Annotation counts reflect entities, clauses, or questions as defined by each benchmark's authors [9], [17], [18], [19].

2.7 *Applied Compliance and Regulatory-Technology Systems*

In addition to benchmark evaluation, there are a number of studies directly using document-intelligence and financial NLP in compliance workflows. The hierarchical attention network consisted of an LSTM encoder and word-level and sentence-level attention was trained to detect financial statement fraud based on financial ratios and MD&A text, with an AUC of 92.64 percent on a class-imbalanced data set of 208 fraudulent and 7,341 non-fraudulent annual reports [21]. In the anti-money-laundering area, a gradient-boosted model, using transaction data from a large Norwegian bank, was trained to determine whether a transaction is suspicious enough to be manually reviewed, and it was able to cut in half the number of transactions that needed to be manually reviewed, while still achieving 80 percent accuracy at identifying confirmed cases of money laundering [22]. Compliance analysis has also been extended to privacy regulation, proposing a BERT-based system to check whether text in a privacy policy meets the notifications requirements of the General Data Protection Regulation (GDPR) where it was previously only possible to do this manually under legal supervision [23]. Although spatial information between candidate table cells in financial documents has been modeled using graph neural networks for structure extraction [24], frameworks like PICK have integrated textual, visual, and layout-graph representations to jointly localize key-value fields in business documents [25]. The innovations listed in Sections 2.1–2.4 are already available in the form of narrow compliance systems that address various observable tasks in the evidence, not the entirety of the evidence's diversity described in Section 1.

3. METHODS

This review, which is not a quantitative meta-analysis, is based on a narrative summary synthesis protocol because the reviewed literature reported a variety of tasks (sentiment classification, named entity recognition, key-information extraction, visual question answering, fraud and anti-money-laundering detection), and metrics (F1 score, accuracy, area under the curve, and Average Normalized Levenshtein Similarity) were heterogeneous, thus making it impossible to pool effect sizes into a summary statistic. The literature reviewed was limited to twenty-five studies that were peer reviewed and archived from 2015 to 2022 across the following venues: Association for

Computational Linguistics and North American Chapter conferences, International Conference on Document Analysis and Recognition (ICDAR), Winter Conference on Applications of Computer Vision (IWAVIS), the Journal of Economics and Business, the Journal of Money Laundering Control, the Journal of Decision Support Systems, the Journal of Finance and Regulation, Contemporary Accounting Research, and the ACM Multimedia, the ACM SIGKDD, the ACM Winter Conference on Computer Vision and Pattern Recognition (IWCPPR), and ICDAR.

All the sources were read in total and the four dimensions outlined below were coded: (a) architectural representation (text-only, layout-aware multi-modal, and graph-based extraction models); (b) task formulation (classification, sequence labeling, span extraction, and visual question answering); (c) reported quantitative performance (where present, the figures were extracted from the results tables in each source); and (d) application domain (finance, legal, and regulation domain versus general natural language-processing benchmarks). To create the comparative progression reported in Section 4.1, studies were cross-referenced when they assessed the same benchmarks, such as the FUNSD dataset that was introduced in one study and then used as an assessment benchmark in three.

The synthesis was done in three steps. First, the architectural lineage was established by following the citation relationship between the encoder-architecture papers, aiming to build the chronological order of the papers: from the contextual word representations to multimodal document transformers. Second, scale and task granularity data were tabulated for each dataset paper to allow for scale and task comparisons across compliance-relevant documents. Third, applied compliance systems were analyzed to find patterns in the architecture and operation of such systems, which were distilled into the generalized extraction pipeline in Section 4.5. No original model training/evaluation was done for this review, all the quantitative values reported in Section 4 are taken from cited primary sources.

4. RESULTS AND DISCUSSION

4.1 Architectural Progression on Form-Understanding Benchmarks

To summarize, the F1 scores of the various entity architectures on the FUNSD benchmark are reported in Table 1, along with the approximate number of parameters that each architecture has. The text-only BERT models get to a 60.30 percent F1 score at the base model, and 65.60 percent F1 score at the large model; the first layout-aware encoder achieves 79.27 percent F1 score with a similar number of parameters [3], [4]. Each successive generation brings a small improvement, as indicated in Table 1, with the last generation, the unified text and image masking objective, achieving 92.08 percent F1. The number of parameters in the large LayoutLMv3 configuration is not significantly higher than that of its predecessor, so the improvement is more attributed to the unification of the masking objectives across modalities, which is consistent with the trend discussed in Section 2.3.

Table 1. Entity-Level F1 Progression on the FUNSD Benchmark Across Encoder Generations

Model	Modality	Approx. Parameters	FUNSD F1 (%)	Source
BERT-Base	Text only	110M	60.30	[3]
BERT-Large	Text only	340M	65.60	[3]
LayoutLM-Base	Text + Layout	~160M	79.27	[4]
LayoutLMv2-Base	Text + Layout + Image	~200M	82.76	[10]
LayoutLMv2-Large	Text + Layout + Image	~426M	84.20	[10]
LayoutLMv3-Large	Unified Text + Image Masking	~368M (approx.)	92.08	[11]

Source: F1 values as reported by the original authors of each cited architecture [3], [4], [10], [11].

4.2 *Benchmark Landscape for Compliance-Adjacent Extraction*

Table 2 and Figure 2 (Section 2.6) provide a summary of the five benchmark datasets most relevant for compliance evidence extraction that were identified in this review. The more complex the form, the longer documents are, and the more they are annotated, the more challenging it is to get the data from FUNSD's brief, single-page forms to Kleister Charity's multi-page scanned financial reports or DocVQA's open-ended question answering. Two observations follow. First, there is no single benchmark in the literature that measures accuracy of both forms, contracts and financial reports, so reported accuracy values cannot be extrapolated to other types of documents without further, domain-specific validation. Second, the scale of DocVQA is smaller than FUNSD and CUAD, which is a methodological choice that has direct consequences on the cost of building compliance-specific training data, since the labeling cost per instance of question-answer pairs is less than that for complete span or entity-level labeling.

Table 2. Benchmark Datasets for Compliance-Relevant Document Understanding

Dataset	Document Type	Scale	Annotated Instances	Source
FUNSD	Scanned forms	199 forms (149 train / 50 test)	9,707 entities	[9]
Kleister NDA	Non-disclosure agreements	540 documents, 3,229 pages	2,160 entities	[17]
Kleister Charity	Charity financial reports	2,788 reports, 61,643 pages	21,612 entities	[17]
CUAD	Commercial legal contracts	510 contracts	13,101 clause labels (41 categories)	[18]
DocVQA	Mixed document images	12,767 images	50,000 questions	[19]
RVL-CDIP	Scanned business documents	400,000 images, 16 classes	320k / 40k / 40k split	[20]

Source: dataset statistics as reported by the original dataset authors [9], [17], [18], [19], [20].

4.3 *Extraction Quality Metrics*

Entity- and clause-level extraction quality throughout the reviewed literature is reported using precision, recall, and their harmonic mean, F1. For a set of predicted spans P and ground-truth spans G:

$$Precision = |P \cap G| / |P| \quad (2)$$

$$Recall = |P \cap G| / |G|$$

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall) \quad (3)$$

Studies on fraud and anti-money-laundering detection also include an area under the receiver operating characteristic curve (AUC) that quantifies the quality of the ranking of the classifiers for all possible decision thresholds, and is preferred over raw accuracy when the positive class is heavily under-represented (as in the fraud item detection dataset of 208 fraudulent items vs. 7,341 non-fraudulent items) [21].

4.4 *Applied Compliance Systems: Comparative Analysis*

Table 3 presents four compliance applied systems found in the review by task, size of the data and the headline result. Both the fraud-detection system [21] and the anti-money-laundering system [22] are based on a structured or semi-structured financial record, with both systems having a natural-language element (MD&A text and transaction metadata respectively) to supplement the tabular data, and compliance evidence is often a mix of both. The privacy-policy compliance system [23] on the other hand works on a single relatively short

document per assessment, showing a document level of unit of analysis instead of a transaction level. The combination of these systems indicates that there is a need for a general-purpose compliance evidence-extraction platform to support at least two units of analysis for compliance evidence extraction, individual document and individual transactions, and at least two types of output — extracted entities and/or clauses, and a downstream compliance score.

Table 3. Applied Compliance and Regulatory-Technology Systems

System / Task	Data	Method	Headline Result	Source
Financial statement fraud detection	208 fraudulent / 7,341 non-fraudulent annual reports	Hierarchical attention network + LSTM on MD&A text and financial ratios	AUC = 92.64%	[21]
Anti-money-laundering transaction triage	33,134 transaction records (Norwegian bank)	Gradient-boosted trees (XGBoost)	−51% manual review; 80% suspicious-case detection	[22]
GDPR Article 13 privacy-policy compliance	Privacy-policy text corpus	BERT-based text classification	Automates legal notification-requirement checks	[23]
Invoice table structure extraction	Invoice documents	Graph neural network over candidate cells	Structural extraction beyond sequence-only methods	[24]
Financial sentiment classification	Financial PhraseBank	FinBERT (fine-tuned BERT)	Accuracy = 86%, F1 = 84%	[15]

Source: results as reported by the original authors of each cited applied system [21], [22], [23], [24], [15].

4.5 A Generalized Extraction Pipeline

Combining the above, the six stages of the pipeline have been distilled from the literature reviewed and shown in Figure 3. A multimodal encoder similar to the LayoutLM family [4], [10], [11] maps text, position, and, if available, image patches; a span identification and typing layer is based on the sequence-labeling foundations of Section 2.4 [12], [13] for the extracted span; a compliance-evidence-mapping layer maps extracted spans to the specific regulatory obligation they evidence, similar to the article-level mapping stage used for privacy-policy notices [23]; a human-in-the-loop review layer is a quality-assurance step and source of corrective feedback for model refinement used in all the applied systems reviewed. This final stage aligns with performance gaps reported above: none of the reviewed systems reached 100 percent human accuracy baseline, reported by DocVQA [19] for the period of review, for fully automated extraction without human verification for compliance grade use in evidentiary documents.

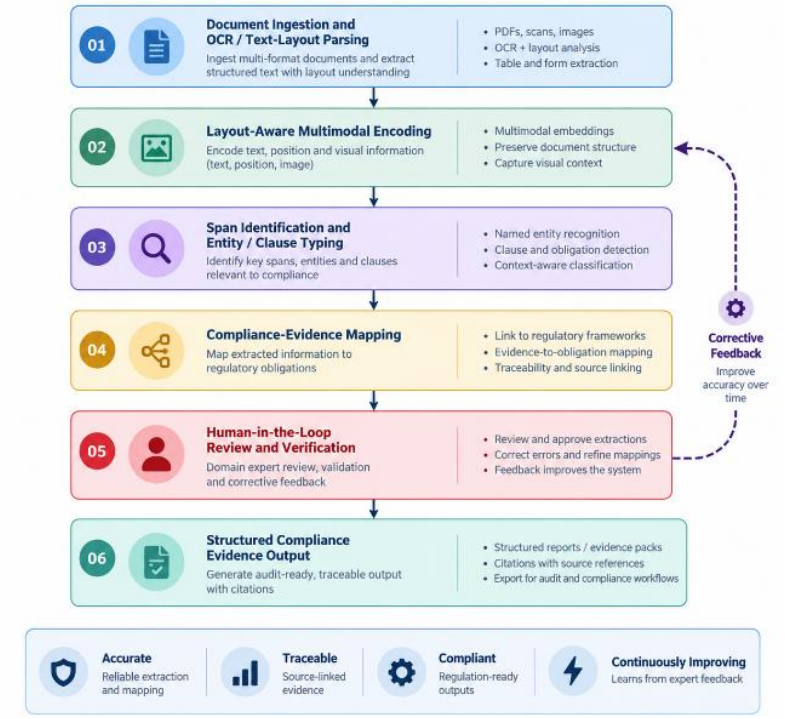

Figure 3. Generalized Context-Aware Pipeline for Compliance Evidence Extraction from Heterogeneous Financial Documents

Note: Pipeline synthesized by the authors from the architectural and applied-system studies reviewed in Sections 2.3, 2.4, and 2.7.

4.6 Limitations and Challenges

There are three recurring limitations in the literature reviewed. By construction, many of the tasks involved in compliance are also imbalanced, such as detecting fraudulent filings, suspicious transactions or non-compliant privacy clauses, which are rare compared to routine cases, where other metrics like AUC work better with imbalanced class distributions [21] and [22]. Second, long-document handling is still relatively underdeveloped when compared to the extraction of short documents; most key-information-extraction benchmarks available up to 2021 focused on single-page forms or receipts [17] and the Kleister benchmark was intentionally designed to be used for long documents with complex layout. Third, there are several architectures that are considered to be high performing but have licensing and/or reproducibility constraints, as some of the checkpoints reviewed in the layout-aware category during this time period came with restrictions on direct use in commercial compliance products.

5. CONCLUSION

The trend in this review has been steady, with the accuracy of entity-level extraction improving from being in the mid-sixties, with text-only contextual encoders, to over ninety percent, with the multimodal transformer including both text and image inputs to the model. And applied compliance systems using simpler architectures already show measurable operational impact, including a documented reduction of over 50 percent in manual anti-money-laundering review volume without proportionately reducing detection coverage. While the benchmark landscape identified in this review is still disjointed between the different types of documents, there is no single dataset that covers the full spectrum of document types — forms, contracts, and financial reports — at the same time, and each system applied in this review still has a stage where

human verification is needed rather than complete automation. The results indicate that such work towards full compliance evidence extraction from diverse financial documents is not solely about scaling up the encoders and increasing their coverage of the benchmark, but also about combining layout-aware extraction with a cross-document-type benchmark coverage, and finally, about bringing structured human-in-the-loop verification into the pipeline of operations, which the proposed generalized one presented here makes a first step towards. Moving forward, and based on the knowledge gained over the time frame of this work, the construction of benchmarks that can be multi-document and multi-type should be considered as a priority.

REFERENCES

[1] I. Anagnostopoulos, "Fintech and regtech: Impact on regulators and banks," J. Econ. Bus., vol. 100, pp. 7–25, 2018, doi: 10.1016/j.jeconbus.2018.07.003.

[2] G. Gasparri, "Risks and opportunities of RegTech and SupTech developments," Front. Artif. Intell., vol. 2, Art. no. 14, 2019, doi: 10.3389/frai.2019.00014.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. 2019 Conf. North American Chapter Assoc. Comput. Linguistics: Human Language Technologies, vol. 1, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[4] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "LayoutLM: Pre-training of text and layout for document image understanding," in Proc. 26th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining, 2020, pp. 1192–1200, doi: 10.1145/3394486.3403172.

[5] M. E. Peters et al., "Deep contextualized word representations," in Proc. 2018 Conf. North American Chapter Assoc. Comput. Linguistics: Human Language Technologies, vol. 1, 2018, pp. 2227–2237, doi: 10.18653/v1/N18-1202.

[6] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv:1907.11692, 2019, doi: 10.48550/arXiv.1907.11692.

[7] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in Proc. 2016 Conf. Empirical Methods in Natural Language Processing, 2016, pp. 2383–2392, doi: 10.18653/v1/D16-1264.

[8] A. Vaswani et al., "Attention is all you need," arXiv:1706.03762, 2017, doi: 10.48550/arXiv.1706.03762.

[9] G. Jaume, H. K. Ekenel, and J.-P. Thiran, "FUNSD: A dataset for form understanding in noisy scanned documents," in 2019 Int. Conf. Document Analysis and Recognition Workshops (ICDARW), vol. 2, 2019, pp. 1–6, doi: 10.1109/ICDARW.2019.10029.

[10] Y. Xu et al., "LayoutLMv2: Multi-modal pre-training for visually-rich document understanding," in Proc. 59th Annu. Meeting Assoc. Comput. Linguistics and 11th Int. Joint Conf. Natural Language Processing, vol. 1, 2021, pp. 2579–2591, doi: 10.18653/v1/2021.acl-long.201.

[11] Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei, "LayoutLMv3: Pre-training for document AI with unified text and image masking," in Proc. 30th ACM Int. Conf. Multimedia, 2022, pp. 4083–4091, doi: 10.1145/3503161.3548112.

[12] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in Proc. 2016 Conf. North American Chapter Assoc. Comput. Linguistics: Human Language Technologies, 2016, pp. 260–270, doi: 10.18653/v1/N16-1030.

[13] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," IEEE Trans. Knowl. Data Eng., vol. 34, no. 1, pp. 50–70, 2022, doi: 10.1109/TKDE.2020.2981314.

[14] S. Francis, J. Van Landeghem, and M.-F. Moens, "Transfer learning for named entity recognition in financial and biomedical documents," Information, vol. 10, no. 8, Art. no. 248, 2019, doi: 10.3390/info10080248.

[15] D. Araci, "FinBERT: Financial sentiment analysis with pre-trained language models," arXiv:1908.10063, 2019, doi: 10.48550/arXiv.1908.10063.

[16] A. H. Huang, H. Wang, and Y. Yang, "FinBERT: A large language model for extracting information from financial text," Contemp. Account. Res., vol. 40, no. 2, pp. 806–841, 2022, doi: 10.1111/1911-3846.12832.

[17] T. Stanisławek et al., "Kleister: Key information extraction datasets involving long documents with complex layouts," in Document Analysis and Recognition – ICDAR 2021, 2021, pp. 564–579, doi: 10.1007/978-3-030-86549-8_36.

[18] D. Hendrycks, C. Burns, A. Chen, and S. Ball, "CUAD: An expert-annotated NLP dataset for legal contract review," arXiv:2103.06268, 2021, doi: 10.48550/arXiv.2103.06268.

[19] M. Mathew, D. Karatzas, and C. V. Jawahar, "DocVQA: A dataset for VQA on document images," in Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (WACV), 2021, pp. 2200–2209, doi: 10.1109/WACV48630.2021.00225.

[20] A. W. Harley, A. Ufkes, and K. G. Derpanis, "Evaluation of deep convolutional nets for document image classification and retrieval," in Proc. 2015 13th Int. Conf. Document Analysis and Recognition (ICDAR), 2015, pp. 991–995, doi: 10.1109/ICDAR.2015.7333910.

[21] P. Craja, A. Kim, and S. Lessmann, "Deep learning for detecting financial statement fraud," Decis. Support Syst., vol. 139, Art. no. 113421, 2020, doi: 10.1016/j.dss.2020.113421.

[22] M. Jullum, A. Løland, R. B. Huseby, G. Ånonsen, and J. Lorentzen, "Detecting money laundering transactions with machine learning," J. Money Laund. Control, vol. 23, no. 1, pp. 173–186, 2020, doi: 10.1108/JMLC-07-2019-0055.

[23] S. Liu, B. Zhao, R. Guo, G. Meng, F. Zhang, and M. Zhang, "Have you been properly notified? Automatic compliance analysis of privacy policy text with GDPR Article 13," in Proc. Web Conf. 2021, 2021, pp. 2154–2164, doi: 10.1145/3442381.3450022.

[24] P. Riba, A. Dutta, L. Goldmann, A. Fornés, O. Ramos, and J. Lladós, "Table detection in invoice documents by graph neural networks," in 2019 Int. Conf. Document Analysis and Recognition (ICDAR), 2019, pp. 122–127, doi: 10.1109/ICDAR.2019.00028.

[25] W. Yu, N. Lu, X. Qi, P. Gong, and R. Xiao, "PICK: Processing key information extraction from documents using improved graph learning-convolutional networks," in 2020 25th Int. Conf. Pattern Recognition (ICPR), 2021, pp. 4363–4370, doi: 10.1109/ICPR48806.2021.9412927.