Automatic Extraction of Key Information from Financial Statements Using a Multi-Task Learning Model

Chunyan Chen
College of Economics & Management, Hubei University of Arts and Science, Xiangyang 441053, Hubei, China
Corresponding author: Chunyan Chen (e-mail: chenchen991018@163.com)

ABSTRACT Financial statement information extraction and technical document data management both face significant challenges due to complex formats, scattered layouts, and intricate proximity-based semantic relations within semi-structured documents. These challenges are particularly evident in engineering enterprises involving electromagnetic devices, antenna systems, radio-frequency equipment, and related technical service activities, where accurate extraction of financial and operational information is required for reliable reporting and performance analysis. This paper proposes an automatic extraction method based on multi-task learning. The model employs a shared encoder, in which text features and layout features are fused to process four related tasks simultaneously: financial named entity recognition, key-value extraction, entity-value relationship classification, and report item classification with DistilRoBERTa. Experiments based on 500 annual reports show that the accuracy of entity recognition reaches 96.8%, the F1 score of value extraction reaches 88.7%, and the F1 score of relationship classification reaches 91.5%. The model performs particularly well in structured sections such as the balance sheet and income statement. Although its performance declines to some extent in the most complex and unstructured sections of financial statements, such as notes, the overall results demonstrate its effectiveness in improving the accuracy and semantic consistency of financial disclosure information extraction. This study provides a technical reference for intelligent document understanding and reliable data extraction in financial reporting scenarios involving engineering-oriented enterprises and electromagnetic application industries.

INDEX TERMS financial statement information extraction, multi-task learning, named entity recognition, layout feature fusion, electromagnetic engineering

I. INTRODUCTION

AS the central bearer of financial condition and operational results, the accuracy and efficiency of extracting information from financial reports and financial reports and enterprise operational performance metrics form the essential basis for investment choices, risk monitoring, and academic research. This data-driven foundation ensures that stakeholders can evaluate both fiscal health and the technical viability of engineering-oriented business assts with precision. With the growing dimensions and complexities of the corporate financial information disclosure, the conventional information extraction techniques are limited to issues like inefficiency, poor scalability and prone to error when operating with different formats and non-standardized structures in the financial statements. Although the research has been done on the correlation between text readability, corporate governance and quality of financial reporting, more research should be conducted on the application of deep learning technology towards automatic and structural extraction of some key information from the financial statements (for example, the company entity, reporting period, financial indicators, and the numerical relationship between them). The main contribution of this paper is the proposal and realization of an end-to-end multi-task learning model designed for the joint and automatic extraction of key information from financial statements and technical specifications. This integrated framework optimizes data processing capabilities to handle the unique semantic structures of both fiscal reporting and engineering-oriented technical documentation. This model is innovative through the integration of textual semantics and page layout features and, with an architecture of shared encoding and task-specific decoding, completes four closely related sub-tasks of entity recognition, numerical extraction, relation classification, and chapter classification at the same time. This is to solve the semantic fragmentation and information inconsistency problem that may exist in single task model, so as to improve the extraction of accuracy and robustness. The structure of this

paper is as follows: Section "RELATED WORK" reviews relevant work in the field of financial statement analysis and information extraction Section "METHOD" elaborates on the problem definition, architecture design, data processing and training methods of the multi-task learning proposed in this paper Section "RESULTS AND DISCUSSION" demonstrates and discusses the model performance through the quantitative experiment and qualitative analysis Section "CONCLUSION" summarizes the entire paper and looks forward to the future research directions.

II. RELATED WORK

The quality, transparency, and economic consequences of financial reporting are core issues in the field of corporate finance and accounting. Seifzadeh et al. [1] used listed companies on the Tehran Stock Exchange in Iran as a sample to collect their financial statement texts and management characteristics data, used text analysis tools to measure readability, and tested the relationship between variables through a multiple regression model. Olayinka [2] sorted out the main methods of financial statement analysis and their application scenarios in investment decision-making. Kaawaase et al. [3] used structural equation modeling to analyze the path relationship between corporate governance mechanisms, internal audit quality, and financial reporting quality. Ilori et al. [4] summarized the strategic methods, technical tools, and management practices adopted by different organizations in the compliance process and conducted comparative analysis. Hasnan et al. [5] analyzed the impact of governance variables such as board independence, audit committee professionalism, and equity structure, as well as characteristic variables such as company size and profitability, on the likelihood of financial restatement. Parmenas et al. [6] analyzed the abnormal stock price fluctuations of listed companies on the Moscow Stock Exchange before and after the release of their annual financial statements, and examined the relationship between key financial indicators such as earnings surprises and cash flow information and market reactions. Hasibuan [7] used multiple regression analysis to examine the relationship between various characteristic variables such as company size, age, profitability, leverage ratio and industry type and the quality of financial reports. Through theoretical analysis and practical cases, Efunniyi et al. [8] explored how to strengthen financial compliance and transparency through multiple approaches such as improving corporate governance structure, strengthening internal control system, adopting compliance technology and cultivating ethical culture. Anik et al. [9] explored the key factors driving sustainable value creation in the emerging economy of Indonesia. Nastiti and Susanto [10] pointed out that in the Indonesian market, the financial characteristics of companies (especially leverage levels and profitability) explain changes in real earnings management behavior better than the formal corporate governance structure examined. This paper proposes an automatic extraction method for key information of financial statements based on a multi-task learning model, which aims to realize the automatic extraction of key information of financial statements and provide a new technical path and research perspective for the intelligent analysis of financial texts.

III. METHODS

A. PROBLEM DEFINITION AND TASK FORMALIZATION

The input to the model is structured output of a single page financial statement that went through an OCR engine with a sequence of text blocks. Each text block includes its plain text content, normalized bounding box coordinates on the page and font size information. The output is a set of prediction results for four related tasks of the text sequence of the page. Task 1 is financial named entity recognition, with the BIOES annotation system. The target entity types include company entities, reporting periods, currency units, and management accounting financial indicators. Task 2 is key numerical item extraction and the objective of this subtask is to extract and standardize some of the numerical values that relate to financial indicators. Its output is a triplet of (indicator mention, value, unit), where value has to be uniformly converted to standard floating-point numbers. Task 3 is entity-value relationship classification. This task is based on prediction results of Task 1 and Task 2, and determines whether a specific value triplet belongs to a certain identified company entity and reporting period. This is a problem of binary classification. Task 4 is the financial statement item classification where text blocks are classified into predefined categories such as balance sheet, income statement, cash flow statement, or notes, which provide contextual information for other tasks [11]. To assess the inherent consistency of the extracted results, this paper defines a set of logical check rules. This check is performed automatically after the Entity-Value Relationship Classification (ENRC), and the specific rules include: (1) Currency unit consistency: The currency unit of the values associated in the same reporting item should be consistent with the reporting currency declared by the entity identified in the context (such as "the Company" or "Consolidated Financial Statements"); (2) Time attribution logic: The reporting period associated with the value must match the accounting period to which the value belongs (for example, the income statement item is associated with "2023", and the balance sheet item is associated with "December 31, 2023"); (3) Basic accounting equation verification: On the balance sheet page, check whether the extracted "Total Assets" and "Total Liabilities and Shareholders' Equity" values are equal within the allowable tolerance.

B. MODEL ARCHITECTURE DESIGN

The shared encoding layer uses DistilRoBERTa as the basic text encoder, whose input is the character sequence of text blocks. To incorporate layout information, the normalized top-left corner coordinates, width, height, and font size of each text block are projected onto the same dimension as the word vectors through independent linear layers,

and then summed to obtain a layout feature vector. This vector is added to the word vector of the first word of the corresponding text block as an augmented input representation for that word. The model adopts a sequence labeling paradigm, concatenating the text of all text blocks into a long sequence in the reading order for encoding. The encoder outputs a shared context representation sequence. In the task-specific decoding layer, the FNER decoder connects a fully connected layer to the shared representation to output the entity label logical value at each position, and then connects to a conditional random field layer for sequence-level optimization. The KVE decoder uses a reading comprehension pointer network, taking the concatenated representation of financial indicator names as the question vector, interacting with the shared sequence representation through attention, and predicting the start and end positions of the text span corresponding to the numerical values. The ENRC decoder concatenates the representations of entity mentions and numerical triples, and performs relation determination through a two-layer feedforward network. The FSIC decoder performs linear classification on the representation of the first word at the corresponding position of each text block.

C. DATA PREPROCESSING AND FEATURE ENGINEERING

First, the PyMuPDF library is used to parse the original PDF file to obtain the text blocks and their precise bounding boxes for each page. Then, OCR post-processing is performed, including merging text lines with segmentation errors (based on rules) and sorting them according to the y and x coordinates of the text blocks. To enhance the robustness of the model to common OCR noise in real-world scenarios, layout feature calculation and text cleaning are performed after preprocessing. Specifically, two strategies are implemented: (1) Text regularization: For financial numbers and currency units, robust regular expression patterns are used for matching and repair. For example, the common OCR error "1,5OO.OO" is corrected to "1,500.00", and "USD" or "UsD" is uniformly normalized to "USD". (2) Noise injection training: During the model training phase, character-level noise (such as random replacement, deletion or insertion of numbers, decimal points and currency symbols) is randomly injected into the input text of the training samples with a certain probability to simulate OCR errors, so that the model can still maintain accurate understanding of semantics and numerical values under noisy input. Table 1 shows the model performance before and after strategy implementation:

Based on the analysis of noise test results, the multi-task learning model proposed in this paper demonstrates excellent OCR noise robustness. Thanks to noise injection training and text regularization strategies, the model exhibits strong tolerance to common OCR errors (such as character obfuscation), and its performance is close to that in noise-free scenarios. Even under moderate garbled text conditions, the extraction and association of key information maintain high accuracy and logical consistency, confirming the practicality and reliability of this method in processing complex real-world financial statements.

D. MODEL TRAINING AND OPTIMIZATION

The model employs an end-to-end joint training approach. The total loss function is defined as the weighted sum of the losses from the four tasks, with an initial weight of 1.0. Dynamic uncertainty weighting is used to automatically adjust the weights of each task during training to balance the learning rate. The optimizer is AdamW with an initial learning rate of 2e-5, using linear warm-up and decay. The batch size in the training steps is 8, and gradient accumulation steps are added to simulate larger batch sizes. To avoid gradient conflicts between tasks, norm pruning is performed on the gradients of the shared encoder after backpropagation, with a pruning threshold of 1.0. This threshold, determined based on preliminary experiments, aims to mitigate gradient direction conflicts while avoiding severe learning stagnation. The training process lasts for 30 epochs, evaluating the overall F1 score on the validation set every 500 steps and saving the optimal checkpoint. The early stopping strategy employs a tolerance value of 5 evaluation intervals. During the validation phase, predictions for all tasks are performed simultaneously; however, during the training phase, the inputs for the ENRC tasks employ a course learning strategy: in the early stages, real labels are used to stabilize the learning of the relation classifier, and in the later stages, the model-predicted labels are gradually switched to simulate real inference conditions and improve generalization ability.

IV. RESULTS AND DISCUSSION

A. QUANTITATIVE RESULTS

To evaluate the effectiveness of the multi-task learning model, we conducted experiments using a dataset containing 500 real-world listed company annual reports (approximately 10,000 pages) and compared it with a baseline model using single-task learning. Evaluation metrics included entity recognition accuracy (Acc), numerical association F1 score, and logical check recall (LCR), defined as the proportion of successfully associated (entity, value) pairs to all correctly associated pairs, used to validate data consistency. Key quantitative results are shown in Tables 2 and 2 below:

Table 2 shows in a very clear way the comprehensive advantages of the multi-task learning model over the single task baseline. In the core task of the Financial Named Entity Recognition (FNER), the improvement was more than 10 percentage points to 96.8%. This tells us that the shared encoding layer by merging for context signals from other tasks (particularly classification of chapter FSIC), significantly improved its ability to discriminate entity boundaries and types. The F1 score for Key Value Extraction (KVE) grew by 6.6% indicating the positive spread of better entity recognition accuracy on numerical localization. The significant improvement of the performance of the Entity-Number

TABLE 1. Performance evaluation of the model under different levels of OCR noise

Noise Level	Description & Example	Test Samples	FNER Accuracy (%)	KVE F1 (%)	ENRC F1 (%)	Logic Check Pass Rate (%)
Low Noise	Simulates common OCR errors (e.g., 0/O, 1/l confusion, missing decimal points). Corrected by text regularization. Example: "Revenue: 1,5OO.OO" -> "Revenue: 1,500.00"	500	98.1	92.3	94.8	97.5
Moderate Noise	Increased random character substitution/deletion, irregular spacing, and broken words, yet overall semantics remain inferable. Example: "Net_lncome w4s 2,3*0 million" -> "Net Income was 2,30 million"	500	95.4	87.6	90.1	93.2
High Noise ("Garbled")	High probability of character-level noise, severely distorting numbers and keywords, nearly unreadable. Example: "C4sh fl0w fr0m oP3rat1on$: 4,5X2"	500	88.7	82.5	85.3	86.9
(Baseline) No Noise	Clean, regularized standard input text.	500	98.5	93.1	95.4	98.0

TABLE 2. Performance comparison between multi-task learning models and single-task baseline models

Model	FNER Acc (%)	KVE F1 (%)	ENRC F1 (%)	FSIC Acc (%)
Single task baseline	86.4	82.1	85.3	90.2
Multi-task model	96.8	88.7	91.5	94.6
Performance improvement	+10.4	+6.6	+6.2	+4.4

Relation Classification (ENRC) (+6.2% F1) is a direct evidence to verify that multi-task joint training allowed the model to learn more consistent entity and numerical semantic representations, thereby enabling a more accurate association. Even accuracy improved in the (relatively independent) Financial Statement Item Classification (FSIC) task, where the fine-grained text understanding offered by other tasks fed back into classification at the chapter level. Table 3 shows the performance of the multi-task model on different types of report pages:

The model works best for the most structured balance sheets and income statements, with F1 scores greater than 90% and logic validation recall approaching or even exceeding 98%. On the more flexible and complex financial statement notes page, although various indicators dropped slightly, the FNER accuracy rate exceeded 95%, the KVE F1 score was 86%, and the logical verification recall rate exceeded 92%, which shows that the model has good processing capabilities for unstructured text. The overall logical verification recall shows that the information retrieved by the model has a high degree of consistency between entities and values, meeting the basic requirements for data reliability in financial analysis.

TABLE 3. Performance of the multi-task model on different types of report pages

Report Page Types	Number of pages	FNER Acc (%)	KVE F1 (%)	LR (%)
Balance Sheet	1500	98.2	91.5	97.8
Income Statement	1500	97.5	90.1	96.5
Cash Flow Statement	1500	96.9	88.3	95.2
Notes to Financial Statements	4000	95.1	86.0	92.1

B. QUALITATIVE ANALYSIS

Quantitative results validated the model's overall performance advantages. To further elucidate its working mechanism and multi-task collaborative benefits in real-world complex scenarios, this section presents a qualitative case analysis. To specifically quantify the model's performance differences under different complex text structures, a fine-grained evaluation was conducted on randomly sampled challenging pages in the test set (covering the various scenarios described below). The results are shown in Table 4.

Table 4 illustrates that the model achieved a high entity recognition F1 score(95.8%) and logic verification pass rate (93.0%) in the "spanning continuous tables" scenario. This is attributed to the common encoding layer combining the layout information and "balance sheet" chapter context provided by the FSIC task to jointly identify them as financial feature of the same company and time period, and an effective association of the cross-page values in the KVE decoder, which reveals the success of combining layout understanding and discourse semantics in a multi-task setting. Another case is that of non-standard descriptions of "mortgage loans," in footnotes. It has been found that in such "financial statement footnotes with mixed paragraphs", the numerical association accuracy rate is 85.2% from the test. The model identified important entities based on the FNER task, extracted a variety of values based on the KVE task and the ENRC task has the capacity to accurately distinguish the loan amount and the value of assets mortgaged with the shared representations and won't misjudge the relationship. The table data further indicates that for more difficult scenarios like "explanatory text with numerical numbers and dots" and "pages with complex footnotes", the model performance has a predictable and reasonable decline which represents the limit of current model abilities. These case studies and fine-grained data prove that it is by joint training that the model realizes knowledge transfer from layout structure perception to deep semantic association, which is flexibly capable of handling the problems of different formats and scattered information in financial reports. At the same time, it has stable performance gradients in scenarios with different levels of structure, which improves the robustness and practicality of the extraction system.

C. DISCUSSION OF LIMITATIONS

While both quantitative and qualitative analyses verify better overall performance in the multi-task model, there are also some limitations of the research, mainly in the aspects of developing understanding of semantic nature of footnote text in financial statements which is unstructured. Although the model preserves the foundation of the basic accuracy of information extraction on text pages (footnotes) through the combination of layout features and multi-task context, this architecture is intrinsically more aligned with the handling of tabular data with regular formats and obvious contextual cues. For the non-standard paragraphs of description that are semantically complex, varied in expression, and implicit in logical sense, such as the role of complex accounting policy changes, the qualitative description of contingent liabilities in the case of management costs and the causal relationship among sentences and sentences in management discussion, the model is mainly based on surface vocabulary and local context of expression, and the model lacks the systematic encoding of domain-specific knowledge and long-distance logical reasoning ability. This might result in misjudgments of relationships or insufficient extraction of information when it comes to the entities and values which need some background knowledge of accounting for proper association. Furthermore, the robustness of this model with multilingual text, extreme layout distortions or the highest compression of text formats needs more verification. Future research needs to go beyond the current paradigm which focuses on layout and shallow semantics, and connect domain knowledge graphs with multi-task learning frameworks. By adding in some structured concept relationships and business rules of accounting, the model can understand and reason about the context and so achieve more reliable and intelligent automated full-text and deep-level intelligent extraction of financial reports data.

V. CONCLUSION

This paper proposes and validates an automatic extraction method based on multi-task learning for extracting key information from financial statements. This method effectively addresses the semantic fragmentation and inconsistency issues of traditional single-task models when processing complex and fragmented financial texts by sharing encoding, fusing textual and layout features, and collaboratively training four related tasks. Experiments show that the model performs excellently in tasks such as entity recognition and numerical association in structured sections, and maintains

TABLE 4. Fine-grained performance of the multi-task model in different complex text scenarios.

Test Scenario Description	Number of test samples	Entity recognition F1 (%)	Numerical association accuracy (%)	Logic verification pass rate (%)
Standard Balance Sheet/Income Statement	30	98.5	96.7	98.2
Continuous Tables Across Pages	25	95.8	90.4	93.0
Notes to Financial Statements with Mixed Paragraphs	25	94.1	85.2	89.6
Description text with numbered and dotted text	12	92.3	82.1	85.0
Pages containing complex footnotes or endnotes	8	88.7	76.5	80.4

usable basic performance in unstructured footnotes, improving the systematic nature of the extraction. However, the model's deep understanding and complex logical reasoning capabilities for unstructured footnote text still need improvement. Future research will explore how to integrate domain knowledge graphs and enhance reasoning mechanisms. For example, to address the limitations of complex layouts such as numbered/dashed lists mentioned in the paper, a more powerful layout encoder or the introduction of visual features could be studied. To address the insufficient deep semantic understanding, an accounting concept knowledge graph could be injected as a priori into the model to enhance its logical association capabilities, thereby further improving the deep intelligent analysis capabilities of financial report texts.

AUTHOR CONTRIBUTIONS

Chunyan Chen designed, collected and analyzed the data, and drafted the manuscript. Chunyan Chen conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Chunyan Chen participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] M. Seifzadeh, M. Salehi, B. Abedini, and M. H. Ranjbar, "The relationship between management characteristics and financial statement readability," EuroMed Journal of Business, vol. 16, no. 1, pp. 108-126, 2021, doi: 10.1108/EMJB-12-2019-0146.

[2] A. A. Olayinka, "Financial statement analysis as a tool for investment decisions and assessment of companies' performance," International Journal of Financial. Accounting, and Management, vol. 4, no. 1, pp. 49-66, 2022, doi: 10.35912/ijfam.v4i1.852.

[3] T. K. Kaawaase, C. Nairuba, B. Akankunda, and J. Bananuka, "Corporate governance, internal audit quality and financial reporting quality of financial institutions," Asian Journal of Accounting Research, vol. 6, no. 3, pp. 348-366, 2021, doi: 10.1108/AJAR-11-2020-0117.

[4] O. Ilori, N. T. Nwosu, and H. N. N. Naiho, "Optimizing Sarbanes-Oxley (SOX) compliance: strategic approaches and best practices for financial integrity: A review," World Journal of Advanced Research and Reviews, vol. 22, no. 3, pp. 225-235, 2024, doi: 10.30574/wjarr.2024.22.3.1728.

[5] S. Hasnan, M. H. Mohd Razali, and A. R. Mohamed Hussain, "The effect of corporate governance and firm-specific characteristics on the incidence of financial restatement," Journal of Financial Crime, vol. 28, no. 1, pp. 244-267, 2021, doi: 10.1108/JFC-06-2020-0103.

[6] N. Parmenas, B. Michaela, B. M. Heinrich, and S. Dmitry, "Stock price reactions to publications of financial statements: Evidence from the Moscow Stock Exchange," Korporativnye financy, vol. 15, no. 1, pp. 19-36, 2021, doi: 10.17323/j.jcfr.2073-0438.15.1.2021.19-36.

[7] A. N. Hasibuan, "The Role of Company Characteristics in the Quality of Financial Reporting in Indonesian," Jurnal Ilmiah Peuradeun, vol. 10, no. 1, pp. 1-12, 2022, doi: 10.26811/peuradeun.v10i1.666.

[8] C. P. Efunniyi, A. O. Abhulimen, A. N. Obiki-Osafiele, O. S. Osundare, E. E. Agu, and I. A. Adeniran, "Strengthening corporate governance and financial compliance: Enhancing accountability and transparency," Finance & Accounting Research Journal, vol. 6, no. 8, pp. 1597-1616, 2024, doi: 10.51594/farj.v6i8.1509.

[9] S. Anik, A. Chariri, and J. Isgiyarta, "The effect of intellectual capital and good corporate governance on financial performance and corporate value: A case study in Indonesia," The Journal of Asian Finance, Economics and Business, vol. 8, no. 4, pp. 391-402, 2021, doi: 10.13106/jafeb.2021.vol8.no4.0391.

[10] M. D. Nastiti and Y. K. Susanto, "Corporate governance, financial ratio and real earnings management in Indonesia stock exchange," Global Financial Accounting Journal, vol. 6, no. 2, pp. 250-264, 2022, doi: 10.37253/gfa.v6i2.6783.

[11] N. Djamil, "Akuntansi Terintegrasi Islam: Alternatif Model Dalam Penyusunan Laporan Keuangan: Islamic Integrated Accounting: Alternative Models in Preparing Financial Statements," JAAMTER: Jurnal Audit Akuntansi Manajemen Terintegrasi, vol. 1, no. 1, pp. 1-10, 2023, doi: 10.5281/zenodo.8384951.