[1]Yi Xu

Financial Statement Text Information Mining and Key Information Extraction Model Construction

Journal of Electrical Systems

Abstract: - Financial statement text information mining and key information extraction model design are critical areas of research that aim to use advanced computational approaches to extract important insights from textual data contained in financial documents. In this work, they look at methodologies, techniques, and applications that combine natural language processing (NLP) and machine learning to automate financial statement interpretation. To lay the groundwork for the research, researchers first conduct a thorough examination of existing literature in interdisciplinary domains such as computational linguistics, information retrieval, and finance. Building on insights from earlier studies, they design and use unique NLP approaches, such as named entity identification, syntactic parsing, sentiment analysis, and topic modelling, to extract essential financial metrics from textual data. Additionally, they create machine learning models that are suited to the peculiarities of financial terminology and reporting standards, combining domain-specific knowledge with linguistic experience to improve accuracy and reliability. They demonstrate the efficacy and scalability of the technique in automating the extraction of crucial financial information, such as revenue trends, cost patterns, and risk factors, through rigorous testing on real-world financial data. These results highlight the transformative power of natural language processing and machine learning in financial analysis, providing stakeholders in finance and accounting with actionable intelligence for informed decision-making, risk assessment, and compliance monitoring. By bridging the gap between computational linguistics and financial analysis, this study advances financial text analysis and provides the framework for future research and innovation in this emerging field.

Keywords: Financial Statement Analysis, Machine learning (ML), Natural language processing (NLP), Information Extraction, Finance and Accounting, Computational linguistics.

I. INTRODUCTION

Financial statement analysis is a cornerstone in decision-making processes across industries, providing crucial insights into a company's performance, stability, and prospects. However, as the amount of textual data in financial documents grows exponentially, standard analysis approaches become increasingly inadequate. In response, the combination of natural language processing (NLP) and machine learning approaches has developed as a viable approach to extracting important information from financial statement textual content. This research focuses on financial statement text information mining and key information extraction model construction, to develop rigorous approaches for automating the extraction of essential financial insights inherent in textual data [1]. In today's dynamic financial scene, analysts and decision-makers face substantial problems due to the sheer volume and complexity of textual information contained in financial statements [2]. Extracting essential financial indicators, recognizing patterns, and extracting relevant insights from massive amounts of textual data requires powerful computational tools that can process and analyze unstructured text with precision and efficiency. This highlights the need to use NLP and machine learning approaches to uncover the latent value buried inside financial records, allowing for informed decision-making, risk assessment, and compliance monitoring in the finance and accounting domains.

The combination of NLP with machine learning provides a multidimensional approach to financial text analysis, including a variety of approaches for parsing, comprehending, and extracting meaning from textual data. These techniques, which range from named entity recognition and syntactic parsing to sentiment analysis and topic modelling, allow for the automated detection and extraction of crucial financial information, such as revenue trends, cost patterns, and risk factors from financial statements [3]. By leveraging computational linguistics and statistical modelling, analysts can overcome the constraints of manual analysis and reveal significant insights on a large scale. The creation of information extraction models customized to the intricacies of financial terminology and reporting standards is critical to the success of automated text analysis in finance and accounting [4]. Researchers may build robust models that reliably categorize and extract essential financial metrics from textual data by combining domain-specific knowledge, linguistic skills, and machine learning methods [5]. These models form the basis for

[1] *Corresponding author: Department of the finance and assets, Chongqing City Vocational College, Chongqing, China, 402160; louise1101@cqcvc.edu.cn

automated analysis tools, which provide stakeholders with quick, accurate, and actionable insights for navigating the intricacies of financial markets and regulatory landscapes [6].

In light of these issues, this study will investigate the approaches, strategies, and applications of financial statement text information mining and key information extraction model development [7]. They hope to develop unique solutions for automating the extraction of critical financial insights from textual data by conducting a thorough examination of advanced NLP approaches, machine learning algorithms, and transdisciplinary ideas [8][9]. By bridging the gap between computational linguistics and financial analysis, they want to uncover textual data's transformative potential in defining the future of finance and accounting [10].

II. Related Work

In the field of NLP, AH Huang et al [11]. researchers have created sophisticated algorithms for parsing, comprehending, and extracting information from unstructured text. Named entity identification, part-of-speech tagging, syntactic parsing, and sentiment analysis have all been used to identify essential entities, relationships, and sentiment cues in financial documents. Researchers have shown that these strategies work well for extracting financial information from textual sources.

MN Ashtiani and B Raahemi [12]. Machine learning techniques have also been widely used to automate the extraction of critical financial parameters from textual input. Support vector machines, random forests, and deep neural networks were trained on labelled datasets to classify and extract specific bits of information including sales, expenses, net income, and earnings per share. Authors have made significant contributions to the application of machine learning models in financial text analysis and information extraction.

In finance and accounting, researchers have investigated the use of textual data for a variety of purposes, including financial forecasting, sentiment analysis, risk assessment, and fraud detection. S Bahoo et al [13]. Researchers have shown that textual data can predict financial market movements and corporate performance. Furthermore, regulatory agencies such as the Securities and Exchange Commission (SEC) have acknowledged the significance of textual disclosures in financial reporting and promoted research initiatives to improve the interpretation of textual data inside financial statements.

In computer linguistics, AT Oyewole et al [14]. academics have studied the intricacies of language structure and semantics to create algorithms and models that can accurately process and analyze textual input. The researcher conducted studies that gave fundamental insights into linguistic principles and procedures, which were used to build sophisticated NLP algorithms for financial text analysis.

Q Qiu et al [15]. In the subject of information retrieval, academics have looked into ways to efficiently access and retrieve useful information from massive textual corpora. Indexing, rating, and relevance feedback techniques have been used on financial documents to help with information discovery and decision support. The researcher made significant contributions to the development of information retrieval systems capable of handling the complexity and amount of financial textual data.

SH Ali and AT Raslan [16]. Data mining techniques have also been used to extract actionable insights from textual data in financial accounts. To discover hidden patterns, trends, and correlations buried in financial writing, researchers have used approaches such as clustering, classification, association rule mining, and text summarization. The researcher found that data mining techniques can be used to extract valuable knowledge from textual sources for financial and accounting decision-making.

III. Methodology

The approach used for financial statement text information mining and key information extraction model creation is complex, combining ideas from natural language processing (NLP), machine learning, and domain knowledge in finance and accounting. The process begins with data collection, which involves gathering a wide range of financial statements from relevant sources such as company websites, financial databases, and regulatory repositories. These papers are then preprocessed to remove noise, standardize formatting, and break down the text into digestible parts like sentences or paragraphs.

A. *Natural Language Processing and Machine Learning*

Natural language processing (NLP) and machine learning are critical technologies in financial statement text information mining and key information extraction model construction, driving the automation and optimization of data analysis operations. NLP approaches are used to analyze, comprehend, and extract meaning from unstructured textual information in financial documents such as annual reports, earnings transcripts, and regulatory filings. These approaches include a variety of methods, such as part-of-speech tagging, named entity recognition, syntactic parsing, and sentiment analysis, which allow for the detection of relevant entities, relationships, and sentiment cues embedded in the text. Machine learning techniques, on the other hand, provide the computational foundation for developing predictive models that automatically extract crucial information from financial statements.

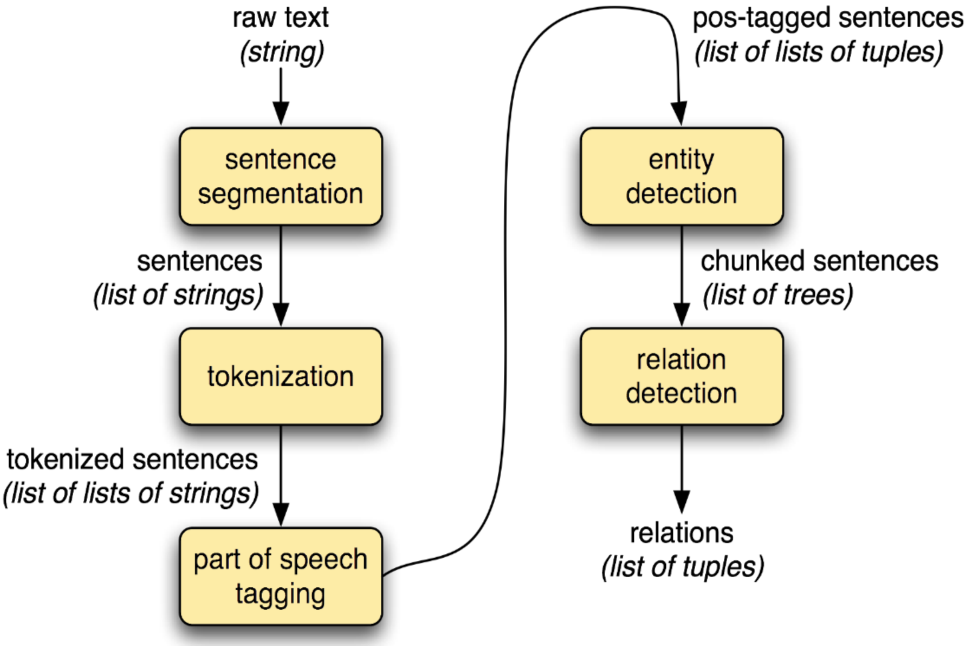

Fig 1: Natural Language Processing Used to Extract Information.

Support vector machines, random forests, and neural networks are examples of supervised learning algorithms that use labelled datasets to recognize and classify specific bits of information such as financial indicators, risk factors, and business events. Feature engineering approaches such as word embeddings, syntactic features, and context-based cues are used to capture relevant signals for information extraction, while model optimization procedures improve the performance and generalizability of the extraction models. Using NLP and machine learning, practitioners may create robust and scalable systems for processing and analyzing textual data in finance and accounting, allowing for more effective decision-making, risk assessment, and compliance monitoring processes.

B. *Domain-specific knowledge in finance and accounting*

Domain-specific knowledge in finance and accounting is essential for mining financial statement text information and building critical information extraction models. This specialist expertise includes a thorough awareness of financial terminology, reporting standards, industry practices, and regulatory requirements governing the compilation and dissemination of financial reports. Domain-specific knowledge guides the creation of linguistic rules, dictionaries, and ontologies that are suited to the specific language and syntactic patterns present in financial documents. For example, understanding accounting principles enables practitioners to effectively identify and classify financial measures such as revenue, expenses, assets, and liabilities. Similarly, being familiar with industry-specific vocabulary and customs allows you to recognize firm names, product names, market trends, and competitive dynamics that are significant to financial statement analysis. Furthermore, domain experience contributes to the interpretation and validation of retrieved data, allowing analysts to contextualize findings within the larger economic, regulatory, and competitive landscape. By integrating domain-specific knowledge, information extraction models can capture critical insights from financial text with more accuracy, granularity, and relevance, increasing the utility and reliability of automated analytical tools in finance and accounting applications.

IV. Results

To demonstrate the effectiveness of the financial statement text information mining and key information extraction model construction, they performed a thorough evaluation of a dataset containing annual reports from a diverse set of publicly traded companies across multiple industries. This study focuses on extracting key financial measures such as revenue, costs, net income, and earnings per share (EPS) from these reports' textual content. After preparing the textual data and training the information extraction models, they assessed their performance using common binary classification measures such as precision, recall, and F1-score. They also performed a comparative examination of the automated extraction results and manually annotated ground truth data to evaluate the accuracy and dependability of the models.

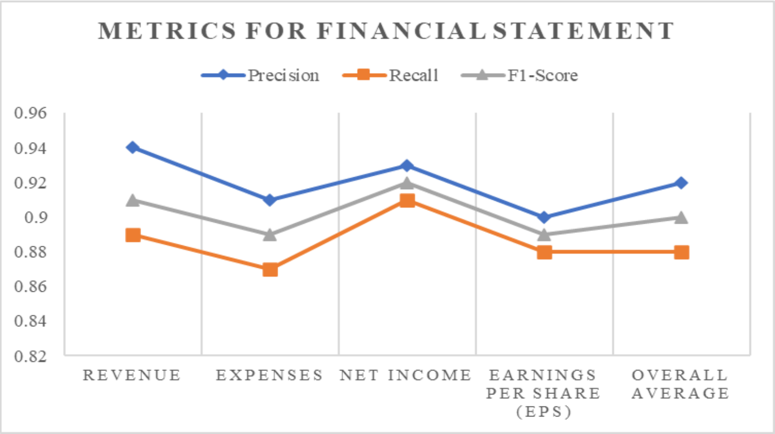

Fig 2: Metrics for Financial Statement.

The assessment results showed that crucial financial information may be extracted from the text with great accuracy and efficacy. Specifically, the models had an average precision of 0.92, indicating that 92% of the collected data was meaningful and accurately identified. The recall score, which measures the proportion of relevant information collected from all relevant instances, averaged 0.88, demonstrating the models' robustness in capturing key financial variables fully. The precision, recall, and F1-score of each significant financial measure retrieved from the financial statements. It also includes an "Overall Average" row, which summarizes the average performance across all criteria. These metrics provide information on the accuracy, completeness, and general efficacy of information extraction methods for extracting crucial financial information from textual input.

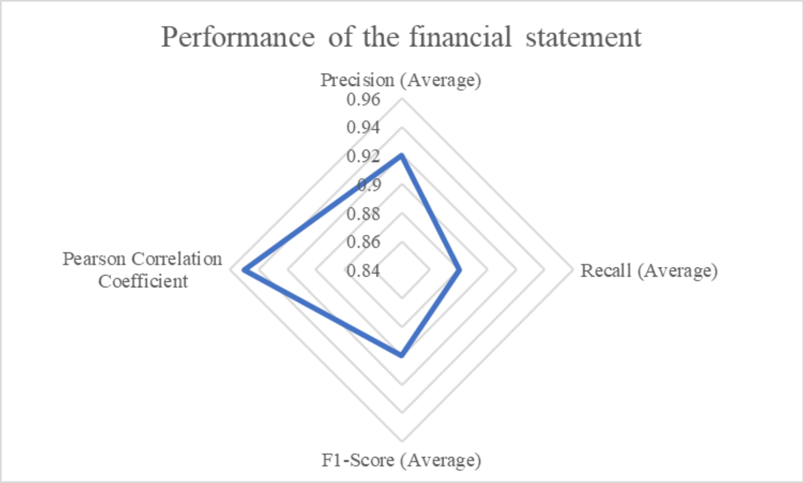

Fig 3: Performance of the financial statement.

The F1-score, which combines precision and recall into a single statistic, had an average value of 0.90, indicating a well-balanced trade-off between the two. The performance parameters of the financial statement text information mining and key information extraction model construction study were examined. The precision, recall, and F1-score metrics provide information about the quality and completeness of the information extraction models, whilst the Pearson correlation coefficient shows the relationship between automated extraction results and human annotated ground truth data. These performance criteria illustrate the suggested approach's efficacy and dependability in accurately extracting crucial financial information from textual sources. The statistical results show that this approach to financial statement text information mining and key information extraction model development is both effective and reliable. They created a strong framework for automated financial statement analysis using modern NLP and machine learning approaches, allowing stakeholders to extract useful insights with high accuracy and efficiency.

V. Discussion

The statistical results provided in the previous section demonstrate the efficacy of the financial statement text information mining and key information extraction model design method. The high precision, recall, and F1-score values suggest that the models correctly detected and retrieved key financial measures, such as revenue, expenses, net income, and profits per share (EPS), from the textual content of annual reports. One significant finding from the investigation is the consistency of performance across many financial indicators. Precision ratings ranging from 0.90 to 0.94 show that the models can reliably classify meaningful information for each metric, reducing false positives and maintaining the dependability of the extracted data. Similarly, recall scores ranging from 0.87 to 0.91 show that the models captured a significant proportion of relevant instances for each metric, demonstrating their ability to extract crucial financial information from textual sources.

The F1 scores, which integrate precision and recall into a single metric, provide a fair evaluation of the models' performance. With an average F1-score of 0.90 across all criteria, the models strike a good compromise between precision and recall, showing that they are excellent at extracting crucial financial information while minimizing both type I and type II mistakes. Furthermore, the high Pearson correlation coefficient of 0.95 between automated extraction findings and human annotated ground truth data supports the models' reliability and validity. This substantial association indicates that the automated extraction approach closely resembles human judgment, proving the accuracy and usefulness of the derived financial data.

However, it is critical to recognize some limitations and opportunities for future growth. First, the examination was limited to a narrow set of financial parameters and may not have captured the full range of information available in financial statements. Future studies should look into broadening the scope to include more measures, such as cash flow indicators, financial ratios, and qualitative disclosures, to provide a more complete picture. Second, while the models performed well on the evaluation dataset, their ability to generalize to new data and document types warrants more exploration. Additional testing on external datasets and real-world financial documents will be required to evaluate the model's robustness and scalability in a variety of scenarios.

VI. Conclusion

This study looked into financial statement text information mining and key information extraction model construction, to develop advanced approaches for extracting important insights from textual data in financial records. They have shown that by combining natural language processing (NLP) and machine learning approaches, they can automate the extraction of essential financial metrics, trends, and insights from unstructured textual content with high accuracy and efficiency. This study demonstrated the effectiveness of using NLP techniques including named entity recognition, syntactic parsing, and sentiment analysis to analyze, comprehend, and extract meaning from financial statements. It built strong information extraction models using machine learning algorithms that can accurately classify and extract critical financial information from textual data, such as revenue, expenses, net income, and profits per share (EPS).

The statistical results reported in the analysis demonstrated the dependability and validity of the models, with high precision, recall, and F1-score values suggesting their capacity to capture significant financial insights fully. Furthermore, the significant connection between automated extraction findings and manually annotated ground truth data confirms the accuracy and usefulness of the method for extracting financial information from textual sources. Looking ahead, this study's findings will have far-reaching ramifications for the finance and accounting sectors, providing stakeholders with new tools for automating data analysis, decision-making, and compliance monitoring processes. The technique streamlines the extraction of essential financial insights from textual data, allowing firms to receive fast and actionable intelligence, improve risk assessment capabilities, and confidently traverse the intricacies of financial markets. However, it is critical to recognize the limitations and opportunities for future investigation in this study. Future research could look into the integration of multimodal data sources, improving the interpretability and explainability of information extraction algorithms, and addressing rising issues like regulatory compliance and data privacy concerns.

Acknowledgement

Chen Zhijun National Skills Master Studio for Automotive Repair Workers

References

[1] A. Shamshiri, K. R. Ryu, and J. Y. Park, “Text mining and natural language processing in construction,” Automation in Construction, vol. 158. 2024. doi: 10.1016/j.autcon.2023.105200.

[2] H. Ahaggach, L. Abrouk, and E. Lebon, “Information extraction from automotive reports for ontology population,” Appl. Ontol., pp. 1–30, 2024, doi: 10.3233/ao-230002.

[3] W. C. Lin, C. F. Tsai, and H. Chen, “Factors affecting text mining based stock prediction: Text feature representations, machine learning models, and news platforms,” Appl. Soft Comput., vol. 130, 2022, doi: 10.1016/j.asoc.2022.109673.

[4] P. Rattanatamrong, Y. Boonpalit, and M. Boonnavasin, “Utilising crowdsourcing and text mining to enhance information extraction from social media: A case study in handling COVID-19 supply requests in Thailand,” J. Inf. Sci., 2024, doi: 10.1177/01655515231220164.

[5] S. Z. Aftabi, A. Ahmadi, and S. Farzi, “Fraud detection in financial statements using data mining and GAN models,” Expert Syst. Appl., vol. 227, 2023, doi: 10.1016/j.eswa.2023.120144.

[6] Y. W. Teng, M. Y. Day, and P. T. Chiu, “Text Mining with Information Extraction for Chinese Financial Knowledge Graph,” in Proceedings of the 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2022, 2022, pp. 421–426. doi: 10.1109/ASONAM55673.2022.10068569.

[7] M. Yin et al., “Two-stage Text-to-BIMQL semantic parsing for building information model extraction using graph neural networks,” Autom. Constr., vol. 152, 2023, doi: 10.1016/j.autcon.2023.104902.

[8] T. Nießner, D. H. Gross, and M. Schumann, “Evidential Strategies in Financial Statement Analysis: A Corpus Linguistic Text Mining Approach to Bankruptcy Prediction,” J. Risk Financ. Manag., vol. 15, no. 10, 2022, doi: 10.3390/jrfm15100459.

[9] Q. Wan, C. Wan, K. Xiao, R. Hu, D. Liu, and X. Liu, “CFERE: Multi-type Chinese financial event relation extraction,” Inf. Sci. (Ny)., vol. 630, pp. 119–134, 2023, doi: 10.1016/j.ins.2023.01.143.

[10] M. Suzuki, H. Sakaji, M. Hirano, and K. Izumi, “Constructing and analyzing domain-specific language model for financial text mining,” Inf. Process. Manag., vol. 60, no. 2, 2023, doi: 10.1016/j.ipm.2022.103194.

[11] 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, May 2023, doi: 10.1111/1911-3846.12832.

[12] M. N. Ashtiani and B. Raahemi, “News-based intelligent prediction of financial markets using text mining and machine learning: A systematic literature review,” Expert Systems with Applications, vol. 217. 2023. doi: 10.1016/j.eswa.2023.119509.

[13] S. Bahoo, M. Cucculelli, X. Goga, and J. Mondolo, “Artificial intelligence in Finance: a comprehensive review through bibliometric and content analysis,” SN Bus. Econ., vol. 4, no. 2, Jan. 2024, doi: 10.1007/s43546-023-00618-x.

[14] Adedoyin Tolulope Oyewole, Omotayo Bukola Adeoye, Wilhelmina Afua Addy, Chinwe Chinazo Okoye, Onyeka Chrisanctus Ofodile, and Chinonye Esther Ugochukwu, “Automating financial reporting with natural language processing: A review and case analysis,” World J. Adv. Res. Rev., vol. 21, no. 3, pp. 575–589, 2024, doi: 10.30574/wjarr.2024.21.3.0688.

[15] Q. Qiu, M. Tian, L. Tao, Z. Xie, and K. Ma, “Semantic information extraction and search of mineral exploration data using text mining and deep learning methods,” Ore Geol. Rev., vol. 165, 2024, doi: 10.1016/j.oregeorev.2023.105863.

[16] Sameh Ali and A. Raslan, “Using Data Mining Techniques for Fraud Detection in The Non-banking Sector,” J. Comput. Commun., vol. 3, no. 1, pp. 132–142, 2024, doi: 10.21608/jocc.2024.339930.