A white background with black text

Description automatically generated

Report on

­Summary of Research Papers related to

Extracting and Summarising Financial Quarterly Resul­t Sheets from PDFs

Prepared By:

Anusha P Hullathi

**Paper 1**

Title: Financial Sentiment Analysis of Quarterly Reports and Stock Performance

Link: <https://nmbu.brage.unit.no/nmbu-xmlui/handle/11250/3021389>

Problem statement:

This work focuses on modeling unstructured data, and the lack of research on sentiment analysis of quarterly reports and the few publications applying the FinBERT model for financial sentiment analysis form the narrowed scope of the study.

problems which are addressed:

* Gap in research on accurate sentiment prediction in financial reports.
* Comparison of FinBERT's prediction accuracy with the general BERT model.
* Limited exploration of the relationship between financial sentiment and stock price movements over time.
* Analysis of the correlation between FinBERT-predicted sentiment and stock performance.
* Unclear variation in sentiment-stock performance correlation between asset-light and asset-heavy companies.
* Study of how asset strategies impact the correlation in the technology, real estate, and energy sectors.

problems which are not addressed:

* Improving the quality of text extraction and cleaning should be the focus of future work in this study.
* Furthermore, grouping and predicting text for different topics, such as sustainability, would be an interesting extension for further domain insight.
* The number of reports, companies, and industries should be increased to provide more reliable results for the relationship between stock performance and sentiment scores.
* An API endpoint for automatic download of reports and dates would be beneficial for this purpose and for a decision-support prototype.

**Paper 2**

Title: Automated Financial Data Extraction – An AI Approach

Link: http://worldcomp-proceedings.com/proc/p2013/ICA3471.pdf

Problem statement:

Creating a compliant XBRL (eXtensible Business Reporting Language) document from financial reports, such as 10-Q filings, is a complex and labor-intensive process. XBRL compliance mandates that all elements, including financial tables, Document and Entity Information (DEI), and footnotes, be tagged with specific standardized tags for proper data analysis. Manual tagging is error-prone, time-consuming, and requires a high level of accuracy. Thus, the need arises for an automated system that can accurately extract and tag these elements from financial documents.

problems which are addressed:

• The manual process of tagging financial documents with XBRL-compliant labels is labor-intensive, error-prone, and requires specialized knowledge.

• Extracting and tagging data from financial tables is difficult due to varying formats and structures.

• Unstructured and varied notes in financial documents present challenges for accurate extraction and XBRL tagging.

• Identifying and tagging Document and Entity Information (DEI) sections is a critical but complex task.

• Capturing and tagging contextual parenthetical data alongside financial statements is often ignored and difficult to automate.

• The manual process of preparing XBRL-compliant documents does not scale well, necessitating a more efficient solution.

problems which are not addressed:

* The paper focuses on standard filings like 10-Q reports but does not address handling non-standard or highly complex reports with unusual formats or multiple data sources.
* The paper does not address the challenges of extracting and tagging financial data from reports in multiple languages or handling multilingual XBRL taxonomies.
* The paper does not address how to update or adapt the system for evolving XBRL standards and new compliance requirements, potentially requiring significant manual intervention.
* While the paper focuses on automated tagging, it does not discuss mechanisms for error detection, correction, or validation feedback loops for human reviewers, especially in cases of ambiguous or uncertain data tagging.

**Paper 3**

Title: An Intelligent information segmentation approach to extract financial data for business valuation

Link: https://www.sciencedirect.com/science/article/abs/pii/S0957417410001739

Problem statement:

Extracting financial data for business valuation from statements, notes, and news is challenging. The manual process is inefficient and prone to errors, with issues like incorrect word segmentation and ambiguous POS tagging complicating accurate data extraction. Therefore, there is a pressing need for systematic and advanced tools to enhance the accuracy and efficiency of financial data extraction for business valuation.

problems which are addressed:

* The existing word segmentation tools and methods sometimes fail to correctly segment financial terms and entities, leading to incomplete or erroneous extraction of financial data.
* There are issues with the part-of-speech (POS) tagging of financial terms, which affects the accuracy of extracting relevant data such as monetary values, percentages, and dates.
* The traditional methods of data extraction from financial documents often result in incomplete or incorrect data due to limitations in current extraction algorithms and tools.
* Financial data is often scattered across various sources (e.g., financial statements, notes, news), and there is a lack of integrated tools that can extract and consolidate data from these diverse sources effectively.
* The extraction process struggles with recognizing and correctly handling named entities such as account names, organization names, and specific financial terms, especially in the context of Chinese text.

problems which are not addressed:

* The paper mentions that the system may still struggle with Chinese article structures, syntax, and meaning, which can lead to incomplete or incorrect data extraction due to the nuances and complexity of the language.
* Although the paper focuses on certain data sources like financial statements, notes, and financial news, it does not extensively address the extraction techniques.
* The methods and results are specific to Chinese financial texts, and the applicability to other languages or financial contexts may not be fully explored or validated.

**Paper 4**

Title: Financial Narrative Summarisation at MultiLing 2019: Data Creation and Task Description

Link: <https://aclanthology.org/W19-8902>

Problem statement:

This work addresses the challenge of summarizing UK annual reports. Issues include the complexity of summarizing unstructured, lengthy reports, difficulties in extracting narrative sections while excluding non-narrative content, and the need for effective evaluation metrics like AutoSummENG and Rouge.

problems which are addressed:

* The annual reports from UK firms are lengthy and have a less standardized structure, making it challenging to summarize the narratives automatically.
* The paper discusses the difficulties in creating a dataset for financial narrative summarisation. It involves handling the unstructured text of annual reports and categorizing sections appropriately.
* Summarizing extensive and complex documents like annual reports while maintaining the accuracy and relevance of the information is a challenging task.
* Evaluating the quality of automatically generated summaries is challenging. The paper uses AutoSummENG and Rouge metrics for evaluation, but these metrics may have limitations in capturing the nuances of financial narratives.

problems which are not addressed:

* The paper focuses on summarizing narrative sections but does not address how to effectively handle and integrate information from non-narrative sections, such as financial statements, which may be crucial for a complete summary.
* The paper highlights the lack of standardization in UK annual reports but does not propose solutions for addressing this issue or improving the consistency of report structures.
* The use of AutoSummENG and Rouge metrics for evaluation may not fully capture the quality and relevance of financial summaries. The paper does not explore alternative or additional metrics that might provide a more nuanced evaluation.

**Paper 5**

Title: Retrieval of Financial Statements and Extraction of Textual Information

Link: https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=2870309

Problem statement:

The problem statement addressed in the paper revolves around the challenges associated with retrieving and analyzing financial statements filed with the SEC using the EDGAR system.

problems which are addressed:

* Many researchers and analysts are not familiar with programming languages like Perl, Python, or R, which are typically used to retrieve data from EDGAR. This complexity makes the process challenging and time-consuming for those without technical expertise.
* Traditional methods often require downloading one filing at a time, which is particularly slow when dealing with large volumes of data. This inefficiency is a significant barrier when trying to access and analyze massive datasets.
* The process of retrieving filings involves algorithms that can be complex and prone to errors, especially when incremental changes are required to accommodate different filing forms or companies.
* Existing methods for extracting textual information from financial statements often rely on specific HTML parsers, making them platform-dependent and difficult to integrate with other tools.
* Existing methods may not provide easily replicable or validated results, making it difficult for researchers to verify and reproduce findings.

problems which are not addressed:

* The paper does not discuss potential issues related to the quality and accuracy of the data extracted from SEC filings. Errors in the original filings or inconsistencies in how data is presented can impact the analysis.
* The paper focuses on the methods for extracting data from filings formatted in HTML and does not address potential issues with other formats or changes in filing standards over time, such as those involving XBRL or newer formats.
* The paper does not address concerns related to data privacy and security, particularly if sensitive financial information is involved or if the data extraction process involves handling personally identifiable information.
* While the paper provides a method for extracting structured textual data, it does not address challenges related to handling unstructured data or integrating extracted data with other sources.
* The performance and scalability of the proposed algorithms for very large datasets or high-frequency data retrieval are not discussed. As the volume of data grows, ensuring that the methods remain efficient and performant can be challenging.
* The paper does not discuss the user interface or usability aspects of the tools and algorithms provided. How easily users can implement and interact with these tools is crucial for practical adoption.
* There is no discussion on how the proposed methods might integrate with other financial data sources or databases, which could be important for comprehensive financial analysis.

**Paper 6**

Title: Extracting Financial Data from Unstructured Sources: Leveraging Large Language Models

Link: https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=4567607

Problem statement:

The research addresses the difficulty of extracting financial data from unstructured PDF files, which complicates the work of accounting researchers, investors, and regulators. The study aims to develop an automated framework using large language models (LLMs) to improve accuracy and efficiency in extracting key financial indicators from these documents.

problems which are addressed:

* The research addresses the difficulty of extracting financial data from unstructured formats, such as PDFs of governmental annual comprehensive financial reports (ACFRs) and environmental, social, and governance (ESG) reports. Existing methods fall short in efficiently processing and standardizing this data.
* Conventional text mining techniques like sentiment analysis and topic modeling struggle with aggregative summarization and fail to isolate granular data effectively. The study aims to overcome these limitations by employing large language models (LLMs) for more precise data extraction.
* The research highlights that manual data collection is still prevalent and inefficient, particularly for documents with inconsistent formatting and descriptions. The proposed LLM-enabled framework offers a faster and more accurate alternative to manual methods.
* The study addresses the issue of the absence of well-developed taxonomies or machine-readable formats for financial data, which complicates data extraction. The LLM-based framework aims to bridge this gap by providing an automated solution.
* The research seeks to provide a cost-effective and accessible method for financial data extraction, which is crucial for both academic and industrial applications. It contrasts its approach with expensive and complex traditional methods like XBRL.

problems which are not addressed:

• Compared with XBRL, the approach lacks a direct one-to-one mapping between extracted data and a clearly defined taxonomy, making data interpretation less intuitive.

• Regulators may be resistant to complex, "black box" extraction processes due to their reliance on specialized expertise and deeper analytical reasoning.

• Future research should focus on implementing traceability within the framework to enhance understanding, credibility, and verifiability of extracted data by providing intrinsic evidence from the context.

• There are concerns about sensitive information leakage when using LLMs. An offline LLM platform for analyzing sensitive data could mitigate this risk.

• The use of LLMs like GPT-4 incurs token-based expenses for batch queries. Alternatives like GPT-3.5 or Facebook’s LLaMA could be more cost-effective or free.

**Paper 7**

Title: :An Exploration of Automatic Text Summarization of Financial Reports

Link: https://aclanthology.org/2021.finnlp-1.1.pdf

Problem statement:

The paper tackles the challenge of automatically summarizing UK annual reports by extracting key sections or sentences.

The goal is to develop models that can effectively summarize the key sections or sentences of the annual reports in a way that captures the essential information, making it easier for stakeholders to digest and understand the contents.

problems which are addressed:

* Annual reports are lengthy and complex, making manual summarization labor-intensive.
* Converting PDFs to text results in noisy and unstructured data, complicating the summarization process.
* Existing summarization methods, especially sentence-based approaches, often produce incoherent summaries.
* There is a scarcity of publicly available datasets with annual reports and corresponding abstractive summaries.

problems which are not addressed:

* In the future, many limitations could be addressed through improving the accuracy of section identification and extraction.
* We also plan to tackle the task of sentence-based summarization in a more adequate setup by exploring the use of a dataset where summaries are built by extracting sentences in context, instead of full sections.
* Another important point is to conduct an extrinsic evaluation by including several Finance experts to evaluate the quality of the summaries generated automatically.

**Paper 8**

Title:  Extraction of Financial Information from Online Business Reports

Link: https://www.researchgate.net/publication/262322560\_Extraction\_of\_Financial\_Information\_from\_Online\_Business\_Reports

Problem statement:

The problem statement of the paper revolves around the challenge of efficiently extracting specific and relevant information from lengthy and complex business reports, particularly 10-Q reports filed with the SEC.

problems which are addressed:

* Manual extraction of specific information from lengthy business reports is time-consuming and labor-intensive.
* Existing systems, like FIRST, are limited in handling unstructured text and extracting relevant semantic information from extensive reports.
* Ensuring high accuracy in retrieving relevant information and addressing issues with precision and recall in manual extraction processes.
* Evaluating user satisfaction and acceptance of automated systems compared to traditional methods, and ensuring that the system is user-friendly and effective.

problems which are not addressed:

* The study's sample size, while adequate, was limited, and a larger sample with financial professionals could provide more robust results. The study mostly involved business students, and results might differ with experienced financial professionals.
* The focus of the study was on financial reports, and there is potential for CAINES to be applied to other domains, such as healthcare. Future work could explore CAINES' effectiveness in different fields.
* While CAINES was effective in the financial domain, further research is needed to explore how well it can be customized and scaled for other types of text and information extraction tasks.
* The potential for CAINES to conduct more profound content analysis beyond the financial domain was not fully explored.

**Paper 9**

Title:  Financial Statement Analysis with Large Language Models

Link: https://bfi.uchicago.edu/wp-content/uploads/2024/05/BFI\_WP\_2024-65.pdf

Problem statement:

The challenge is to determine whether large language models (LLMs) can perform financial statement analysis and make informed predictions about future earnings with the same accuracy and reliability as human financial analysts.

problems which are addressed:

• Test GPT-4 Turbo’s performance in analyzing financial statements and predicting earnings changes based solely on numerical data.

• Compare the LLM’s predictions with those of professional human analysts to determine if the LLM offers additional value or matches human performance.

• Evaluate the effectiveness of a “Chain-of-Thought” (CoT) prompt versus a “simple” prompt in improving financial analysis results.

• Compare the LLM’s predictions with those from logistic regression and artificial neural networks (ANNs) to assess relative strengths and weaknesses.

• Design a trading strategy based on LLM predictions and compare its performance (Sharpe ratios and alphas) with strategies based on other machine learning models.

• Explore the limitations of LLMs in financial analysis and their potential to democratize financial information processing for investors and regulators.

problems which are not addressed:

* The limitations of LLMs in performing deep numerical reasoning and complex financial judgments.
* The impact of integrating textual context, such as management discussions or industry insights, on the accuracy of financial analysis.
* Practical challenges in scaling LLM-based analysis for real-time financial decision-making.
* The long-term effectiveness and reliability of LLMs in financial predictions and their impact on market dynamics.

**Paper 10**

Title: Automating the extraction of Financial data

Link: https://www.diva-portal.org/smash/get/diva2:1709489/FULLTEXT01.pdf

Problem statement:

The problem in this thesis is that there are no fast and efficient way of extracting financial data of European companies compared to their American counterparts. The process of extracting financial data is most likely done manually which is why it is a time consuming process. This process, at the moment, cannot be automated or done efficiently. The difficulty of accessing financial data in the European market compared to American counterparts deters investors from investing in European markets.

problems which are addressed:

* Unlike the United States, where financial reports are stored in a central database (EDGAR), European companies do not have a central database for financial reports. This makes it harder and time-consuming for investors, especially retail investors, to access financial data.
* Currently, the process of collecting financial data from European companies is largely manual, which is inefficient and slow compared to the automated data extraction systems available for American companies.
* The paper addresses the need for automated systems that can collect financial reports and extract relevant financial data from them in a fast and efficient manner.
* The paper discusses the difficulty of building a web scraper that can handle different design patterns and technologies used on European companies’ websites to collect financial reports effectively.
* European companies structure their websites differently, making it difficult for a single system to effectively scrape investor relations pages and collect all necessary financial reports.

problems which are not addressed:

* Creating webscraper that can navigate to the correct investor relations pages and collect financial reports. One suggestion of how the webscraper could do this is to create a controlled web scraper controlled by an Artificial Intelligence(AI) model.
* The web scraper could visit a website, then take The AI model could use deep learning techniques to decide where to click and send that information back to the web scraper, which follows the instructions of the AI model.
* The separation of financial reports could be done with Machine Learning (ML). The ML model could be trained with an appropriate data set to classify a Financial report from other PDF files. If the ML model can classify a Financial report, then a brute-force algorithm could be implemented with the webscraper; this ML model described could ”clean” the data.