Aalto-yliopisto
Sähkötekniikan
korkeakoulu

Master's Programme in Data Science

Automating Information Extraction from Non-Standard Financial Reports Using Large Language Models

Enhancing Efficiency through Format-Aware Extraction with Large Language Models

Gabriel Gomes Ziegler

Master's Thesis
2024

Aalto-yliopisto
Sähkötekniikan
korkeakoulu

Author Gabriel Gomes Ziegler

Title Automating Information Extraction from Non-Standard Financial Reports Using Large Language Models — Enhancing Efficiency through Format-Aware Extraction with Large Language Models

Degree programme Data Science

Major ICT Innovation

Supervisor Prof. Bo Zhao

Advisor MS Manne Larsson (MSc)

Collaborative partner Datia

Date 21 September 2023 **Number of pages** 14+1 **Language** English

Abstract

The abstract is a short description of the essential contents of the thesis, usually in one paragraph: what was studied and how and what were the main findings.

For a Finnish thesis, the abstract should be written in both Finnish and English; for a Swedish thesis, in Swedish and English. The abstracts for English theses written by Finnish or Swedish speakers should be written in English and either in Finnish or in Swedish, depending on the student's language of basic education. Students educated in languages other than Finnish or Swedish write the abstract only in English. Students may include a second or third abstract in their native language, if they wish.

The abstract text of this thesis is written on the readable abstract page as well as into the pdf file's metadata via the `\thesisabstract` macro (see comment in this TEX file above). Write here the text that goes onto the readable abstract page. You can have special characters, linebreaks, and paragraphs here. Otherwise, this abstract text must be identical to the metadata abstract text.

If your abstract does not contain special characters and it does not require paragraphs, you may take advantage of the `\abstracttext` macro (see the comment in this TEX file below).

Keywords For keywords choose, concepts that are, central to your, thesis

Aalto-yliopisto
Sähkötekniikan
korkeakoulu

Tekijä Gabriel Gomes Ziegler

Työn nimi Opinnäyteen otsikko — Opinnäytteen mahdollinen alaotsikko

Koulutusohjelma Elektroniikka ja sähkötekniikka

Pääaine Sopiva pääaine

Työn valvoja Prof. Pirjo Professori

Työn ohjaajat TkT Alan Advisor, DI Elsa Expert

Yhteistyötaho Yhtiön tai laitoksen nimi (tarvittaessa)

Päivämäärä 21.9.2023 **Sivumäärä** 14+1 **Kieli** englanti

Tiivistelmä

Tiivistelmä on lyhyt kuvaus työn keskeisestä sisällöstä usein yhtenä kappaleena: mitä tutkittiin ja miten sekä mitkä olivat tärkeimmät tulokset. Suomenkielisen opinnäytteen tiivistelmä kirjoitetaan suomeksi ja englanniksi ja ruotsinkielisen vastaavasti ruotsiksi ja englanniksi. Suomen- tai ruotsinkielisten opiskelijoiden, joiden opinnäytteen kieli on englanti, tulee kirjoittaa tiivistelmänsä englanniksi ja koulusivistyskielellään. Muiden kuin koulusivistyskieleltään suomen- tai ruotsinkielisten tulee kirjoittaa tiivistelmänsä vain englanniksi. Opiskelija voi halutessaan lisätä opinnäytteeseensä toisen tai kolmannen tiivistelmän omalla äidinkielellään. Tämän opinnäytteen tiivistelmäteksti kirjoitetaan opinnäytteen luettavan osan lomakkeen lisäksi myös pdf-tiedoston metadataan. Kirjoita tähän metadataan kirjoitettavaa teksti. Metadatatekstissa ei saa olla erikoismerkkejä, rivinvaiho- tai kappaleenjakomerkkiä, joten näitä merkkeja ei saa käyttää tässä. Jos tiivistelmäsi ei sisällä erikoimerkkejä eikä kaipaa kappaleenjakoa, voit hyödynttää makroa abstracttext luodessasi lomakkeen tiivistelmää (katso kommentti tässä TeX-tiedostossa alla). Metadatatiivistelmatekstin on muuten oltava sama kuin lomakkeessa oleva teksti.

Avainsanat Vastus, resistanssi, lämpötila

Aalto-yliopisto
Sähkötekniikan
korkeakoulu

Författare Gabriel Gomes Ziegler

Titel Arbetets titel — Opinnäytteen mahdollinen alaotsikko

Utbildningsprogram Electronik och electroteknik

Huvudämne Sopiva pääaine

Övervakare Prof. Pirjo Professori

Handledare TkD Alan Advisor, DI Elsa Expert

Samarbetspartner Company or institute name in Swedish (if relevant)

Datum 21.9.2023	**Sidantal** 14+1	**Språk** engelska

Sammandrag

Sammandraget är en kort beskrivning av arbetets centrala innehåll: vad undersöktes, hur undersöktes det och vilka var de viktigaste resultaten?

I lärdomsprov som skrivs på svenska skrivs sammandraget på svenska och engelska, på motsvarande sätt skrivs sammandraget på finska och engelska i lärdomsprov på finska. Finsk- eller svenskspråkiga studerande som skriver sitt lärdomsprov på engelska ska skriva sammandraget på engelska och på sitt skolutbildningsspråk. Studerande vars skolutbildningsspråk inte är svenska eller finska skriver sammandraget endast på engelska. Den studerande kan om hen så önskar lägga till ett andra eller tredje sammandrag på sitt eget modersmål. Sammandraget fungerar då ofta som mognadsprov och bör i så fall vara minst 300 ord långt. Information om mognadsprov på svenska finns på MyCourses:
`https://mycourses.aalto.fi/course/view.php?id=26872`.

Nyckelord Nyckelord på svenska, temperatur

Preface

Thanks notes

Otaniemi, 31 August 2024

Eddie E. Engineer

Contents

NLP	Natural Language Processing	9
PDFs	Portable Document Formats	10
OCR	Optical Character Recognition	10
LLMs	Large Language Models	7
LLM	Large Language Model	7
GPT	Generative Pre-trained Transformer	8
BERT	Bidirectional Encoder Representations from Transformers	8
KPI	Key Performance Indicator	10

1 Introduction

1.1 Background of the Field of Study

The field of data extraction from financial reports has evolved significantly with advancements in text processing and machine learning technologies. Historically, this task involved manual data entry or rule-based systems that were labor-intensive and prone to errors. The emergence of LLMs, such as Generative Pre-trained Transformer (GPT) and Bidirectional Encoder Representations from Transformers (BERT), has revolutionized this domain. These models have the ability to understand and extract complex financial information from unstructured data, thereby increasing accuracy and efficiency. Recent studies have demonstrated the potential of LLMs in automating financial data extraction, highlighting improvements in processing time and data accuracy over traditional methods.

1.2 General Objective

This study aims to extend the current capabilities of data extraction systems by incorporating advanced LLMs and exploring novel methodologies in the field. The primary goals include: ellaborating a comprehensive comparison of methods for extracting information from financial reports, with a focus on non-standard reports,

enhancing the precision and efficiency of data extraction from financial reports, developing a scalable system capable of processing large volumes of data, and comparing the effectiveness of various LLMs and extraction techniques. By achieving these goals, the study seeks to contribute to the broader understanding of automated data extraction and its application in financial analysis.

1.3 Research Question and Sub-Problems

The primary research question of this study focuses on: "LLMs be optimized for more accurate and efficient extraction of financial data from unstructured reports?" Sub-problems in this line of inquiry include: identifying the most effective LLM architectures for financial data recognition, developing methodologies for context-aware data extraction, enhancing the system's ability to handle diverse report formats, and evaluating the impact of training data quality and volume on model performance. These sub-problems are essential for understanding the intricacies of applying LLMs to financial data extraction and for developing a comprehensive solution.

Scope and Constraints

The scope of this study is limited to the extraction of financial data from English-language reports, focusing on publicly available annual and quarterly financial statements. Key constraints include the variability in report formats, the complexity of financial terminology, and the inherent limitations of current LLM technologies in understanding domain-specific contexts. The study primarily revolves around the use of GPT and BERT models, considering their widespread adoption and state-of-the-art performance in text processing tasks. Main concepts involved include Natural Language Processing (NLP), machine learning, data extraction, and financial analysis, with a particular emphasis on the adaptation and optimization of LLMs for specialized data extraction tasks.

2 Literature review

Ever since Portable Document Formats (PDFs) were created by Adobe in 1993, they have been used to store and share information. These document standard quickly became a way of companies reporting their financial information for the public as well as Key Performance Indicator (KPI)s and other important information internally. This has led to a large amount of information being stored in PDFs, which has led to a need to extract information from these files. A series of professions have arised from this need, such as data entry, data extraction, and data analysis. The extraction of information from PDFs has been a manual process for most of the tasks until recent years, when Optical Character Recognition (OCR) and NLP technologies have been developed to automate processes involving processing PDFs.

Extracting information from a document, recently referred to "Document AI" is a complex problem that often involves cross-modal interactions where information is represented in both text and visual form. This is particularly true for financial reports, where information is often presented in tables, charts, and text. The problem is further complicated by the fact that financial reports are often not standardized, and the information is presented in diverse range of formats. methods:

vector search gpt4 multimodal vision Q&A with RAG

2.1 Structure of the thesis

The thesis is composed by a comprehensive comparison of methods for extracting information from financial reports, with a focus on non-standard reports. The thesis is structured as follows:

1. Introduction
2. Literature review
3. Research material and methods
4. Results / Findings
5. Discussion
6. Summary / Conclusions
7. References

3 Financial Reports Dataset

Dataset used to benchmark different methods.

4 Extracting information from financial reports

In this section, we define the metrics, methods and processes used to extract information from financial reports

4.1 Metrics and Evaluation Criteria

4.2 System Specifications

Define the system specifications and requirements used to run the experiments

4.3 LLM to make sense of text

LLM model using only text to identify key indicators reported in PDF files.

4.4 Multimodal LLMs to extract information from images

4.5 Multimodal LLMs to extract information from images and text

5 Results

Present the results of your study here and answer the research questions, asked earlier in the thesis (in the introduction, perhaps), this study strives to answer. The scientific value of your work is measured by the results you obtain along with the arguments you give to back the answers to your research questions.

Be critical of the significance of your results. You may critically scrutinise the results and your interpretation of the results here, or you may do so later in the chapter with the discussion of your work or in the conclusions part.

This part should discuss how reliable the data used in the study are. You may discuss the reliability of the conclusions drawn from the study either in this chapter or later in the discussions part. You may have the discussion in a chapter of its own, separate from the summary or conclusions.

6 Summary/Conclusions

This is where you tie up any loose ends. Tell your reader briefly and clearly what you have done, what you have discovered, and the value of your discovery in the context of similar work done earlier. Draw clear conclusions regarding the research problem, sub-problems or hypotheses. You also discuss future lines of study and new questions your study might have posed.

As the author of the thesis, you alone are responsible for ensuring that the layout, form and structure of your thesis adheres to the guidelines outlined by your school. This template aims to help you meet these requirements.

References

This is the list of references to the sources cited in appendix **??**. The list more or less follows the Vancouver style (IEEE). See appendix **??** for a detailed exposition on cross-referencing and bibliography styles. Follow the description there.

[1] Citation Guide: Making a bibliography, *Aalto University Learning Centre*. Online article. Available `https://libguides.aalto.fi/c.php?g=410674&p=2797572` (accessed on 14.7.2021)

A Contents of an appendix