University of Manchester
School of Computer Science
Project Report 2023

Extractive Summarisation of UK Annual Reports

Author: Vladislav Yotkov

Supervisor: Dr. Jonathan Shapiro

Abstract

Extractive Summarisation of
UK Annual Reports

Author: Vladislav Yotkov

Although there has been considerable progress in Natural Language Processing (NLP) over the years, it has not been quite integrated in the Accounting and Finance (AF) industry. In the meantime, companies worldwide produce vast amounts of textual data as part of their reporting packages to comply with regulations and inform shareholders of their financial performance. The glossy annual report is such an example, widely read by investors but also quite long. Inspired by the Financial Narrative Summarisation (FNS) 2021 Task, we will design an Automatic Text Summarisation (ATS) system for the narrative parts of UK financial annual reports. With this goal in mind, we will implement and explore the following models for Extractive Text Summarisation (ETS): 1. custom Recurrent Neural Network (RNN), 2. fine-tuned FinBERT. In terms of evaluation, we will use the ROUGE metric to compare the performance of these models against standard ATS baselines: TextRank, and LexRank.

Supervisor: Dr. Jonathan Shapiro

0.1 Introduction

0.1.1 Financial Reports

Due to international regulations, companies are obliged to report their periodic performance (annual, bi-annual, quarterly) to various regulatory authorities[1] and other users (e.g., corporate stakeholders, investors, customers, suppliers, etc.). These reports contain essential information about the operations and finances of a business and are crucial for making informed decisions (from user perspective), but differ in regulatory forms. For example,

1. 10-K reports filed to the SEC[2] and accessible through their Electronic Data Gathering, Analysis, and Retrieval[3] (EDGAR) system (only for US registrants). They follow a standardised template and are plain text, which makes them particularly easy for automated large-scale research ([EHAR+19]). Also, the contents of these reports is quite strict, requiring solely the five information sections: (a) Business Overview (b) Risk Factors (c) Management's Discussion and Analysis of Financial Condition and Results of Operations (MD&A) (d) Financial Statements (e) Supplementary Disclosures.

2. UK annual reports, the regulation of which is overseen by the Financial Reporting Council (FRC). Unlike the 10-K, they are glossy and more stakeholder-oriented, and enjoy unlimited discretion over non-mandated content ([EHAR+19]) (e.g., photography and company brand material, non-mandatory narrative sections, etc.). These are more challenging for automated processing, due to their variable section structure, formatting and rich visual representations (e.g., infographics).

As outlined in [Ell98], investors' trust in the accountability of businesses would be based no longer as much on just the financial statements, but also on more descriptive narratives that define strategy and planning of resource use.

[1] Regulation authorities worldwide:

- Securities and Exchange Commission (SEC) in the USA
- European Securities and Markets Authority (ESMA) in Europe
- Financial Reporting Council (FRC) in the UK
- International Financial Reporting Standards (IFRS) in 167 jurisdictions worldwide

[2] `https://www.sec.gov`

[3] `https://www.sec.gov/edgar`

0.1.2 NLP in Accounting and Finance

The relevance of this project should also be understood from the perspective of the development of Natural Language Processing (NLP) in the Accounting and Finance (AF) domain [EHRW$^+$19]. report that this industry is doubtful and cynical about the application of Computational Linguistic (CL) methods in analysing financial market disclosures. Furthermore, they also observe that AF researchers rely extensively on bag-of-words models, which are *not sufficient to encode complex contextual and semantic meaning* (especially in a domain with such *specialized language*). As for ATS [CHW19], is said to be the single AF study into disclosure summarisation and it demonstrates that machine-generated summaries are less likely to positively bias investors' decisions compared to managerial ones. This only confirms that there is a wide gap in NLP applications in Accounting research, and this further motivates our work.

Bibliography

[CHW19] Eddy Cardinaels, Stephan Hollander, and Brian White. Automatic summarization of earnings releases: attributes and effects on investors' judgments. *Review of Accounting Studies*, 24, 09 2019.

[EHAR+19] Mahmoud El-Haj, Paulo Alves, Paul Rayson, Martin Walker, and Steven Young. Retrieving, classifying and analysing narrative commentary in unstructured (glossy) annual reports published as pdf files. *Accounting and Business Research*, 2019. Forthcoming.

[EHRW+19] Mahmoud El-Haj, Paul Rayson, Martin Walker, Steven Young, and Vasiliki Simaki. In search of meaning: Lessons, resources and next steps for computational analysis of financial discourse. *Journal of Business Finance & Accounting*, 46(3-4):265–306, 2019.

[Ell98] Robert K Elliott. Accounting in the 21st century. 1998.