

Analysing UK Annual Report Narratives using Text Analysis and Natural Language Processing

Mahmoud El-Haj
SCC, Lancaster University



Corporate Financial Information Environment

RESEARCH QUESTIONS

- Can we advance research on the lexical properties and narrative aspects of corporate disclosures using NLP?
- Can we publish what's already freely available?
- Are firms honest with us?
- Could the news tell us something?
- Who writes corporate disclosures?
- What if we enforced a standard?

Things to know:

- Work on analysing corporate disclosures ranges from manual analysis up to using WordSmith. (concordance, wordlist, keywords count)
- Crowdsourcing is what some firms use to quickly build their databases

ANALYSING UK ANNUAL REPORT NARRATIVES

- Project background
- Before our work...
- Work so far... in progress...in the future...
- UK vs USA Filings (Current Research)
- Overview of our research tool
- Answering research questions
- Demo of research tool

- Questions

BACKGROUND & OBJECTIVES



- Part of an ESRC- and ICAEW-funded project examining the *Corporate Financial Information Environment (CFIE)*
 - Martin Walker, *Manchester Business School*
 - Steven Young, *Lancaster University Management School*
 - Paul Rayson, *Lancaster University School of Computing & Communications*
 - Mahmoud El-Haj, *Lancaster University School of Computing & Communications*
 - Vasiliki Athanasakou, *London School of Economics*
- Project seeks to analyse UK financial narratives, their association with financial statement information, and their informativeness for investors
- Automated, large sample analysis of UK annual report narratives represents a cornerstone of the project
 - Develop a tool for general use by academics

BEFORE...

- Researchers worked on document level
- Section level analysis had to be done manually
- There was no clear relationship between documents
- Analysis didn't go beyond word level
- And therefore it wasn't possible to automate any cross sectional analysis



SO FAR... IN PROGRESS...IN THE FUTURE...

- So far:
 - NLP suit to aid researcher in quickly analysing financial disclosures that goes beyond simple word level analysis.
 - Built financial disclosures section-based (reference) corpus
 - Used machine learning to train a system to detect attribution and tone in financial sentences.
 - Used heuristic approaches to define Performance and Strategic sections.
- In progress:
 - Study the relationship and similarities between press releases and media articles using text reuse.
 - Use NLP and machine learning to detect strategic sections in annual reports
 - Use web as a corpus to harvest freely available financial documents.
 - Study text reuse in financial disclosures.
- Future
 - Work with Annual Reports in Chinese and some European languages
 - Using crowdsourcing to help build lexicons for each language

UK vs USA FILINGS

WHY'S USA LEADING?



CURRENT RESEARCH IS ON USA FILINGS!

- Majority of large sample analysis of annual report narratives has been conducted on US filings
- Usually 10-Ks (aka Annual Reports)
- those available via EDGAR (Free access to more than 20 million filings)
- Analysis of 10-K filings in EDGAR is relatively straightforward
 - Plain text files with consistent structure
 - Use HTML parser to identify section(s) and extract text



10-K ANNUAL FORM

Each 10-K contains 4 parts and 15 items

- **PART I**
- **ITEM 1.** Description of Business
- **ITEM 2.** Description of Properties
- **ITEM 3.** Legal Proceedings
- **ITEM 4.** Mine Safety Disclosures
- **PART II**
- **ITEM 5.** Market for Registrant's Common Equity....
- **ITEM 6.** Selected Financial Data
- **ITEM 7.** Management's Discussion and Analysis....
- **ITEM 8.** Financial Statements and Supplementary Data
- **ITEM 9.** Changes in and Disagreements
- **PART III**
- **ITEM 10.** Directors, Executive Officers and Corporate Governance
- **ITEM 11.** Executive Compensation
- **ITEM 12.** Security Ownership of Certain Beneficial Owners....
- **ITEM 13.** Certain Relationships and Related Transactions....
- **ITEM 14.** Principal Accounting Fees and Services
- **PART IV**
- **ITEM 15.** Exhibits, Financial Statement Schedules....



10-K ANNUAL (STARBUCKS VS. MCDONALD'S)

Starbucks Corporation

PART I

- Item 1 [Business](#)
- Item 1A [Risk Factors](#)
- Item 1B [Unresolved Staff Comments](#)
- Item 2 [Properties](#)
- Item 3 [Legal Proceedings](#)
- Item 4 [\(Removed and Reserved\)](#)

PART II

- Item 5 [Market for the Registrant's C](#)
- Item 6 [Selected Financial Data](#)
- Item 7 [Management's Discussion an](#)
- Item 7A [Quantitative and Qualitative I](#)
- Item 8 [Financial Statements and Sup](#)
- Item 9 [Report of Independent Regist](#)
- Item 9A [Changes in and Disagreemen](#)
- Item 9B [Controls and Procedures](#)
- Item 9B [Other Information](#)

PART III

- Item 10 [Directors, Executive Officers](#)
- Item 11 [Executive Compensation](#)

McDONALD'S CORPORATION

Part I.

- Item 1 [Business](#)
- Item 1A [Risk Factors and Cautionary](#)
- Item 1B [Unresolved Staff Comments](#)
- Item 2 [Properties](#)
- Item 3 [Legal Proceedings](#)
- Item 4 [Mine Safety Disclosures](#)

Part II.

- Item 5 [Market for Registrant's Comn](#)
- Item 6 [Selected Financial Data](#)
- Item 7 [Management's Discussion ar](#)
- Item 7A [Quantitative and Qualitative I](#)
- Item 8 [Financial Statements and Su](#)
- Item 9 [Changes in and Disagreemer](#)
- Item 9A [Controls and Procedures](#)
- Item 9B [Other Information](#)

Part III.

- Item 10 [Directors, Executive Officers](#)
- Item 11 [Executive Compensation](#)



UK FILINGS

- Content and structure varies across firms.
- Management have more discretion over what, where, and how much information on topics such as risk, strategy, performance, etc. is reported.
- UK annual reports pose more significant challenges to researchers
 - Normally supplied as *.pdf
 - No consistent template



UK ANNUAL REPORTS SAMPLE

Financial highlights

Sales

+6.8%

Sales (including VAT, including fuel)

Underlying operating profit

£789m

Underlying operating profit up 6.9%

Underlying profit before tax

£712m

Underlying profit before tax up 7.1%

Return on capital employed

11.1%

Return on capital employed

Underlying basic earnings

28.1p

Underlying basic earnings per share up 6.0%

Contents

Business review

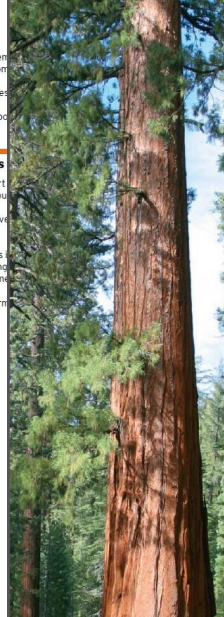
Financial highlights	1
Chairman's letter	2
Chief Executive's letter	4
Market overview	6
Key performance indicators	8
Our strategy	10
Great food	12
Compelling general merchandise & clothing	14
Complementary channels & services	16
Developing new business	18
Growing space & creating property value	20
Operational excellence	22
Our values make us different	24
Financial review	26

Governance

Board of Directors	
Operating Board	
Directors' report	
Corporate governance statement	
Corporate Responsibility Committee	
Audit Committee	
Principal risks & uncertainties	
Remuneration report	
Statement of Directors' responsibilities	

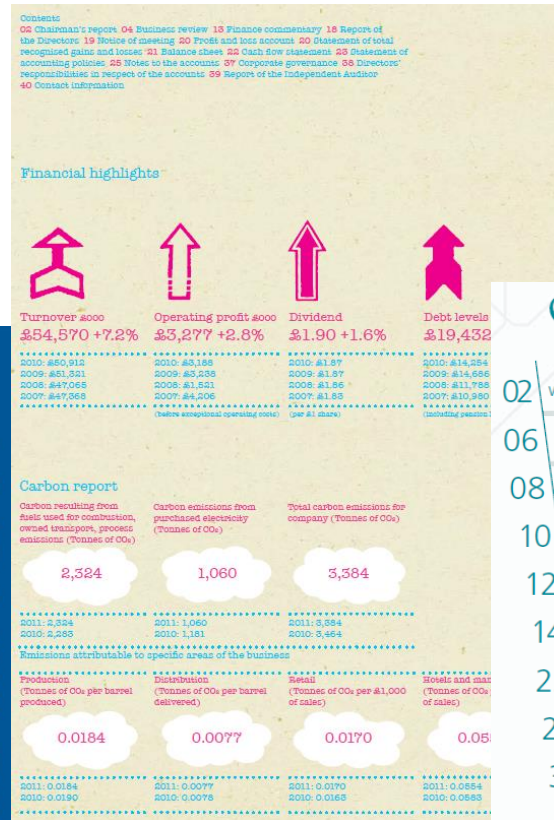
Financial statements

Independent auditors' report to the members of J Sainsbury plc	
Group income statement	
Statements of comprehensive income	
Balance sheet	
Cash flow statements	
Group statement of changes in equity	
Company statement of changes in financial statements	
Notes to the financial statements	
Five year financial record	
Additional shareholder information	
Financial calendar	
Glossary	



Contents

Spirax Sarco at a glance	6
Chairman's statement	8
Business review	10
Market overview	10
Performance review	15
Board of Directors	28
Directors' report	31
Corporate governance	34
Corporate social responsibility	38
The Directors' remuneration report	42
Statement of Directors' responsibilities	50
Financial statements	51
Report of the independent auditor	51
Group income statement	52
Balance sheets	53
Statements of recognised income and expense	54
Cash flow statements	55
Notes to the accounts	56
Financial summary	86
Officers and advisers	88



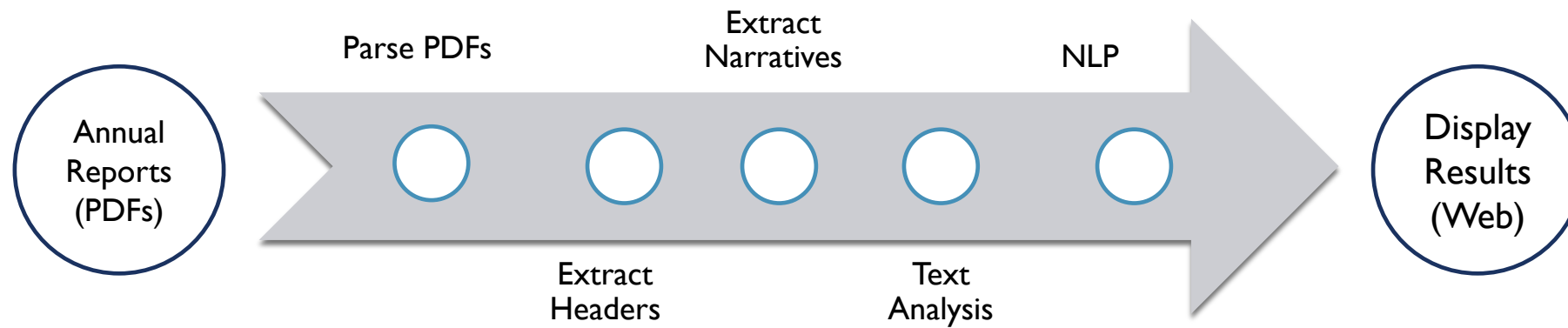
02 Who we are and what we do	42 Corporate governance
06 24 hours in the life of Arriva	46 Statement of directors' responsibilities
08 Our growth story	47 Independent auditors' report on group financial statements
10 Our markets	48 Financial statements
12 Chairman's statement	52 Accounting policies
14 Chief executive's review	56 Notes to the accounts
22 Financial review	82 Five-year financial performance
26 Corporate responsibility	83 Parent company financial statements
32 Board of directors	90 Statement of the parent company
34 Directors' report	91 Independent auditors' report on the parent company financial statements
37 Directors' remuneration report	92 Financial calendar, registered office and advisors



DATASET

- >14,000 searchable financial annual reports
- of around 500 of the largest UK firms listed on the LSE
- between the years 2002 and 2014.
- Annual reports were automatically downloaded using a Perl script and some Java code.
- >150,000 media articles and press releases (US firms)

CFIE ANALYSIS PIPELINE



EXTRACTION PROCESS

WHAT ARE WE LOOKING TO EXTRACT?



HEADERS AND THEIR SECTIONS

- Front End

1. Chairman's statement
2. CEO Review
3. Corporate Government Report
4. Directors Remuneration Report
5. Directors Report and Business Review
6. Directors Responsibilities Statement
7. Directors Report
8. Financial Review
9. Key Performance Indicator
10. Operational Review
11. Highlights

- Back End

- Financial statements (financial numbers)

UK ANNUAL REPORT TOOL: EXTRACTION

- Use contents page to extract text by section from digital pdf
- Steps in extraction process:
 - Detect contents page
 - Parse contents page
 - Detect page numbering to determine section start/end
 - Add headers as bookmarks to pdf
 - Extract text for each section
- Analyse extracted text by section and for entire document

Contents

02	Chairman's Statement
05	Directors' Report
09	Directors' Remuneration Report
14	Corporate Governance Report
18	Auditors' Report
20	Consolidated Profit and Loss Account
21	Consolidated Balance Sheet
22	Company Balance Sheet
23	Consolidated Cash Flow Statement
24	Notes to the Financial Statements
36	Notice of Annual General Meeting
39	Directors and Advisors

HOW?

Not consistent across ARs

Contents

Spirax Sarco at a glance 6

Chairman's statement 8

Business review

Market overview

Performance review.....

Board of Directors.....

Directors' report.....

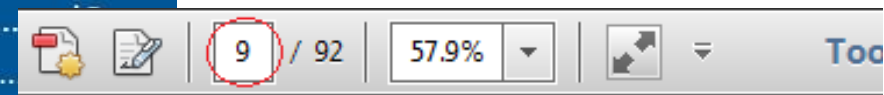
Corporate governance

Corporate social responsibility

The Directors' remuneration report

Statement of Directors' responsibilities.....50

Doesn't always refer to the correct page



to find out more, visit marksandspencer.com/annualreport2010

Chairman's overview

by Sir Stuart Rose



I) DETECTING THE CONTENTS PAGE

- Experts gold–standard
- highest matching score is the potential contents page
- Search by structure – e.g. *Chairman’s Statement 13*.

2) PARSING THE CONTENTS PAGE

- Use regEx for parsing
- Use linguistic algorithms to avoid dates/addresses/figures (e.g. 77 million)
- Use synonyms and heuristic approaches for header's type (e.g. Chairman Statement)
- To tackle the problem of broken headers we concatenate sentences that end or begin with prepositions such as 'of', 'in' ...etc.
- The algorithm also concatenates sentences ending with singular or plural possessives, symbolic and textual connectors (e.g. 'and', 'or', '&'...etc), and sentences ending with hyphenations.

Corporate Governance Report

46	Directors and Secretary
48	Shareholders and Share Capital
50	Other Statutory Information
52	Corporate Governance Statement
58	Remuneration Report

Group Financial Statements

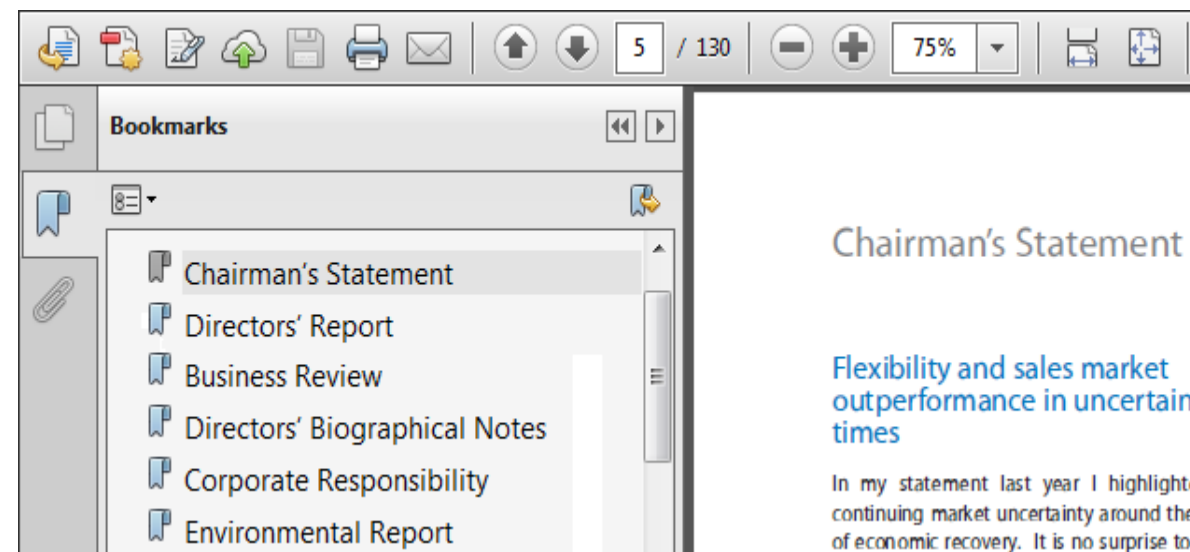
67	Independent Auditor's Report
68	Group Income Statement
69	Group Balance Sheet
70	Group Cash Flow Statement
71	Group Statement of Recognised Income and Expense
72	Accounting Policies
78	Notes to the Accounts

3) DETECTING PAGE NUMBERING

- The page numbers appearing on the contents page do not usually match with the actual page numbers in the pdf files.
- Created a simple page detection tool that crawls through all the numbers in the pdf pages and try to detect a pattern of consecutive number with an increment of 1 (e.g. 31, 32, 33). Comparing that to the actual pdf page number the tool then picks the difference with highest vote.
- Running this process we got an accuracy rate of 98%.
- Manually examining a sample of the 2% showed 1) encoding, 2) formatting and 3) design (booklet).

4) ADDING HEADERS AS BOOKMARKS

- Using the headers and their correct page numbers we implemented a tool to insert the extracted contents page headers as bookmarks (hyperlinks) to sample PDFs.



5) EXTRACTING HEADERS' NARRATIVES

We get everything but mainly interested in front end:

1. Chairman's statement
2. CEO Review
3. Corporate Government Report
4. Directors Remuneration Report
5. Directors Report and Business Review
6. Directors Responsibilities Statement
7. Directors Report
8. Financial Review
9. Key Performance Indicator
10. Operational Review
11. Highlights

RESEARCH TOOL OVERVIEW

- In addition to performing text extraction, the tool provides a range of text analysis options:
 - Readability metrics
 - Word counts using pre-determined lists (e.g., forward looking, uncertainty, tone, etc.)
 - Word counts based on user-defined wordlists
 - Part of speech tagging
 - Semantic tagging
 - Comparison with reference corpus (word, parts of speech and semantic level)
 - Concordance and collocates
 - Word clouds
 - Search for word in context
 - ...etc

ANSWERING RESEARCH QUESTIONS



Can we advance research on the lexical properties and narrative aspects of corporate disclosures using NLP?

- Yes
- Speed up processing time
- Provide extra level of analysis than just keyword and wordlist levels.
- Help relate profit/loss in firms to tone/attribution and other linguistic features.
- Help analyse management narratives and for example check number of times management shifts from singular to plural pronouns or using past and present verbs.

...etc

CAN WE PUBLISH WHAT'S ALREADY FREELY AVAILABLE?

- No for the files and Yes for the output/results
- Even though most of the financial disclosures are freely available we still cannot publish files collected and organised by other database companies.
- Solutions:
 - Use web as a corpus
 - Standardise data (firm codes, report layout, sections ...etc)

ARE FIRMS HONEST WITH US?

- Companies will try to hide or downplay their losses – this can be sometimes automatically detected by measuring the tone and the use of forward looking statements (compared to previous reports).
- Analyst are usually behind uncovering these information but takes them a huge amount of time to do so (e.g. 90 reports in 6 months).

COULD THE NEWS TELL US SOMETHING?

- Part of the work we are conducting analyses press releases (by firms) and media articles (by news).
- The work studies the relationship between what a company reports and what media says about it.
- We collect media articles at a period following the release of a press release.

WHO WRITES CORPORATE DISCLOSURES?

- Some companies reuse their previous reports and add/edit/delete sections.
- It's hard to tell who writes an annual report - interviewing an expert in writing annual reports we found that:
 - Up to 20 people could participate in writing annual reports
 - Chairman usually don't write their own section but will read it and ask for edits.
 - They hire a company to do so.
- We used 3 text reuse methods for comparing and measuring text similarity and derivation in sets of texts.:
 - Word Error Rate (WER)
 - Cosin Similarity
 - Texas (identifying text reuse based on text alignment)

WHAT IF WE ENFORCED A STANDARD?

- This will upset many database companies.
- But will benefit analyst, and researchers in the field of accounting and finance as well as investors.
- How?
 - Having a standard format for wiring annual reports will help in restricting what firms report
 - Help in better understanding of certain sections or where the company stands in the current market.
 - Until now there is not a single identifier to link all the databases available which makes it difficult to study for example annual reports and press releases in one go and that's what we provide researchers with.

UK ANNUAL REPORT TOOL: DEMO

- <https://cfie.lancaster.ac.uk:8443/>

QUESTIONS

CFIE UREL: <http://ucrel.lancs.ac.uk/cfie>

