Financial Narrative Processing (FNP)

Mahmoud El-Haj
Lancaster University

DocElhaj

http://www.lancaster.ac.uk/staff/elhaj/
Financial Narrative Projects

- CFIE
  - ESRC, ICAEW, UCREL
  - 2012

- CorpComm
  - ESRC
  - 2014

- FNP
  - ICRA, UCREL, LUMS, FRC
  - 2016
Project Team

CFIE ¹

CorpComm ²

FNP ³

SCC
Dr Mahmoud El-Haj¹,²,³
Dr Paul Rayson¹,²,³

LUMS
Prof Steve Young¹,²,³

MBS
Prof Martin Walker¹,²,³
Dr Thomas Schleicher¹

CASS
Prof Tony McEnery²
Dr Andrew Hardie²

LSE
Dr Vasiliki Athanasakou¹
What is it about?

The projects analyse U.K. financial narratives, their association with financial statement information, and their informativeness for investors.
Contributions

• Develop an automated method for extracting narrative content and structure from UK annual reports provided as PDFs

• First large-sample tests of the incremental predictive ability of UK annual report narratives

• First study to examine the incremental and differential predictive ability of alternative annual report narrative sections

• First study to model disagreement between preparers’ and to examine the impact of disagreement on the predictive properties of narratives
Why Financial Narratives

a doubling of the median word count over the sample period
Why Financial Narratives

a doubling of the median page count over the sample period
FNP: Fields of Study

- NLP
- Discourse Analysis
- Computational Linguistics
- Corpus Linguistics
- Text Analysis
- Machine Learning
Annual Reports: how big?

Report Count

Word Count

Number of reports

word count in Millions

10,443

312
Financial Media: how big?

**News Articles Count**

- Media Articles: 282
- Press Releases: 86

**Word Count**

- Media Articles: 179
- Press Releases: 138

<table>
<thead>
<tr>
<th></th>
<th>UK</th>
<th>USA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Articles</td>
<td>54</td>
<td>44</td>
</tr>
<tr>
<td>Words</td>
<td>30</td>
<td>109</td>
</tr>
</tbody>
</table>
Previous Work

Can we apply NLP methods used in prior US studies to UK annual reports?
US Filings

• US companies must submit:
  1. 10-K: Annual
  2. 10-Q: Quarterly
  3. 8-K: Special Events (between 10-K and 10-Q)
  4. Annual Report
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 Annuals (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 Common Equity and Related Stockholder Matters
Item 6 Selected Financial Data
Item 7 Management’s Discussion and Analysis of Financial Condition and Results of Operations
Item 7A Quantitative and Qualitative Disclosure About Market Risk
Item 8 Financial Statements and Supplementary Data
Item 9 Changes in and Disagreements with Accountants on Accounting and Financial Disclosure
Item 9A 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 Statement Regarding Forward-Looking Information
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 Common Equity and Related Stockholder Matters
Item 6 Selected Financial Data
Item 7 Management’s Discussion and Analysis of Financial Condition and Results of Operations
Item 7A Quantitative and Qualitative Disclosure About Market Risk
Item 8 Financial Statements and Supplementary Data
Item 9 Changes in and Disagreements with Accountants on Accounting and Financial Disclosure
Item 9A Controls and Procedures
Item 9B Other Information

Part III.
Item 10 Directors, Executive Officers
Item 11 Executive Compensation
UK Annual Reports

- Free style (no standard structure)
- Use of images, text, hyperlinks, ...etc.
- PDF format
  - 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.

This makes the extraction and analysis task more challenging; but it provides research opportunities.
CFIE-FRSE Tool

Annual Reports (PDFs)

Parse PDFs

Extract Headers

Extract Narratives

Text Analysis

NLP

Display Results (Web)
UK Annual report tool: Extraction

Steps in extraction process:

- Detect contents page
- Parse contents page
- Extract section
- Detect section type
- Reorder section
### 2005 financial highlights

<table>
<thead>
<tr>
<th>IFC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Who we are</td>
</tr>
<tr>
<td>01 A short voyage around our business</td>
</tr>
<tr>
<td>02 Number one UK ports operator</td>
</tr>
<tr>
<td>04 Long-term, blue chip customers</td>
</tr>
<tr>
<td>06 Almost a quarter of the UK’s seaborne trade</td>
</tr>
<tr>
<td>07 Our markets</td>
</tr>
<tr>
<td>08 Investment programme</td>
</tr>
<tr>
<td>14 Generating revenue from coal</td>
</tr>
<tr>
<td>16 Immingham Outer Harbour</td>
</tr>
<tr>
<td>18 Hull shortsea container terminal</td>
</tr>
<tr>
<td>20 Sizeable returns from smaller investments</td>
</tr>
<tr>
<td>22</td>
</tr>
<tr>
<td>2. How we have performed</td>
</tr>
<tr>
<td>24 Chairman’s statement</td>
</tr>
<tr>
<td>26 Group Chief Executive’s review of strategy</td>
</tr>
<tr>
<td>28 Operating and financial review</td>
</tr>
<tr>
<td>32</td>
</tr>
<tr>
<td>3. How our results add up</td>
</tr>
<tr>
<td>50 Group income statement</td>
</tr>
<tr>
<td>52 Balance sheets</td>
</tr>
<tr>
<td>53 Cash flow statements</td>
</tr>
<tr>
<td>54 Statement of recognised income and expense</td>
</tr>
<tr>
<td>55 Notes to the financial statements</td>
</tr>
<tr>
<td>56</td>
</tr>
<tr>
<td>4. How we behave</td>
</tr>
<tr>
<td>98 Board of directors</td>
</tr>
<tr>
<td>100 Operational management team</td>
</tr>
<tr>
<td>102 Statement of directors’ responsibilities</td>
</tr>
<tr>
<td>104 Independent auditors’ report</td>
</tr>
<tr>
<td>105 Directors’ report</td>
</tr>
<tr>
<td>106 Corporate governance</td>
</tr>
<tr>
<td>108 Remuneration report</td>
</tr>
<tr>
<td>118 Shareholder analysis</td>
</tr>
<tr>
<td>129 Corporate social responsibility</td>
</tr>
<tr>
<td>130 Notice of meeting</td>
</tr>
<tr>
<td>134 Five-year summary</td>
</tr>
<tr>
<td>136 Company information</td>
</tr>
<tr>
<td>138 Glossary</td>
</tr>
<tr>
<td>140</td>
</tr>
</tbody>
</table>
2006 financial highlights

1. Who we are
   01
   A short voyage around our business 02
   Number one UK ports operator 04
   Long-term, blue chip customers 06
   Almost a quarter of the UK’s
   seaborne trade 07
   Our markets 08
   Investment programme 14
   Generating revenue from coal 16
   Immingham Outer Harbour 18
   Hull shortsea container terminal 20
   Sizeable returns from
   smaller investments 22

2. How we have performed
   24
   Chairman’s statement 26
   Group Chief Executive’s review
   of strategy 28
   Operating and financial review 32

3. How our results add up
   50
   Group income statement 52
   Balance sheets 53
   Cash flow statements 54
   Statement of recognised income
   and expense 55
   Notes to the financial statements 56

4. How we behave
   98
   Board of directors 100
   Operational management team 102
   Directors’ report 106
   Corporate governance 108
   Remuneration report 118
   Corporate social responsibility 130

2006 financial highlights

1. Who we are
   01
   A short voyage around our business 02
   Number one UK ports operator 04
   Long-term, blue chip customers 06
   Almost a quarter of the UK’s
   seaborne trade 07
   Our markets 08
   Investment programme 14
   Generating revenue from coal 16
   Immingham Outer Harbour 18
   Hull shortsea container terminal 20
   Sizeable returns from
   smaller investments 22

2. How we have performed
   24
   Chairman’s statement 26
   Group Chief Executive’s review
   of strategy 28
   Operating and financial review 32

3. How our results add up
   50
   Group income statement 52
   Balance sheets 53
   Cash flow statements 54
   Statement of recognised income
   and expense 55
   Notes to the financial statements 56

4. How we behave
   98
   Board of directors 100
   Operational management team 102
   Directors’ report 106
   Corporate governance 108
   Remuneration report 118
   Shareholder analysis 129
   Corporate social responsibility 130
Techniques

- Automatically detect 8 key headers
- Reorder headers to front and rear
- String matching, regex
- Levenshtein Distance / text similarity
- Machine learning to classify headers
- Crowdsourcing
- Text Reuse
- Hand crafted keyword lists by experts
- Bottom up and top down word lists:
  - Computer based approaches
  - By Linguists
Report Classification

• Heuristic Approach
• Machine Learning

• Classify narratives (front) component into the following generic categories to facilitate cross-sectional and temporal analysis:
  – Chairman’s statement
  – Performance commentary (incl. CEO review, strategic review, finance director’s review, operating review, business review, etc.)
  – Governance statement (incl. chairman’s introduction, separate committee statements, statement on internal control, etc.)
  – Remuneration report
  – Residual commentary (incl. overview, highlights, CRS report, principal risks and uncertainties, directors’ report, etc.)
Output

• NLP publicly available tool (CFIE-FRSE)
• First large scale study of UK annual reports structure.
• First ever published disclosure scores for UK annual reports.
• Approaches used help speed up the analysis process and close the gap between firms and investors.
• Leads to better understand corporate financial decisions and corporate financial performance.
SSRNPaper

• Heterogeneous Narrative Content in Annual Reports Published as PDF Files: Extraction, Classification and Incremental Predictive Ability

Output Evaluation

• Manual evaluation
  – Precision (Type I errors) $\rightarrow$ false positives
  – Recall (Type II errors) $\rightarrow$ false negatives
  – Overall accuracy ($F_1$) $\rightarrow$ harmonic mean of precision and recall

• Compare extracted section headers with table of contents

• Assigned page numbers with actual page numbers from table of contents

• Examine accuracy of section classification

• Evaluations based on 586 reports selected at random (approx. 5% of initial population)
  – 11,720 annual report sections
Output Evaluation

**Panel A: Section extraction**

<table>
<thead>
<tr>
<th></th>
<th>N actual</th>
<th>N extracted</th>
<th>Type 1</th>
<th>Type 2</th>
<th>Precision</th>
<th>Recall</th>
<th>F₁ score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pooled annual report</td>
<td>11,009</td>
<td>10,820</td>
<td>286</td>
<td>475</td>
<td>97.47</td>
<td>95.69</td>
<td>96.57</td>
</tr>
<tr>
<td>Narratives component</td>
<td>5,237</td>
<td>5,233</td>
<td>216</td>
<td>220</td>
<td>96.04</td>
<td>95.80</td>
<td>95.92</td>
</tr>
</tbody>
</table>

**Panel B: Page number synchronization**

<table>
<thead>
<tr>
<th></th>
<th>N errors</th>
<th>Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pooled annual report</td>
<td>736</td>
<td>93.20</td>
</tr>
<tr>
<td>Narratives component</td>
<td>500</td>
<td>90.45</td>
</tr>
</tbody>
</table>
### Panel B: Page number synchronization

<table>
<thead>
<tr>
<th></th>
<th>Type I errors for section extraction treated as incorrect pagination</th>
<th>Type I errors for section extraction not treated as incorrect pagination</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>N errors</td>
</tr>
<tr>
<td>Pooled annual report</td>
<td>10,820</td>
<td>736</td>
</tr>
<tr>
<td>Narratives component</td>
<td>5,233</td>
<td>500</td>
</tr>
</tbody>
</table>
**Panel C: Document classification**

<table>
<thead>
<tr>
<th></th>
<th>N actual</th>
<th>N classified</th>
<th>Error frequencies</th>
<th>Retrieval performance (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Type 1</td>
<td>Type 2</td>
</tr>
<tr>
<td>Narratives component</td>
<td>4,929</td>
<td>4,846</td>
<td>88</td>
<td>83</td>
</tr>
<tr>
<td>Financials component</td>
<td>5,434</td>
<td>5,536</td>
<td>83</td>
<td>88</td>
</tr>
</tbody>
</table>
## Output Evaluation

### Panel C: Document classification

<table>
<thead>
<tr>
<th>Component</th>
<th>N actual</th>
<th>N classified</th>
<th>Error frequencies</th>
<th>Retrieval performance (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Type 1</td>
<td>Type 2</td>
</tr>
<tr>
<td>Narratives component</td>
<td>4,929</td>
<td>4,846</td>
<td>88</td>
<td>83</td>
</tr>
<tr>
<td>Financials component</td>
<td>5,434</td>
<td>5,536</td>
<td>83</td>
<td>88</td>
</tr>
<tr>
<td>By section category:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chairman’s letter</td>
<td>521</td>
<td>517</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>CEO review</td>
<td>280</td>
<td>273</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>CFO review</td>
<td>328</td>
<td>309</td>
<td>12</td>
<td>19</td>
</tr>
<tr>
<td>Governance statement</td>
<td>491</td>
<td>477</td>
<td>27</td>
<td>14</td>
</tr>
<tr>
<td>Remuneration report</td>
<td>406</td>
<td>397</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Highlights</td>
<td>276</td>
<td>275</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>
Reporting growing over Time

how some of the main types of content have grown over the sample period
Text Reuse (Boilerplating)

Governance statements: Distribution of similarity
Limitations

• Designed for large sample analysis
  – Accuracy of extraction process estimated to be around 90%
  – Extraction process can result in errors at the individual firm level (which are removed or diluted in large sample work)
  – Cannot capture the richness of how information is presented (e.g., graphics, charts, tables, etc.)
  – No attempt (yet) to parse text in the financial statements
Summary & Conclusions

• Develop and validate a method for extracting the content and structure of UK annual reports published as PDF files
  – Retaining the structure of the report creates opportunities for research examining new features of disclosure

• Provide the first large-sample evidence regarding the predictive ability of UK annual report narratives
  – Narratives in their entirety are incrementally predictive for earnings beyond financial statement data
  – Difference annual report sections are associated with different predictive qualities
  – Abnormal managerial optimism is associated with lower predictive ability and independent chairman commentaries help to negate managerial bias
## Collaboration / Future Work

<table>
<thead>
<tr>
<th>Organization</th>
<th>Areas of Interest</th>
</tr>
</thead>
</table>
| FRC          | • Funding  
               • Help understand UK annual reports |
| Financial Firm, London | • Fraud Detection  
                             • Management Change. Management vs Analysts |
| Financial Services Provider, London | • Interest in the tool  
                                           • Automatic Structuring, Indexing Annual Reports |
| Big Four in Accounting | • Interest in the tool  
                                   • Provided free-to-publish annual reports |
| Big Data company, London | • Interest in working together |
| KSU, Saudi Arabia | • Arabic Financial Narratives (Book Chapter) |
| MultiLingual Analysis | • Portuguese and Spanish (Book Chapter) |
| Sunway University, Malaysia | • Malay Financial Narratives |
| National Natural Science Foundation of China (NSFC) | • Chinese Financial Narratives |
Thanks

More about the projects:  http://ucrel.lancs.ac.uk/cfie/

CFIE-FRSE - WEB:  https://cfie.lancaster.ac.uk:8443/

CFIE-FRSE – Desktop:  https://drelhaj.github.io/CFIE-FRSE
Tutorial

• CFIE-FRSE - WEB:
  https://cfie.lancaster.ac.uk:8443/

• Sample Annual Reports:
  Or

• Wmatrix Tutorial:
  http://ucrel.lancs.ac.uk/wmatrix/tutorial/