Financial Narrative Processing: a Corpus-based NLP approach

Mahmoud El-Haj

CASS, Lancaster University
Research Projects

- CorpComm • ESRC
- CFIE • ESRC
- SenseSourcing • ESRC/UCREL
- VardSourcing • Jisc
- Arabic Dialects • UCREL
- Strategic Contents • ICRA/UCREL/LUMS
Financial Narrative Projects

Corporate Financial Information Environment (CFIE) → Understanding Corporate Communications (CorpComm) → Extend for 18 months
Project Team

CorpComm¹

CASS
Prof Tony McEnery¹
Dr Mahmoud El-Haj¹,²
Dr Andrew Hardie¹

SCC
Dr Paul Rayson¹,²

LUMS
Prof Steve Young¹,²

MBS
Prof Martin Walker¹,²
Dr Thomas Schleicher²

CFIE²

LSE
Dr Vasiliki Athanasakou²
What is it about?

Project analyses U.K. financial narratives, their association with financial statement information, and their informativeness for investors.
Why Financial Narratives

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

- NLP
- Discourse Analysis
- Computational Linguistics
- Corpus Linguistics
- Financial Disclosures
- Text Analysis
Big Data

Financial Narratives
- Annual Reports
- PEAs
- Conference Calls

Financial News
- Press Releases
- Media Articles
Annual Reports: how big?

Report Count

<table>
<thead>
<tr>
<th>Number of reports</th>
<th>Word Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>10,443</td>
<td>312</td>
</tr>
</tbody>
</table>

word count in Millions

0 50 100 150 200 250 300 350
Output

• NLP publicly available tool (Wmatrix-Import)
• 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.
Wmatrix-Import

Annual Reports (PDFs)

Parse PDFs
Extract Narratives
NLP

Extract Headers
Text Analysis

Display Results (Web)
# UK Annual Reports Sample

## Financial highlights

<table>
<thead>
<tr>
<th>Sales</th>
<th>+6.8%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Underlying operating profit</strong></td>
<td><strong>£789m</strong></td>
</tr>
<tr>
<td><strong>Underlying profit before tax</strong></td>
<td><strong>£712m</strong></td>
</tr>
<tr>
<td><strong>Return on capital employed</strong></td>
<td><strong>11.1%</strong></td>
</tr>
<tr>
<td><strong>Underlying basic earnings</strong></td>
<td><strong>28.1p</strong></td>
</tr>
</tbody>
</table>

## Contents

- **Turnover anno** £54,370 +7.2%
- **Operating profits anno** £3,279 +12.8%
- **Dividend**
- **Debt levels anno**

---

**Editorial**

1. Editorial introductions
2. Executive summary
3. Financial highlights
4. Business review
5. Shareholders' report
6. Corporate governance
7. Financial statements
8. Corporate responsibility
9. Board of directors
10. Directors' report
11. Directors' remuneration report

---

**ARRIVA**

1. Who we are and what we do
2. 24 hours in the life of Arriva
3. Our growth story
4. Our markets
5. Chairman's statement
6. Chief executive's review
7. Financial review
8. Corporate responsibility
9. Board of directors
10. Directors' report
11. Directors' remuneration report

---

**Intelligent Driving**

12. Statement of directors
13. Independent auditors' report on the parent company financial statements
14. Financial calendar, registered office and advisors

---

**Notes to the accounts**

15. Accounting policies
16. Notes to the accounts
17. Five-year financial statements
18. Parent company financial statements
19. Independent auditors' report on the parent company financial statements

---

**Contact**

20. For further information, please contact us:
21. Website: www.arriva.com
22. Telephone: +44 (0)161 922 8282
23. Email: info@arriva.com

---

**Governance**

25. Corporate governance framework
26. Board structure and responsibilities
27. Shareholder rights
28. Remuneration of directors
29. Shareholder engagement
30. Financial statements
31. Business review
32. Shareholder communications
33. Directors' remuneration report
UK Annual report tool: Extraction

Steps in extraction process:

- Detect contents page
- Parse contents page
- Extract section
- Detect section type
- Reorder section
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
how some of the main types of content have grown over the sample period
## Collaboration

<table>
<thead>
<tr>
<th>Organization</th>
<th>Collaboration Areas</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRC</td>
<td>• Funding&lt;br&gt;• Help standardise UK annual reports</td>
</tr>
<tr>
<td>Financial Firm, London</td>
<td>• Fraud Detection&lt;br&gt;• Management Change&lt;br&gt;• Management vs Analysts</td>
</tr>
<tr>
<td>Financial Services Provided, London</td>
<td>• Interest in the tool&lt;br&gt;• Automatic Structuring&lt;br&gt;• Indexing Annual Reports</td>
</tr>
<tr>
<td>Big Four in Accounting</td>
<td>• Interest in the tool&lt;br&gt;• Provided free-to-publish annual reports</td>
</tr>
<tr>
<td>Big Data company, London</td>
<td>• Interest in working together</td>
</tr>
<tr>
<td>KSU, Saudi Arabia</td>
<td>• Arabic Financial Narratives (Book Chapter)</td>
</tr>
<tr>
<td>Sunway University, Malaysia</td>
<td>• Malay Financial Narratives (consultancy)</td>
</tr>
<tr>
<td>National Natural Science Foundation of China (NSFC)</td>
<td>• Chinese Financial Narratives (consultancy)</td>
</tr>
</tbody>
</table>
Thanks

More about the projects: [http://ucrel.lancs.ac.uk/cfie/](http://ucrel.lancs.ac.uk/cfie/)

Wmatrix-Import: [https://cfie.lancaster.ac.uk:8443/](https://cfie.lancaster.ac.uk:8443/)