Retrieving, Classifying and Analysing Narrative Commentary in Unstructured (Glossy) Annual Reports Published as PDF Files

Mahmoud El-Haj[†]

Paulo Alves[#]

Paul Rayson[†]

Martin Walker[*]

Steven Young[‡]

This draft: February 2019

[‡] Corresponding author: Lancaster University Management School, Lancaster University, U.K., Tel. ++44 1524 594242, email s.young@lancaster.ac.uk. [†]School of Computing and Communication, Lancaster University, U.K. [#]Universidade Católica Portuguesa, Portugal and International Centre for Research in Accounting, Lancaster University, U.K. [*]Alliance Manchester Business School, Manchester University, U.K. We are grateful for comments and advice from Vasiliki Athanasakou, Ozlem Arikan, Helena Isidro, Thomas Schleifer, Richard Taffler, Stephen Zeff, two anonymous reviewers, the editor (Mark Clatworthy) and workshop participants at Coimbra, Lancaster, Instituto Universitário de Lisboa ISCTE, Manchester, Nottingham, Rice, Warwick, the Portuguese Accounting Research Network, the 2014 Financial Reporting and Business Communication conference, the 2015 South West Accounting Group Annual Conference, the 2016 Portuguese Xmas Accounting Workshop, the European Accounting Association Annual Congress 2017, the 2017 Portuguese Accounting Network Conference, and the 10th Internal Conference Católica Porto Business School. Research support was provided by José Costa and Stephen Wattam. Funding for this project was provided by the Economic and Social Research Council (contracts ES/J012394/1, ES/K002155/1, and ES/R003904/1), the Research Board of the Institute of Chartered Accountants in England and Wales, and the International Centre for Research in Accounting. The software tool described in this paper is copyrighted to Lancaster University and is available for download as a desktop application (for non-commercial use) at <insert doi>. The dataset of annual report scores is available at <insert doi> and the annual report section corpora are available at <insert doi>. Further details of these and related resources are available at http://ucrel.lancs.ac.uk/cfie/. Elements of this paper previously formed part of a draft circulated under the title “Heterogeneous Narrative Content in Annual Reports Published as PDF Files: Extraction, Classification and Incremental Predictive Ability”.

Retrieving, Classifying and Analysing Narrative Commentary in Unstructured (Glossy) Annual Reports Published as PDF Files

Abstract

We provide a methodological contribution by developing, describing and evaluating a method for automatically retrieving and analysing text from digital PDF annual report files published by firms listed on the London Stock Exchange (LSE). The retrieval method retains information on document structure, enabling clear delineation between narrative and financial statement components of reports, and between individual sections within the narratives component. Retrieval accuracy exceeds 95% for manual validations using a random sample of 586 reports. Large-sample statistical validations using a comprehensive sample of reports published by non-financial LSE firms confirm that report length, narrative tone and (to a lesser degree) readability vary predictably with economic and regulatory factors. We demonstrate how the method is adaptable to non-English language documents and different regulatory regimes using a case study of Portuguese reports. We use the procedure to construct new research resources including corpora for commonly occurring annual report sections and a dataset of text properties for over 26,000 U.K. annual reports.

Retrieving, Classifying and Analysing Narrative Commentary in Unstructured (Glossy) Annual Reports Published as PDF Files

1. Introduction

Annual reports provide important information to support decision-making (EY 2015: 6, CFA Society U.K. 2016).[1] Extant large sample automated analysis of annual report commentaries focuses almost entirely on Form 10-K filings for U.S. registrants accessed through the Securities and Exchange Commission's (SEC) EDGAR system (El-Haj et al. 2019). Several features make 10-Ks amenable to automated large-sample research including batch retrieval provisions, plain text formatting, and a standardized reporting template. However, 10-Ks are only part of U.S. registrants' annual report disclosure package. Many registrants also publish a glossy report containing graphics, photographs and supplementary narratives such as the letter to shareholders (Dikolli et al. 2017). These documents are typically provided as a digital PDF file and outside the U.S. they represent the primary annual reporting vehicle. Barriers to large-scale automated analysis nevertheless mean that little is known about this ubiquitous reporting channel. We provide a methodological contribution by developing, describing and evaluating an automated procedure for retrieving and classifying the narrative component of glossy annual reports presented as digital PDF files.

A typical annual report comprises two broad elements: a narrative component (often presented in the front portion of the document) and the mandatory financial statements, footnotes and other statutory information (often presented in the rear portion of the document). The

[1] Respondents to the CFA Society U.K. survey ranked annual reports ahead of industry-standard databases such as Bloomberg. Research by Black Sun also found that 84% of long-term investors use the annual report to provide insight into corporate strategy and 53% of long-term investors use it to monitor management credibility and assess whether the senior team has delivered on its promises (https://www.blacksunplc.com/en/insights/blogs/annual-reports-are-really-very-important-investors-say-so.html).

narrative component usually contains management commentary on financial performance during the period, together with supplementary information such as a letter to shareholders, information about principal risks and governance arrangements, corporate social responsibility policy, etc.

Lack of a standardized cross-sectional and temporal reporting template represents the main challenge to large-sample automated analysis of annual report narratives. Most regulatory regimes lack the rigid document structure that characterizes annual reports filed on Form 10-K in the U.S. Although glossy annual report content is typically shaped by legal mandate and securities market requirements, management enjoys a high level of discretion over document content and structure. In particular, regulations do not normally: prescribe the order in which information is presented; mandate the precise format in which disclosures must be provided (e.g., running text versus tables versus infographics); require use of standard titles for mandatory sections; or impose upper limits on the type and degree of non-mandatory disclosures. Not surprisingly, reporting approaches vary significantly across firms and over time for the same reporting entity. Inconsistent document structure is a significant barrier to automated processing, which is further compounded by the PDF file type used for distributing digital reports.[2]

Lang and Stice-Lawrence (2015) conduct the first large sample analysis in the accounting literature of PDF annual reports. Lang and Stice-Lawrence (2015) approach the challenge of analysing unstructured PDF reports by converting files to ASCII format using proprietary software and then isolating running text with a Perl script. While the method facilitates analysis of text at the aggregate level, it does not capture the location of commentary within the document. Lang and Stice-Lawrence (2015) are therefore unable to distinguish narrative

[2] PDF (Portable Document Format) files were designed to be portable across platforms irrespective of hardware, operating system, graphics standards, application software used to create the original document, foreign character language sets, etc. They can also offer compression benefits and they satisfy the legal requirements for admission in a court of law because they cannot be altered without leaving an electronic footprint. A consequence of these features, however, is that PDF content cannot be easily accessed and manipulated.

commentary from financial statement disclosures (e.g., footnotes) or isolate distinct sections of the narrative component. Pinpointing commentary associated with a specific report element is nevertheless a requirement for many research applications, particularly where themes and language properties vary across sections (Dyer et al. 2017).

We propose and evaluate a procedure for retrieving text and document structure from digital PDF annual reports published by firms listed on the London Stock Exchange (LSE). Our method uses JavaScript and iText libraries to locate the report table of contents, synchronize page numbers in the native report with page numbers in the corresponding PDF, and then retrieve content separately for each section listed in the table of contents. For reports where we are unable to detect the table of contents, we use pre-existing document bookmarks to retrieve text by section. The script is packaged as a desktop application to support academic research.

The ability of our text retrieval method to return information on report structure represents an important contribution over Lang and Stice-Lawrence (2015) because it facilitates more granular classification of text by report section and theme. Specifically, section headings from tables of contents and bookmarks are used to partition retrieved text into the audited financial statements component of the report and the “front-end” narratives component, with the latter further subclassified into a set of core sections that feature regularly in automated analyses of 10-Ks and manual analyses using PDF files, including the chair’s letter (Clatworthy and Jones 2006, Dikolli et al. 2017), management commentary (Li 2008 and 2010, Loughran and McDonald 2011), and remuneration reports (Laksmana et al. 2012, Hooghiemstra et al. 2017). Unlike Lang and Stice-Lawrence (2015) whose retrieval approach relies on proprietary software, our method is fully autonomous and unconstrained by researchers’ software resources.

We validate the accuracy of our retrieval and classification procedure using manual and statistical procedures. Manual tests on over 11,000 sections extracted from 589 processed reports selected at random compare section titles and adjusted page numbers from retrieved tables of contents with corresponding details from the native PDF files, as well as evaluating the accuracy of section classification procedures. Precision and recall statistics (Manning and Schütze 1999) for section retrieval, page synchronization, and section classification generally exceed 95%.

Manual validations are complemented by analyses that test for predictable intra- and inter-report variation in the length, tone and readability of narrative commentary using a sample of over 11,500 documents published between 2003 and 2014. Cross-sectional tests confirm extant evidence that document length is increasing in firm size, business complexity, and intangible assets (Lang and Stice-Lawrence 2015, Dyer et al. 2017). Report length also varies predictably with changes in disclosure regulations. In particular and consistent with Lang and Stice-Lawrence (2015), we show how annual report length increased for LSE Main Market (Alternative Investment Market) firms following mandatory adoption of International Financial Reporting Standards in 2005 (2007). As an extension to Lang and Stice-Lawrence (2015)**,** we confirm expectations that these increases are concentrated in the financial statement component of the report (Morunga and Bradbury 2012).

Findings for net tone are also consistent with predictions and prior research. Like Henry and Leone (2016), we find net report tone is increasing in reported earnings and decreasing in the book-to-market ratio and stock return volatility. Further tests using a within-report design that controls for omitted variable bias confirms expectations that net tone is more positive in performance commentary sections compared to mandated, compliance-focused sections such as the governance statement and remuneration report where scope for managerial optimism is more limited. Finally, readability tests also provide some evidence of expected intra- and inter-report variation in the Fog index (Gùnning, 1968), although we are unable to replicate some of the associations reported by Li (2008).

An important residual question is whether our method is applicable to reports published in other languages and regulatory settings. Since much of our tool is regime and language independent, it is possible to adapt the method to other settings without making changes to the JavaScript code. The primary adaptations involve: substituting the keyword list used to identify the document table of contents in U.K. reports with a comparable keyword list optimized for the chosen reporting language and regulatory setting; and developing new synonym lists that serve as inputs to our section classification algorithms to replace those optimized for U.K. reports. We illustrate the tailoring process directly using a case study of annual reports published in Portuguese by firms listed on Euronext Lisbon, and report retrieval and classification accuracy rates similar to those obtained for our sample of U.K. reports.

Our study provides several methodological contributions to the literature. We present and validate a method for retrieving content from unstructured annual reports distributed as PDF files. Distinct from Lang and Stice-Lawrence (2015), our method facilitates analysis of content by section. Our method is also packaged as a software tool available for use and development by other researchers. Our approach opens the door to new research on annual reports such as the role of document structure, and the determinants and impact of international differences in narrative reporting policy. Nevertheless, our inability to capture aspects of disclosure format such as the relative position of text on the page and the presence and content of tables, charts and other infographics means that our tool cannot be used to explore many important questions relating to disclosure effectiveness.

In addition to our methodological contribution, we also provide a unique dataset of structure and content for over 26,000 annual reports for fiscal year-ends January 2003 through December 2017 published by 4,131 financial and non-financial firms listed on the LSE Main Market and Alternative Investment Market (AIM). The dataset provides researchers with the first opportunity to undertake large-sample analysis of annual report narrative disclosures that are not constrained by the SEC's 10-K reporting template. We also provide an annual report corpus consisting of nearly 200 million words, together with a set of corpora for common annual report sections including the chair's letter, governance statements, remuneration reports, risk reports, and audit reports.

The remainder of the paper is organised in six sections. Section 2 reviews relevant research and summarizes regulations governing annual reporting. Section 3 describes our extraction and classification procedure. Section 4 reports details of our manual and large sample statistical validity tests, while section 5 presents details of annual report data resources created to support future research. Section 6 demonstrates how our procedure can be adapted to analyse non-English language reports published outside the U.K. Conclusions are presented in section 7.

2. Background and overview

The annual report and accounts represents a key disclosure in the corporate reporting cycle. Annual reports are a legal requirement for publicly traded firms in most jurisdictions and although shareholders are the legislative focal point, these disclosures are used by a range of stakeholders including financial analysts, prospective investors, customers and suppliers, lobby groups, regulators, journalists, and academics. The majority of automated textual analysis research on annual reports focuses on 10-K filings due to their accessibility, amenable file format, and standardized reporting template with regularized schedule titles (El-Haj et al. 2019).

Many U.S. registrants complement their statutory 10-K filing with a brochure-style annual report distributed as a digital PDF file in which summary information is combined with additional disclosures.[3] Outside the U.S., these glossy brochure-style PDF reports represent the primary format in which firms' mandatory annual report and accounts are available (Lang and Stice-Lawrence 2015).[4] The International Accounting Standards Board (IASB) does not provide a formal definition of either financial reporting or the annual report.[5] Instead, specific components of the annual report have evolved in practice (Financial Reporting Council 2012: 8), with significant local variation from a mandated core. For example, European Union Directive 2013/34/EU requires annual financial reports of public-interest entities traded on a regulated market of any Member State to include: a management report, a corporate governance statement, and the financial statements. Corporate laws and securities regulations in individual Member States further refine and supplement these baseline requirements. At a more primitive level, the typical PDF annual report file can be decomposed into two distinct elements: a narrative component (often presented in the front portion of the document) and the mandatory financial statements, footnotes and other statutory information (often presented in the rear portion of the document). The narrative component usually contains management commentary, together with

[3] SEC rules require companies to supply shareholders with an annual report prior to annual meetings involving election of boards of directors. While some companies send their 10-K filing to shareholders in lieu of a separate annual report, many produce a separate document that contains a summary of the 10-K plus additional content such as infographics and a letter to shareholders from the CEO. A significant fraction of registrants incorporate much of their mandatory 10-K filing by reference to these separate annual reports (Loughran and McDonald 2014: 98).

[4] Electronic filing and retrieval systems are rare outside the U.S. Examples include TSX SecureFile and the System for Electronic Document Analysis and Retrieval (SEDAR) in Canada, and the Data Analysis and Retrieval Transfer (DART) system in South Korea.

[5] International Standard on Auditing (ISA) 720 (Revised) describes the annual report as "a document, or combination of documents… An annual report contains or accompanies the financial statements and the auditors' report thereon and usually includes information about the entity's developments, its future outlook, a risks and uncertainties statement by the entity's governing body, and reports covering governance matters." (International Auditing and Assurance Standards Board 2015: 7, para. 12a). The annual report is not to be confused with firms' annual reporting package which the IASB describes as including annual financial statements, management commentary, press releases, preliminary announcements, investor presentations, and information for regulatory filing purposes (IASB 2017a, para 19B).

supplementary information such as a letter to shareholders and reviews of strategy, risk, corporate governance, and executive remuneration policy. Text is often augmented with photographs, tables and infographics aimed at improving disclosure quality.

Glossy annual reports supplied as PDF files lack the consistent, linear structure of the 10-K. Instead, management enjoys significant discretion over the information disclosed, the order in which information is presented, and the labels used to describe individual sections. Discretion over content, placement and nomenclature helps management tailor commentary to their firm's particular circumstances (Institute of Chartered Secretaries and Administrators 2015). However, inevitable variation in report structure across firms and over time renders automated document processing a significant challenge (Dikolli et al. 2017). Research examining these documents is therefore scarce and limited primarily to manually-coded samples involving individual report sections (Merkl-Davies and Brennan 2007).[6] The lack of large sample evidence on the properties of these documents is startling given the degree of regulatory scrutiny they attract, coupled with high preparation costs and their enduring status as a key element of corporate communication.

Lang and Stice-Lawrence (2015) conduct the first large sample analysis of English-language annual reports using more than 87,600 PDF files for over 15,000 non-U.S. firms from 42 countries for calendar years 1998 through 2011. Results reveal how text attributes correlate predictably with regulatory features and managers' reporting incentives, and how higher quality disclosures are associated with positive stock market outcomes. They extract text from unstructured PDF English-language reports by converting files to ASCII format using Xpdf and QPDF proprietary software and then construct aggregate measures of the entire textual content of glossy annual reports. While these aggregate measures are reasonable for the research questions

[6] Notable exceptions include Schleicher et al. (2007), Grüning (2011), Lang and Stice-Lawrence (2015), Hooghiemstra et al. (2017) and Athanasakou et al. (2019).

examined by Lang and Stice-Lawrence (2015), the inability to associate narratives with specific annual report sections is inconsistent with the majority of extant research that studies narrative content at a more granular level (e.g., Clatworthy and Jones 2006, Li 2010, Loughran and McDonald 2011, Campbell et al. 2013, Dyer et al. 2017, Dikolli et al. 2017).

3. Document processing procedure

This section summarizes our procedure for: retrieving text and document structure from PDF annual report files; partitioning reports into the "front-end" narratives component (hereinafter *Narratives*) and the "back-end" mandatory financial statements and footnotes component (hereinafter *Financials*); and classifying the *Narratives* component into core sections that are cross-sectionally and temporally comparable.

Annual report structures vary significantly across reporting regimes and therefore to make the initial development task feasible we focus on reports for a single reporting regime. We select the U.K. due to the LSE's position as one of the largest equity markets by capitalization outside the U.S. The extraction process is nevertheless designed to be generalizable insofar as reports published in other reporting regimes and languages can be analysed by modifying the language- and regime-dependent aspects of our procedure without editing the underlying JavaScript. (See section 6 for further details and an application to Portuguese annual reports.)

3.1 Retrieval

Our procedure for retrieving text and document structure from digital PDF reports involves the following four steps:[7]

[7] Image-based PDF files cannot be processed reliably using our procedure. We convert image-based PDFs to digital equivalents using Adobe X Pro's optical character recognition (OCR) facility. Unfortunately, OCR methods rarely produce annual report files with a well-structured table of contents in our experience and as a result our procedure is

1) Detect the page containing the annual report table of contents. The contents table serves as the map by which we navigate the remainder of the report. Information in the table of contents is used to identify individual sections and the pages on which they begin and end. Lack of a common location and format for the table of contents, together with the absence of regularized section headers makes detecting the contents page a nontrivial task.[8] Our approach involves identifying a set of common section titles and associated synonyms based on an initial sample of 50 reports selected at random. We use this provisional list of headers to identify the contents page by matching the text on each page of the document against our key-phrase list. This provisional list is augmented through several iterations where we extract tables of contents from 1,000 reports selected at random in each cycle and then use the results to update our list for frequently occurring headers and synonyms based on manual review. The final "gold standard" list is presented in an online appendix.

   To further improve detection accuracy and minimize Type 1 errors, we match gold standard headers to lines of text that follow a contents page-like style (i.e., gold standard phrases preceded or followed by alphanumeric characters representing a number). Each page in the PDF is matched against the gold standard header set and the page with the highest similarly score (Levenshtein 1966) is identified as the potential contents page;

2) Isolate the report table of contents and discard co-located material. Our algorithm involves matching each line of text in the candidate contents page against a regular expression command that extracts any line of text starting or ending with an alphanumeric

not guaranteed to extract content reliably. Although we have processed image-based PDFs, we do not include the results in our final dataset due to validity concerns.

[8] Many firms present information such as highlights, overview, etc. prior to the contents page. Tables of contents also take a variety of styles in addition to a standard two-column tabular format. The contents table may also appear in isolation on a page or co-located with other text such as highlights and "at a glance" information. Finally, the contents may be disaggregated across multiple pages.

representation of a number between one and the number of pages in the annual report. To be classified as a valid table of contents for use in retrieval steps (3)-(4) described next, results must satisfy conditions detailed in the appendix;

3) Synchronize page numbers in the digital PDF file with page numbers in the valid table of contents. Pagination in the PDF file rarely corresponds to pagination in the native annual report because the front and inside front cover pages, which are almost always included in the PDF, are not normally paginated in the actual report. We develop a page detection algorithm that crawls through a dynamic set of three consecutive pages with the aim of detecting a pattern of sequential numbers with increment one (e.g. 31, 32, 33). The extracted sequence is then used to calibrate page numbers across the entire PDF file;

4) Use synchronized page numbers to determine the start and end of each section in the annual report table of contents, insert bookmarks into the PDF for each section based on the page mapping, and extract annual report content section by section using these bookmarks.[9] All retrievable text is captured including text from tables and infographics. The absence of HTML-type tags in PDF files means we are currently unable to isolate tables and charts, capture different font styles and sizes, and pinpoint the relative position of text on the page.

Steps (2) - (4) are tested and refined using multiple iterations for samples of 1,000 reports selected at random from years 2004 through 2010, with manual evaluation of precision and recall performed at each step (Manning and Schütze 1999).

Step (2) distinguishes between valid and invalid candidate tables of contents. We apply an alternative retrieval procedure based on bookmarks assigned by the PDF originator for reports where the candidate table of contents is classified as invalid in step (2). We create a flag for such

[9] Pre-existing bookmarks are overwritten. The majority of annual report sections start on a new page. In the rare cases where sections end and start mid-page, our retrieval procedure double-counts commentary because all content associated with the transition page is attributed to both adjacent sections.

reports indicating that document structure and section-level text retrieval is based on document bookmarks rather than the report table of contents.[10]

3.2 Classification: *Narratives* and *Financials*

Most applications involving annual report narratives require researchers to distinguish between content from the *Narratives* and *Financials* components of the annual report. The absence of a standardized reporting format means that management are free to present individual report sections in any order, and therefore *Narratives* and *Financials* components are often not delineated clearly and consistently. Isolating these two generic elements of the report is therefore a non-trivial task. We use a two-step classification procedure based on section headers in the table of contents (or bookmarks where a valid table of contents is not detected). Step one involves applying a binary split based on the naïve linear document structure represented in Figure 1, with the delineating point set at either the audit report or directors' statement of responsibilities (whichever occurs first).[11] Sections occurring before this cut-off point in the table of contents are allocated to *Narratives_Null* while sections including and following the cut-off point are allocated to *Financials_Null*. Step two of the process adjusts both components for sections misclassified in the first pass. Specifically, we search all section headings in *Narratives_Null* for character strings associated with standard section headers expected to form part of *Financials* (e.g., consolidated statement of net income, consolidated statement of

[10] Most digital PDF annual reports published since 2012 contain bookmarks that either replicate sections from the table of contents or provide additional granularity beyond headers listed in the table of contents. Inconsistency across reports in the mapping from the table of contents to bookmarks creates comparability problems for analyses requiring the complete report structure, hence our preference for basing retrieval on the published table of contents. Retrieval based on bookmarks represents a reliable second-best option where the report table of contents cannot be identified reliably. Retrieval using bookmarks does not impair the reliability of report partitioning and classification of core narrative sections. Further details are provided in the appendix.

[11] Figure 1 is a representative annual report based on a combination of the median structure of all documents reviewed and the template provided by the Institute of Chartered Secretaries and Administrators (2015). Note, however, that relatively few documents follow this exact structure, hence the need to apply a second-stage adjustment as part of the classification procedure.

financial position, notes to the accounts, etc.) and reallocate these sections to the *Financials* component. Analogously, we search all section headings allocated to *Financials_Null* for strings associated with headers expected to form part of *Narratives* in a U.K. annual report (e.g., chairman's statement, CEO review, financial review, business review, remuneration report, corporate governance statement, etc.) and reallocate these sections to the *Narratives* component.

3.3 Classification: *Narratives* subcomponents

Analysing the entire textual content of *Narratives* provides a useful starting point for exploring the properties of annual report disclosures. However, more granular analysis of common subcomponents such as management commentary is the norm in most applications. The 10-K filing template makes this decomposition relatively straightforward for U.S. registrants because reports contain a prescribed list of standardized schedules. Unstructured PDF annual reports lack such standardization, with content varying significantly across firms and time. Different naming conventions are also used to describe the same report section.[12] We approach this classification problem by identifying a set of core report categories based on Institute of Chartered Secretaries and Administrators (2015) and Financial Reporting Council (2014) guidance, coupled with manual review of reports selected at random. Our final category list includes the following elements: performance highlights, statement from the board chair, management commentary (including CEO review, operating review, business review, strategic review, CFO review, financial review, etc.), governance statements (including internal control),

[12] For example, the annual letter to shareholders has 33 distinct labels in our dataset after controlling for minor string differences, any variations including the term "CEO", and chairs' overview of corporate governance. The list expands to over 250 when these differences are considered.

and remuneration reports. Remaining sections are allocated to an aggregate residual category.[13] A synonym list for each category is developed and used as the basis for a search algorithm that crawls through the table of contents classifying sections.[14] Synonyms for each core section are presented in the appendix.

3.4. Text processing

Retrieved text is processed automatically by our procedure and outputs are provided in a spreadsheet (.csv format).[15] We provide aggregate scores for the *Narratives* and *Financials* components, along with scores for each section in the *Narratives* component. Default metrics comprise: total word count; total page count; Fog index of readability (Gǜnning, 1968) computed using a version of Fathom (Svoboda 2013); Flesch-Kincaid readability index (Kincaid et al., 1975); and counts for positive and negative words from Henry (2006, 2008) and Bill McDonald's webpage (http://www3.nd.edu/~mcdonald/Word_Lists.html), forward-looking words drawn from prior research, strategy-related words (Athanasakou et al., 2019), uncertainty words from Bill McDonald's webpage (http://www3.nd.edu/~mcdonald/Word_Lists.html), and causal reasoning words from based on an author-defined list. (Details of wordlist elements are provided in the appendix.) Our tool also offers users the option of uploading and applying their own bespoke wordlists in addition to our default lists. Raw text retrieved at the section level is

[13] The following four additional core report sections are classified in version of the tool available at the date of publication: risk report, corporate social responsibility, chair's governance overview, and group audit report. See section 5 and the appendix for further details.

[14] Preliminary synonym lists were based on a sample of 1,500 annual reports selected at random. To address the problem of variable word ordering and the presence of stop words (e.g. "the", "of", "and", etc.) in the header title we used Levenshtein distance to compare header strings (Levenshtein 1966). The Levenshtein distance between two words is the minimum number of single-character edits (insertion, deletion, or substitution) required to change one word into the other. To work at the phrase level we modified the algorithm to deal with words instead of characters. All headers with a Levenshtein distance value less than six were manually reviewed and used to create revised lists. The process was repeated two further times to determine the final gold standard synonym lists.

[15] The text scoring procedures described below can be applied to plain text files containing textual content derived from any document source. Specifically, users have the option of bypassing the retrieval and classification steps and instead uploading a pre-processed text file for scoring and further analysis along the lines described in section 3.4.

also saved as a .txt file for further analysis in software packages such as Diction, WordSmith, AntConc and WMatrix (Rayson, 2008: http://ucrel.lancs.ac.uk/wmatrix/).

4. Evaluations

This section reports results of tests designed to evaluate the performance of our retrieval and classification procedure. Section 4.1 reports results for manual comparisons of extracted text against source PDF files while section 4.2 presents additional large-sample validity tests that correlate the length, tone and readability of retrieved narratives with expected determinants.

4.1 Manual evaluation

Manual evaluations are performed to assess the following four aspects of retrieval and classification performance: (i) detecting and extracting the annual report table of contents; (ii) synchronizing page numbers for each section reported in the annual report table of contents with corresponding page numbers in the PDF file; (iii) classifying the annual report into *Narratives* and *Financials*; and (iv) classifying *Narratives* into generic categories.

Evaluations are based on a random sample of 586 reports that were not used to implement and refine steps (1)-(4) described earlier. This sample represents approximately five percent of reports for non-financial firms with year-ends from January 2003 through September 2014 collected from Perfect Information in March 2015 and processed by our procedure. Extraction performance is assessed by comparing all sections listed in the table of contents for each report with headers extracted by our tool, and by identifying instances where assigned page numbers marking the start and end of each section differ from actual start and end pages in the native PDF file. Classification accuracy is assessed by identifying sections incorrectly classified as *Narratives* (*Financials*), and by identifying errors classifying *Narratives* into core sections.

We use precision and recall constructs to evaluate extraction and classification performance (Manning and Schütze, 1999). Precision measures the fraction of retrieved instances that are relevant (or the incidence of Type I errors) and is viewed as a measure of exactness or quality, while recall measures the fraction of relevant instances that are retrieved (or the incidence of Type II errors) and reflects a measure of completeness or quantity:

$$precision = \frac{N(tp)}{N(tp) + N(fp)} \tag{1a}$$

$$recall = \frac{N(tp)}{N(tp) + N(fn)}, \tag{1b}$$

where *N(tp)* is the number of true positives, *N(fp)* is the number of false positives (Type I errors), and *N(fn)* is the number of false negatives (Type II errors). We also compute the F_1 score, defined as the harmonic mean of precision and recall, as an overall measure of retrieval and classification accuracy (Van Rijsbergen 1979):[16]

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}. \tag{2}$$

Table 1 reports evaluation results computed at the section-level. Results are presented for the pooled annual report (*Narratives* plus *Financials*) as well as separately for the *Narratives* component. Panel A of Table 1 presents results for retrieval accuracy. Our random sample of 586 processed annual reports contains 11,009 individual sections in aggregate as listed in the tables of contents. Our tool extracts 10,820 headers in total, of which 10,534 sections are correct. The 286 Type I errors (10,820 – 10,534) correspond to conditional retrieval precision of 97.4 percent, while the 475 Type II errors (11,009 – 10,534) correspond to a conditional recall rate of 95.7 percent. Overall conditional retrieval accuracy as measured by the F_1 score is 96.5 percent.

[16] The F score is derived such that F_β measures the effectiveness of retrieval with respect to an individual who attaches β times as much importance to precision as recall. The F_1 score places equal weight on precision and recall, whereas the F_2 ($F_{0.5}$) score weights recall (precision) higher than precision (recall).

Results for *Narratives* are quantitatively similar with precision, recall and F_1 scores equal to 95.9 percent, 95.8 percent, and 95.8 percent, respectively.

Panel B of Table 1 reports conditional error rates for page number assignment. Findings presented in columns 2-4 treat Type I extraction errors from Panel A as incorrectly assigned page numbers and as such provide a lower bound assessment of pagination performance. Findings reported in columns 5-7 are computed using the subsample of 10,534 sections extracted correctly and therefore represent an upper bound on pagination accuracy. Pagination error rates in Panel B for the entire annual report range from 93.2 percent for the more restrictive test to 95.7 percent using the more lenient test. Similar findings are reported in the second row of Panel B for *Narratives*.

Results for document partitioning and header classification are presented in Panel C of Table 1. Evaluations are conducted using the subsample of 10,534 sections extracted correctly by our system. The first two rows in Panel C report results of partitioning reports into the *Narratives* and *Financials* components. The total number of misclassified sections is 171 (1.6 percent), of which 88 (83) are *Financials* (*Narratives*) misclassified as *Narratives* (*Financials*). These results translate into conditional precision and recall scores of approximately 98 percent for *Narratives* (*Financials*).

The final six rows in Panel C present evidence on classification accuracy for generic sections in the *Narratives* component. Classification accuracy as reflected in the F_1 score is highest for chair's statements (99.3 percent), remuneration reports (98.6 percent) and summary highlights (98.3 percent). CFO reviews and governance statements are associated with the lowest F_1 scores of 96.8 percent and 94.3 percent, respectively. Nevertheless, with all bar one F_1 scores exceeding 95 percent in Panel C, results support the conclusion that our classification method provides a valid approach for large samples of documents. Accuracy rates are particularly encouraging given the complex, highly unstructured nature of PDF annual reports.

4.2 Statistical evaluation

This section evaluates extraction and classification performance by examining correlations between properties of annual report disclosures and known or expected determinants thereof. We focus on three properties of annual report narratives that have featured prominently in prior accounting research and policy debates: length, tone, and readability. Tests are conducted using a sample of annual reports published in calendar years 2003 through 2014 by non-financial firms listed on the LSE. Our tool processed 20,446 reports from an initial set of 24,142 available reports (85%). Non-processed reports comprise 1,700 image-based PDF files (7%) and 1,996 other reports (8%). Processed reports are filtered further to exclude: 609 booklet style reports comprising two annual report pages on a single PDF page (2.5%); non-English language reports (one case); and 28 regulatory filings including reports containing 20-F reconciliations (0.1%). The resulting 19,808 reports for 3,302 financial and non-financial firms are matched with firm identifiers and fiscal year-ends from Thomson Reuters Datastream to permit collection of accounting and market data.[17] Further analysis suggests that these criteria do not introduce material selection bias into the final annual report sample, apart from the image-based file condition which tends to result in a disproportionate loss of reports for small firms and

[17] PDF annual report filenames do not contain a unique firm identifier. Instead, filenames typically use a standard naming convention comprising firm name and publication year. We use filenames as the basis for a fuzzy matching algorithm that pairs firm names extracted from the PDF filename with firm names provided by Thomson Reuters Datastream. Matching on name is problematic because firms can change their name over the sample period. The matching procedure must therefore track name changes. To address this problem, we combine firm registration numbers and archived names from the London Share Price Database with Datastream's firm name archive in our fuzzy matching algorithm. For those cases where our algorithm fails to find a sufficiently reliable match, we perform a second round of matching by hand. Further details of the matching procedure, including a copy of the algorithm and a -guide to implementing the matching procedure in SAS are available at http//cfie.Lancaster.ac.uk.8443/. Our dataset contains a unique firm identify code that adjusts for name changes and ensures time series continuity of reports published by a given entity. Licensing restrictions prevent direct publication of proprietary identifiers.

fiscal years prior to 2006 in the U.K. setting. Excluding matching errors, missing Datastream accounting and market data, fiscal years greater (less) than 15 (nine) months, and financial firms reduces the sample to 11,856 non-financial firm-years, although some tests use fewer observations where additional data restrictions apply.

Report length

Factors identified in prior research as correlating with longer annual report commentary include: firm size, because larger firms tend to disclose more (Watts and Zimmerman 1986); organisational complexity, because more complex businesses and business models are likely to have more complex annual reports (Li 2008, Dyer et al. 2017); accounting losses, because poor financial results are harder to explain (Bloomfield 2008) or involve more management obfuscation (Li 2008); return volatility, because communication to investors is likely to be more complicated for firms with more volatile operations (Li 2008); and intangible assets proxied by the book-to-market ratio, because narratives provide information about assets and future revenue streams that extend beyond the scope of financial statements (Dyer et al. 2017).[18] We test for similar relations in our data. We also expect firms listed on the LSE Main Market to have longer reports than their AIM counterparts because Main Market firms face more extensive disclosure requirements. Finally, International Financial Reporting Standards (IFRS) (IASB 2010 para. 23) and ISA 720 (Revised) require auditors provide assurance on the degree of consistency between the *Financials* and *Narratives* components of the annual report. This consistency requirement is expected to induce a positive association between the volume of information presented in these two components. We therefore estimate the following OLS regression:

[18] While Dyer et al. (2017) find evidence consistent with their prediction, Li (2008) finds that intangible rich firms' reports are shorter.

$$\begin{aligned} Length^{k}_{it} = {} & \beta_0 + \beta_1 Size_{it} + \beta_2 Loss_{it} + \beta_3 BTM_{it} + \beta_4 ReturnVol_{it} + \beta_5 Segments_{it} \\ & + \beta_6 Main_{it} + \beta_7 Length\_Financials_{it} + \phi + \xi_{it} \end{aligned} . \quad (3)$$

Length is report length, where *k* indicates either word count (scaled by 10^3 to simplify reporting) or page count for firm *i*'s report in fiscal year *t*. We estimate model (3) using both management commentary (i.e., MD&A-equivalent sections) and the entire *Narratives* component. *Length* is positively skewed and so we report results using both raw values and log-transformed values for completeness. Covariates in model (3) are defined as follows: *Size* is the natural logarithm of total assets; *Loss* is an indicator for firm-years where earnings from continuing operations are negative; *BTM* is book-to-market ratio and proxies for intangible assets; *ReturnVol* is the standard deviation of monthly stock returns computed over fiscal year *t*; *Segments* is number business segments and proxies for organisational complexity; *Main* is an indicator variable equal to one if firm *i* is listed on the LSE Main Market in fiscal year *t* and zero otherwise; *Length_Financials* is the number of words in the *Financials* component of the annual report; ϕ represents industry fixed effects; and ξ is the regression residual. Based on prior research we test β_1, $\beta_{2,}$ β_4, β_5, β_6 and $\beta_7 > 0$ and $\beta_3 < 0$.

All accounting and market data required to estimate model (3) are obtained from Thomson Reuters Datastream. All continuous (lower-bounded) variables are winsorized at the top and bottom (top) percentile. Coefficient estimates and model summary statistics are presented in Table 2. Findings are broadly consistent with expectations. *Size, BTM, Segments, Main* and *Length_Financials* all load significantly and with the expected sign in all specifications. Coefficient estimates for *Loss* and *ReturnVol* also provide support for the predicted positive association although conclusions are more sensitive to model specification. Specifically, while loss firms' annual report narratives are associated with a higher word count as expected, the number of pages is unrelated to the sign of reported earnings. Firms with high stock return volatility also have longer annual report commentaries using raw word count, whereas results are insignificant for log word count and all specifications using page count. Overall, we interpret results in Table 2 as evidence that our retrieval and classification procedure extracts annual report text reliably.[19]

In addition to the variables included in model (3), prior research highlights a link between annual report length and financial disclosure regulations (Dyer et al. 2017; Lang and Stice-Lawrence 2015). We therefore conduct supplementary validity tests by extending model (3) to capture the impact of key regulatory developments predicted to affect annual report length. These tests also address endogeneity concerns by exploiting phased adoption of regulations.

Lang and Stice-Lawrence (2015) document an increase in report length for an international sample of firms following mandatory adoption of IFRS. We test for a positive impact of IFRS adoption on annual report length using an identification strategy that exploits staggered IFRS adoption by LSE firms. Specifically, while Main Market firms adopted IFRS for fiscal years beginning on or after January 1, 2005, mandatory IFRS adoption was delayed for AIM firms until January 1, 2007. Accordingly, we expect to observe a structural increase in report length for Main Market (AIM) firms after 2005 (2007). Further, because IFRS relate primarily to financial statements and accompanying footnote disclosures, the IFRS-related impact on disclosure length should centre on the *Financials* component of the annual report.

More generally, Dyer et al. (2017) show how FASB and SEC compliance requirements have increased the length of 10-K disclosures. We therefore use the introduction of enhanced

[19] In supplementary tests we replaced *Loss* in model (3) with a vector of indicator variables corresponding to ROA quintiles to provide evidence on variation within profit and loss groups. The benchmark quintile is q5 (i.e., highest ROA partition). No obvious pattern across quintiles is evident in the results. There is weak evidence that reports are longer for firms in the lowest quintile of ROA. Negative coefficients on the indicator for the fourth quintile also suggest relatively longer reports for firms in the very highest ROA quintile.All significance levels and conclusions for other covariates in the model are consistent with those described in the main text.

compliance requirements on corporate governance reporting for Main Market firms post-2007 as an additional setting in which to validate our extraction and classification procedure. Specifically, implementation of European Directive 2006/46/EC in 2008 increased annual report disclosure requirements on corporate governance for Main Market firms with a registered office in the European Community. Additional governance- and remuneration-related disclosure requirements were also mandated for Main Market firms following revisions to the U.K. Corporate Governance Code in 2008 and 2010. Crucially, these requirements relate exclusively to *Narratives* and do not apply to AIM firms. Contrary to the relative increase in the post-2007 length of *Financials* for AIM firms following IFRS adoption, we therefore expect to observe a decline in *Narratives* length for AIM firms post-2007 relative to their Main Market counterparts.

We test the above predictions by estimating the following extended version of model (3):

$$\begin{aligned} LengthAR^{p}_{it} &= \gamma_0 + \gamma_1 Post2005 + \gamma_2 Main_{it} \times Post2005 + \gamma_3 Post2007 + \gamma_4 AIM_{it} \times Post2007 \\ &+\gamma_5 Size_{it} + \gamma_6 Loss_{it} + \gamma_7 Main_{it} + \gamma_8 BTM_{it} + \gamma_9 Segments_{it} + \gamma_{10} LengthAR^{q}_{it} \qquad , (4) \\ &+ \phi + \mu_{it} \end{aligned}$$

where *LengthAR* is either the number of words or the number of pages for the p^{th} (q^{th}) annual report component for firm *i* and fiscal year *t* (*p = Narratives, Financials; q = Financials, Narratives*); *Main* is an indicator variable for LSE Main Market firms in year *t* and *AIM* is the converse of *Main; Post2005* and *Post2007* are indicator variables for fiscal years beginning on or after 1 January 2005 and 1 January 2007, respectively; other variables are as defined in model (3); and μ is the regression residual.[20] We test γ_1, γ_2, and $\gamma_4 > 0$ for $LengthAR^{Financials}$, $\gamma_3 > 0$ for both $LengthAR^{Financials}$ and $LengthAR^{Narratives}$, and $\gamma_4 < 0$ for $LengthAR^{Narratives}$.

[20] The Stable Unit Treatment Value Assumption (SUTVA) applies to equation (4). The SUTVA requires that the treatment status of the treated group does not affect the outcomes of the control population and vice versa. In our context, the SUTVA is violated if IFRS adoption by Main Market firms influences annual reporting trends among AIM firms. Assuming positive spillover effects are most likely among AIM firms, γ_2 and γ_4 will be downward-biased estimates and results will underestimate the reporting effect of mandated IFRS adoption.

Findings for model (4) in Table 3 are consistent with expectations. Columns 4 and 5 are estimated using word counts for *Narratives* and *Financials*, respectively, while columns 6 and 7 are estimated using page counts. As predicted, *Post2005* loads positively for *Financials*, reflecting IFRS-adoption effects. Similarly, *Post2007* loads positively for *Narratives* and *Financials*, reflecting the concurrent impact of expanded disclosure rules on corporate governance (for Main Market firms) and IFRS adoption (for AIM firms). Consistent with mandatory IFRS adoption increasing financial statement disclosures for Main Market firms, *Main×Post2005* loads positively for *Financials* in column 5 (word count) and column 7 (page count). A similar effect is evident following mandatory adoption by AIM firms in 2007: coefficients on *AIM×Post2007* are positive in columns 5 and 7 for *Financials* reflecting a relative increase in financial statement disclosures for AIM firms post-IFRS implementation. Since IFRS adoption effects are likely concentrate in the *Financials* component of the annual report, *AIM×Post2007* is not expected to load positively in columns 4 and 6 when the model is estimated for *Narratives*. Indeed and as expected, *AIM×Post2007* loads with a negative coefficient in columns 4 and 6 for *Narratives* reflecting the relative increase in governance reporting requirements imposed on Main Market firms post-2007. Finally and in sharp contrast to the results for *Financials*, the increase in *Narratives* for Main Market firms post-2005 (i.e., -0.535 + 1.098 for word count and -1.431 + 2.257 for page count) is statistically indistinguishable from zero at the 0.05 level, consistent with the view that the disclosure impact of mandatory IFRS adoption centred primarily on the financial statements. Findings collectively provide further support for the validity of our text retrieval and classification method.

Report tone

Our second large sample validation test focuses on net tone, defined as the number of positive words minus the number of negative words, scaled by the sum of positive and negative words (Henry and Leone 2016). We test for predictable variation in tone using both cross-sectional and within-report approaches.

Within-report tests exploit predictable variation in tone across different sections from *the same annual report*. Examining within-document variation in tone helps mitigate endogeneity concerns regarding omitted variable bias because firm- and time-specific factors affecting reporting style and content are held constant. Tests compare tone for governance statements and remuneration reports with tone in the management commentary and the letter from the board chair. Governance statements and remuneration reports are mandatory disclosures for Main Market firms, with content shaped by compliance considerations that limit scope for relentless management optimism. In contrast, management face few constraints on the form and content of key performance-related commentaries such as the letter to shareholders and management's commentary (MD&A). Consistent with management exploiting their reporting discretion to present a favourable view of periodic performance, evidence of systematic positive reporting bias has been widely reported for management performance commentaries generally (Merkl-Davies and Brennan 2007, Li 2010) and for U.K. annual report commentaries in particular (Clatworthy and Jones 2006). Accordingly, we expect performance-focused sections such as the chair's letter and management commentary to be associated with more positive tone than governance statements and remuneration reports *in the same annual report*.[21]

[21] Dikolli at el. (2017) use similar arguments to motivate their within-firm comparison between the MD&A and the letter to shareholders.

We compute the within-report difference in net tone between Main Market firms' k^{th} performance section and their p^{th} mandatory governance-related section, where *k* is equal to the chair's letter or management commentary and *p* is equal to the governance statement or remuneration report. We expect $Tone^{k} - Tone^{p}$ to be positive. Findings for the resulting four pairwise combinations are reported in models (1)-(4) in Table 4. Consistent with expectations, intercept coefficients capturing the pairwise difference in tone are consistently positive and significant at the 0.01 level. The average chair's letter is over five (four) times more positive than the corresponding governance statement (remuneration report), while the average management commentary section is over four (three) times more positive than the corresponding governance statement (remuneration report). These within-document tests suggest our classification method is capable of reliably identifying key annual report sections.

Cross-sectional validity tests assess the replicability of established correlations between annual report tone and firm characteristics. Henry and Leone (2016, Table 8) report a robust positive correlation between MD&A tone and reported earnings, and robust negative associations with the book-to-market ratio and contemporaneous stock return volatility (due to lower growth options and higher uncertainty, respectively). Building on Henry and Leone (2016), we also expect annual report tone to have been less positive during the global financial crisis when valuations declined and economic forecasts looked bleak. Similar to Henry and Leone (2016), we estimate the following OLS regression:

$$\begin{aligned} Tone\_MD\,\&\,A_{it} = {} & \delta_0 + \delta_i Earn_{it} + \delta_2 BTM_{it} + \delta_3 ReturnVol_{it} + \delta_4 Crisis + \delta_5 Return_{it} \\ & + \delta_6 Size_{it} + \delta_7 ACC_{it} + \phi + \varepsilon_{it} \end{aligned} \quad . \qquad (5)$$

Variable definitions are as follows: *Tone_MD&A* is the aggregate number of positive minus negative words (scaled by the number of positive plus negative words) for the management commentary sections of the annual report; *Earn* is earnings per share from continuing operations scaled by lagged price; *ReturnVol* is the standard deviation of monthly stock returns in the 12 months prior to the fiscal year-end date; *Crisis* is an indicator variable equal to one for reports published during the financial crisis period (June 2007 through December 2010);[22] *Return* is cumulative stock returns for the fiscal year; *ACC* is earnings from continuing operations minus cash from operations, scaled by total assets; and *Size*, *BTM* and ϕ are as defined in model (3). Following Henry and Leone (2016) we test $\delta_1 > 0$, and δ_2 and $\delta_3 < 0$. We also test $\delta_4 < 0$ based on the prediction that management commentary was systematically less optimistic during the financial crisis. We treat *Returns*, *Size*, and *ACC* as control variables in equation (5) because findings reported by Henry and Leone (2016) for these covariates differ across tone measures.

Results for regression (5) are presented in the final column of Table 4, with all continuous (lower-bounded) variables winsorized at the top and bottom (top) percentile. *Earn*, *BTM* and *ReturnVol* load significantly with the expected signs. The estimated coefficient on *Crisis* is also negative at the 0.1 level. Management tone also correlates positively with contemporaneous 12-month stock returns which is intuitive despite not being evidenced robustly by Henry and Leone (2016). Finally, we note that tone is increasing in firm size although no prediction is offered for this variable. Evidence that the tone of management commentary varies cross-sectionally in ways predicted by prior research provides further support for the validity of our retrieval and classification procedure.

[22] The start of our financial crisis window coincides with U.S. congressional testimony on 1 June 2007. The end of our crisis window is 31 December 2010 following announcements on 1 December 2010 by the Federal Reserve (details of actions taken to stabilize markets since the start of the crisis) and 7 December 2010 by the U.S. Treasury Department (sale of remaining stake in Citigroup). We set *Crisis* equal to one for fiscal years ending after 1 March 2007 and before 31 March 2011) to allow a three months publication lag for the annual report. See https://www.stlouisfed.org/financial-crisis/full-timeline for a comprehensive timeline of events associated with the financial crisis.

Report readability

Our third large sample validation test focuses on document readability measured using the Fog index. Consistent with our analysis for net tone, we test for predictable variation in the Fog index using both within-report and cross-sectional approaches.

Our document-level approach tests for predictable disparity in readability across different sections of *the same annual report*. We expect narratives linked to regulatory compliance to be characterized by more complex language due to a higher incidence of jargon and a more legalistic writing style. Governance statements and remuneration reports are two U.K. annual report sections where content is determined by prevailing regulations to a large degree. In contrast, the chair's letter to shareholders is a voluntary disclosure designed specifically to provide a concise, accessible overview of firm performance and corporate milestones. We therefore expect the average chair's letter to display higher readability (lower Fog index) compared with governance statements and remuneration reports contained in the same report.

We compute the within-report difference in Fog index between the chair's letter and the p^{th} governance-related section, where p is equal to the governance statement or remuneration report. We expect $Readability^{Chair} – Readability^{p}$ to be negative.[23] Findings for pairwise comparisons are reported in columns (1) and (2) in Table 5. Intercept coefficients capturing the pairwise difference in Fog are negative and significant at the 0.01 level. The average chair's letter requires 1.9 years less education to read compared with the typical governance statement

[23] Descriptive statistics for readability reveal a high number of extreme values. For example, the minimum Fog index value for the chair's letter is zero and 95th percentile value is 30. We address this issue by trimming at the one and 95 percentiles. Results using raw readability scores are generally not significant.

and 2.4 years less training relative to the average remuneration report. (Untabulated descriptive statistics reveal that the chair's letter is associated with a Fog index of 19.7.)[24]

Cross-sectional validity tests for readability follow Li (2008) who predicts the Fog index for management commentary is increasing in weak earnings performance and transitory losses (due to managerial obfuscation), the market-to-book ratio (because growth options require more complex disclosures), firm size and the number of business segments (because disclosures tend to be more complicated for larger firms with more complex operations), and stock return volatility and earnings volatility (because high business and operating uncertainty are associated with more complex disclosures). Results reported by Li (2008, Tables 2 and 3) broadly support the predicted associations, although size and number of business segments do not load as expected. Following Li (2008) we estimate the following OLS regression:

$$\begin{aligned} Fog\_MD\&A_{it} = \lambda_0 + \lambda_1 Earn_{it} + \lambda_2 NonRec_{it} + \lambda_3 BTM_{it} + \lambda_4 Segments_{it} \\ + \lambda_5 Size_{it} + \lambda_6 ReturnVol_{it} + \lambda_7 EarnVol_{it} + \varphi + \phi + \upsilon_{it} \end{aligned} \quad , \qquad (6)$$

Variable definitions are as follows: *Fog_MD&A* is the Fog index (Gùnning 1968) for the management commentary section of firm *i*'s annual report published in year *t*, computed using Svoboda's (2013) algorithm; *Loss* is equal to one where reported earnings are negative and zero otherwise; *NonRec* is equal to one where GAAP earnings include negative exceptional items and zero otherwise; *EarnVol* is the standard deviation of earnings per share for the three-year period ending in year *t*; φ represents calendar year fixed effects; and all other variables are as defined in models (4) and (5). Following Li (2008) we test λ_1, λ_2, $\lambda_{4,}$ λ_5, λ_6 and $\lambda_7 > 0$ and $\lambda_3 < 0$.

Results for regression (6) are presented in models (3) and (4) of Table 5, with all continuous (lower-bounded) explanatory variables winsorized at the top and bottom (top)

[24] Loughran and McDonald (2016) note that differences in readability are often economically small although statistically significant (e.g., Lang and Stice-Lawrence. 2015). This is also true in our case, although not so extreme. We document differences equal to approximately two years of education, which is more material.

percentile, and *Fog_MD&A* trimmed at the one and 95 percentiles. With the exception of *Earn*, all explanatory variables in model (3) load with their expected signs, and *Segments*, *Size*, *ReturnVol* and *EarnVol* are significant at conventional levels. Similar results are evident in model (4) when the regression is extended to include time and industry fixed effects, with the exception that *Segments* is no longer significant. Lang and Stice-Lawrence (2015) also report mixed results using the Fog index. Our evidence suggests that caution is necessary when using readability scores for annual report text retrieved by our procedure. The mixed findings are also consistent with concerns about the Fog index as a measure of financial readability (Loughran and McDonald 2016) and evidence reported by El-Haj et al. (2019, Appendix) that award winning U.K. annual reports are not associated with reliably lower Fog scores. Collectively however, results reported in Tables 2-5 support conclusions from manual validation tests which suggest that our retrieval and classification procedure provides a reliable means of measuring textual content and document structure for large-sample analyses.

5. Annual report data resources

This section provides brief details of annual report narrative resources constructed using our procedure to support further research in this area. The first data resource is a comprehensive dataset of U.K. annual report features designed to support large-sample research into the properties and usefulness of glossy annual report narratives. The starting point for the dataset is reports published in calendar years 2002 through 2017 by firms listed on the LSE. The sample at the date of publication comprises 26,284 reports for 4,131 financial and non-financial firms. We use information from Datastream and the London Share Price Database to construct a unique, time-invariant firm identifier to account for name changes in an entity’s annual report time series. The dataset contains a range of narrative features including length, tone, readability and uncertainty for key report sections, and for the aggregate *Narratives* and *Financials* components. The dataset is available at <doi>, along with variable definitions, full details of the sampling procedure, and instructions on how to match reports with firm identifiers from Thomson Reuters Datastream.[25]

The second data resource is a set of annual report corpora designed to support corpus-based approaches to studying financial report narratives (Hardie, 2015). Using the subsample of 15,883 reports processed using the table of contents, we pool text from the k^{th} annual report section across all reports containing section *k*, where *k* is equal to the following generic categories: letter from the board chair, business review, CEO review, finance director review, operating and financial review, governance statement, remuneration report, risk report, corporate social responsibility disclosures, and the group audit report. (We also pool business reviews, CEO reviews, finance director reviews, and operating and financial reviews into a single management commentary category.) The *K* section corpora are available at <doi> for further analysis. Summary details for the corpora are presented in Table 6 and further details regarding corpus construction are provided on the appendix.

6. Extension to non-English language and reporting regime

This section provides evidence on the generalizability of our retrieval and classification procedure to non-English language annual reports published in regulatory settings other than the U.K. (See the appendix for more detailed guidelines.) We select Portuguese annual reports for because the authors have good knowledge of the Portuguese language and reporting environment, Portuguese is a structurally different language to English and therefore presents

[25] The dataset is revised on an annual basis. Old versions of the dataset are archived on GitHub. See appendix for further details of archiving strategy.

new linguistic challenges that help shed additional light on the robustness of our method, and the Portuguese regulatory environment governing annual reports differs significantly from the U.K. PDF reports published in Portuguese by firms listed on Euronext Lisbon are retrieved from Perfect Information for calendar years 2006 through 2015. The final sample of consists of 606 digital PDF reports for 77 firms (ranging from 64 firms in 2011 to 38 firms in 2015).

While much of our retrieval and classification procedure is independent of language and reporting regime, key elements rely on domain-specific gold standard wordlists and detailed knowledge of local reporting norms and therefore manual intervention is unavoidable. The two areas where manual intervention is required are: (a) constructing the list of section headers used to identify the report table of contents; and (b) developing new synonym lists that serve as inputs to our section classification algorithms.

We create the gold standard list of section headers for Portuguese annual reports by extracting all section titles from the contents table of 67 reports selected at random. The initial set contains 2,053 headers, which collapses to 694 after screening for duplicates and extraction errors. The resulting list contains multiple synonyms for the same section.[26] For example, our list contains 12 different titles for chair's letter to shareholders and 35 versions for the auditor's report. The complete list of synonyms is included in the appendix.

Synonym lists used as inputs to our section classification algorithm are constructed using the same approach as described in section 3. We start by reviewing Portuguese reporting rules and practices to determine a set of core sections that appear in the *Narratives* component of the representative report. We identify the following generic sections: chair's letter, CEO review, and performance commentary. All *Narratives* sections not classified into one of these three generic

[26] We retain commas and hyphens which leads us to treat two otherwise identical headers.as distinct elements of a synonym list. All other forms of punctuation are removed and ignored.

categories are allocated to a residual catch-all category (other). (Generic sections identified for the *Financials* component are audit report and financial statements.) We also identify performance commentary as the section that most frequently delineates the *Narratives* component of the annual report from the *Financials* component. (The equivalent to Figure 1 for Portuguese reports is presented in the appendix.) Next we run our retrieval algorithm over all reports to recover a comprehensive list of section headers from the tables of contents and then review the list manually to construct final synonym lists for our three generic *Narratives* categories. These lists are used as inputs to our classification algorithm that compares section headers in the table of contents with elements from the synonym lists. (Character string comparisons are performed after removing all spacing and punctuation from both table of contents headers and elements in the synonym lists.) Synonym lists are refined through several iterations where classified sections are reviewed manually to identify and fix errors.

Our procedure processed 396 reports via the table of contents, representing 65% of the 606 documents in the initial sample. Further analysis reveals that problems detecting or reading the table of contents are the primary reason why reports are not processed. (The majority of such reports can be processed using bookmark-based extraction.) Specifically, 62 reports do not contain a table of contents; 52 reports contain a table of contents that is not detected; 45 reports contain a table of contents that is unreadable due to unconventional formatting; 39 reports' table of contents do not contain page numbers; and 12 reports contains a table of contents spread over two or more pages.

We validate extraction and classification performance using a sample of 100 reports selected at random from the 396 processed documents. The validation process follows the same procedure described in section 4.1. Precision, recall and F_1 scores reported in Table 7 are very similar to those presented in Table 1 for U.K. reports. Panel A presents error rates relating to section identification. The overall accuracy rate as indicated by the F_1 score is 95.9%, compared with 96.5% reported for U.K. annual reports in Table 1. Our procedure correctly identifies 2,628 of the 2,682 actual sections in the 100 reports analysed, equating to a recall rate of 98%. The precision rate, although lower at 94% (169 type I errors), is nevertheless respectable in absolute terms. Results are broadly identical if we focus exclusively on the *Narrative* report component.

Page number synchronization rates reported in Panel B and document classification rates reported in Panel C are above 95% in all cases with the exception of performance commentary classification (94.7%). These rates are also consistent with results reported for U.K. annual reports in Table 1. Collectively, these findings confirm that the retrieval and classification method developed for U.K. annual reports is generalizable to non-English language annual reports published in regulatory settings other than the U.K.

7. Summary and conclusions

We develop, describe and evaluate a procedure for automatically retrieving and analyzing textual content in digital PDF annual report files. Extant large-sample research examining annual report content is confined primarily to 10-K filings prepared by U.S. registrants (El-Haj et al. 2019). However, most firms also publish an unstructured, glossy annual report containing additional disclosures and graphics. These documents are typically distributed as PDF files and represent the normal annual reporting method outside the U.S.

Our procedure for analyzing PDF annual report files involves detecting and retrieving the document table of contents, synchronizing page numbers in the native report with page numbers in the corresponding PDF file, and then using the synchronized page numbers to extract and analyse text separately for each section listed in the contents table. We retrieve text using bookmarks added by the PDF originator for reports where a valid table of contents cannot be identified. Our method retains information on document structure, facilitating delineation between narrative and financial statement components of reports, and between individual sections in the narratives component.

Manual and large-sample validity tests confirm the procedure provides a reliable means of capturing and classifying unstructured narrative disclosures. While the method is implemented using U.K. reports published in English, tests on Portuguese reports confirm that the procedure is generalizable to annual reports published in other languages and regulatory environments. The tool is available for researchers to use. At the date of publication, a dataset of text properties for over 26,000 annual reports published by 4,131 LSE-listed financial and non-financial firms between 2002 and 2017 is also available, together with a suite of annual report corpora derived from almost 16,000 reports.

An important limitation of our method that is mirrored in the large-sample text processing literature more generally is the failure to capture important aspects of disclosure format. The IASB's disclosure initiative outlines the features of effective communication, which include use of tables and infographics (IASB 2017b, para 2.21). The absence of content tags in the PDF file type means that we are unable to directly identify the presence and content of tables and infographics, and to distinguish text contained therein from that in the main narrative. We are also unable to measure the relative position and format (e.g., font size) of text on any given page. We doubt whether automated methods are capable of shedding significant light on questions relating to disclosure format and presentation and as such we view the large sample opportunities provided by our tool and dataset as complementing rather than replacing the need for careful small sample manual analysis.

References

Athanasakou, V., M. El-Haj, P. Rayson, M. Walker and S. Young (2019). Large sample evidence on the properties and impact of strategic commentary in annual reports. Available at: http://ssrn.com/abstract=3212854

Bloomfield, R. (2008). Discussion of ''Annual report readability, current earnings, and earnings persistence". *Journal of Accounting and Economics* 45: 248– 252

Campbell, J. Chen, H., Dhaliwal, D., Lu, H., Steele, L. (2013). The information content of mandatory risk factor disclosures in corporate filings. *Review of Accounting Studies* 19(1): 396-455

CFA Society U.K. (2016). *CFA UK annual survey on Financial Reporting and Analysis*. Available at: https://secure.cfauk.org/assets/1345/Analysis_of_FRAC_survey_2015.pdf

Clatworthy, M. and M. Jones, 2006. Differential patterns of textual characteristics and company performance in the chairman's statement. *Accounting, Auditing and Accountability Journal* 19 (4): 493-511

Dikolli, S., T. Keusch, W. Mayew and T. Steffen (2017). Using shareholder letters to measure CEO integrity. Available at SSRN: http://ssrn.com/abstract=2131476

Dyer, T., M. Lang and L. Stice-Lawrence (2017). The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics* 64: 221-245

El-Haj, M., P. Rayson, V. Simaki, M. Walker and S. Young (2019). In search of meaning: Lessons, resources and next steps for computational analysis of financial discourse. *Journal of Business Finance and Accounting* forthcoming

EY (2015). Annual Reporting in 2014: *Reflections on the Past, Direction for the Future*. Available at: http://www.ey.com/Publication/vwLUAssets/EY_-_Annual_reporting_in_2014_reflections_on_the_past_direction_for_the_future/$FILE/EY-Annual-reporting-in-2014.pdf

Financial Reporting Council (2012). *Thinking About Disclosures in a Broader Context: A Road Map for a Disclosure Framework*. London: FRC Publications. Available at: https://frc.org.uk/Our-Work/Publications/Accounting-and-Reporting-Policy/Thinking-about-disclosures-in-a-broader-contex-File.pdf

Financial Reporting Council (2014). *Guidance on the Strategic Report*. London: FRC Publications. Available at: https://www.frc.org.uk/Our-Work/Publications/Accounting-and-Reporting-Policy/Guidance-on-the-Strategic-Report.pdf

Gunning, R. (1968). *The Technique of Clear Writing*. McGraw-Hill

Grüning, M. (2011). Artificial intelligence measurement of disclosure (AIMD). *European Accounting Review* 20(3): 485–519

Hardie, A. (2015). Corpus linguistics. In: *The Routledge Handbook of Linguistics*: 502-515. London: Routledge.

Henry, E. (2006). Market Reaction to verbal components of earnings press releases: Event study using a predictive algorithm. *Journal of Emerging Technologies in Accounting* 3: 1-19

Henry, E. (2008). Are investors influenced by how earnings press releases are written? *Journal of Business Communication* 45(4): 363-40

Henry, E. and A. Leone (2016). Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone. *The Accounting Review* 91(1): 153-178

Hooghiemstra, Y. F. Kuang and B. Qin (2017). Does obfuscating excessive CEO pay work? The influence of remuneration report readability on say-on-pay votes. *Accounting and Business Research* 47(6): 695-729

Institute of Chartered Secretaries and Administrators (2015). *Guidance note Contents List for the Annual Report of a UK Company*. London, ICSA. Available at: https://www.icsa.org.uk/assets/files/free-guidance-notes/contents-list-for-the-annual-report-of-a-uk-company.pdf

International Accounting Standards Board (2017a). *Exposure Draft: Improvements to IFRS 8 Operating Segments (Amendments to IFRS 6 and IAS 34)*. IFRS Foundation. Available at: https://www.ifrs.org/-/media/project/improvements-to-ifrs-8-operating-segments/exposure-draft/published-documents/ed-proposed-amendments-ifrs8-ias34.pdf

International Accounting Standards Board (2017b). *Disclosure Initiative - Principles of Disclosure (Discussion Paper DP/2017/1)*. March 2017. International Accounting Standards Board: London

International Accounting Standards Board (2010). *IFRS Practice Statement: Management Commentary. A Framework for Presentation*. IFRS Foundation. Available at http://www.ifrs.org/Current-Projects/IASB-Projects/Management-Commentary/IFRS-Practice-Statement/Documents/Managementcommentarypracticestatement8December.pdf

International Auditing and Assurance Standards Board (2015). *International Standards on Auditing (ISA) 720 (Revised): The Auditor's Responsibilities Relating to Other Information*. Available at: http://www.ifac.org/publications-resources/international-standard-auditing-isa-720-revised-auditor-s-responsibilities--0

Kincaid, J. P., R. P. Fishburne, R. L. Rogers, and B. S. Chissom (1975). Derivation of new readability formulas (automated readability index Fog count and flesch reading ease formula) for navy enlisted personnel. In *Research Branch Rep*: pp. 8–75. Memphis, TN

Laksmana, I., W. Tietz and Y-W Yang (2012). Compensation discussion and analysis (CD&A): Readability and management obfuscation. *Journal of Accounting and Public Policy* 31(2): 185-203

Lang, M. and L. Stice-Lawrence (2015). Textual Analysis and International Financial Reporting: Large Sample Evidence. *Journal of Accounting and Economics* 60 (2-3): 110-135

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10 (8): 707–710

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics* 45 (2-3): 221-247

Li, F. (2010). Textual analysis of corporate disclosures: A survey of the literature. *Journal of Accounting Literature* 29: 143-165

Loughran, T. and B. McDonald (2011), When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance* 66(1): 35–65

Loughran, T., and B. McDonald (2014). Regulation and financial disclosure: The impact of plain English. *Journal of Regulatory Economics* 45: 94-113

Loughran, T., and B. McDonald (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research* 54 (4): 1187-1230

Manning C. and H. Schütze (2008) *Foundations of Statistical Natural Language Processing*, MIT Press. Cambridge, MA

Merkl-Davies, D. and N. Brennan (2007). Discretionary disclosure strategies in corporate narratives: Incremental information or impression management? *Journal of Accounting Literature* 26: 116-194

Morunga, M. and M. E. Bradbury (2012). The Impact of IFRS on annual report length, *Australasian Accounting, Business and Finance Journal* 6(5): 47-62

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics* 13 (4): 519-549

Schleicher, T., K. Hussainey and M. Walker (2007). Loss firms' annual report narratives and share price anticipation of earnings. *British Accounting Review* 39(2): 153-171

Svoboda, R. (2013). Framework and API for assessing quality of documents and their sources. Available at: http://epublications.uef.fi/pub/urn_nbn_fi_uef-20130301/urn_nbn_fi_uef-20130301.pdf and https://github.com/ogrodnek/java_fathom

Van Rijsbergen, C. J. (1979). *Information Retrieval* (2nd ed.). Butterworth

Watts, R. and J. Zimmerman (1986). *Positive Accounting Theory*. Prentice Hall Englewood Cliffs, New Jersey

Table 1. Manual evaluation of annual report extraction and classification performance

Panel A: Section extraction

			Error frequencies		Retrieval performance (%)		
	N actual	N extracted	Type 1	Type 2	Precision	Recall	F_1 score
Pooled annual report	11,009	10,820	286	475	97.36	95.69	96.52
Narratives component	5,237	5,233	216	220	95.87	95.80	95.83

Panel B: Page number synchronization

	Type I errors for section extraction treated as incorrect pagination			Type I errors for section extraction not treated as incorrect pagination		
	N	N errors	Precision (%)	N	N errors	Precision (%)
Pooled annual report	10,820	736	93.20	10,534	450	95.73
Narratives component	5,233	500	90.44	5,017	248	95.06

Panel C: Document classification

			Error frequencies		Retrieval performance (%)		
	N actual	N classified	Type 1	Type 2	Precision	Recall	F_1 score
Narratives component	4,929	4,934	88	83	98.18	98.32	98.25
Financials component	5,434	5,429	83	88	98.50	98.38	98.44
By section category:							
Chairman’s letter	521	520	3	4	99.42	99.23	99.32
CEO review	280	283	10	7	96.34	99.23	97.76
CFO review	328	321	12	19	96.12	97.50	96.80
Governance statement	491	504	27	14	94.34	94.21	94.27
Remuneration report	406	397	0	9	100.00	97.15	98.55
Highlights	276	278	3	1	98.91	97.78	98.34

The analysis is based on 11,009 (10,820) actual (retrieved) sections for 586 digital PDF annual reports processed according to the table of contents and selected at random from reports published by London Stock Exchange-listed non-financial firms during the period January 2003 through September 2014. Panel A presents evidence on the retrieval of section headers listed in the table of contents. Retrieval performance is measured by comparing all sections listed in the table of contents for each randomly selected annual report with headers extracted by our procedure. Type 1 errors (false positives) reflect instances where the procedure retrieves information that is not a valid section listed in the corresponding annual report table of contents. Type 2 errors (false negatives) reflect instances where the procedure fails to retrieve valid sections listed in the table of contents. Retrieval performance is assessed using three criteria. Precision measures the fraction of retrieved instances that are relevant (i.e., frequency of Type I errors) and is viewed as a measure of exactness or quality. Recall measures the fraction of relevant instances that are retrieved (i.e., frequency of Type 2 errors) and reflects a measure of completeness or quantity. F_1 scores represent the harmonic mean of Precision and Recall, and reflect an overall measure of retrieval and classification accuracy. Panel B presents evidence on the performance of the pagination algorithm for aligning page numbers in the PDF file with pages numbers listed in the table of contents. Performance is assessed by identifying instances where page numbers assigned by the procedure to mark the start and end of sections differ from actual start and end pages in the native PDF file. Results are presented for two tests: analyses reported in columns 2-4 use the full sample of extracted sections (i.e., including false positives) and classifies Type 1 extraction errors as instances of incorrect pagination; analyses reported in columns 5-7 use the sample of extracted sections that are valid (i.e., excluding false positives) and therefore does not classify Type 1 extraction errors as instances of incorrect pagination. Panel C reports information on document classification accuracy based on the sample of correctly extracted sections. The first two rows present evidence on the accuracy with which the algorithm partitions annual reports into the *Narrative*s and *Financials* components of the annual report. The remaining rows in Panel C present information on classification accuracy associated with core sections of the *Narratives* component of annual reports.

Table 2. Coefficient estimates and model summary statistics for OLS regressions explaining annual report length. Two-tailed probability values are reported in parentheses.

		Word count				Page count			
Variable	*Expected sign*	*MD&A*	Log *MD&A*	*Narratives*	Log *Narratives*	*MD&A*	Log *MD&A*	*Narratives*	Log *Narratives*
Intercept	?	-10.104	4.865	-26.605	6.200	-20.886	-0.915	-44.630	0.604
		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Size	+	1.322	0.274	3.445	0.251	2.754	0.254	6.207	0.212
		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Loss	+	0.232	0.048	1.012	0.043	-0.140	0.009	0.528	0.014
		(0.05)	(0.05)	(0.01)	(0.01)	(0.60)	(0.70)	(0.25)	(0.35)
BTM	–	-0.589	-0.114	-1.683	-0.090	-1.325	-0.113	-2.972	-0.072
		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
ReturnVol	+	0.862	0.126	2.282	0.071	1.049	0.082	2.223	0.055
		(0.06)	(0.25)	(0.01)	(0.37)	(0.30)	(0.39)	(0.16)	(0.40)
Segments	+	0.240	0.034	0.510	0.032	0.470	0.035	1.005	0.032
		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Main	+	-0.545	-0.059	-3.336	-0.175	-1.047	-0.054	-5.182	-0.119
		(0.01)	(0.04)	(0.01)	(0.01)	(0.01)	(0.06)	(0.01)	(0.01)
Length_Financials	+	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Industry fixed effects		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted-R^2 (%)		47.53	43.22	65.55	61.30	40.40	43.75	57.8	67.65
N		9,863	9,863	11,332	11,332	9,875	9,875	11,340	11,340

The dependent variable is annual report length. Annual report length is measured as either the number of words scaled by 10^3 (columns 3-6) or the number of pages (columns 7-10). Regressions for columns headed *MD&A* are estimated using word count and page count for the management commentary (MD&A) section of the annual report, while columns headed *Narratives* are estimated using word count and page count for the entire *Narratives* component of the annual report. Separate results are presented using raw counts and logged values of word and page counts. Explanatory variables are defined as follows: *Size* is the natural logarithm of total assets (WC02999); *Loss* is an indicator variable equal to one for firm-years where net income (WC01706) is less than zero, and zero otherwise; *Segments* is the number of business segments (count of WC19501 to WC19591); *BTM* is book value of shareholders' funds (WC03995) plus the book value of debt (WC03255) divided by the market value of equity (MV) plus the book value of debt; *ReturnVol* is the standard deviation of monthly stock returns computed over fiscal year *t*; *Main* is an indicator variable equal to one if firm *i* is listed on the LSE Main Market in fiscal year *t* and zero otherwise; *Length_Financials* is the number of words in the *Financials* component of the annual report; *Industry fixed effects* is a vector of indicator variables for Datastream level-4 sectors. All continuous (lower-bounded) variables are winsorized at the top and bottom (top) percentile. Two-tailed probability values reported in parentheses are computed using standard errors clustered by firm and adjusted by (N–1)/(N–P)× G/(G–1) to obtain unbiased estimates for finite samples, where N is the sample size, P is the number of independent variables, and G is the number of clusters.

Table 3. Coefficient estimates and model summary statistics for OLS regressions examining the impact of regulation on annual report length. Two-tailed probability values are reported in parentheses.

	Expected sign for:		Word count		Page count	
Variable	*Narratives*	*Financials*	*Narratives*	*Financials*	*Narratives*	*Financials*
Intercept	?	?	-28.840	3.064	-49.288	-39.748
			(0.011)	(0.55)	(0.01)	(0.01)
AIM	?	–	0.714	2.283	2.397	4.558
			(0.06)	(0.01)	(0.01)	(0.01)
Post2005		+	-0.535	4.121	-1.431	1.795
			(0.03)	(0.01)	(0.01)	(0.01)
Main × Post2005		+	1.098	5.004	2.257	12.322
			(0.01)	(0.01)	(0.01)	(0.01)
Post2007	+	+	5.855	0.123	11.365	2.304
			(0.01)	(0.83)	(0.01)	(0.01)
AIM × Post2007	–	+	-4.888	5.313	-9.078	6.326
			(0.01)	(0.01)	(0.01)	(0.01)
Size	+	+	3.458	0.337	6.254	6.348
			(0.01)	(0.09)	(0.01)	(0.01)
Loss	+	+	1.071	1.127	0.565	2.101
			(0.01)	(0.01)	(0.20)	(0.01)
BTM	–	–	-2.032	0.198	-3.697	-2.100
			(0.01)	(0.48)	(0.01)	(0.01)
Segments	+	+	0.395	0.595	0.778	1.717
			(0.01)	(0.01)	(0.01)	(0.01)
Length_Financials	+	+	0.000		0.000	
			(0.01)		(0.01)	
Length_Narratives	+	+		0.001		0.000
				(0.01)		(0.01)
Industry fixed effects			Yes	Yes	Yes	Yes
Adjusted-R^2 (%)			67.99	32.12	60.61	43.78
N			11,856	11,856	11,856	11,856

The dependent variable is either the total word count scaled by 10^3 (columns 4 and 5) or the total page count (columns 6 and 7) for the *k*th component of the annual report, where *k* = *Narratives* or *Financials*. Explanatory variables are defined as follows (Datastream variable names in parentheses): *Main* is an indicator variable equal to one if firm *i* in year *t* is listed on the LSE Main Market and zero otherwise; *Post2005* is an indicator variable equal to one for fiscal years beginning on or after January 1, 2005 and zero otherwise; *Post2007* is an indicator variable equal to one for fiscal years beginning on or after January 1, 2007, and zero otherwise; *Size* is the natural logarithm of total assets (WC02999); *Loss* is an indicator variable equal to one for firm-years where net income (WC01706) is less than zero, and zero otherwise; *Segments* is the number of business segments (count of WC19501 to WC19591); *BTM* is book value of shareholders' funds (WC03995) plus the book value of debt (WC03255) divided by the market value of equity (MV) plus the book value of debt; *Length_(k)* is the total word count (columns 4 and 5) or the total page count (columns 6 and 7) for the *k*th annual report component; *Industry fixed effects* is a vector of indicator variables for Datastream level-4 sectors. The column headed "Expected sign for" presents predicted coefficient signs for regressions where the dependent variable is the *Narratives* (*Financials*) component of the annual report: Null indicates cases where no association is predicted and ? indicates cases where the expected sign is indeterminate. Two-tailed probability values reported in parentheses are computed using standard errors clustered by firm and adjusted by (N–1)/(N–P)× G/(G–1) to obtain unbiased estimates for finite samples, where N is the sample size, P is the number of independent variables, and G is the number of clusters.

Table 4: Coefficient estimates and model summary statistics for OLS regressions for net tone. Two-tailed probability values are reported in parentheses.

		Pairwise difference				Cross-sectional
		$Tone^{Chair}$ –	$Tone^{Chair}$ –	$Tone^{MD\&A}$ –	$Tone^{MD\&A}$ –	
	Expected	$Tone^{Gov}$	$Tone^{Rem}$	$Tone^{Gov}$	$Tone^{Rem}$	*Tone_MD&A*
Variable	sign	(1)	(2)	(3)	(4)	(5)
Intercept	+	0.588	0.461	0.440	0.314	-0.078
		(0.01)	(0.01)	(0.01)	(0.01)	(0.09)
Earn	+					0.000
						(0.01)
BTM	–					-0.047
						(0.01)
ReturnVol	+					-0.335
						(0.01)
Crisis	–					-0.009
						(0.07)
Return	?					0.000
						(0.01)
Size	?					0.019
						(0.01)
ACC	?					0.000
						(0.12)
Industry fixed effects		No	No	No	No	Yes
Adjusted-R^2 (%)						20.41
N		4,922	4,768	5,291	5,129	9,867

Net tone is equal to the number of positive words minus the number of negative words, divided by the sum of positive and negative words. Positive and negative word counts are constructed using the word lists available on Bill McDonald's webpage. Results for models 1-4 (headed "Pairwise difference") test for the difference in net tone for sections *k* and *p* from the same annual report, where *k* is equal to either the chair's letter or the management commentary section and *p* is equal to either the governance statement or the remuneration report. Results for model 5 (headed "Cross-sectional") present results explaining variation in net tone for the management commentary section of the annual report. Explanatory variables are defined as follows (Datastream variable names in parentheses): *Earn* is net income available to common shareholders (WC01751) before transitory items scaled by lagged the market value of equity at the fiscal year-end; *BTM* is book value of shareholders' funds (WC03995) plus the book value of debt (WC03255) divided by the market value of equity (MV) plus the book value of debt; *ReturnVol* is the standard deviation of monthly stock returns computed over fiscal year *t*; *Crisis* is an indicator variable equal to one for reports published during the financial crisis period (June 2007 through December 2010); *Return* is cumulative contemporaneous stock returns for fiscal year *t*; *Size* is the natural logarithm of total assets (WC02999); *ACC* is net income from continuing operations (WC01751 – max(WC01505, WC18200, WC01269) minus cash from operations (WC04860), scaled by total assets; *Industry fixed effects* is a vector of indicator variables for Datastream level-4 sectors. The column headed "Expected sign" presents predicted coefficient signs for regressors. Two-tailed probability values reported in parentheses are computed using standard errors clustered by firm and adjusted by (N–1)/(N–P)× G/(G–1) to obtain unbiased estimates for finite samples, where N is the sample size, P is the number of independent variables, and G is the number of clusters.

Table 5: Coefficient estimates and model summary statistics for OLS regressions for readability. Two-tailed probability values are reported in parentheses.

	Expected	Pairwise difference		Cross-sectional	
		$Fog^{Chair} - Fog^{Gov}$	$Fog^{Chair} - Fog^{Rem}$		
	sign	(1)	(2)	(3)	(4)
Intercept	–	-1.862 (0.01)	-2.404 (0.01)	18.587 (0.01)	18.389 (0.01)
Earn	–			0.000 (0.79)	0.000 (0.21)
NonRec	+			0.169 (0.18)	0.018 (0.88)
BTM	–			-0.007 (0.93)	-0.092 (0.23)
Segments	+			0.112 (0.01)	0.031 (0.40)
Size	+			0.117 (0.01)	0.187 (0.01)
ReturnVol	+			2.457 (0.01)	1.555 (0.01)
EarnVol	+			0.000 (0.01)	0.000 (0.01)
Year fixed effects		No	No	No	Yes
Industry fixed effects		No	No	No	Yes
Adjusted-R^2 (%)				0.67	5.85
N		7,574	6,856	7,991	7,991

Readability is the Fog index (Gǜnning, 1968) computed using a version of Fathom (Svoboda 2013). Results presented for models 1-2 (headed "Pairwise difference") test for the difference in the Fog index for sections *k* and *p* from the same annual report, where *k* is equal to the chair's letter and *p* is equal to either the governance statement or the remuneration report. Results presented models 3-4 (headed "Cross-sectional") present results explaining variation in the Fog index for the management commentary section of the annual report. Explanatory variables are defined as follows (Datastream variable names in parentheses): *Earn* is net income available to common shareholders (WC01751) before transitory items scaled by lagged the market value of equity at the fiscal year-end; *NonRec* is equal to one where net income available to common shareholders includes negative exceptional items and zero otherwise; *BTM* is book value of shareholders' funds (WC03995) plus the book value of debt (WC03255) divided by the market value of equity (MV) plus the book value of debt; *Segments* is the number of business segments (count of WC19501 to WC19591); *Size* is the natural logarithm of total assets (WC02999); *ReturnVol* is the standard deviation of monthly stock returns computed over fiscal year *t*; *EarnVol* is the standard deviation of earnings per share (WC05202) computed over the three-year period ending in year *t*; *Year fixed effects* is a vector of indicator variables for calendar year; *Industry fixed effects* is a vector of indicator variables for Datastream level-4 sectors. The column headed "Expected sign" presents predicted coefficient signs for regressors. Two-tailed probability values reported in parentheses are computed using standard errors clustered by firm and adjusted by (N–1)/(N–P)× G/(G–1) to obtain unbiased estimates for finite samples, where N is the sample size, P is the number of independent variables, and G is the number of clusters.

Table 6: Summary statistics for annual report corpora

Annual report corpora	Number of reports	Number of firms	Number of words
Letter from board chair	14,032	2,752	15,389,643
Management commentary	11,507	2,261	49,644,028
Comprising:			
CEO review	7,160	1,640	13,947,211
Financial review	8,460	1,686	20,013,680
Operating review	2,819	794	7,008,451
Business review	2,689	795	8,674,686
Principal risks and uncertainties	4,715	1,090	11,781,738
Governance commentary	12,844	2,513	45,033,431
Comprising:			
Governance statement	12,766	2,500	43,695,127
Chair’s governance introduction	1,137	430	1,338,304
Remuneration report	12,725	2,269	39,668,122
Corporate social responsibility disclosures	6,630	1,148	12,948,932
Highlights	11,099	2,082	3,750,407
Group audit report	15,038	2,884	19,036,357
Entire *Narratives* component (excluding audit report)	15,883	2,925	178,216,301
Entire *Narratives* component (including audit report)	15,883	2,936	197,252,658

Corpora are constructed from an initial sample of 31,464 annual reports published between 2002 and 2017 by firms listed on the London Stock Exchange (LSE), or which our tool processed 26,284 reports, The final sample includes reports published by financial and non-financial firms listed on either the LSE Main Market or the Alternative Investment Market (AIM). The document table of contents (TOC) forms the basis of extraction for 15,883 reports (approximately 60%); pre-existing document bookmarks are used to process the remaining 10,401 reports. Corpora are constructed using the pooled set of reports processed using TOC to ensure classification consistency across reports. Corpora are available at <insert doi>. Each corpus is disaggregated by report calendar year and where necessary each year is further decomposed into separate files each comprising approximately million words. Archived versions of corpora will are available at https://github.com/drelhaj/CFIE-FRSE.

Table 7: Manual evaluation of annual report extraction and classification performance for Portuguese annual reports.

Panel A: Section extraction

			Error frequencies		Retrieval performance (%)		
	N actual	N extracted	Type 1	Type 2	Precision	Recall	F_1 score
Pooled annual report	2,682	2,797	169	54	93.95	97.99	95.93
Narratives component	2,313	2,409	147	51	93.90	97.80	95.81

Panel B: Page number synchronization

	Type I errors for section extraction treated as incorrect pagination			Type I errors for section extraction not treated as incorrect pagination		
	N	N errors	Precision (%)	N	N errors	Precision (%)
Pooled annual report	2,797	85	96.96	2,628	84	96.80
Narratives component	2,409	71	97.05	2,262	76	96.64

Panel C: Document classification

			Error frequencies		Retrieval performance (%)		
	N actual	N classified	Type 1	Type 2	Precision	Recall	F_1 score
Narratives component	2,313	2,409	147	51	93.90	97.80	95.81
Financials component	369	388	22	3	94.33	99.19	96.70
By section category:							
Chairman	39	40	1	0	97.50	100.00	98.73
CEO	31	31	1	1	96.77	96.77	96.77
Performance	1,520	1,625	135	30	91.69	98.03	94.75
Auditor	166	160	3	9	98.13	94.59	96.33
Corporate Governance	167	164	0	3	100.00	98.20	99.09
Other	390	389	7	8	98.20	97.95	98.07

The analysis is based on 2,682 (2,797) actual (retrieved) sections for 396 digital PDF annual reports selected at random from reports published by Lisbon Stock Exchange-listed non-financial Portuguese firms during the period January 2006 through December 2015. Panel A presents evidence on the retrieval of section headers listed in the table of contents. Retrieval performance is measured by comparing all sections listed in the table of contents for each randomly selected annual report with headers extracted by our procedure. Type 1 errors (false positives) reflect instances where the procedure retrieves information that is not a valid section listed in the corresponding annual report table of contents. Type 2 errors (false negatives) reflect instances where the procedure fails to retrieve valid sections listed in the table of contents. Retrieval performance is assessed using three criteria. Precision measures the fraction of retrieved instances that are relevant (i.e., frequency of Type I errors) and is viewed as a measure of exactness or quality. Recall measures the fraction of relevant instances that are retrieved (i.e., frequency of Type 2 errors) and reflects a measure of completeness or quantity. F_1 scores represent the harmonic mean of Precision and Recall, and reflect an overall measure of retrieval and classification accuracy. Panel B presents evidence on the performance of the pagination algorithm for aligning page numbers in the PDF file with pages numbers listed in the table of contents. Performance is assessed by identifying instances where page numbers assigned by the procedure to mark the start and end of sections differ from actual start and end pages in the native PDF file. Results are presented for two tests: analyses reported in columns 2-4 use the full sample of extracted sections (i.e., including false positives) and classifies Type 1 extraction errors as instances of incorrect pagination; analyses reported in columns 5-7 use the sample of extracted sections that are valid (i.e., excluding false positives) and therefore does not classify Type 1 extraction errors as instances of incorrect pagination. Panel C reports information on document classification accuracy based on the sample of correctly extracted sections. The first two rows present evidence on the accuracy with which the algorithm partitions annual reports into the *Narrative*s and *Financials* components of the annual report. The remaining rows in Panel C present information on classification accuracy associated with generic categories of the *Narratives* component of annual reports.

Figure 1: Representative U.K. annual report structure used as a basis for document classification

Item	Section	Regulatory status	Component	Description
1.	Introduction	Non-mandatory	Narratives	Summary of the business and its main markets; often includes or comprises a section titled Highlights summarizing performance for the reporting period and/or a section titled At a Glance.
2.	Chairman’s statement	Non-mandatory	Narratives	Summary of periodic performance, events occurring between the fiscal year-end and the annual report publication date, and outlook. (Signed by Chairman of the Board)
3.	Management commentary	Mixed	Narratives	Similar in focus to 10-K item 7A (MD&A) for U.S. registrants. It typically contains multiple sections including one or more of the following (or synonyms thereof): strategic review, CEO review, review of operations, business review, financial or CFO review. Further, the Companies Act 2006 requires that the business review must contain a description of the principal risks and uncertainties facing the company. Nomenclatures for this element of the report vary dramatically across firms and time. Prior to October 2013, U.K. company law required Main Market firms to present a Business Review containing commentary on operational and financial aspects of performance. Formally firms were required to use the term “business review” but practice varied widely, with some firms aggregating all performance-related commentary into single section titled Business Review, while other firms used the term to refer to a set of distinct sections such as the CEO and CFO reviews. Post-October 2013, U.K. company law was revised to require Main Market firms to present a Strategic Review in place of the Business Review, with additional mandatory disclosure requirements relating to strategy and business model. As was previously the case with the Business Review, practical implementation of this requirement varies dramatically across firms and involves a range of different nomenclatures referring to the same underlying content. (All or part therefore signed by CEO; financial reviews are signed by CFO.)
4.	Principal risks and uncertainties	Mandatory	Narratives	The Companies Act 2006 414C(2)(b) requires that the strategic report contains a description of the principal risks and uncertainties facing the company (previously Companies Act 2006 417). Further, Provision C.2.2 of the UK Corporate Governance Code 2014 requires that when taking account of the company’s current position and principal risks, directors should explain in the annual report how they have assessed the prospects of the company, over what period they have done so and why they consider that period to be appropriate. The intention of C.2.2 is for companies to apply the provision in two stages: directors first assess the prospects of the company and then make a statement of its viability.
5.	Other sections	Non-mandatory	Narratives	Various; commonly occurring content includes corporate, social and environmental responsibility reports, employee case studies and awards, senior management structure, details of corporate advisors, etc.

Figure 1 *continued*

6.	Director's biographies	Mandatory	Narratives	Brief biographical details of executive (inside) and non-executive (outside) directors and information on committee membership
7.	Directors' report	Mandatory	Narratives	Contains a range of statutory information (or cross-references to other sections of the report where information is presented) including principal activities, distributions to shareholders, directors' and their interests, directors' indemnities, political donations, share capital, substantial shareholdings, research and development, employee involvement, creditor payment policy, going concern, post-balance sheet events, auditor information, annual general meeting, etc. (Signed by Company Secretary)
8.	Corporate governance statement	Mandatory*	Narratives	Includes information on compliance with the U.K. Corporate Governance Code (including information on internal controls), together with reports from key monitoring committees such as audit and nomination. Increasingly common for (large) firms to subdivide this element in separate subsections. (Signed by Chairman of the Board; individual committee reports signed by respect committee Chairman.)
9.	Remuneration report	Mandatory*	Narratives	Details of remuneration policy and directors' compensation for the reporting period. The report comprises an audited element and a non-audited section. (Signed by Chairman of Remuneration Committee.)
10.	Statement of directors' responsibilities	Mandatory	Financials	Order is interchangeable with Auditor's Report
11.	Auditor's report	Mandatory	Financials	Order is interchangeable with Statement of Directors' Responsibilities
12.	Primary financial statements	Mandatory	Financials	As required by IAS 1 (U.K. GAAP prior to IFRS adoption); group and parent company where appropriate.
13.	Notes to the financial statements	Mandatory	Financials	As required by IAS 1 (U.K. GAAP prior to IFRS adoption)
14.	Miscellaneous disclosures	Mixed	Financials	Various including both statutory shareholder information and discretionary disclosures; common examples of disclosures in this category include, three- or five-year review, information on subsidiaries and operating locations, classification of shareholdings, notice of annual general meeting (with details of resolutions proposed), corporate address and contact information, financial calendar, glossary of terms, etc.

Column 1 contains generic titles for the most common sections presented in annual reports published by firms listed on the London Stock Exchange (LSE) Main Market and Alternative Investment Market (AIM). In reality firms are free to use whatever naming conventions they wish for annual report sections. Column 2 indicates whether a given section is required or voluntary. Mandatory identifies section(s) required by either U.K. company law or the U.K. Corporate Governance Code issued by the Financial Reporting Council; Mandatory* indicates the section(s) is mandatory for LSE Main Market firms but discretionary for AIM registrants; Non-mandatory identifies sections that are not required under prevailing reporting regulations. Mixed indicates a broad class of sections involving statutory and voluntary disclosures. Column 3 classifies the section as either part of the financial statements component of the report (*Financials*) or part of the Narratives component of the report (*Narrative*).