Towards a Multilingual Financial Narrative Processing System

Mahmoud El-Haj[1], **Paul Rayson**[1], **Paulo Alves**[2], **and Steven Young**[3]
[1]School of Computing and Communications, Lancaster University, UK
[2]Management School, Lancaster University, UK
[3]Universidade Católica Portuguesa, Portugal
{[1]m.el-haj, p.rayson}@lancaster.ac.uk, [2]palves@porto.ucp.pt, [3]s.young@lancaster.ac.uk

Abstract

Large scale financial narrative processing for UK annual reports has only become possible in the last few years with our prior work on automatically understanding and extracting the structure of unstructured PDF glossy reports. This has levelled the playing field somewhat relative to US research where annual reports (10-K Forms) have a rigid structure imposed on them by legislation and are submitted in plain text format. The structure extraction is just the first step in a pipeline of analyses to examine disclosure quality and change over time relative to financial results. In this paper, we describe and evaluate the use of similar Information Extraction and Natural Language Processing methods for extraction and analysis of annual financial reports in a second language (Portuguese) in order to evaluate the applicability of our techniques in another national context (Portugal). Extraction accuracy varies between languages with English exceeding 95%. To further examine the robustness of our techniques, we apply the extraction methods on a comprehensive sample of annual reports published by UK and Portuguese non-financial firms between 2003 and 2015.

Keywords: Financial Narrative Processing, NLP, annual reports, Information Extraction, Multilingual

1. Introduction

There are a number of different financial reporting requirements and legislative frameworks for national and international companies in terms of how they must report to their shareholders, potential investors and the financial markets. Companies produce a variety of reports containing both textual and numerical information at various times during their financial year, including annual financial reports, quarterly reports, preliminary earnings announcements and press releases. Additionally, conference calls with analysts are transcribed and made available publicly, and other sources of information such as media articles and online social media are employed by companies, analysts and the general public. This creates a vast financial information environment which can be impossible to keep track of. Previous academic research in accounting and finance areas has tended to focus on numerical information, or small scale manual studies of textual information. Over the last few years, we have been able to contribute to the scaling up of the textual analysis component by applying Information Extraction (IE), Natural Language Processing (NLP) and Corpus Linguistics (CL) methods to the data. [1] have focussed on the UK context where annual financial reports are released in glossy PDF format with a variety of different looser structures, and these have made it harder to apply normal research methods on a large scale. In contrast, much of the previous research has been targetted at the US context where annual 10-K forms are required to follow a rigid structure with a standard set of headings, and are written in plain text. A standard format enables more straightforward selection of relevant sections for further analysis. In this paper, we describe not only the structure detection and extraction process that we have designed and implemented for English annual reports, but also our initial work to extend this research to another national context, in this case to Portugal. We report on our experiments to port the system from English annual reports to those published in Portuguese, and describe the adaptations made to the system to enable this. Our methods extract information on document structure which is needed to enable a clear distinction between narrative and financial statement components of annual reports and between individual sections within the narratives component. The resulting software is made freely available for academic research.

[1]For more details, see the CFIE projects described at http://ucrel.lancs.ac.uk/cfie/

2. Related Work

Previous related work on financial narrative analysis has taken place in a number of areas including accounting and finance research, natural language processing and corpus linguistics. Some early approaches in the accounting and finance literature employed painstaking manual approaches and were therefore limited in scale due to time constraints. Further studies have become larger scale but are still using manually constructed word lists for detecting features without considering local context for disambiguation purposes or more advanced machine learning methods. Well known studies include one by Li (2010) which considered forward-looking statements in 10-K (annual) and 10-Q (quarterly) filings in the US and found a link between positive statements and better current performance and other indicators. Li also found that general content analysis dictionaries (such as Diction, General Inquirer and LIWC) are not helpful in predicting future performance. Loughran and McDonald (2011) also found that negative words in the general purpose Harvard Dictionary were not typically considered as negative in financial contexts, and so were less appropriate than domain specific versions. They also considered US 10-K reports for their study. Schleicher and Walker (2010) found that companies with impending performance downturns bias the tone in outlook sections of the financial narrative. A good survey of text analysis methods in accounting and finance research was recently published by Loughran and McDonald (2016).

In the natural language processing research area, previous research has been carried out to extract document structure mainly from scientific articles and books (Doucet et al., 2009; Teufel, 2010; McConnaughey et al., 2017). Other than this, there has been much recent work in using text mining and sentiment analysis, in particular to Twitter, with the goal of predicting stock market performance (Devitt and Ahmad, 2007; Schumaker, 2010; Im et al., 2013; Ferreira et al., 2014; Neuenschwander et al., 2014) although presumably any really successful methods would not be published. From the other end of the language analysis spectrum, in linguistics, there has been a large amount of research on the language of business communication. Merkl-Davies and Koller (2012) introduced the Critical Discourse Analysis (CDA) approach to the accounting community and showed how it can be used to systematically analyse corporate narrative documents to explore how grammatical devices can be used to obfuscate and guide interpretations. Brennan and Merkl-Davies (2013) considered communication choices and devices which contribute to the phenomena of impression management, where individuals or companies use language to present themselves favourably to others.

3. Dataset

In our work we focus on UK and Portuguese annual reports for large firms listed on the stock exchange market in each country. The number of UK annual reports exceeds 10,000 of mostly UK non-financial firms listed on the London Stock Exchange. The annual reports cover a period of years in the range 2003 and 2014. The extraction methods have been tested and evaluated on English annual reports and were later adapted to work with other languages. We collected 627 Portuguese annual for 77 firms for the period for the period 2006-2015. All firms are listed on the Portuguese Stock Exchange. The annual reports were collected automatically from Perfect Information[2].

3.1. Description of Dataset

We first start with explaining an annual report is. An annual report is an analysis and assessment of the financial trend of the business over the past year. An annual report consists of a description of the accounting activities seen within the report. For example, a description of the principles used for determining the accounting items in both the income statement and the balance sheet. An annual report could also include information on the events that have influenced the company's accounting throughout the year, a statement from management showing an accurate picture of the company's economic standing and development, and an auditor's report.
It was not until legislation was enacted after the stock market crash in 1929 that the annual report became a regular component of corporate financial reporting. Typically, an annual report will contain the following sections:

- Financial Highlights
- Letter to the Shareholders
- Narrative Text, Graphics and Photos
- Management's Discussion and Analysis
- Financial Statements
- Notes to Financial Statements
- Auditor's Report
- Summary Financial Data
- Corporate Information

Most of the published annual reports are in PDF file format. The different variation of annual reports' formatting makes it difficult to automatically extract relevant information or even detect the report's structure. The annual reports vary in respect to their style and number of pages. In the US firms are required to disclose their annual reports by following and filling a preset template by the US Securities and Exchange Commission (SEC). This allows a standard structure to be followed by each company making it easy to extract information and easily detect structure. In contrast to the US, stock exchange-listed firms in UK and Portugal do not present their financial information and accompanying narratives in a standardised format when creating annual reports. Firms in the aforementioned countries have much more discretion regarding the structure and content of the annual report. Added to this is the problem of nomenclature: no standardised naming convention exists for different sections in UK annual reports so that even firms adopting the same underlying structure and content may use different terminology to describe the same section(s).
Table 3.1. shows the dataset size in words in addition to the number of reports for each language.

Language	Reports	Words
English (UK)	11,009	300M
Portuguese	396	7.50M

Table 1: Dataset Size

4. Extraction Methods

To extract information from our dataset of PDF annual reports we used Information Extraction (IE) and Natural Language Processing (NLP) methods and techniques. This helps in extracting sections and their narratives. The methods automatically detect the annual report's table of contents, synchronise page numbers in the native report with page numbers in the corresponding PDF file, and then use the synchronised page numbers to retrieve the textual content (narratives) for each header (hereinafter section) listed in the table of contents. The extraction methods rely on the table of contents by using section heading presented in the table of content to partition into the audited financial statements component of the report and the "front-end" narratives component, with the latter sub-classified further into a set of generic report elements including the letter to shareholders, management commentary, the governance statement, the remuneration report, and residual content.

[2]http://www.perfectinfo.com

4.1. Structure Extraction Process

This section explains in details the steps and process needed to be able to detect the structure of the PDF annual reports of both UK and Portuguese datasets. The process was first applied to the 10,000 UK annual reports (Section 4.1.), we then applied the same process to the smaller Portuguese dataset.

As mentioned in Section none of the UK or Portuguese annual reports follow a standard reporting template as in the US Stock Exchange. Firms and management in the UK have more discretion when it comes to the the format, structure and the contents of the annual reports. On the other hand the US Securities and Exchange Commission forces firms to follow a standard format and a pre–labeled annual reports template which they publish in HTML file format. This has helped in creating a reporting standard making it easy for investors, firms and analysts to access and acquire information automatically from a bulk of annual reports. This is different in the UK where firms tend to publish their annual reports in PDF file format. Despite being cross-platform and a portable file format it is deemed a difficult task to automatically extract information from PDF annual reports since companies' reports vary significantly especially when it comes to the contents and the section headers. In order to automatically analyse a large dataset of UK annual reports we first needed to automatically detect the structure of the PDF annual reports so we can extract the information needed.

To detect and extract the structure of the annual reports each PDF file goes through the following five steps: 1) detecting the contents-page, 2) parsing the detected contents-page and extracting the sections, 3) detecting page numbering, 4) adding the extracted sections to the annual report PDFs as bookmarks, and 5) using the added bookmarks to extract the narrative sections under each heading.

4.1.1. Detecting the Contents Page

An annual report contents page includes information about the main sections of the report and its associated page numbers. Information in the contents page helped us detect the structure of the annual reports. However, detecting the contents page was not a straightforward task. We created a list of gold–standard section names extracted manually from the contents page of a random sample of 50 annual reports. We filtered the gold–standard keywords by removing duplicates and preserving the structure of how they appeared in the annual reports. We matched each page in the annual report against the list of section names in gold–standard, then we selected the page with the highest matching score as the *potential* contents page. The score was calculated by an increment of 1 for each match. To improve the matching process and avoid false positives, we match the gold–standard keywords against lines of text that follow a contents–page–like style (e.g. a section name followed by a page number, such as "Governance Report 22").

4.1.2. Parsing the Contents Page

In order to get the structure of the annual report we automatically parse the selected contents page by extracting the name of each section and its associated page number. To do this we matched each line of text in the selected contents page against a regular expression commands that will extract any line starting or ending with a number between 1 and the number of pages of the annual report.

We built a simple filtering tool that filters out any block of text that matches our regular expression commands. This is done by removing text containing addresses, dates, and postal codes. The filtering tool can also detect email addresses, websites, references to branches and locations using regular expression commands and a gazetteer.

We differentiate between dates and actual page numbers to avoid extracting incorrect section headers. However, lines containing text such as an address (e.g., 23 Robert Avenue) might still be confusing for the tool. We tackled this problem by matching the list of extracted sections against a list of gold–standard section synonyms which we explain in more details in Section 4.1.5..

The structure of the PDF files makes it difficult to extract text in its actual format. Extracting plain text from PDFs results in many line breaks being added in between the text. This makes extracting a section name that is split into two lines a difficult task. To tackle the problem of broken sections (i.e., appearing on two lines or more), we implemented an algorithm to detect broken section headers and fix them by concatenating lines that end or begin with prepositions such as 'of', 'in' ...etc. The algorithm also concatenates sentences ending with singular or plural possessives, symbolic and textual connectors (e.g. 'and', 'or', '&'...etc), and sentences ending with hyphenations. This method was also adapted to Portuguese prepositions and other stop-words needed to concatenate lines of text by forming a list of most common stop-words for each language.

4.1.3. Detecting Page Numbering

The page numbers appearing on the contents page do not usually match with the actual page numbers in the PDF files. For example, page 4 in the annual report could refer to page 6 in the PDF file, which may lead to incorrect extraction[3]. We address this problem by creating a page detection tool that crawls through annual report pages taking three consecutive pages in each iteration. The tool aims to extract a pattern of sequential numbers with an increment of 1 (e.g. 16, 17, 18) but with the complex structure of the PDF files this has been proven to be a difficult task. The tool starts by reading the contents three pages at a time starting from the report's number of pages minus one. For example, assume we are trying to detect the page numbering pattern for a report of 51 pages. The tool starts by extracting text from pages 48, 49 and 50. A regular expression command is then used to extract all the numbers in each page contents that is made up of maximum three digits creating a vector of numbers for each page. Figure 1 shows a sample of 3 vectors for the pages 48, 49 and 50. As shown in Figure 1 the algorithm will only keep numbers that are within a range of 10 pages those linked with small double arrows. The algorithm will then try to form a pattern of sequential

[3]The algorithm responsible for extraction of sections uses start and end page numbers to locate the text and therefore accurate page numbers are required.

numbers with an increment of 1. Figure 1 shows that the pattern 49, 50 and 51 (dark circles) has been found which is equivalent to a one page difference (*page–increment*) between the reports page numbering and those found in the PDF file. The tool will repeat the same process for all the pages in the annual report until it reaches pages 1, 2 and 3 where it stops.

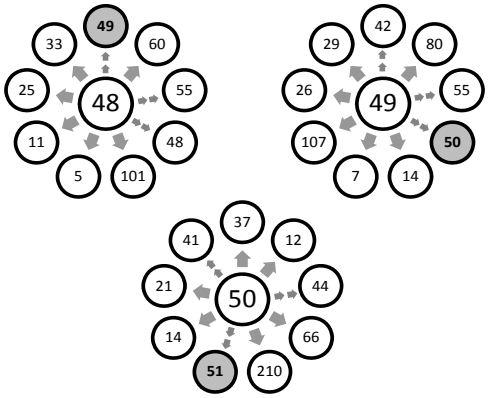

Figure 1: Detecting Page Numbering

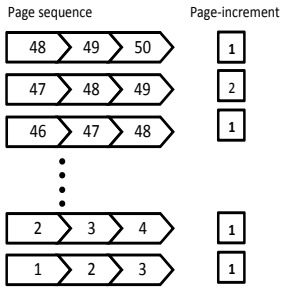

Figure 2: Popular Page Increment

As shown in Figure 2 for each 3 vectors the tool will store the page–increment in an array of numbers and at the end of the process the most popular (most frequent) page–increment will be selected as the difference between the annual report and the PDF numbering.
This process on the sample yielded an accuracy rate of more than 95%. Manual examination of the remaining less than 5% revealed the following reasons for non-detection:

- Encoding Error, unrecognised text
- Images or empty pages interrupting the sequence of pages
- Page numbers appeared on even or odd pages only
- Unusual numbering format (e.g. "001001001029" refers to page 29).
- Page numbers appeared in a written format (e.g. Twenty One)
- Page numbers restarted on each section
- Some pages had no page numbers available
- Every other page has two numbers (e.g. 26/27) with no numbers available on the next page
- Two pages on each PDF page
- Some other errors due to formatting

4.1.4. Adding Section Headers as Bookmarks

Using the sections and their correct page numbers from Sections 4.1.1. and 4.1.3. we implemented a tool to insert the extracted contents page sections as bookmarks (hyperlinks) to sample PDFs. This process helped in extracting narratives associated with each section for further processing (see Section 4.1.5. below).

4.1.5. Extracting Sections' Narratives

We implemented an automatic extraction algorithm to crawl through the data collection and, for each PDF file, extract all inserted bookmarks and their associated pages. Since UK firms do not follow a standard format when creating annual reports, a long list of synonyms are possible for a single section header. For example the section header "Chairman's Statement" may also appear as "Chairman's Introduction", "Chairman's Report" or "Letter to Shareholders". The same case applies to Portuguese as well. To solve this problem, we semi-automatically and by the help of experts in accounting and finance, created a list of synonyms for each of the generic annual report sections (see the list below). This was done by extracting all sections containing "Chairman", "Introduction", "Statement", "Letter to"...etc from a sample of 250 annual reports of 50 UK firms (the quoted unigrams were selected by the same experts). We refined the list by removing redundancies. The accounting experts then manually examined the list and deleted irrelevant or incorrect sections. We used the refined list as gold–standard synonyms to extract all the sections related to each of our generic sections (e.g. all sections about the "Chairman's Statement"). To overcome the problem of different word–order or additional words included in the headline (e.g. "The Statement of the Chairman"), we used *Levenshtein Distance* string metric algorithm (Levenshtein, 1966) to measure the difference between two sections. The Levenshtein distance between two words is the minimum number of single-character edits (insertion, deletion, substitution) required to change one word into the other. To work on a sentence level we modified the algorithm to deal with words instead of characters. All the sections with a Levenshtein distance of up to five were presented to the accounting expert.
We used the above process to create gold–standard synonym lists for the following 8 generic section headers that we wished to extract for further analysis:

1. Chairman Statement
2. CEO Review
3. Corporate Government Report
4. Directors Remuneration Report

5. Business Review
6. Financial Review
7. Operating Review
8. Highlights

Having detected and extracted section headers (or their gold–standard synonyms) and their sections, we then extract the sections' narratives using iText[4], an open source library to manipulate and create PDF documents (Lowagie, 2010), to apply our text analysis metrics, which include readability measurement and counting word frequencies using financial domain hand–crafted word lists.

5. Extraction Tools

We used the extraction methods described in Section 4. to create publicly available web and desktop tools for users to automatically and freely analyse annual reports in different languages. The tools deal with multilingual annual reports of firms within the UK and Portugal written in either English or Portuguese and distributed in PDF file format[5].
The tool is called CFIE-FRSE standing for Corporate Financial Information Environment (CFIE) -Final Report Structure Extractor (FRSE). The tool is available as a web application[6] or as desktop application, which is freely available on GitHub[7]. The tools detect the structure of annual reports by detecting the key sections, their start and end pages in addition to the narrative contents. This works for both languages. The tools provide further analysis for reports written in English such as readability metrics, section classification and tone scores. This is because the tool was built to analyse UK annual reports where we have a large dataset to train the system to provide an extra level of analysis.
The extra level of analysis will be made available for Portuguese at a later stage. For now we do not have enough reports for Portuguese to be able to train the system. As explained earlier the aim of this paper is to show that our extraction methods can be applied to a second language, a vital step towards fully analysing reports in other languages in the future.

6. Multilingual Extraction

In this section we explain the process we followed to extracting sections from annual reports in both English and Portuguese.

6.1. English

As mentioned earlier the work was first designed to analyse UK English annual reports (El-Haj et al., 2014). We automatically harvested more than 10,000 annual reports for firms listed on the London Stock Exchange (LSE). Prior to analysing the annual reports we first worked on sorting them by firm and we created our own unique report identifier which we called "LANCS_ID". Sorting annual reports was done semi-automatically were we used a Java tool to match firm names and extract the reports' years. This was followed by manual post editing to make sure the matching was correct. Firms without a match could be firms that do not exist anymore or firms with a new name due to merging with another firm, those had to be manually matched. PDF filenames do not contain a unique firm identifier. For example, reports collected from Perfect Information 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 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 step-by-guide to implementing the matching procedure in SAS are available at `http//cfie.lancaster.ac.uk.8443/`. Licensing restrictions prevent direct publication of proprietary identifiers.
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 UK due to the LSE's position as largest equity market by capitalisation outside the US. The extraction process is nevertheless generic insofar as reports published in other reporting regimes and languages can be analysed by modifying the language- and regime-dependent aspects of our tool without editing the underlying Java source code. Further guidance will be provided in an online appendix, together with full technical details of our method, in due course.
Table 6.1. shows the structure detection and extraction accuracy for UK annual reports.

Number of downloaded annual reports	11,009
Number of reports analysed	10,820
% of correctly retrieved table of contents	98.28
% of correctly retrieved pages	95.00
% of correctly retrieved text from sections	95.00

Table 2: UK Annual Reports Analysis

As shown in the table the tool analysed more than 98% of the downloaded annual reports. Firms management in the UK have more discretion over what, where, and how much information on topics such as risk, strategy, performance, etc. is reported, this lead the reports to vary significantly in terms of structure and design. Despite the dissimilarity between the structure of the downloaded annual report, our methods were able to accurately analyse the majority of the reports. Those failing the analysis process were due to one

[4] `http://itextpdf.com/api`
[5] For now only the Desktop version of the tool can work with multilingual annual reports
[6] `https://cfie.lancaster.ac.uk:8443`
[7] `https://github.com/drelhaj/CFIE-FRSE`

of the following reasons:

1. The file does not allow the text to be extracted (image-based documents). This problem is more common in the early years of our sample (i.e. 2000-2005), as some of the annual reports were poor quality scanned files. Reports from the more recent years tend not to be of this type.

2. Reports with a table of contents that could not be read due to the limitation imposed by how the table was designed. For example where a table of contents is designed with numbers and text in two different columns, or where the table of contents is split into two pages which causes problems for the PDF library.

3. Absence of page numbers.

6.2. Portuguese

The adaptation of our software to other languages must deal with problems that are both specific to the financial reporting environment and to the language itself. As in most countries, Portuguese market regulations allow a certain degree of flexibility in relation to the content and structure of the annual report[8] and concerning a firm's governance structure. For instance, the board of directors (or its equivalent) and the Fiscal Committee can adopt different structures and names. As an example, we detected 7 alternative titles for the CEO's message, 12 different titles for the chairman's letter and 35 alternatives for the auditor and related governance mechanisms. We believe that this will be a common problem across the different language implementations. The approach we adopted was to list all the alternatives, create a list of synonymous and assign a unique classification to each alternative. On the other hand, the language related issues are specific to each language. During the implementation of the Portuguese version, we identified several different problems. Firstly, the English language is one of the few western languages that does not use phonetic modifications of common characters, such as "À", "Á", "Ã", "Â" and "Ç". These phonetic modifications are common in other languages and can also vary across countries. Secondly, Portuguese is a gender-based language, which increases the complexity in developing a list of stop-words to deal with the line breaking. One such example is the proposition "of", which can be translated as "de", "do", "da", "dos" and "das", depending on the gender of the following word. Thirdly, Portugal signed the Portuguese Language Orthographic Agreement of 1990[9]. This agreement changed the spelling of some words (e.g. the word "Accionistas", is now spelt "Acionistas",). It also allowed an alternative spelling for some words (e.g. the word "Sector" can also be spelt as "Setor",). During the transition period, which ended in 2015, the adoption of the new spelling was voluntary and different firms used different spelling variations for some words. As a result, the algorithm must recognise all spelling variations. To test the adaption of the software to Portuguese Annual Reports, we retrieved from Perfect Information all annual reports published in Portuguese by firms listed on Euronext Lisbon for the period 2006 to 2015, totaling 627 reports for 77 firms (Table 6.2.).

Year	2005	2006	2007
# Downloaded Reports	51	52	60
# Processed	23	26	38
%	%45	%5	%63
Year	2008	2009	2010
# Downloaded Reports	61	61	62
# Processed	37	38	40
%	%61	%62	%65
Year	2011	2012	2013
# Downloaded Reports	64	62	58
# Processed	44	43	42
%	%69	%69	%72
Year	2014	2015	Total
# Downloaded Reports	59	37	627
# Processed	36	29	396
%	%61	%78	%63

Table 3: Number of Reports Per Year

The software was able to process 396 reports (63%) of the annual reports. We then focused on understanding the reason for the non-processed reports and the accuracy of the processed reports. The software failed to process 231 reports:

- Table of contents does not exist: 62 reports
- Table of contents could not be detected: 52 reports
- Table of contents presented in an unconventional format: 45 reports
- Table of contents without page numbers: 39 reports
- Table of contents with more than one page: 12 reports
- Image based file: 21 reports

The adaptation to Portuguese was based on 2 steps. We started by listing all table of contents entries for 67 randomly selected annual reports. This procedure produced a list of 2,053 entries that, after cleaning for errors and minor differences, included 694 different table of contents entries.

[8] The Companies Act (Código das Sociedades Comerciais) and Portuguese market regulations require a firm's Annual Report to include, amongst other items a review of the firm's activities, performance and financial position, a description of the main risks and uncertainties, financial risk management goals and policies, including details of hedging operations and risk exposures, a description of subsequent events, the expected evolution of the firm and a proposed net income allocation and dividends. In addition, firms are required to submit a Corporate Governance Report. Firms can opt to include this report in the Annual Report or to submit a separate document. Disclosure requirements are summarised in Circular sobre Contas Anuais – 9th February 2017.

[9] An international agreement aiming at the creation of a unified orthography for the Portuguese language across all the countries with Portuguese as their official language.

This variety reflects the lack of standardisation of the structure of the annual report that is common to most countries. To deal with this problem, for the second step, we assigned each entry to a pre-defined section (Chairman, CEO, Performance, Auditor, Financial Statements and Other), which reflects the common structure of a Portuguese annual report at a very basic level. We also tested the accuracy of the adaptation to Portuguese by manually checking 100 annual reports processed and we concluded that the software performs with an accuracy comparable to the English implementation (Tables 6.1. and 6.2.).

7. Conclusion

The methods reported in this paper demonstrate the adaptability of our extraction and classification procedures to non-English annual reports published in regulatory settings other than the UK, and we examine Portuguese reports in this paper. This adaptation was achieved by developing methods that are language independent where the extraction process relies on the structure of the annual reports rather than the deep language characteristics. The methods will still require dictionaries and word-lists to be in the same language as the annual reports but the extraction process remains the same across languages. The reported work paves the way for investors, firms and analysts to access and acquire information automatically from a large volume of annual reports in languages other than English.

8. Acknowledgements

We acknowledge support for this research in three projects. The first Corporate Financial Information Environment (CFIE) project was funded (2012-14) by the Economic and Social Research Council (ESRC) (reference ES/J012394/1) and The Institute of Chartered Accountants in England and Wales (ICAEW). This work also continued in the Understanding Corporate Communications sub-project funded as part of the ESRC Centre for Corpus Approaches to Social Science (CASS) (reference ES/K002155/1). Most recently, the research is funded under the new project which started in January 2018 Analysing Narrative Aspects of UK Preliminary Earnings Announcements and Annual Reports (reference ES/R003904/1).

9. Bibliographical References

Brennan, N. M. and Merkl-Davies, D. M. (2013). Accounting narratives and impression management. In *The Routledge Companion to Communication in Accounting*.

Devitt, A. and Ahmad, K. (2007). Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 984–991.

Doucet, A., Kazai, G., Dresevic, B., Uzelac, A., Radakovic, B., and Todic, N. (2009). Icdar 2009 book structure extraction competition. In *Proceedings of the Tenth International Conference on Document Analysis and Recognition (ICDAR'2009)*, pages 1408–1412, Barcelona, Spain, July.

El-Haj, M., Rayson, P., Young, S., and Walker, M. (2014). Detecting document structure in a very large corpus of UK financial reports. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, pages 1335–1338.

Ferreira, J. Z., Rodrigues, J., Cristo, M., and de Oliveira, D. F. (2014). Multi-entity polarity analysis in financial documents. In *Proceedings of the 20th Brazilian Symposium on Multimedia and the Web*, WebMedia '14, pages 115–122, New York, NY, USA. ACM.

Im, T. L., San, P. W., On, C. K., Alfred, R., and Anthony, P. (2013). Analysing market sentiment in financial news using lexical approach. In *Open Systems (ICOS), 2013 IEEE Conference on*, pages 145–149, Dec.

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

Li, F. (2010). The information content of forward-looking statements in corporate filings—a naïve bayesian machine learning approach. *Journal of Accounting Research*, 48(5):1049–1102.

Loughran, T. and McDonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65.

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

Lowagie, B. (2010). *iText in Action*. Covers iText 5. Manning Publications Company.

McConnaughey, L., Dai, J., and Bamman, D. (2017). The labeled segmentation of printed books. In *Proceedings of the EMNLP 2017 conference*.

Merkl-Davies, D. and Koller, V. (2012). 'Metaphoring' people out of this world: a critical discourse analysis of a chairman's statement of a UK defence firm. *Accounting Forum*, 36(3):178–193, 9.

Neuenschwander, B., Pereira, A. C., Meira, W., and Barbosa, D. (2014). Sentiment analysis for streams of web data: A case study of Brazilian financial markets. In *Proceedings of the 20th Brazilian Symposium on Multimedia and the Web*, WebMedia '14, pages 167–170, New York, NY, USA. ACM.

Schleicher, T. and Walker, M. (2010). Bias in the tone of forward-looking narratives. *Accounting and Business Research*, 40(4):371–390.

Schumaker, R. P. (2010). An analysis of verbs in financial news articles and their impact on stock price. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, WSA '10, pages 3–4, Stroudsburg, PA, USA. ACL.

Teufel, S. (2010). *The structure of scientific articles: Applications to Citation Indexing and Summarization*. CSLI Studies in Computational Linguistics. Center for the Study of Language and Information, Stanford, California.