

Identifying Financial Risk through Natural Language Processing of Company Annual Reports

Mini-dissertation by

Jacques Lamont Theron

(u19333634)

Submitted in partial fulfillment of the requirements for the degree
Master in Information Technology (Big Data Science)
in the Faculty of Engineering, Built Environment and Information Technology /
School of Information Technology
University of Pretoria, Pretoria

October 2020

Supervisor : Dr. V. N. V. Marivate
Department : Department of Computer Science
Degree : Master in Information Technology

Publication data:

Jacques Lamont Theron. Identifying Financial Risk through Natural Language Processing of Company Annual Reports.
Master dissertation, University of Pretoria, Department of Computer Science, South Africa, Pretoria, October 2020.

Electronic, hyperlinked versions of this thesis are available online, as Adobe PDF files, at:

<http://upetd.up.ac.za/UPeTD.htm>

Identifying Financial Risk through Natural Language Processing of Company Annual Reports

by

Jacques Lamont Theron

E-mail: u19333634@tuks.co.za

Abstract

A pipeline was developed to source annual reports of South African banks and convert them into a novel corpus. Plain text was extracted from unstructured reports whilst maintaining lineage to its coordinates in the original Portable Document Format (PDF). Initial experiments with Natural Language Processing (NLP) and machine learning classification aim at exposing financial risk inherent in the text as opposed to analysing the numerical financial values. Failed financial or governance events related to banks in the public domain were used to label annual reports as high risk. The balance of the reports were annotated as low risk to formulate a binary classification problem for machine learning. Bag of words and word embedding techniques were applied and supplemented with linguistic features like tone, uncertainty and causality based on available wordlists. Classifiers were built using traditional logistic regression and Support Vector Machine (SVM), as well as modern Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) deep learning models. The corpus and initial findings provide a baseline for further research. Applications include an early warning system for regulators as well as question answering based on the content.

Keywords: financial risk, company annual reports, natural language processing, machine learning, classification, closed domain question answering.

An interesting quotation of a risk event that negatively impacted a South African company and its stakeholders

“More than forty years ago the most daring heist in South Africa’s banking history took place. A group of industrious thieves, led by the elusive Mister Nightingale, skillfully tunneled their way some twenty-one meters underground into the vault of the Standard Bank in Krugersdorp. The bold bank robbers withdrew the princely sum of R400 000. They disappeared with the loot and no arrests were ever made.

I have, for the past five months, investigated the sorry affairs of the VBS Mutual Bank. My report will reveal that the perpetrators of the heist at VBS made away with almost R2 billion. And they certainly did not put in anything like the hard work and effort of Mister Nightingale and his team.

I trust that, in this case, arrests will be made.”

Advocate Terry Motau SC assisted by Werksmans Attorneys, The Great Bank Heist
(2018) [\[50\]](#)

Acknowledgements

Special thanks and appreciation are extended to those below for their support and contribution to the completion of the dissertation:

- Dr. V. N. V. Marivate (University of Pretoria) for his enthusiasm in natural language processing and passion to apply data science for social impact;
- Mr. D. Masiya (South African Reserve Bank) for his interest in, and sponsorship of, the research topic as a catalyst for innovating new ways of supervision;
- Mr. J. Sefara for his review comments that assisted in finalising the dissertation report;
- The Corporate Financial Information Environment (CFIE) Team at Lancaster University for their work in creating a corpus of annual reports from United Kingdom (UK) companies which inspired this work in the South African (SA) context [\[15\]](#).

Contents

List of Figures	vii
List of Algorithms	ix
List of Tables	x
List of Acronyms	xiii
1 Introduction	1
1.1 Problems and motivation	1
1.2 Research questions	2
1.3 Objectives	4
1.4 Summary of methodology	4
1.5 Contributions	5
1.6 Ethical considerations	6
1.7 Dissertation outline	7
2 Literature review	10
2.1 Macro and micro risk identification	10
2.2 Numeric and text data	11
2.3 Data set	11
2.4 Feature extraction	12
2.4.1 Content features	12
2.4.2 Linguistic features	13
2.5 Classification	14

2.6	Gaps and opportunities	15
2.7	Summary	16
3	Methodology	17
3.1	Feasibility test	19
3.2	Data collection	21
3.3	Data reference and labelling	22
3.4	Data preparation and analysis	23
3.4.1	Booklet identification	24
3.4.2	Text extraction	25
3.4.3	Word count	30
3.4.4	Wordlist counts	30
3.4.5	Readability score	32
3.4.6	Header identification	33
3.4.7	Header classification	35
3.4.8	Section aggregation	36
3.4.9	Document aggregation	37
3.4.10	Validation	37
3.4.11	Training and test data sets	37
3.4.12	Feature selection	38
3.5	Data processing and analysis	39
3.5.1	Bag of words	40
3.5.2	Word embedding	41
3.5.3	Feature scaling	41
3.5.4	Topic analysis	43
3.6	Classification and prediction	44
3.6.1	Logistic regression	44
3.6.2	Support vector machine	47
3.6.3	Long short term memory recurrent neural network model	47
3.6.4	Convolutional neural network model	50
3.6.5	Multi-layer perceptron model	53
3.7	Performance and application	53

3.7.1	Classifier prediction	53
3.7.2	Classifier interpretation	54
3.7.3	Question answering	54
3.8	Summary	55
4	Pre-processing	56
4.1	Feasibility test	56
4.2	Data collection	57
4.3	Data reference and labelling	60
4.3.1	Word count validation	60
4.3.2	Company mapping and labelling	60
4.3.3	Document mapping and labelling	62
4.4	Data preparation and analysis	64
4.4.1	Booklet identification	64
4.4.2	Text extraction	64
4.4.3	Word count	65
4.4.4	Wordlist count	65
4.4.5	Readability score	66
4.4.6	Header identification	70
4.4.7	Header classification	71
4.4.8	Section aggregation	73
4.4.9	Document aggregation	77
4.4.10	Validation	78
4.4.11	Training and test data sets	79
4.5	Data processing and analysis	79
4.5.1	Bag of words	79
4.5.2	Topic analysis	80
4.6	Summary	81
5	Modelling	83
5.1	Goals	83
5.2	Results	84

5.3	Application	88
5.4	Interpretation	89
5.5	Question answering	93
5.6	Summary	97
6	Discussion	98
6.1	Pre-processing	98
6.1.1	Feasibility test	98
6.1.2	Data collection	99
6.1.3	Data labelling	99
6.1.4	Data preparation	100
6.2	Modelling	101
6.3	Summary	103
7	Conclusion	104
7.1	Summary of Conclusion	104
7.2	Future work	105
7.2.1	Pre-processing	105
7.2.2	Modelling	107
	Bibliography	109
	Appendices	117
A	Exploratory Data Analysis	118
A.1	Data profiling	118
A.1.1	Word count and pages	120
A.1.2	Uncertain and causal word counts	122
A.1.3	Performance, strategy and forward-looking word counts	126
A.2	Feature importance	130
A.2.1	Chi-Squared	130
A.2.2	Information Gain	133
A.2.3	ANOVA	133

A.3	Topic analysis	134
A.3.1	2016 annual report	135
A.4	Visualisation	136
A.4.1	Principal Component Analysis	137
A.4.2	t-distributed Stochastic Neighbour Embedding	138
A.5	Summary	140
B	Supporting information	141
B.1	Feasibility test	141
B.2	Data collection	148
B.3	Data preparation	149
B.3.1	Text extraction	149
B.3.2	Readability score	151
B.3.3	Header identification	154
B.3.4	Section aggregation	158
B.3.5	Document aggregation	159
B.3.6	Validation	160
B.3.7	Training and test data sets	162
B.3.8	Bag of words	165
B.4	Classifier performance	165
B.4.1	Grid search on TFIDF	165
B.4.2	TF and TFIDF compared	168
B.4.3	Data sets compared	169
B.4.4	N-gram comparison	169
B.4.5	Word normalisation compared	171
B.4.6	Linguistic features compared	172
B.4.7	Using embedding layer of LSTM and CNN models	172
B.5	Question answering	177
B.6	Summary	178
C	Symbols	179

List of Figures

3.1	Methodology overview	18
3.2	Feasibility test overview	19
3.3	Data collection overview	21
3.4	Reference data overview	22
3.5	Data preparation methods	23
3.6	Text extraction for <i>pageRatio=0</i>)	27
3.7	Text extraction for <i>pageRatio=1</i>)	28
3.8	Text extraction for <i>pageRatio=2</i>)	29
3.9	Data processing methods	39
4.1	Unigram cloud (VBS Bank Audit committee report 2016)	80
4.2	2- and 3-gram clouds (VBS Bank Audit committee report 2016)	81
5.1	Performance on additional linguistic features	87
5.2	SP-LIME explanation of Africa Bank 2014 report	90
5.3	SP-LIME explanation of VBS Mutual Bank 2016 report	91
5.4	SP-LIME explanation of Africa Bank 2018 report	92
5.5	SP-LIME explanation of Africa Bank 2019 report	93
A.1	Word count histogram	120
A.2	Uncertain word count scaled histogram	124
A.3	Causal word count scaled histogram	124
A.4	CausalM word count scaled histogram	125
A.5	Performance word count scaled histogram	127
A.6	Strategic word count scaled histogram	129

A.7	Forward-looking word count scaled histogram	130
A.8	Chi-Squared on TFIDFs and select counts and indices	131
A.9	Chi-Squared on TFIDFs and all counts and indices	132
A.10	Information Gain on all counts and indices	133
A.11	Information Gain on all counts and indices	134
A.12	4- and 5-gram clouds for one topic (VBS Audit committee 2016)	135
A.13	4- and 5-gram clouds for two topics (VBS Audit committee 2016)	136
A.14	PCA for the first 1000 words per class)	137
A.15	t-SNE for the first 1000 words per class)	138
A.16	t-SNE for 80 TFIDF tokens per class)	139
B.1	VBS 2015 report bookmarks	146
B.2	VBS 2016 report booklet	147
B.3	Text extraction across columns	149
B.4	Journey layout of contents	156
B.5	Contents for subset of report	157
B.6	'Out-of-context' text within section	158
B.7	Foreign language in Canara Bank annual report	159
B.8	Warning on opening Canara Bank reports	160
B.9	Performance on limited TFIDFs (normalised)	166
B.10	Performance on limited TFIDFs (standardised)	167
B.11	Performance on 40 unigrams (normalised)	168
B.12	Performance on 40 unigrams (standardised)	168
B.13	Performance across data sets using 40 TFIDF (normalised)	169
B.14	Performance across TFIDF n-grams (normalised)	170
B.15	SVMa performance for word normalisation and embedding	171
B.16	Performance on limited TFIDFs plus word counts (normalised)	172
B.17	cdQA filter paragraph impact	177

List of Algorithms

3.1	Page Reading Algorithm	26
3.2	Header Identification Algorithm	34

List of Tables

3.1	PDF page ratios	24
3.2	Page ratio example	25
3.3	Text extraction flags	26
3.4	Wordlists	31
3.5	Readability metrics	32
3.6	Header types	35
3.7	Tone variables	36
3.8	Regular expression patterns	40
3.9	Model parameters	44
3.10	Word count features	45
3.11	Feature parameters	47
3.12	LSTM models	48
3.13	LSTM model parameters	49
3.14	CNN models	50
3.15	CNN1 model parameters	51
3.16	CNN2 model parameters	52
3.17	CNN3 model parameters	52
3.18	MLP model parameters	53
4.1	Annual reports downloaded	58
4.2	Company mapping file	61
4.3	Document mapping file	63
4.4	Document count by page ratios	64
4.5	Top 5 hard reading reports (r_gf)	66

4.6	Top 5 hard reading reports (ts_gf)	67
4.7	Top 5 hard reading reports (r_f)	67
4.8	Top 5 hard reading reports (ts_f)	68
4.9	Top 5 hard reading reports (ts_ts)	68
4.10	Top 5 easy reading reports (ts_ts)	69
4.11	Header text classification	71
4.12	Header type frequency	72
4.13	Tone by section	74
4.14	Scaled counts (UK reference)	75
4.15	Scaled counts (VBS Mutual Bank 2015)	76
4.16	Scaled counts (VBS Mutual Bank 2016)	76
4.17	Tone for selected SA reports	77
4.18	Scaled counts for selected SA reports	78
4.19	Most common unigrams, bigrams and trigrams	79
5.1	Recursive Feature Elimination	86
5.2	SVMa top 5 risk reports	88
5.3	Logistic Regression top 5 risk reports	88
5.4	Example question and answers on reports	94
A.1	Feature correlation	119
A.2	Top 5 documents with most wordCount	121
A.3	Top 5 documents with least wordCount	121
A.4	Top 5 documents with most pages	122
A.5	Top 5 documents with least pages	122
A.6	Top 5 documents with most uncertainty	123
A.7	Top 5 documents with least uncertainty	123
A.8	Top 5 documents with most causal words	125
A.9	Top 5 documents with least causal words	125
A.10	Top 5 documents with most causalM words	126
A.11	Top 5 documents with least causalM words	126
A.12	Top 5 documents with most performance words	126

A.13 Top 5 documents with least performance words	127
A.14 Top 5 documents with most strategic words	128
A.15 Top 5 documents with least strategic words	128
A.16 Top 5 documents with most forward-looking words	129
A.17 Top 5 documents with least forward-looking words	129
B.1 UK report header classification	142
B.2 UK report header start and end identifiers	143
B.3 UK report header word counts	145
B.4 Multi-document annual reports	148
B.5 Text extraction flags	150
B.6 Text block extraction warnings	150
B.7 Readability scores validated	151
B.8 Readability scores by textstat	152
B.9 Readability scores compared	152
B.10 Non-header text bookmarks	154
B.11 Header counts for tables of content	154
B.12 Word count validation	160
B.13 Causal wordlist count validation	161
B.14 Test data set	163
B.15 Training data set positives	165
B.16 LSTM training	173
B.17 LSTM training performance	173
B.18 LSTM test performance	174
B.19 LSTM1 confusion matrix	174
B.20 CNN training	174
B.21 CNN training performance	175
B.22 CNN test performance	176

List of Acronyms

Acronyms used in this dissertation report are described in alphabetical order.

BoW	Bag of Words.
CBOW	Continuous Bag of Words.
cdQA	closed domain Question Answering.
CFIE	Corporate Financial Information Environment.
CFIE–FRSE	Corporate Financial Information Environment - Final Report Structure Extractor.
CFS	Correlation-based Feature Selection.
CNN	Convolutional Neural Network.
CPU	Central Processing Unit.
CV	cross validation.
EDA	Exploratory Data Analysis.
FiR	Fisher Ratio.
GPU	Graphics Processing Unit.
JSE	Johannesburg Stock Exchange.
LDA	Latent Dirichlet Allocation.
LIME	Local Interpretable Model-agnostic Explanations.
LIWC	Inquiry and Word Count.
LR	Logistic Regression.
LSTM	Long Short-Term Memory.

MD&A	Management Discussion & Analysis.
MLP	Multi-Layer Perceptron.
NB	Naive Bayes.
NLP	Natural Language Processing.
NLTK	Natural Language Toolkit.
PDF	Portable Document Format.
POS	Part of Speech.
RBF	Radial Basis Function.
RFE	Recursive Feature Elimination.
RNN	Recurrent Neural Network.
SA	South Africa.
SG	Skip Gram.
SP-LIME	Sub modular Pick LIME.
SVM	Support Vector Machine.
TF	Term Frequency.
TFIDF	Term Frequency Inverse Document Frequency.
TP	true positive.
TPU	Tensor Processing Unit.
UK	United Kingdom.
US	United States.

Chapter 1

Introduction

An electronic engineer fulfilling an enterprise architecture function for a financial regulator and studying toward a master degree in big data science is testament to the multi-disciplinary nature of the modern workplace and the need for lifelong learning. Likewise, inspiration for this research emanated from two departments at the University of Pretoria - Electrical, Electronic & Computer Engineering (EECE) for a presentation on the application of big data science techniques for early warning of a looming banking crisis and Computer Science for NLP approaches to understanding fake news.

Previous work on risk identification has focused on quantitative analysis of financial statements and not the text within annual reports [55, 38, 10]. Of the papers applying computational linguistics to annual reports, more than 80% have focused solely on United States (US) companies [17]. These reports are more structured and contain less freestyle text than those of non-US companies, like the United Kingdom (UK) and South Africa (SA). This research is toward closing this gap in qualitative analysis with a view to better understand SA financial institutions and their risk profile.

1.1 Problems and motivation

Various stakeholders have a vested interest in the financial wellbeing of a company including shareholders, investors, customers and regulators. The annual report published

by these companies is a critical input to the decisions that these stakeholders make, especially for listed companies trading on exchanges like the Johannesburg Stock Exchange (JSE). Shareholders may wish to invest or disinvest based on prior or projected performance. Prudential regulators continuously supervise financial institutions to assess their risk and compliance with their license conditions. A company annual report can augment other statutory submissions upon which a risk assessment is made.

VBS Mutual Bank [50] and Steinhoff International Holdings [60] are just two recent high profile examples of commercial fraud resulting in billions of rand loss to the South African economy. Both had financial statements and earnings reports involving disinformation, misinformation or non-disclosure. Had inconsistencies been detected earlier, then the extent of damages could have been curtailed for these companies, their shareholders and customers.

Regulators, analysts and researchers have traditionally focused on quantitative analysis of company annual reports or they have focused on the financial statements and not the overarching narrative. The mundane and onerous nature of manually reading, reviewing and analysing text within these reports may be a contributing factor. Yet financial statements alone are insufficient predictors of risk. As an example, some business of VBS Mutual Bank was deemed fraudulent even though its 2017 financial statements had been audited.

Despite much corporate information being disclosed qualitatively, there are few examples of text within annual reports being studied [23]. Also, unlike the US and the UK, there is also no collection of SA annual reports that can serve as a research corpus.

1.2 Research questions

Given the relative under-evaluation of text within annual reports and the dire consequences of company financial failure, the main research question was formulated as:

Can risk be identified in unstructured company annual reports using Natural Language Processing (NLP)?

If NLP can be used to identify risk, it may serve as an early warning for regulators to investigate further or as an additional input into their professional judgement of the situation. Three subordinate questions were considered for investigation that would contribute to addressing the main research question. Firstly, to understand:

Can text be extracted from annual reports of South African companies?

With no annual report data set available for South African companies, the number of documents to be collected and the amount of text extracted was constrained by the time available for the research. The alternative was to use the data set of UK annual reports sponsored by CFIE which was aggregated by report section and not by company [18]. Assuming that appropriate annual report text would be accessible for the study, a subsequent research question was:

Which features can be extracted from annual reports to distinguish companies at risk?

The null hypothesis for this study was that companies who deliberately set out to mislead or manipulate have left markers in the narrative that distinguish them from the rest. Machine learning models were trained to identify these patterns and use them as predictors of risk. This led to a related research question:

Which classifiers perform best at identifying risk using the text features of annual reports?

Traditional approaches require careful feature engineering in order to train and tune classifiers based on Logistic Regression (LR) and Support Vector Machine (SVM) where as modern algorithms like Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) have demonstrated the ability to learn representations that previously had to be engineered as features. Despite the lack of interpretability of these models, there is still utility for regulators if it has superior classification performance. It will allow regulators to investigate these potential risks, thus optimising the use of their resources. Both traditional and more modern approaches to classification were investigated.

1.3 Objectives

Considering the problems and research questions discussed, the aim of the study was to develop a pipeline for reading, processing, labelling and classifying South African annual reports as high risk. Achievement of the goal would result in a foundation that can be expanded upon and used to build insightful and stable classifiers for real-world application.

To this end, specific objectives of this research were identified as follows:

- Create a text corpus of annual reports from 2009 and 2018 of South African companies;
- Label a subset of the reports as high risk;
- Predict financial risk through NLP of the reports;
- Determine the best performing classifiers of reports from companies at risk;
- Identify features of reports that distinguish companies at risk;
- Identify topics in annual reports of labelled companies that are indicators of risk;
- Demonstrate potential applications of the data and algorithms developed through this study.

1.4 Summary of methodology

The feasibility of reusing the Corporate Financial Information Environment - Final Report Structure Extractor (CFIE-FRSE) tool was assessed upfront. The tool was viable if it could process South African annual reports completely and accurately. It was also feasible if the source code could be obtained from the developers and adapted. Failing this, Python code was developed utilising a generic PDF extractor and performing page and word counts.

A single source of South African company annual reports was not found leading to the onerous task of sourcing annual reports by company, by year through internet

searches. The banking sector was prioritised and annual reports downloaded within a 11-year window from 2009 to 2019. Reports were mapped to companies, years and a binary risk flag. The first method identified headers such that text could be extracted by section. The second method extracted text by page and then aggregated by section or by document. Wordlists were adapted for regular expression searches and fuzzy matching. Keywords for each wordlist were counted to extract linguistic features like uncertainty, causality and negativity. A UK annual report was processed so that word counts per section and per report could be validated against the output of the CFIE–FRSE tool.

Pre-processing converted text to lower case. Stemming and lemmatization created five alternative word normalisation features. Additional readability features were calculated using two Python packages. Risk documents in the training data set were resampled to address the class imbalance and allow cross validation. Test data was held out for evaluating the generalisation performance of the classifiers. Term Frequency (TF) and Term Frequency Inverse Document Frequency (TFIDF) token counts created two bag of words (BoW) feature sets. Five further content features were created using word embedding and different transforms. In addition to the raw values, features were scaled using normalisation and standardisation.

Both traditional and more modern approaches to classification were compared. A LR and two SVM classifiers were trained on a bag of words to establish a performance baseline. Training on BoW increased n-grams ranges and the feature limit in a grid search manner. Modern algorithms included CNN and LSTM. Training on word vectors increased minimum word counts and vector size in a grid search manner. The F1 metric was used for five-times, five-fold cross validation. One of the LR classifiers was used in Recursive Feature Elimination to filter for the most important features, and in Sub modular Pick LIME (SP-LIME) to interpret the model. Pre-trained models were used for question answering to demonstrate another use case for the annual report corpus.

1.5 Contributions

Contributions that are novel and specific to this work include:

- Creation of a SA company annual report corpus starting with commercial banks;

- This research leverages previous work in which a corpus of annual reports of UK companies was created [18].
- Text extraction by company by page;
 - The UK corpus was split by section and concatenated by company. Identifying the start and end of narratives relating to each company was challenging given no explicit delimiters;
 - The SA corpus circumvents this challenge in extracting text by page allowing aggregation and analysis by company, report or section.
- Text extraction code;
 - Python scripts can be adapted and adopted to continue similar research.
- Text extraction for booklets;
 - The CFIE–FRSE tool used to create the UK corpus discarded PDF documents in booklet format i.e. two actual pages published as one PDF page.
- Algorithms for header identification in unstructured PDF documents.
 - The UK corpus discarded PDF documents without bookmarked contents;
 - Majority of the SA annual reports were not bookmarked. The text extraction algorithm caters for two forms of tables of content.

Items listed in this section will be made available on GitHub¹.

1.6 Ethical considerations

Privacy, property, access and accuracy were considered in the handling of information collected, analysed and inferred in this study [45].

Annual reports by their nature are not private in so far as companies publish this information to their stakeholders including shareholders for publicly listed companies.

¹For GitHub repository, see <https://github.com/dsfsi/bankrisk>.

Annual reports of banks trading in South Africa were also sourced from the public domain and not through filings. Had information internal to a regulator been used, or should it be used to continue this work, then it would have to be handled in accordance with its security classification. This information may be watch lists, risk ratings, financial institution board packs or self assessments. Confidential or secret information cannot be disclosed.

As the annual report is property of the company concerned, its own corporate website was used as a primary source of the annual report. If it could not be accessed directly from the company then public search results were investigated. Reports were downloaded from online share trading and investment portals that were not subject to payment. Proprietary information was limited to company names, identifiers and peer groupings.

Since watch lists, risk assessments, risk ratings and company-confidential data were not accessible for this study, two banks placed under curatorship were used as risk events with which to label preceding annual reports with elevated risk. Should future work require confidential or secret data, then access should be subject to non-disclosure.

The accuracy with which predictions and classifications were made was curtailed by several factors. The positive risk label was attributed solely to those two risk events. Annual reports for the year in which the institution was placed under curatorship and prior years were labeled as high risk. Since curatorship intends to rehabilitate the institution (e.g. by addressing liquidity or solvency) annual reports for subsequent years were labeled as low risk. Irrespective of sound model design, the machine learning was framed by this rudimentary initial annotation. Machine generated probability or classification of risk requires validation from an authority that can exercise appropriate judgement. Considering these limitations and the need to safeguard company reputations, names other than the two curated banks were anonymised in the prediction and classification results.

1.7 Dissertation outline

Chapter 2 focuses on the most recent and relevant literature. It begins with a broad review of risk identification and compares research involving either numeric or text data.

The nature of data sets created for risk identification is considered together with features extracted and risk classification performance using these features. The chapter concludes with gaps and opportunities identified in the literature review.

The method used for addressing the problem and research gaps is covered in Chapter 3. The approach to collecting annual reports and referencing previous results is explained. Mapping of reports to companies and years as well as risk to companies and reports allows increments of collected reports to be appended to previously processed reports. Two methods of text extraction are set out as well as techniques used for processing and analysing prior to classification. The compilation and training of machine learning models is explained inclusive of LR, SVM, LSTM and CNN.

Pre-processing outcomes are shared in Chapter 4 starting with the feasibility test results using a text and feature extraction tool from previous work [18]. Annual reports collected are quantified by company and financial year. Listings are given of companies mapped to reference data and labelled with risk. Similarly, an extract of the first ten reports mapped to companies is provided. The number of booklets within the corpus is given as it impacted the text extraction mechanism. Partial success in identifying and categorising headers allowed comparison of risk-positive VBS Mutual Bank annual reports with a UK report by section. The occurrence of forward-looking, uncertain, causal, performance and strategic keywords as well as tone were used as metrics. The comparison is repeated at document level against two risk-negative South African reports. Finally, scores are validated against the UK reference and the holdout of test data is quantified.

Modelling results are given in Chapter 5 including parameters for building the models and reproducing the results. Performance of the classifiers are compared to propose the best performing model. The primary application of the annual report corpus is demonstrated by predicting the top five risky documents. A secondary application of the corpus is then demonstrated by asking questions and applying pre-trained models to predict answers from the extracted text.

Research questions and objectives introduced earlier are discussed in Chapter 6. Conclusions drawn from this research and recommendations for further work are given in Chapter 7 to finalise the report.

For completeness, the following appendices are provided at the end of the report:

- **Appendix A** shows the results of initial Exploratory Data Analysis (EDA) profiling some of the data, testing the significance of features and plotting additional wordclouds.
- **Appendix B** contains additional information elaborating upon the results reported in Chapter 4.
- **Appendix C** lists and defines the mathematical symbols and algorithm variables used in this work, categorised according to the relevant chapter in which they appear.

The next chapter reviews literature related to the problems and research questions.

Chapter 2

Literature review

This chapter is a critical review of the literature and related work.

- Section 2.1 starts with the identification of broad macro systemic risk and then drops into the micro risk related to individual institutions;
- Section 2.2 compares numeric and qualitative research in analysing data;
- Section 2.3 reviews the nature of data sets created for risk identification;
- Section 2.4 looks at features extracted from text as predictors of risk;
- Section 2.5 notes classifier performance in detecting risk from text;
- Section 2.6 highlights gaps and opportunities from the literature review;
- Section 2.7 summarises findings of this chapter.

2.1 Macro and micro risk identification

Identification of risk can act as an early warning system for various stakeholders including regulators. Machine learning techniques to identify risk often focus on the systemic risk that a national economy or the global market collapses. Macro or systemic risk may be due to the financial performance or failure of multiple companies [13, 66, 9]. Dabrowski

et al. [13] used time-series data incorporating events like banking crises and indicators like house prices, mortgage rates and consumer price index.

There are fewer studies into micro risks i.e. the identification of risk of an individual company failing. Huang et al. [33] consider the share price and credit default swaps (CDS) in proposing indicators of a financial institution at risk. Hajek and Henriques [27] used revenue forecasts provided by capital market analysts together with the company's financial statements to detect fraud.

2.2 Numeric and text data

Majority of previous related work has the disadvantage of having focused on the quantitative financials and not including the text. Perols' [55] analysis of machine learning techniques to detect fraudulent financial statements, is that it limited the predictors to quantitative financial figures. Kotsiantis et al. [38] as well as Chen [10] also focused their predictive models solely on quantitative data.

Analysis of annual report text within the literature is in the minority compared with numeric analysis. However, advances in NLP over text mining and manual text analysis [19] has led to more papers considering explicit and latent features of the text within annual reports as indicators of risk. Both Hajek and Henriques [27] as well as Minhas and Hussain [48] could distinguish the language used by companies guilty of disinformation. Hajek and Henriques [27] found that ensemble methods were best at predicting true positives and Bayesian Belief Networks (BBN) were best at predicting true negatives. Moreover, financial statements and annual reports allude to quantitative and qualitative data, respectively.

2.3 Data set

Majority [23] of the related research has been based on the Management Discussion & Analysis (MD&A) section of the 10-K form used by US [21, 23, 48] or Canadian companies [58]. El-Haj et al. [15] sponsored by the Corporate Financial Information Environment (CFIE) created a data set of UK annual reports. These annual reports

have the advantage of being more unstructured compared with the US and a closer match to what South African companies use. The data set used by Minhas and Hussain [48] was limited to 408 annual reports and that of Fissette [21] to 1600.

2.4 Feature extraction

To understand *Which features can be extracted from annual reports to distinguish companies at risk?* the literature review revealed that previous work in similar settings used explicit content features as well as inherent linguistic features, specifically the style or choice of words.

Goel et al. [23] as well as Fissette [21] used Chi-Squared [41] independence between features and the target class to inform their forward feature stepwise selection approach. The Chi-Squared test may be more suited to categorical input. For each value of the feature, it counts the number of instances of that value as it sums across each feature value and counts the number of instances of that value in the class.

Since most content and linguistic features are converted into continuous values there may be more appropriate tests. Information gain can be used to theoretically measure mutual information between an input feature and the output class. Else, statistical tests like Correlation-based Feature Selection (CFS) and analysis of variance (ANOVA) [40] could be applied to understand the most important features. An alternative to these filter feature selection methods is to use a supervised feature selection in which the classifier is used to select feature subsets. Embedded methods use the classifier structure for feature selection whilst wrapper methods assess every combination of feature on the classifier's training performance [40].

2.4.1 Content features

El-Haj et al. [17] studied 149 finance papers using NLP and found that half looked solely at content features like word counts. Moffitt [49] chose 'lexical bundles' [6] as phrases of four to ten words, found at least 20 times in a corpus for every million words. These lexical bundles were chosen above unigrams for studying formal narratives like financial reports. The meaning of unigrams was seen to be more ambiguous than the

lexical bundles which provide more context for interpretation in a specific domain like finance. When limiting lexical bundles to those appearing in at least 15% of the class, Moffitt found that they occurred at different frequencies when comparing annual reports of fraudulent and non-fraudulent companies. Fissette’s [21] first model considered the TF and the TFIDF for solely unigrams, solely bigrams and combining unigrams with bigrams. Other content features were descriptors of report length, diversity of words and degree to which numbers were reported. For the CNN, Fissette transformed the text into word embeddings using the Word2Vec [47] model.

2.4.2 Linguistic features

Fissette’s [21] second model added linguistic features for complexity, grammar, readability and psychology. The proprietary Linguistic Inquiry and Word Count (LIWC) tool was used to identify Part of Speech (POS) for the grammatical features and categories relating to the psychological process. The conviction with which companies provide feedback in their annual report was measured using the LIWC categories for tentativeness and certainty. Eight English readability features were calculated using the Python `textstat` package. Two scores (LIX and RIX [2]) were considered as they can be used with non-English languages. The Stanford natural language parser was used to identify relations between any two words in a sentence like determiner, nominal subject and modifiers [36]. It is surprising that only bigrams improved performance and not the linguistic features.

In contrast, Goel et al. [23] improved their model accuracy by 18% in adding linguistic features to the BoW. Minhas and Hussain [48] used readability and tone. El-Haj et al. [15] constructed linguistic features from tone, uncertainty, causal and forward looking wordlists. They also computed readability indices, specifically Gunning Fog [25] and Flesch Reading Ease [22]. Equation (2.1) gives the Gunning Fog readability index:

$$gf(sentences) = ((wordsPerSentence) + (complexWordPercentage)) \times 0.4 \quad (2.1)$$

where *wordsPerSentence* : number of words divided by number of sentences;
complexWordPercentage : number of *complexWords* divided by number of words x 100;
complexWords : words with > 2 syllables excluding proper nouns, jargon and suffixes.

The index value indicated the number of years formal education required to understand the narrative. A value of 12 is representative of an 18-year-old school pupil in the US. Lower values imply that the narrative is easier to understand. The Flesch Reading Ease is given in Equation (2.2).

$$f(\text{sentences}) = 206.835 - (1.015 \times \text{wordsPerSentence}) - (84.6 \times \text{syllablesPerWord}) \quad (2.2)$$

where *syllablesPerWord* : the number of syllables divided by the number of words;
wordsPerSentence : the number of words divided by the number of sentences.

Index values of less than 10 means that the text is almost unreadable. Values less than 30 imply that understanding the text requires a college degree. Text with Flesch Reading Ease values above 60 are deemed ‘Plain English’ whilst values above 90 imply that the text is very easy to read only requiring grade 5 of US schooling to read the narrative.

Moffitt [49] used denounced earnings management practices to suggest that certain lexical bundles had loaded connotations. For example, a high frequency of ‘long lived assets and’ may suggest a company is disposing these assets as a last resort to get out of financial trouble. Companies using the phrases ‘the fair value of,’ and ‘in foreign currency exchange’ may be exploiting the subjectivity of fair value accounting over cost-based accounting to inflate earnings.

2.5 Classification

The literature review set out to uncover both traditional and modern models that have performed well in similar settings. The basic and more complex algorithms provided candidates for answering *Which classifiers perform best at identifying risk using the text features of annual reports?*

Traditional models, specifically ensemble methods, SVM, CNN and Naive Bayes (NB) have shown good results using the recall metric. Since recall can readily be achieved at the expense of precision, additional or other measures of performance should be consider. For example, precision could be used to complement the recall metric or the F1 metric

should be used to account for both recall and precision.

Goel et al. [23] achieved an F1 score of 87% using an SVM in detecting fraudulent financial statements. SVM's as well as Random Forests performed best in Minhas and Hussain's [48] evaluation.

Fissette [21] showed that a single convolutional layer CNN detected twice as many fraud cases than the NB classifier. Unfortunately, more false positives were predicted by the CNN model too. It should be noted that there was minimal tuning and that the model only ran on 40% of the data set due to system constraints and it had the disadvantage of being hard to interpret. Similar system constraints were experienced in this study.

Another modern, deep learning model used in text classification, is the Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN) [3]. These models predict a class based on a sequence of word tokens. The words can be considered sequenced over space or time. The sequences are a list of lists with integer identifiers for the words in the order in which they occur. The sequence is multiplied by a matrix in the embedding layer to effectively select columns to propagate through subsequent layers. The error is back propagated to the embedding later to adjust weights and improve classification. The LSTM model can be trained to memorise long-term relationships across the sequence of tokens [8].

Marivate et al. [43] showed that a Multi-Layer Perceptron (MLP) [5] performed well in classifying non-English news text. Headlines written in two of South Africa's official languages, Setswana and Sepedi, were categorised into 10 topics.

2.6 Gaps and opportunities

This study extends research into the use of unstructured text within individual annual reports to identify companies potentially at risk of imminent failure. The literature review revealed an opportunity to create a data set comprising annual reports of South African companies. The unique context of a developing economy and South Africa's language and cultural differences may require adjusted wordlists and may result in different predictors of risk compared with the US and the UK.

2.7 Summary

The review of literature in this chapter showed that most studies into risk identification (Section 2.1) have focused on identifying macro systemic risk with fewer studies aiming to identify risk relating to an individual company. This study bolsters work on micro risk by focusing on individual banks operating in SA.

In Section 2.2 it was evident that previous work on the annual financial statements outweigh studies of the unstructured narrative in annual reports. The approach taken with this study was to source annual reports and not stand-alone annual financial statements. It also cited advancements in NLP that has led to more papers considering features within annual reports as risk predictors, which is the focus of this study.

Section 2.3 contrasted studies based on the more structured 10-K form with data sets based on unstructured annual reports of companies outside the US. This presented an opportunity for this work to initiate a data set of annual reports from SA companies.

A limitation of the Chi-Squared text was noted in Section 2.4 and alternative tests and feature selection methods proposed. Previous use of both content and linguistic features were reviewed. Content features included n-grams, frequencies and word embeddings whilst linguistic features included frequencies relative to tokens within wordlists, Part of Speech and readability indices. Formulae were provided for Gunning Fog and Flesch Reading Ease readability indices.

Section 2.5 compared results from using traditional models like SVM with more modern approaches like CNN deep learning. The inappropriateness of recall as the sole performance measure was also discussed.

The chapter concludes with a review of gaps and opportunities in Section 2.6 identified through the literature review. It positioned this study at the intersection of financial risk identification, unstructured text, company annual reports and South Africa's unique setting. The following chapter lays out a method for addressing the problem and research gaps from collecting, preparing and processing the data to the conducting of experiments and analysis of results.

Chapter 3

Methodology

This chapter describes the approach taken to this study as illustrated in Figure [3.1](#) on the following page.

- Section [3.1](#) describes how previous work was re-used to assess feasibility;
- Section [3.2](#) explains how annual reports were collected;
- Section [3.3](#) is an approach to referencing previous results, mapping reports to companies and years as well as allocating risk to companies and reports;
- Section [3.4](#) sets out two methods of text extraction in preparing the data;
- Section [3.5](#) gives techniques used for processing and analysing the extracted data prior to classification;
- Section [3.6](#) is an explanation of the machine learning models and how they were trained for classification;
- Section [3.7](#) describes how models were assessed, applied and interpreted, and how the data was used to predict answers to questions;
- Section [3.8](#) summarises the methodology given in this chapter.

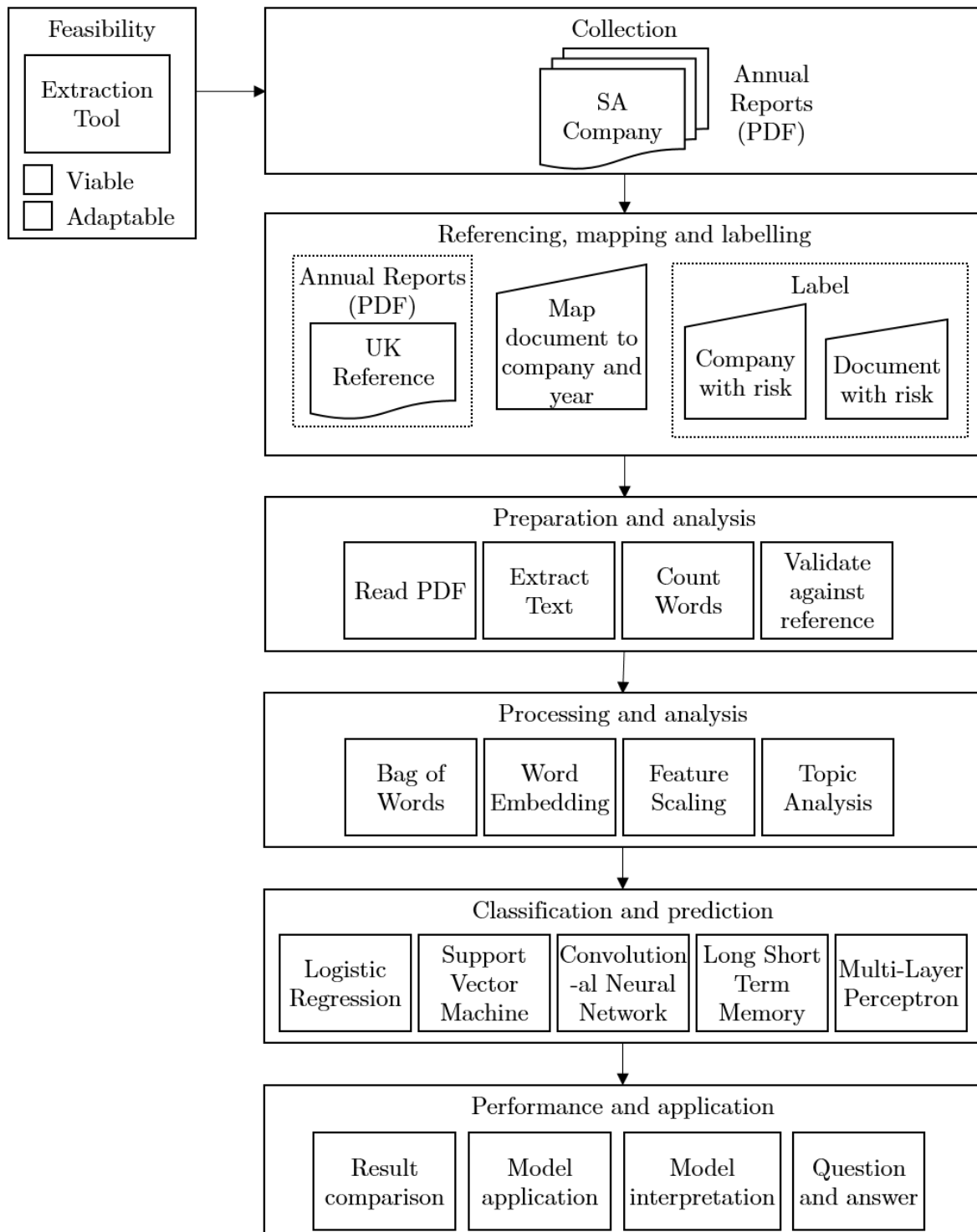


Figure 3.1: Overview of the end-to-end methodology.

3.1 Feasibility test

Figure 3.2 illustrates the flow of the feasibility test.

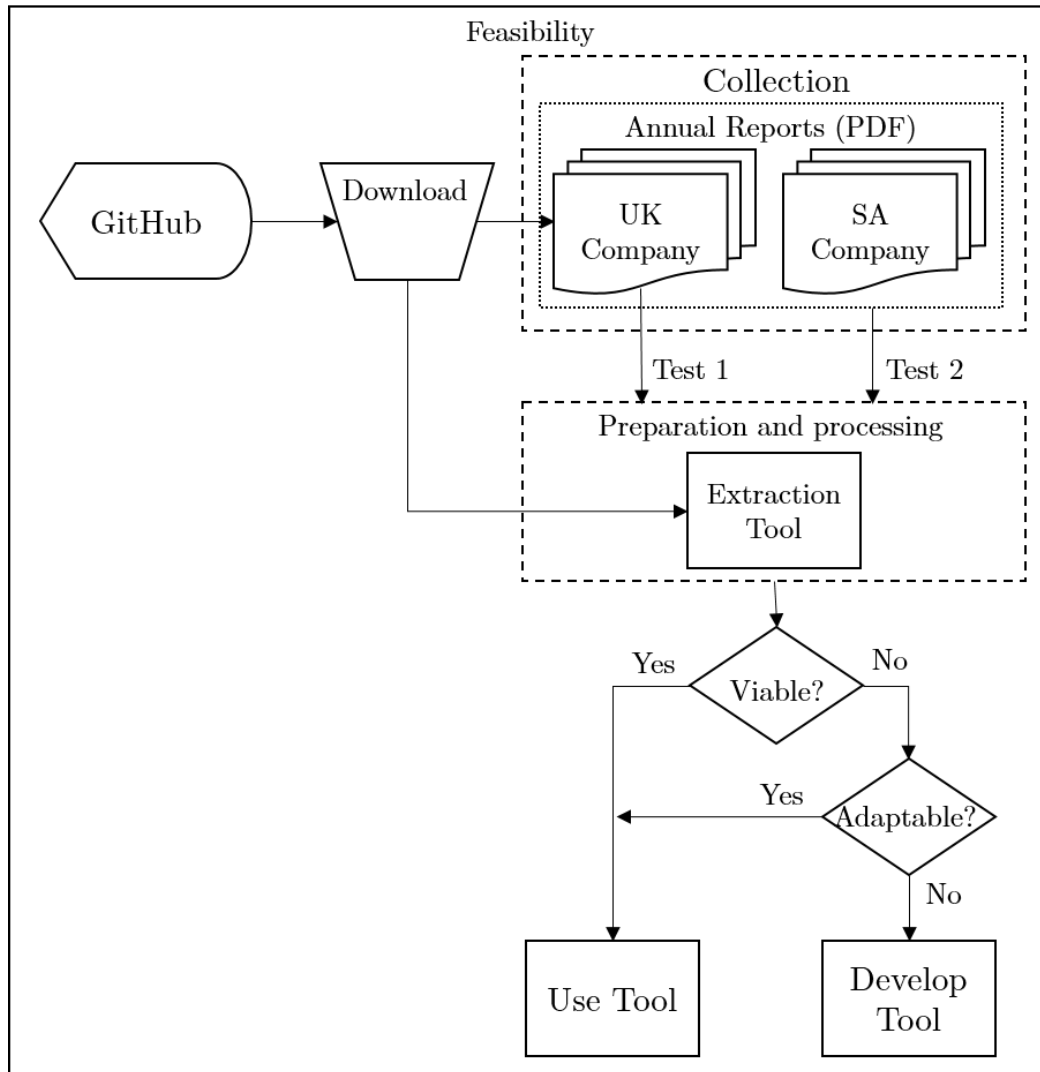


Figure 3.2: Overview of the feasibility test.

To pre-test the viability of using or adapting the Corporate Financial Information Environment - Final Report Structure Extractor (CFIE-FRSE) tool, version 2.0 of the Java executable application was downloaded from GitHub [16]. The tool was set up on an Ubuntu virtual machine (VM) that already had Java installed. It was hosted on an

Intel Core i7 2.6GHz laptop with 20GB of RAM.

Test 1, the first run of the tool, was on a sample of UK reports to ensure compatibility of the reports, code and system environment. Results were manually scanned and compared with what was expected, namely:

- Header identification and classification;
- Page start and end identifiers per header;
- Overall word counts and wordlist-specific word counts;
- Table of content and booklet flags.

Test 2, the second run of the tool, included the 2015 and 2016 annual reports of the SA company, VBS Mutual Bank.

As an alternative to using or adapting the CFIE–FRSE tool, a new tool had to be developed with similar functionality using Python notebooks and generic PDF mining packages. Testing the new tool on a UK company annual report served as a reference control for text extraction and word counts. Once the CFIE–FRSE results were approximated, the new tool was used on the SA company annual reports.

3.2 Data collection

An overview of the data collection approach is given in Figure 3.3.

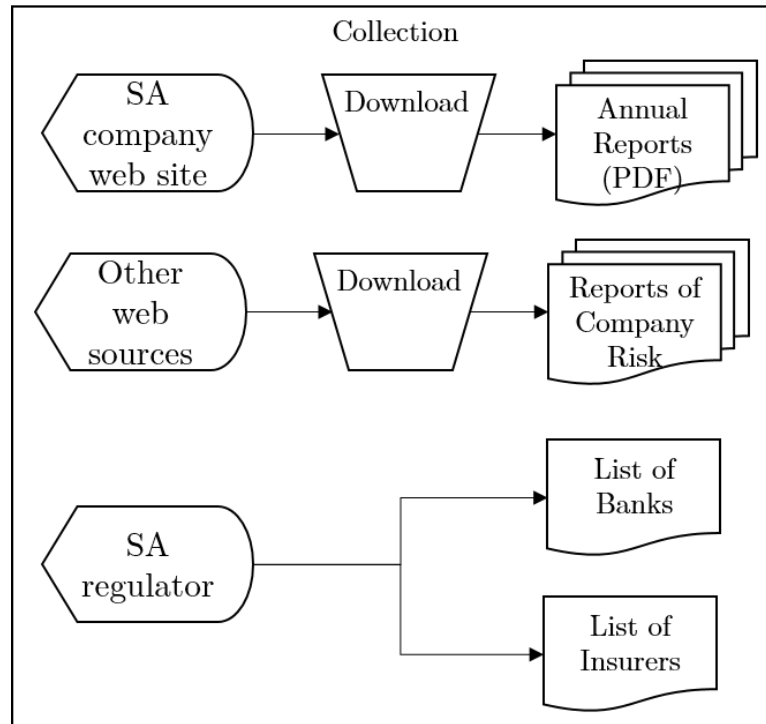


Figure 3.3: Overview of data collection.

Financial institutions are not required to submit annual reports to the SA prudential regulator although financial statements and insurer Own Risk Solvency Assessments (ORSA) are requested on a case-by-case basis. Since annual reports are ad hoc and not stored centrally, they were downloaded from the institutions' websites. When not found on the company website, online databases and portals were investigated as potential sources of the annual report.

A list of licensed financial institutions was sourced from a national regulator to serve as master data and as a checklist for downloading the annual reports. This comprised a list of banks and a list of insurers.

3.3 Data reference and labelling

The flow of this section is depicted in Figure 3.4.

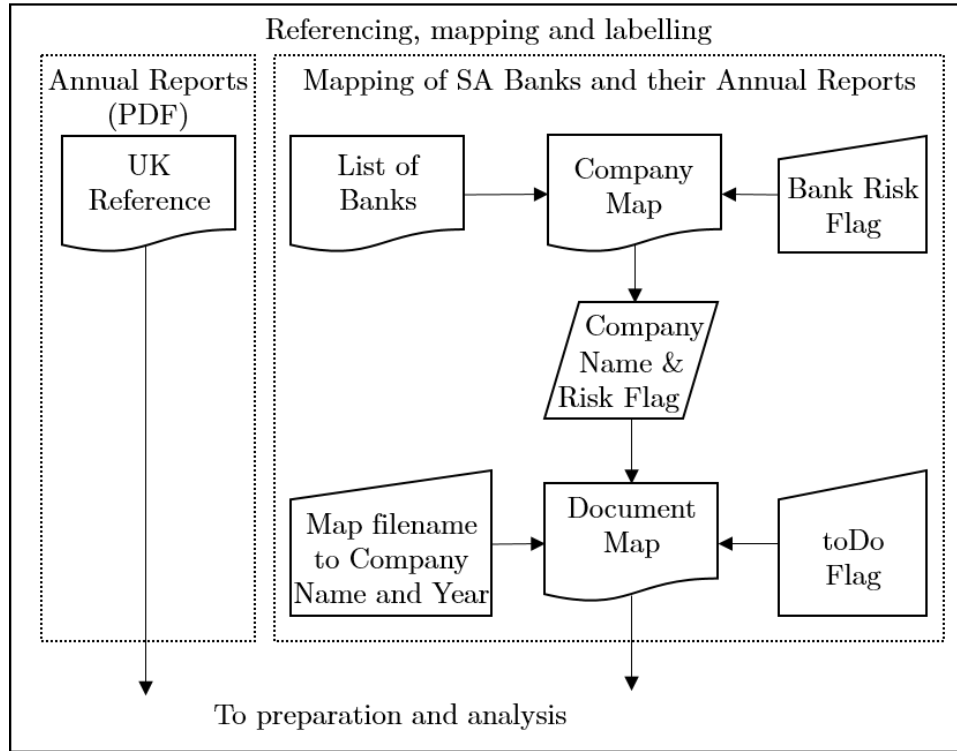


Figure 3.4: Overview of referencing, mapping and labelling.

In addition to the SA companies and annual reports, one UK company and annual report was included to provide a reference against which word counts could be validated. The selected annual report is referred to as the ‘*UK reference report*’ in this study.

A CSV file was created to map company identifiers and names to the financial institution master data. Risk ratings for South African companies are not public. For a fee, commercial services may screen a company against sanctions lists or offer a company risk rating based on the service provider’s research and algorithms. This study did not have access to regulator risk ratings and did not pay for similar commercial offerings. Instead, banks were labelled with a binary risk flag based on issues reported in public.

The initial approach for assigning metadata to the downloaded documents was to extract the year from the filename using regular expressions. This quickly proved not

to be a scalable or reliable approach due to inconsistent file names. To extract year and company name programmatically, files would have to be consistently renamed e.g. ‘[Company Name][CCYY][Report Name].pdf’. A simpler approach was to use a CSV mapping file. Years were captured during data collection or by manually reading the filename or document. A `todo` binary flag was used to configure documents to be processed. Documents were labelled by merging risk from the company mapping file using the company identifier.

3.4 Data preparation and analysis

Figure 3.5 shows the two methods followed in preparing data for classification.

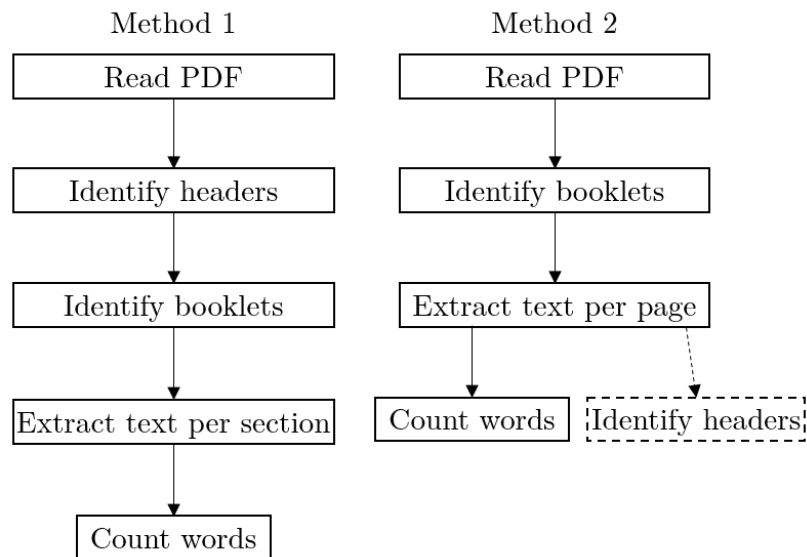


Figure 3.5: Methods of data preparation.

In method 1, an attempt was made to identify headers so that text could be extracted per section. This placed dependence on header identification. If headers could not be identified then text was not extracted and words could not be counted. Method 2 dealt with booklets upfront. This allowed text to be extracted per published page including booklets having two published pages per PDF page. Method 2 attempted header identification last and removed it from the critical path.

3.4.1 Booklet identification

In method 1, a booklet was identified by virtue of the total number of PDF pages being less than the actual page number of the last header. As an example, the 2015 annual report of VBS Mutual Bank had 56 PDF pages. ‘Notes to the annual financial statements’ was the last header of that report. The number extracted from the table of contents for that header was 26. With the last header’s actual page number being less than the 56 PDF pages, the 2015 annual report was not a booklet. The same header formed the last section of the 2016 annual report of VBS Mutual Bank with an actual page number of 58 yet the report had only 47 PDF pages. It was concluded that the 2016 annual report of VBS Mutual Bank was a booklet.

Since Method 2 was not dependent on header identification another mechanism was needed to identify booklets. Equation (3.1) shows how the width of the first two pages were used to calculate a ratio of page sizes.

$$pageRatio = \frac{page1.rect.x1}{page0.rect.x1} \quad (3.1)$$

where page0 : first PDF page;

page1 : second PDF page;

rect : rectangular coordinates of the page (x0, y0, x1, y1).

Since the first horizontal coordinate of the page x_0 was 0, the second horizontal coordinate of the page x_1 equated to the page width. The *pageRatio* can only have 4 distinct values. The meaning of each value is given in Table 3.1.

Table 3.1: Ratios of PDF pages.

Value	Meaning
-1	Less than 2 pages
0	<i>page0</i> is cover consisting of front and back pages
1	<i>page0</i> and <i>page1</i> are the same width
2	<i>page1</i> is twice as wide as <i>page0</i> i.e. is booklet

The -1 value highlighted erroneous files that were then replaced. Table 3.2 illustrates

the use of Equation (3.1) on annual reports of VBS Mutual Bank.

Table 3.2: Example of booklet identification using *pageRatio*.

Year	Page	rect	rect.x1	pageRatio	Booklet
2015	0	(0.0, 0.0, 606.6, 853.2)	606.6	$1 = \frac{606.6}{606.6}$	No
	1	(0.0, 0.0, 606.6, 853.2)	606.6		
2016	0	(0.0, 0.0, 595.3, 841.9)	595.3	$2 = \frac{1190.6}{595.3}$	Yes
	1	(0.0, 0.0, 1190.6, 841.9)	1190.6		

3.4.2 Text extraction

Various Python packages for PDF text extraction were explored and considered including `textract`¹, `PyPDF2`², `PDFminer`³ and `PyMuPDF`⁴. `PyMuPDF` version 1.18 was selected because it had the distinguishing capability of extracting headers from PDFs with tables of content.

In method 1, `PyMuPDF`'s default text extraction was used whereas method 2 extracted text blocks. Blocks are a list of items with the following format:

$$(x0, y0, x1, y1, \text{"lines in block"}, block_no, block_type)$$

where

$x0, y0, x1$ and $y1$: coordinates of the rectangular block;

$block_no$: sequential number (starting with 0) of block occurrence on the page;

$block_type$: 0 denotes text (1 denotes images).

Extraction experiments were run for different combinations of flags as listed in Table 3.3.

¹For information on `textract`, see <https://textract.readthedocs.io/en/stable>.

²For information on `PyPDF2`, see <https://pythonhosted.org/PyPDF2/1>.

³For information on `PDFminer`, see <https://pdfminersix.readthedocs.io/en/latest1>.

⁴For information on `PyMuPDF`, see <https://pymupdf.readthedocs.io/en/latest/index.html>.

Table 3.3: Flags for text extraction.

Flag	Value	Effect
<i>TEXT_PRESERVE_LIGATURES</i>	1	‘ffi’ instead of ‘f f I’
<i>TEXT_PRESERVE_WHITESPACE</i>	2	‘V B S M u t u a l B a n k’
<i>TEXT_INHIBIT_SPACES</i>	8	See Table B.5

Text was extracted per block using Algorithm 3.1 which processes both normal PDF documents ($pageRatio = 1$) and booklets ($pageRatio = 2$) simultaneously.

```

for all pages in PDF document do
  if  $pageRatio = 0$  and first page then
    Read rect0 as last extracted page
    Read rect1 as first extracted page
  else if  $pageRatio = 2$  and page between first and last then
    Read rect0 as extracted page
    Increment extracted page number
    Read rect1 as extracted page
    Increment extracted page number
  else
    Read rect as extracted page
    Increment extracted page number
  end if
end for

```

Algorithm 3.1: Algorithm for reading pages of documents and booklets.

To better understand Algorithm 3.1, Figures 3.6-3.8 illustrate the actual pages extracted relative the first and second PDF pages of reports with $pageRatio = 0, 1$ and 2 respectively.

Figure 3.6: Text extraction for report with $pageRatio=0$

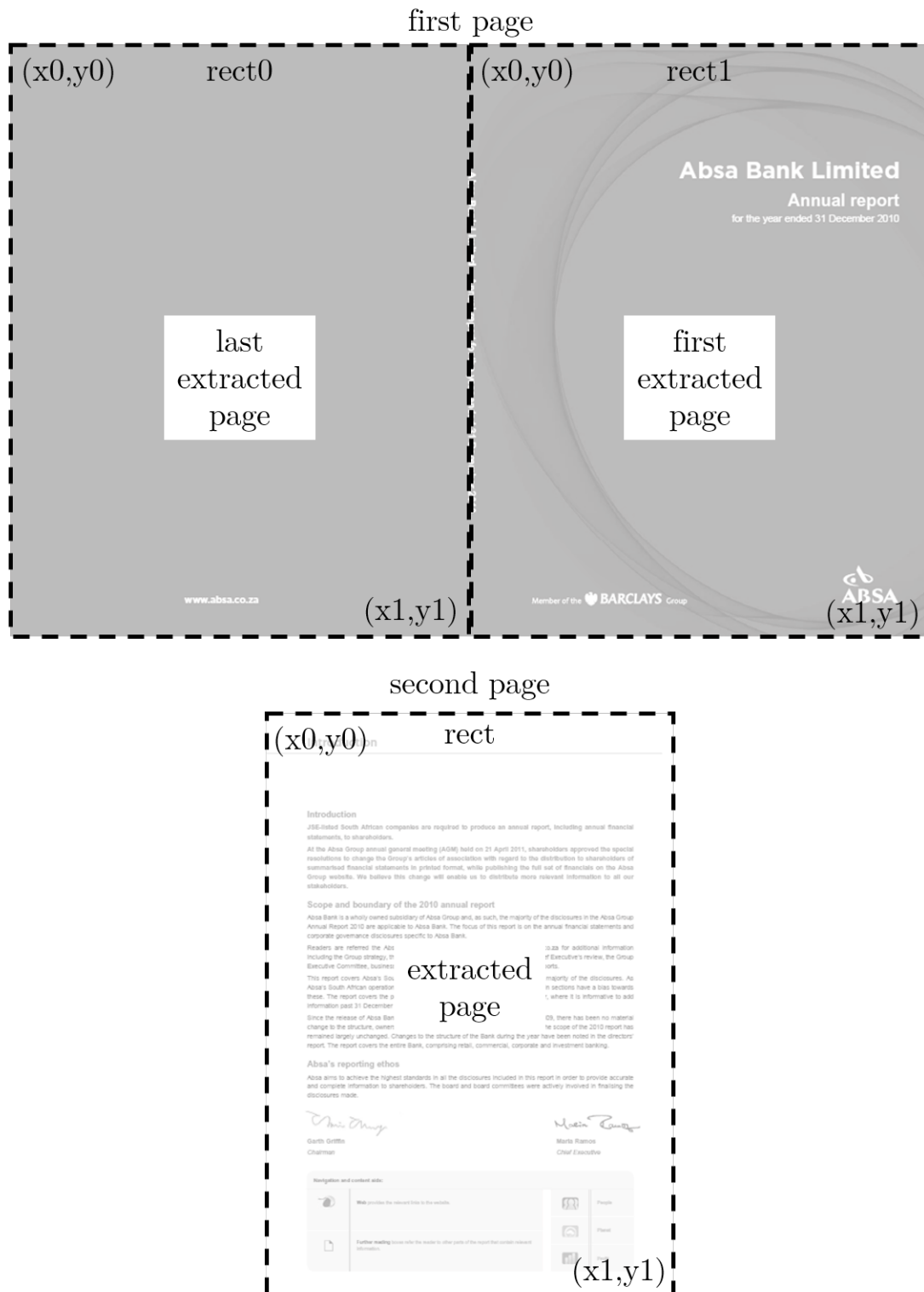


Figure 3.7: Text extraction for report with *pageRatio*=1

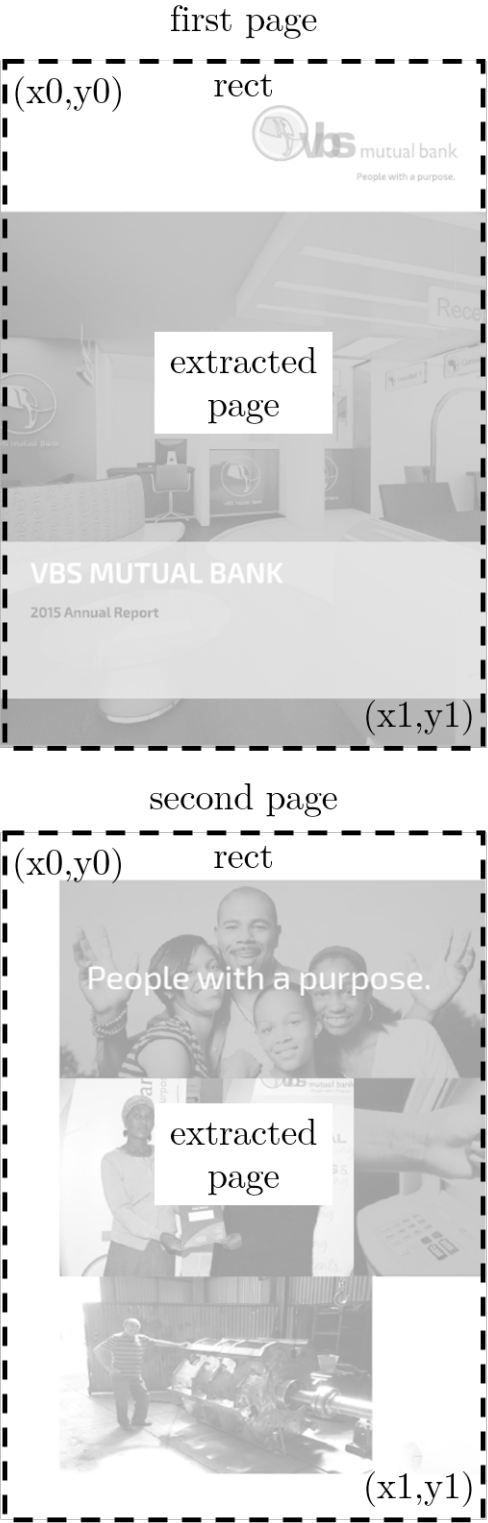
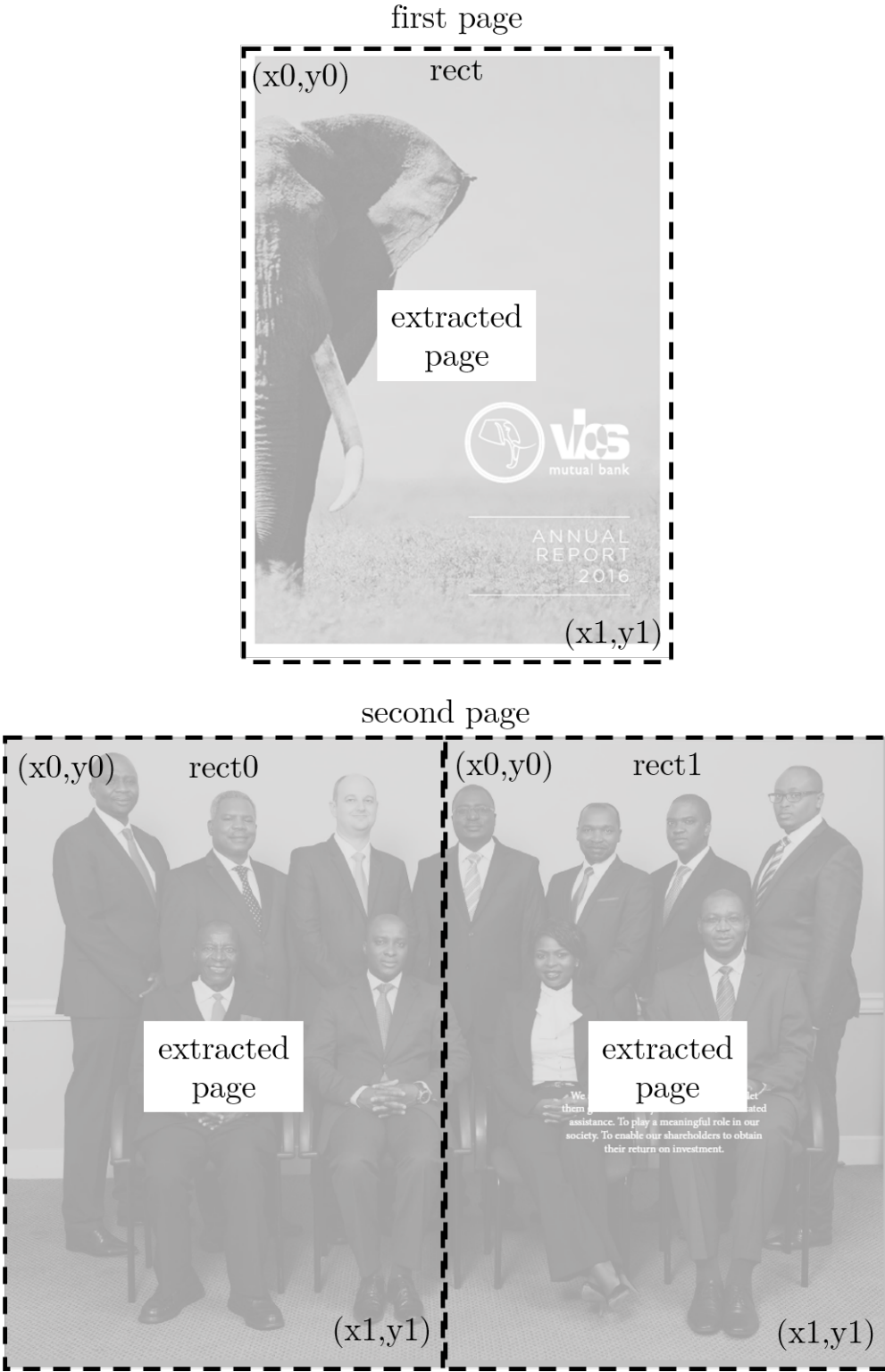


Figure 3.8: Text extraction for report with *pageRatio*=2



Resultant text blocks were saved to a CSV file. *year*, *coID*, *coName* and *risk* columns from the document and company mapping files were merged with the page blocks for ease of reference and aggregation. This made block coordinates available for subsequent header identification.

3.4.3 Word count

The final method concatenated blocks of text with a space and then split on spaces. Other methods tested were:

- Count page words extracted into a list by `PyMuPDF.getText('words')`;
- Regular expression `re.findall(r'\w+', text)` on page text;
- Split on spaces in page text;
- Split on spaces in page blocks extracted into a list by `PyMuPDF.getText('blocks')`.

The CFIE word count results were aggregated by section delimited by the header text. To allow validation in method 2 without the need to identify headers, sections of the reference report comprising no more than a single page were selected for validating the word count. This approach eliminated variances that could be attributed to header identification and not word count.

3.4.4 Wordlist counts

Table 3.4 shows the wordlists used to extract linguistic features from the text. Lists were obtained from two sources:

- CFIE-FRSE tool downloaded from GitHub [16];
- Data set CFIE-FRSE May 2019 extracted as `Elhaj_ABR_paper_files` [18].

Table 3.4: Wordlists variables and text files explained.

No	Variable	Text File	Comment
1	forward1	forwardLooking	Forwardlooking wordlist based on the list proposed by Hussainey et al. [35]
2	forward2	forwardLooking New	Forwardlooking wordlist based on an updated version of the list proposed by Hussainey et al. [35, 18]
3	LMneg	LMnegative	Negative wordlist based on Loughran and McDonald [42]
4	LMpos	LMpositive	Positive wordlist based on Loughran and McDonald [42]
5	H6neg	HenryNeg2006	Negative wordlist based on Henry [30]
6	H6pos	HenryPos2006	Positive wordlist based on Henry [30]
7	H8neg	HenryNeg2008	Negative wordlist based on Henry [29]
8	H8pos	HenryPos2008	Positive wordlist based on Henry [29]
9	uncert	Uncertainty	Uncertainty based on Loughran and McDonald [42]
10	causal	causal	Causal reasoning wordlist relating to performance commentary, based on a composite wordlist [18]
11	causalM	causalMartinAll	Causal reasoning wordlist [18]
12	causalM50	causalMartin50	Causal reasoning wordlist [18]
13	perf	performance	For identifying performance-related commentary [18]
14	strat	newStrategy	For identifying strategy-related commentary [18]

The Python package **Fuzzywuzzy**⁵ (version 0.18.0) was used for matching of words in the wordlists with words in the annual report page. The Python package **Python-Levenshtein**⁶ (version 0.12.0) was installed to prevent **Fuzzywuzzy** from using the slow pure-python **SequenceMatcher**.

Each substring in each wordlist was pre-processed for use within a regular expression

⁵For information on **Fuzzywuzzy**, see <https://pypi.org/project/fuzzywuzzy/>

⁶For information on **Python-Levenshtein**, see <https://pypi.org/project/python-Levenshtein/>

(re) pattern as follows:

- Remove leading space e.g. ‘ eps’ in performance.txt;
- Replace ‘#’ with ‘[a-zA-Z]*’ i.e. find words starting with substring when substring has suffix #;
- Replace <0:2> with ‘\s*’ i.e. allow zero or more spaces.

Loughran and McDonald’s [42] negative wordlist (LMnegative) has substring variants (‘abandon’, ‘abandoned’, ‘abandoning’, ‘abandonment’, ‘abandonments’, ‘abandons’). To prevent the regular expression pattern ‘abandon’ overlapping with the other variants and overstating the frequency, the substring was surrounded with raw word boundaries `r‘\b’`.

The function was a cumulative count of the length of `re.findall(r‘\b’+substring+r‘\b’, text)` and specifying that case should be ignored. Both the concatenated text (Section 3.4.9) and word counts were saved to a CSV file.

3.4.5 Readability sore

The Python package `py-readability-metrics`⁷ (*r*) was installed to import the Readability function. Text extracted from the UK reference report was passed to the function for each page. Scores were compared against the CFIE results. Given some discrepancies, the Python package `textstat`⁸ (*ts*) was also installed and compared. Table 3.5 shows the metrics used.

Table 3.5: Metrics used as indicators of readability.

No	Metric	Code	Package
1	Gunning Fog	gf	r, ts
2	Flesch Reading Ease	f	r, ts

Continued on next page

⁷For information on `py-readability-metrics`, see <https://py-readability-metrics.readthedocs.io/en/latest>

⁸For information on `textstat`, see <https://textstat.readthedocs.io/en/stable>

Table 3.5 – *Continued from previous page*

No	Metric	Code	Package
3	Flesch-Kincaid Grade Level	fk	r, ts
4	Dale-Chall	dc	r, ts
5	Automated Readability Index	ari	r, ts
6	Coleman-Liau	cl	r, ts
7	Linsear Write	lw	r, ts
8	SMOG	sm	r, ts
9	SPACHE	sp	r
10	Text Standard	ts	ts

The `textstat` package did not have the SPACHE metric. Instead it had a Text Standard as a proposed consensus value based on the scores of metrics 1-8. The Text Standard estimates the US school grade needed to understand the text. High values are therefore indicative of texts that are hard to read.

All metrics were calculated again when pages were aggregated into sections (Section 3.4.8) and documents (Section 3.4.9). The Gunning Fog metric was used to select the top and bottom five documents and sections.

3.4.6 Header identification

The method used for header identification is a slight improvement upon the CFIE-FRSE tool which processed only PDF's that were bookmarked. Algorithm 3.2 shows the multiple phases of header identification implemented.

When tables of content were not properly bookmarked in the PDF file, Algorithm 3.2 was limited by a number of assumptions. The table of contents header was assumed to be displayed at the top of the page. Experiments with manual checks showed that most table of content headers appeared within the first five page blocks. The fifth page block was used as an early stop condition to avoid elements of the body text being proposed as header text. A second assumption was that the table of contents page would appear at the beginning of the document. Experiments with manual checks showed that most table of content pages appeared within the first 20% of pages in the document. The

```
Extract bookmarks using PyMuPDF getToC()
    Find documents with 'contents' in a bookmark
    Extract table of contents using bookmarks for the selected documents
Read the blocks previously extracted using PyMuPDF
Identify the contents page
Sort blocks top-to-bottom using the y0 coordinate
repeat
    Read rect1 as extracted page
    Replace white space within the block text
    Lookup block text against items in wordlists
until fuzzy match or fifth page block or 20% of document read
Match number and text blocks using y-coordinate
    Identify blocks that start and end with numbers, and treat others as text
if no blocks contain solely numbers then
    Split the text between alpha and numeric
end if
```

Algorithm 3.2: Algorithm for identifying headers in documents.

20% page threshold was used as an early stop condition to avoid other tables of content pages being proposed e.g. notes to financial statements (see [B.3.3](#)). A third limitation was that the algorithm only catered for two table of content styles.

Start and end pages between headers were identified to define the section. As with the UK corpus, headers on the same page were concatenated and the words for each section were determined by counting between headers. For the last header, words were counted to the end of the document.

The Python notebook was developed incrementally as follows:

1. The UK reference report;
2. A SA bank positively labelled for risk;
3. A SA bank negatively labelled for risk;
4. All SA banks.

3.4.7 Header classification

The CFIE–FRSE pre-configured sections in Table 3.6 were used to classify the headers.

Table 3.6: Header types defined by CFIE–FRSE.

Header Type	Header Type Description
0	Indicates any other section that doesn't belong to the sections between 1 and 12
1	Letter from board chair which works with synonyms as well, e.g. Chairman's Statement
2	CEO review
3	Governance statement
4	Remuneration report
5	Business review
6	Financial review
7	Operating review
8	Highlights
9	Auditors report
10	Risk management
11	Chairman's governance
12	CSR disclosures

Substrings used in the CFIE–FRSE tool's IF–THEN clauses were used as the basis for wordlists. Exceptions were as follows:

2 CEO review: repeated 'review', 'report', 'statement' and 'interview' suffixes for

- 'ceo';
- 'chiefexecutives';
- 'chiefexecutiveofficer';
- 'managingdirector';
- 'executivechairman'; and
- 'executivechairmans'.

- 5 Business review: Created new wordlist;
- 6 Financial review: Created new wordlist;
- 7 Operating review: Created new wordlist;
- 8 Highlights: Created new wordlist;
- 9 Auditors report: Removed ‘uk’;
- 11 Chairman’s governance: Created new wordlist.

Keywords in the wordlists were used in fuzzy lookups against the header text to associate a header type. Header texts were also filtered by combinations of keyword to compare the the header types of Table 3.6 with the corpus of SA bank annual reports.

3.4.8 Section aggregation

Words between headers were counted as the section length. A limitation of this approach is that ‘out-of-context’ text may interlace the text of interest (see Appendix B.3.4).

Wordlist counts were scaled relative to the number of words and a net tone was calculated using Equation (3.2).

$$tone = \frac{positiveWordCount - negativeWordCount}{positiveWordCount + negativeWordCount} \quad (3.2)$$

The net tone was the proportion by which positive counts exceeded negative counts out of the sum of positive and negative words as listed in Table 3.7.

Table 3.7: Net tone calculated from wordlist count variables.

No	Tone	No	Negative	No	Positive
29	<i>LMtone</i>	3	<i>LMneg</i>	4	<i>LMpos</i>
30	<i>H6tone</i>	5	<i>H6neg</i>	6	<i>H6pos</i>
31	<i>H8tone</i>	7	<i>H8neg</i>	8	<i>H8pos</i>

Table 3.4 provides the text files and references related to the positive and negative wordlists used.

For Exploratory Data Analysis (EDA), tone values were tabulated and assigned graded colours based on the scale -1 (cyan) to +1 (magenta). Visual comparisons were made between measures and sections for the UK reference and select SA reports.

3.4.9 Document aggregation

Text was joined with a space and word counts summed per page, in grouping by document. Wordlist counts were scaled in the same manner as section aggregation (Section 3.4.7).

Tone values for each page or each section cannot be simply summed as this gives an incorrect result. Tone values were recalculated using Equation (3.2).

3.4.10 Validation

The output of the CFIE–FRSE tool from the feasibility test on UK reports (Section 3.1) was used as a reference to check the text extraction and word counts of the Python notebook.

3.4.11 Training and test data sets

Two annual reports with a positive risk flag were held out for testing together with all 2018 and 2019 reports. The data set was shuffled and a random state was set to ensure reproducibility.

Machine learning models were trained on three variations of the data set. The first iteration attempted training on the original data despite its class imbalance. Given the risk that positive documents would not appear within random folds of cross validation (CV) [67], random oversampling [28] was applied to positive training documents to balance the classes 50:50 i.e. an equal ratio of positive to negative documents. Several authors noted that their classifiers become too specific and overfit on the training data due to duplication of the minority class samples, especially for a small minority with many replicas [4, 46]. This resulted in generalisation performance on unseen test data being poorer than the observed training performance [7]. The third iteration tried to address the risk of overfitting by limiting positive documents to a 80:20 minority.

3.4.12 Feature selection

Given the many word counts and readability indices added to token frequencies or word embedding, initially a wrapper approach was taken. The affect of adding a subset of word counts to content features on classifier performance was assessed. Only scaled and not absolute counts were used together with net tone. Since longer annual reports may result in larger absolute word counts than smaller annual reports, the counts were scaled by the total number of words and selected as predictors instead of the absolute counts. Gunning Fog (Equation 2.1) and Flesch Reading Ease (Equation 2.2) were selected as two of the three readability indices as they had been applied successfully in previous studies (see Section 2.4.2.) The Text Standard (Section 3.4.5) readability index was also added as an experiment to determine whether this consensus of other readability indices was a more important predictor.

Experiments to assess the effect of each individual feature is resource intensive and does not account for the affect of multiple linguistic features being used in combination. The wrapper approach was refined using Recursive Feature Elimination (RFE) [26]. The RFE function of sklearn's feature_selection package removes a feature and then builds the model on the reduced feature set. Cross validation training performance is tracked and the process is repeated until all feature combinations are tested and their relative influence on the classification. Training and test performance was assessed to check if the select features reduced overfitting on training data and whether it improved test performance.

3.5 Data processing and analysis

Pre-processing and analysis conducted before performing the machine learning is shown in Figure 3.9.

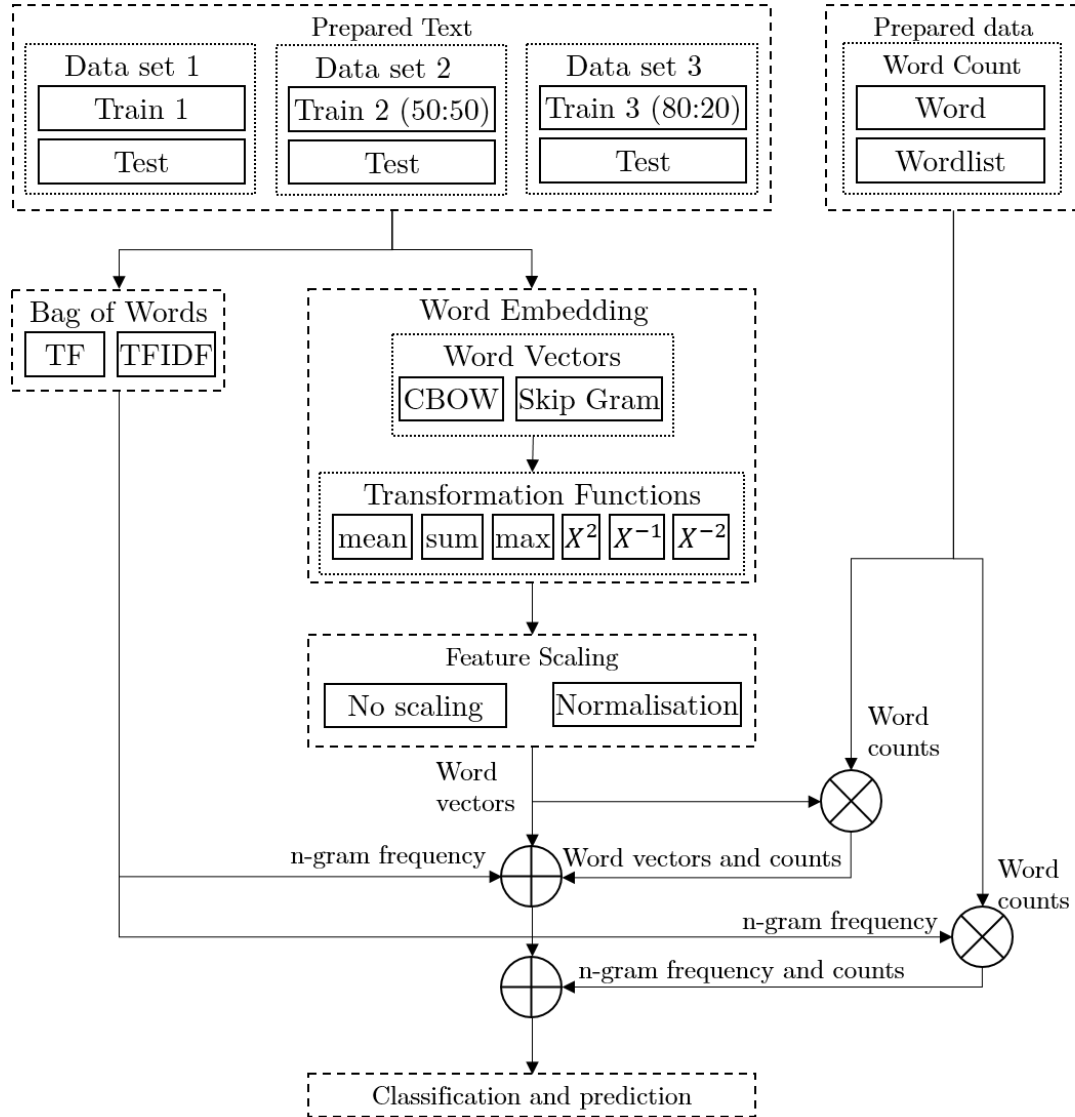


Figure 3.9: Data processing in the context of the experimental setup.

3.5.1 Bag of words

TF and TFIDF functions were used. Both of which converted the text into lower case and split on space and punctuation to tokenise into words. English stopwords were also removed.

Table 3.8 provides details of the two regular expressions used in the experiments.

Table 3.8: Regular expression patterns for tokenisers.

No	Pattern	Elements	Description
1	<code>r'[A-Za-z]{2,}'</code>	A-Z: upper case A-z: lower case {2,}: 2 or more characters	Strictly alpha characters
2	<code>r'\b[^\d\W]+\b'</code>	\b: word boundary []: character class ^\d: not digit \W: not a word character +: One or more	Remove punctuation and numbers

The chosen token pattern of strictly alpha (option 1) also removed numeric and punctuation characters. Option 2, which excluded non-word characters and digits and numbers, was compared with the first option by observing the number of tokens extracted from the text.

Features used a zero or one word normalisation technique as follows:

- x: raw text without normalisation;
- ps: Porter Stemmer⁹ [56];
- ls: Lancaster Stemmer¹⁰ [51];

⁹For information on `nlk.stem.porter`, see https://www.nltk.org/_modules/nltk/stem/porter.html.

¹⁰For information on `nlk.stem.lancaster`, see https://www.nltk.org/_modules/nltk/stem/lancaster.html.

- ss: Snowball Stemmer¹¹ [1];
- wl: WordNet Lemmatizer¹² [72].

3.5.2 Word embedding

The NLTK `RegexpTokenizer`¹³ was used to tokenise the text in preparation for word embedding. Regular expression patterns in Table 3.8 were compared and the strictly alpha pattern selected as the number of tokens selected were comparable with the vectoriser used for the BoW. Tokens were converted to lowercase.

Gensim `Word2Vec` was used to represent the tokens as Continuous Bag of Words (CBOW) and Skip Gram (SG) vectors. `Word2Vec` embeds the words into a high dimensional space where word similarity is denoted by word vectors being in close proximity to each other. Six mathematical functions were used to transform the word vectors for each document. The functions used were mean, sum, max, power of 2, power of -1 and power of -2.

3.5.3 Feature scaling

As SVMs use distance between vectors to determine their similarity, their performance is affected by the range of feature values. To avoid the model assigning greater weight to features with wider ranges, features were scaled. Min-max scaling was used to fit values between 0 and 1 as shown in Equation (3.3).

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (3.3)$$

¹¹For information on `nltk.stem.snowball`, see https://www.nltk.org/_modules/nltk/stem/snowball.html.

¹²For information on `nltk.stem.wordnet`, see http://www.nltk.org/_modules/nltk/stem/wordnet.html#WordNetLemmatizer.

¹³For information on `RegexpTokenizer`, see <https://www.kite.com/python/docs/nltk.RegexpTokenizer>.

where

x' : scaled feature value;

x : raw feature value.

Min-max scaling will reduce the range of values to ensure that outliers fall within the range [0,1]. A standard scaler was also applied to compare the effect of outliers. The standard scaler implements z-score normalisation as given in Equation (3.4).

$$x' = z = \frac{x - \mu}{\sigma} \quad (3.4)$$

where

x' : scaled feature value;

z : z-score;

x : raw feature value;

μ : mean of feature values;

σ : standard deviation of feature values.

With normalisation, values are scaled so that the standard deviation is unity either side of a zero mean. This implies that the range of each feature may differ even though they are in the same order of magnitude. Also, outliers have the effect of increasing the range as opposed to the compressed values of the min-max scale.

Neural Networks [24] may perform better with the min-max scale as they make no assumption on the distribution of the data. The Gaussian Radial Basis Function (RBF) kernel [73] used in some SVMs assume feature values are distributed either side of zero and may perform better with the standard scaler.

To ensure that all models were applied consistently to training and test data sets, the same scaler transform was used on both training and test features i.e. only the training documents were used to determine the scaler transform.

3.5.4 Topic analysis

Topic analysis was performed on specific sections of annual reports flagged with risk, namely:

- Independent external auditor report;
- Internal audit and risk committee report.

The assumption was that topic clusters of these annual reports may highlight words or patterns that distinguish them from reports that weren't annotated with risk.

Previously identified headers (Section 3.4.5) were merged with text extracted by page (Section 3.4.2). The text was then concatenated in grouping the pages by section header. A custom tokenizer removed string punctuation. It also removed words found in the English stopwords of the NLTK corpus or tokens of 1 or 2 characters. Inflected words were normalised into their root form using the WordNet Lemmatizer. The `TfidfVectorizer` converted text into lower case and removed accents found in some of the foreign texts (see Figure B.7) [65]. TFIDF was calculated for all features across all document sections. Performing the calculation solely on the selected section is equivalent to a TF calculation; however the hypothesis assumed distinguishing features in the section of interest. Given this assumption, TFIDF was more appropriate.

The top twenty features were visualised by generating wordclouds from the frequencies i.e. TFIDF n-grams of strictly n and not a range of 1 to n [52]. Experiments were conducted for values of 1 through 5. Latent Dirichlet Allocation (LDA) was also performed on the TFIDF tokens to cluster features based on the assumed number of topics per section. Again, experiments were conducted for values of 1 through 5. Random state 0 was used to ensure results were reproducible. Bilinear interpolation of the pixels was used for a clear image [71, 34].

When 1 topic was assumed for the section, wordclouds were produced using TFIDF and compared with the LDA frequencies.

3.6 Classification and prediction

Table 3.9 summarises the parameters used with each model to experiment on different feature types.

Table 3.9: Codes and parameters of models used.

Code	Model	Parameters
LR	Logistic Regression	solver='liblinear'
SMVa	Support Vector Machine	kernel='rbf', gamma='auto', C=1
SVMs	Support Vector Machine	kernel='rbf', gamma='scaled', C=1

The following subsections provide details for these classifiers.

3.6.1 Logistic regression

Scikit Learn's¹⁴ Logistic Regression was selected as a simple classifier to establish a performance baseline. Notebooks were developed on a local Anaconda¹⁵ client as well as the Google Colab cloud platform. These environments used Scikit Learn version 0.20.3 and version 0.22.2 respectively.

Models were trained on the following combination of features:

- Bag of Words
- Bag of Words + word counts
- Bag of Words + word counts + readabilities indices
- Word Embedding
- Word Embedding + word counts
- Word Embedding + word counts + readabilities indices

¹⁴For information on Scikit Learn, see <https://scikit-learn.org/stable>.

¹⁵For information on Anaconda, see <https://docs.anaconda.com/>.

Word count features comprised the overall *wordCount* (Section 3.4.3), counts of keywords in wordlists (Section 3.4.4) scaled relative to the *wordCount* as well as net tone (Section 3.4.8). Table 3.10 lists these word count features.

Table 3.10: Variables for word count features.

No	Variable	Description
0	wordCount	Count of all words
15	forward1_scaled	Scaled count of words in the forwardLooking wordlist
16	forward2_scaled	Scaled count of words in the orwardLookingNew wordlist
17	LMneg_scaled	Scaled count of words in the LMnegative wordlist
18	LMpos_scaled	Scaled count of words in the LMpositive wordlist
19	H6neg_scaled	Scaled count of words in the HenryNeg2006 wordlist
20	H6pos_scaled	Scaled count of words in the HenryPos2006 wordlist
21	H8neg_scaled	Scaled count of words in the HenryNeg2008 wordlist
22	H8pos_scaled	Scaled count of words in the HenryPos2008 wordlist
23	uncert_scaled	Scaled count of words in the HenryPos2008 wordlist
24	causal_scaled	Scaled count of words in the causal wordlist
25	causalM_scaled	Scaled count of words in the causalMartinAll wordlist
26	causalM50_scaled	Scaled count of words in the causalMartin50 wordlist
27	perf_scaled	Scaled count of words in the performance wordlist
28	strat_scaled	Scaled count of words in the newStrategy wordlist
29	LMtone	Net tone using Loughran and McDonald (2011)
30	H6tone	Net tone using Henry (2006)
31	H8tone	Net tone using Henry (2008)

The first run had no feature scaling and used 5-fold cross validation from the Scikit Learn model selection package. Recall as given in Equation (3.5) is the ratio of true positives (TP) predicted to actual positives.

$$recall = \frac{TP}{TP+FN} \quad (3.5)$$

where

TP : true positives;

FN : false negatives.

Recall is easily achieved with imbalanced classes, as is the case with this data set where the positive class is the minority and the negative class is the majority (see 4.4.11). A model with a good recall result may still predict many false positives. To address this, the F1 score as given in Equation (3.8) was used as the performance metric.

The random probabilities used in cross validation led to variability in comparing results for the same experiments on the small data set. The second run normalised the input to the range [0,1] as well as standardising the data to one standard around a zero mean. 5-times, 5-fold cross validation was then implemented for more consistent results.

$$F1 = \frac{2 \times precision \times recall}{precision + recall} = \frac{TP}{TP + 0.5 \times (FN + FP)} \quad (3.6)$$

where

TP : true positives;

FP : false positives;

FN : false negatives.

The precision metric is the ratio of true positives (TP) predicted to all predicted positives including false positives (FP) as reflected in in Equation (3.7).

$$precision = \frac{TP}{TP+FP} \quad (3.7)$$

where

TP : true positives;

FP : false positives.

Models were trained for each feature type as set out in Table 3.11.

Table 3.11: Feature parameters used in training models.

Feature Type	Parameter	Start	End	Increment
Bag of Words	N-gram range	1	3	1
	Maximum features	20	4000	variable
Word Embedding	Minimum word count	1	8	1
	Dimensions	100	500	100

3.6.2 Support vector machine

Two Support Vector Machines [69] were built by using the default RBF kernel. SVMa had gamma (γ) specified as ‘auto’. SVMs had γ specified as ‘scaled’ (SVMs) Equation (3.8) shows the use of γ in the exponential of the RBF [65].

$$y'(x) = e^{-\gamma\|x-x'\|^2} \quad (3.8)$$

$$\text{where for SVMa : } \gamma = \frac{1}{\text{numberOfFeatures}}$$

$$\text{SVMs : } \gamma = \frac{1}{\text{numberOfFeatures} \times \text{var}(x)}$$

The SVMs model incorporates the variance of the feature values ignored by SVMa model. Default values were used for all other parameters e.g. penalty parameter C=1. SVM models were trained on the same combination of features using the same cross validation as the Linear Regression models. Models were trained for each feature type as set out in Table 3.11.

3.6.3 Long short term memory recurrent neural network model

The Long Short-Term Memory (LSTM) [32] Recurrent Neural Network (RNN) [62] model was used to predict a risk class based on the sequence of word tokens. The input sequence was attempted for the longest annual report in the training data as well as half that length.

Given the class imbalance of the original data set and the data set resampled to 20:80 (positive:negative), LSTM weights for each training class were set as the inverse frequency. For the data set resampled to 50:50, this obviously led to equal weights e.g. 0: 0.00490, 1: 0.0049.

Table 3.12 describes the LSTM models used in the experiments and Table 3.13 shows the layer configuration. Parameters were based on Sawarn's [64] LSTM model used for multi-label classification of movie plots.

Table 3.12: Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN) models.

Code	Description
LSTM1	2-layer LSTM with 50% dropout)
LSTM2	2-layer LSTM with 20% dropout)

The dropout parameter defines the proportion of hidden layer neurons to be randomly disabled by the model. Randomly disabling hidden layer neurons effectively produces an ensemble of models based on the originally defined model.

The risk target remained a single label with binary value 0 or 1. This allowed the numbers of classes to be set to 1 which resulted in one output neuron. A sigmoid activation function was specified to shape the log-odds output towards the extreme 0 and 1 values. The risk target could also have been one-hot encoded to categorical values $[[1,0],[0,1]]$ in which case two classes are specified. In that case, a softmax activation function can be used to ensure a probability distribution of one across both labels.

Since a sigmoid function is used in the LSTM, the input was normalised to the same scale of 0-1 to see if performance improved.

Vectors of length 50 were selected to represent words in the embedding layer. Neurons in the LSTM layers provide 128 and 64 memory units for layers 2 and 4. 50% dropout was added to layers 3 and 5 to reduce overfitting [8]. Since this is a binary classifier, the final dense layer only requires a single neuron for the sigmoid function to predict risk of 0 or 1.

Table 3.13: Parameters used in building LSTM models.

Model:		LSTM1	LSTM2
Layer	Type	Parameters	Parameters
1	Embedding	Vocab_size output_dim=50 input_length	Vocab_size output_dim=50 input_length
2	LSTM	units=128 return_sequences=True	units=128 return_sequences=True
3	Dropout	50%	20%
4	LSTM	units=64 return_sequences=True	units=64 return_sequences=True
5	Dropout	50%	20%
6	Dense	N_classes=1 Sigmoid activation	N_classes=1 Sigmoid activation

The following inputs were used:

1. input sequence;
2. TFIDF matrix.

The LSTM was compiled on a local anaconda environment with 20GB RAM as well as Google Colab with CPU, GPU and TPU runtimes. Numerous adjustments were needed to compile the model to run on Colab's TPU runtime. A cluster resolver had to reference the Colab TPU address. The resolver was used to connect to the host, to initialise the TPU system and to configure a tensorflow distribution strategy. Eager execution was disabled before creating, compiling and fitting the LSTM model with the strategy.

The model was compiled with the Adam optimiser [37] and the binary_crossentropy function to perform log loss using the F1 metric [8]. The model was trained on five epochs updating weights after a batch size of 32 .

Experiments were conducted on two of the three training data sets, namely:

1. The original unbalanced data set;

2. The 50:50 (risk:no risk) resampled data set.

In the final implementation, the embedding layer of these models were removed due to processing challenges (Appendix B.4.7). Instead, the same types of features as prepared for LR and SVM were used as input into the first LSTM layer.

3.6.4 Convolutional neural network model

Convolutional Neural Network (CNN) models were built and adapted based on published models for similar use cases, namely:

1. Model used for multi-label text classification with 4 convolution layers [64];
2. Model with max pooling and LSTM used for movie sentiment classification [8];
3. Model with 3 convolution and max pooling layers used with news classification [11].

The CNN models used in the experiments are described in the following table.

Table 3.14: CNN models used in the experiments.

Code	Model Description
CNN1a	4 convolution layers with sigmoid activation (no dropout)
CNN1b	4 convolution layers with ReLU activation (no dropout)
CNN1c	4 convolution layers with ReLU activation (20% dropout)
CNN1d	4 convolution layers with no activation (no dropout)
CNN2	1 convolution layer with max pooling and LSTM layers
CNN3	3 convolution layers with max pooling

The structure of the models used in the experiments as well as the layer parameters are given in Tables 3.15 to 3.17.

The models were compiled to use the Adam optimiser and binary cross-entropy for loss calculation. Precision, recall, F1 and area under the curve (AUC) were monitored as training metrics using the default threshold of 0.5.

Table 3.15: Description of CNN1 with 4 convolution layers.

Model:		CNN1a	CNN1b	CNN1c	CNN1d
No	Type	Parameters	Parameters	Parameters	Parameters
1	Embedding	Vocab_size output_dim:50 input_length	Vocab_size output_dim:50 input_length	Vocab_size output_dim:50 input_length	Vocab_size output_dim:50 input_length
2	Convolution	Filter:64 kernel_size:3 Sigmoid	Filter:64 kernel_size:3 Sigmoid	Filter:64 kernel_size:3 ReLU	Filter:64 kernel_size:3 No activation
3	Convolution	Filter:100 kernel_size:3 Sigmoid	Filter:100 kernel_size:3 Sigmoid	Filter:100 kernel_size:3 ReLU	Filter:100 kernel_size:3 No activation
4	Convolution	Filter:100 kernel_size:3 Sigmoid	Filter:100 kernel_size:3 Sigmoid	Filter:100 kernel_size:3 ReLU	Filter:100 kernel_size:3 No activation
5	Dropout	0%	50%	50%	0%
6	Convolution	Filter:48 kernel_size:3 Sigmoid	Filter:48 kernel_size:3 Sigmoid	Filter:48 kernel_size:3 ReLU	Filter:48 kernel_size:3 No activation
7	Dense	N_classes:1 Sigmoid	N_classes:1 Sigmoid	N_classes:1 Sigmoid	N_classes:1 Sigmoid

Batch sizes of 16 and 32 were used. The training data was split 80:20 for cross validation. The models were trained on raw sequences as well as scaled sequences using normalisation $[0,1]$ and standardisation.

As with the LSTM, the final implementation of the CNN models had the embedding layer removed due to processing challenges (Appendix B.4.7). Instead, the same types of features as prepared for and SVM were used as input into the first LSTM layer.

Table 3.16: Description of CNN2 with LSTM layer.

Layer	Type	Parameters
1	Embedding	Vocab_size, output_dim:50, input_length
2	Convolution	Filter:32, kernel_size:3, Same padding, ReLU activation
3	Max Pooling	pool_size:2
4	LSTM	units:100
5	Dense	N_classes:1, Sigmoid activation

Table 3.17: Description of CNN3 with 3 convolution and max pooling layers.

Layer	Type	Parameters
1	Embedding	Vocab_size, output_dim:50, input_length
2	Convolution	Filter:128, kernel_size:5, ReLU activation
3	Max Pooling	pool_size:5
4	Convolution	Filter:128, kernel_size:5, ReLU activation
5	Max Pooling	pool_size:5
6	Convolution	Filter:128, kernel_size:5, ReLU activation
7	Max Pooling	pool_size:35
8	Flatten	
9	Dense	N_classes:1, Sigmoid activation

3.6.5 Multi-layer perceptron model

The MLP model was introduced as an alternative to the custom-built LSTM and CNN models. The **Scikit Learn** package has a MLP classifier that can be trained in the same manner as the LR and CNN.

Three hidden layers, each with eight neurons were specified as per Table 3.18.

Table 3.18: Parameters used in building MLP models.

Model:		MLP
Layer	Type	Parameters
1	Input	input_length
2	Hidden	units=8, ReLU activation
3	Hidden	units=8, ReLU activation
4	Hidden	units=8, ReLU activation
5	Dense	N_classes=1, Sigmoid activation

The training samples only specify the output value that the model must approximate. A learning algorithm is needed to calculate the weights of the other hidden layers to approximate the output. As with training the CNN, the weights of each layer are updated using gradient descent to minimise a loss function. In particular, the weight change is based on the derivative of the loss function and present weight value. The loss function is the negative log-likelihood described as the cross-entropy between the expected training data and the calculated values of the model. This locates the same optimum weights as the mean squared error between the expected and calculated values of each output [24].

3.7 Performance and application

3.7.1 Classifier prediction

As with training, the F1 score (Equation (3.8)) was used as the performance metric in testing the classifiers on unseen test data held out as explained in Section 3.4.11. The maximum number of features and dimensions were reduced incrementally (Table 3.11).

The classifiers responding with the best test performance for the smallest number of features were identified. This approach was effectively a simplified wrapper feature selection method. Experiments involving the best performance classifiers were assessed for the presence or absence of the additional linguistic features.

Training and test performance of these classifiers were compared across different word embedding transforms. The effect of scaling on features was also measured.

Test performance of the modern neural network classifiers was assessed using various thresholds for the range of output values. To gain insight into the F1 scores, recall and precision were also measured.

Two of the top performing models were selected for applying a probability score to the test data. The probability scores of the top 3 negative reports were compared with the 2 positive reports.

3.7.2 Classifier interpretation

Local Interpretable Model-agnostic Explanations (LIME) learns an interpretable model locally around the prediction and can be used to explain the probability prediction of classifiers. Sub modular Pick LIME (SP-LIME) was used to describe the model more globally by picking predictors across a broader value range and returning a set of predictions and explanations [61]. The model selected for interpretation was a Logistic Regression (LR) using 80 TFIDF tokens without word normalisation and a combination of unigrams and bigrams.

3.7.3 Question answering

Although the risk prediction in Section 3.7.1 is the use case relating to the main research question *Can risk be identified in unstructured company annual reports using Natural Language Processing (NLP)?* closed domain Question Answering (cdQA) was identified as a useful byproduct of the corpus. Risk prediction and cdQA are thus two potential applications of the annual report data set developed through this study and proposed to meet the last objective of Section 1.3.

The `cdQA-Suite` was installed for closed domain Question Answering (cdQA). The lo-

cal anaconda environment required an installation of `PyTorch` version 1.2.0 and `TorchVision` version 0.4.0.

Text extracted by page (Section 3.4.2) was concatenated in grouping pages by document. The document data frame and text had to be transformed for compatibility with the cdQA suite. In particular, the document name was selected as the title and text as the paragraphs. Paragraph text was tokenised into sentences using NLTK. New line characters ‘\n’ were replaced with a space ‘ ’. Paragraphs were filtered using the cdQA utility.

Pre-trained BERT [14] and DistilBERT [63] reader models were downloaded and two distinct cdQA pipelines constructed from them. Paragraphs were fitted to both pipelines which were then used to predict the 3 best answers to questions.

3.8 Summary

This methodology chapter described how the feasibility of using the CFIE–FRSE tool was assessed (Section 3.1) followed by how the data was collected (Section 3.2). The files used for validating word counts as well as mapping and annotation companies and their annual reports was explained in Section 3.3. Two methods for preparing the data were illustrated in Section 3.4. The methods include booklet identification, text extraction, word counts and header identification. Section 3.5 explained processing of the text into BoW and embedding words into vectors followed by various scaling options and some topic analysis.

Section 3.6 described the LR and SVM models built as traditional classifiers as well as the construction of more modern neural network models. Finally, Section 3.7 describes steps taken to test and interpret classifier performance. Application of the models to predict the top five risk reports as well as application of the data for were described.

Chapter 4

Pre-processing

Having provided the methodology in Chapter 3, this chapter shares the outcomes of data preparation, processing and analysis prior to modelling.

- Section 4.1 assesses the feasibility of reusing text and feature extraction tools from previous work;
- Section 4.2 lists and quantifies the annual reports collected;
- Section 4.3 shows data referenced from previous work as well as documents and companies mapped and labelled in this study;
- Section 4.4 provides results of booklet and header identification as well as text and feature extraction in preparation for the machine learning;
- Section 4.5 gives results from processing the extracted text into BoW as well as visualisation of some assumed topics.

See Appendix B for more detailed results that complement the tables and figures in this chapter.

4.1 Feasibility test

The CFIE–FRSE repository downloaded from GitHub had a ‘pdfs’ folder containing five annual reports of UK companies.

The first run of the CFIE–FRSE tool successfully identified and classified headers for the five UK reports. Words were correctly counted for each section. The frequency of keywords from the wordlists appearing in each section were also correctly counted. These values could be reconciled against the original output file for the corpus.

The output of the CFIE–FRSE tool was text by report header as well as text across these headers for each annual report. Having the tool split text by report was suitable for this study whereas the overall corpus was split by header and not by report.

The second run of the tool returned not applicable values for the 2015 annual report of VBS Mutual Bank. The PDF document was published without a table of contents which was noted by the tool.

The 2016 annual report of VBS Mutual Bank was in booklet format. As a result, it was ignored and not processed at all.

4.2 Data collection

For banks licensed to take deposits in South Africa, there was a potential 297 annual reports to be analysed for the 11 financial years from 2009 to 2019. Of these, 258 were sourced on the internet and downloaded. The difference of 39 was attributed to the report not being found or the bank not being operational in that year. The files were organised into 41 folders holding 1.67Gb of raw PDF data.

When not found on the company website, then either the ShareData online portal [57] or the Vault investment research outsourcing site was used [70]. The 2010 annual report of Habib Overseas Bank Ltd could not be found on any of the three sources. A database focusing on global banks reported to have access to this report but required a payed subscription [68]. An obscure finding was sourcing the 2009 and 2011 annual reports of African Bank Limited on a Swiss online magazine publishing platform [74].

Table 4.1 reflects the list of licensed banks and the financial years for which annual reports were downloaded. VBS Mutual Bank was added to the list as an example of a previously failed financial institution. No annual reports could be found for Bank of India and Bank of Taiwan. Since these are foreign banks with branches operating in SA, the omission is assumed acceptable for the initial corpus.

Table 4.1: Annual reports downloaded by bank and year.

Peer	Institution	Years	Reports
1G	ABSA BANK LTD	2009-2010	2
1	ABSA BANK LTD	2009-2010	2
1	FIRSTRAND BANK LIMITED	2009-2019	11
1G	INVESTEC BANK LTD	2009-2019	11
1	INVESTEC BANK LTD	2009-2019	11
1G	NEDBANK LTD	2010-2019	10
1	NEDBANK LTD	2009-2018	10
1G	THE STANDARD BANK OF S A LTD	2009-2019	11
1	THE STANDARD BANK OF S A LTD	2009-2019	11
2	BNP PARIBAS SOUTH AFRICA BRANCH	2009-2015, 2018-2019	9
2	CITIBANK N.A	2009-2019	11
2	DEUTSCHE BANK AG	2009-2019	11
2	HSBC BANK PLC - JOHANNESBURG BRANCH	2009-2019	11
2	JPMORGAN CHASE BANK (JOHANNESBURG BRANCH)	2009-2019	11
2	SOCIETE GENERALE JOHANNESBURG BRANCH	2009-2019	11
2	STANDARD CHARTERED BANK	2009-2019	11
3	ALBARAKA BANK LTD	2009-2019	11
3	BIDVEST BANK LIMITED	2009-2019	11
3	ITHALA LIMITED	2012-2018	7
3	MERCANTILE BANK LTD	2009-2019	11
3	SASFIN BANK LTD	2013-2018	7
3	THE S A BANK OF ATHENS LTD	2010-2015, 2017-2018	11

Continued on next page

Table 4.1 – *Continued from previous page*

Peer	Institution	Years	Reports
4	BANK OF BARODA	2010-2018	9
4	BANK OF CHINA LTD JHB BRANCH (T/A BANK OF CHINA JHB BRANC	2015, 2017-2019	4
4	BANK OF INDIA - JOHANNESBURG BRANCH	None	0
4	BANK OF TAIWAN SOUTH AFRICA BRANCH	None	0
4	CANARA BANK	2009-2018	10
4	CHINA CONSTRUCTION BANK CORPORATION - JHB BRANCH	2009-2018	10
4	ICICI BANK LIMITED	2009-2018	10
4	STATE BANK OF INDIA	2012, 2015-2019	6
5	AFRICAN BANK LIMITED (N)	2016-2019	4
5	AFRICAN BANK LIMITED (O)	2009-2015	7
5	CAPITEC BANK	2009-2012, 2014-2019	10
5	GRINDROD BANK	2009-2019	11
5	HABIB OVERSEAS BANK LTD	2011-2019	9
5	HBZ BANK LTD	2014-2018	5
5	UBANK LIMITED	2009-2019	11
5	VBS Mutual Bank	2015-2016	2

Company structures, trading practices and branding complicated the collection. For example, the South African Bank of Athens Limited (“SABA”) traded as Grobank after the acquisition by GroCapital Holdings Proprietary Limited (“GroCapital Holdings”). Peer group 1 has banks like ABSA, Investec, Nedbank and Standard Bank that are subsidiaries of a group company resulting in two sets of annual reports. In those cases,

institutions in Table 4.1 were duplicated. The G distinguishes the group company from the bank. Peer group 2 are South African branches of banks based in other countries (e.g. BNP Paribas South Africa Branch), in which case, it is an annual report of a European company.

Seven annual reports comprised multiple documents (see Appendix B.2). These exceptions were treated as individual annual reports.

4.3 Data reference and labelling

4.3.1 Word count validation

The AI Claim Solutions 2007 report was selected as the UK company document to be used to validate word counts against. It was selected due to being the first report out of the five reports analysed in ascending alphabetical order and for having two occurrences of multiple headers on the same page. The annual report has the filename ‘AI CLAIM SOLUTIONS PLC REPORT AND ACCOUNTS 2007.pdf.’ This study refers to the report as the ‘*UK reference report*’.

Three of the five UK reports did not have multiple headers on the same page. The 2004 annual report of Biocompatibles International had one occurrence of multiple headers on the same page and could have been selected as an alternative to the ‘*UK reference report*’.

4.3.2 Company mapping and labelling

Table 4.2 shows the mapping of company identifiers and names to the master list as well as company type, subtype and risk attributes.

Table 4.2: Companies by type, subtype and risk.

No	Company Name	No	Master Company Name	Type	Sub Type	Risk
0	AI Claim Solutions Plc			Other		0
1	VBS Mutual Bank		VBS MUTUAL BANK	Bank	Peer 1	1
2	Absa Group			Bank	Group	0
3	Absa Bank Ltd	I0034118	ABSA BANK LTD	Bank	Peer 1	0
4	Firststrand Ltd			Bank	Group	0
5	Firststrand Bank Ltd	I0416053	FIRSTSTRAND BANK LIMITED	Bank	Peer 1	0
6	Investec Group			Bank	Group	0
7	Investec Bank Ltd	I0025054	INVESTEC BANK LTD	Bank	Peer 1	0
8	Nedbank Group			Bank	Group	0
9	Nedbank Ltd	I0416088	NEDBANK LTD	Bank	Peer 1	0
10	Standard Bank Group			Bank	Group	0
11	Standard Bank of S A Ltd	I0416061	THE STANDARD BANK OF S A LTD	Bank	Peer 1	0
12	Bnp Paribas South Africa Branch	I0478636	BNP PARIBAS SOUTH AFRICA BRANCH	Bank	Peer 2	0
13	Citibank N.A	I0165565	CITIBANK N.A	Bank	Peer 2	0
14	Deutsche Bank AG	I0333778	DEUTSCHE BANK AG	Bank	Peer 2	0
15	HSBC Bank Plc - Johannesburg Branch	I0535761	HSBC BANK PLC - JOHANNESBURG BRANCH	Bank	Peer 2	0
16	JP Morgan Chase and Co	I0331899	JPMORGAN CHASE BANK (JOHANNESBURG BRANCH)	Bank	Peer 2	0
17	Societe Generale Johannesburg Branch	I0025046	SOCIETE GENERALE JOHANNESBURG BRANCH	Bank	Peer 2	0
18	Standard Chartered Bank	I0535788	STANDARD CHARTERED BANK	Bank	Peer 2	0
19	Albaraka Bank Ltd	I0110728	ALBARAKA BANK LTD	Bank	Peer 3	0
20	Bidvest Bank Ltd	I0025070	BIDVEST BANK LIMITED	Bank	Peer 3	0
21	Ithala Ltd	I0329568	ITHALA LIMITED	Bank	Peer 3	0
22	Mercantile Bank Ltd	I0416096	MERCANTILE BANK LTD	Bank	Peer 3	0
23	Sasfin Bank Ltd	I0335487	SASFIN BANK LTD	Bank	Peer 3	0
24	Fairfax Africa			Bank	Group	0
25	S A Bank of Athens Ltd	I0416134	THE S A BANK OF ATHENS LTD	Bank	Peer 3	0
26	Bank of Baroda	I0331562	BANK OF BARODA	Bank	Peer 4	0

Continued on next page

Table 4.2 – *Continued from previous page*

Co	Company Name	Ref	Master Company Name	Type	Sub Type	Risk
27	Bank of China Ltd Jhb Branch	I0337889	BANK OF CHINA LTD JHB BRANCH (T/A BANK OF CHINA JHB BRANC	Bank	Peer 4	0
28	Bank of India - Johannesburg Branch	I0477842	BANK OF INDIA - JOHANNESBURG BRANCH	Bank	Peer 4	0
29	Bank of Taiwan South Africa Branch	I0148520	BANK OF TAIWAN SOUTH AFRICA BRANCH	Bank	Peer 4	0
30	Canara Bank	I0154679	CANARA BANK	Bank	Peer 4	0
31	China Construction Bank Corporation - Jhb Branch	I0341037	CHINA CONSTRUCTION BANK CORPORATION - JHB BRANCH	Bank	Peer 4	0
32	ICICI Bank Ltd	I0159158	ICICI BANK LIMITED	Bank	Peer 4	0
33	State Bank of India	I0170798	STATE BANK OF INDIA	Bank	Peer 4	0
34	African Bank Ltd (N)	I0160571	AFRICAN BANK LIMITED (N)	Bank	Peer 5	0
35	African Bank Ltd (O)	I0024988	AFRICAN BANK LIMITED (O)	Bank	Peer 5	1
36	Capitec Bank	I0333107	CAPITEC BANK	Bank	Peer 5	0
37	Grindrod Bank	I0155438	GRINDROD BANK	Bank	Peer 5	0
38	Habib Overseas Bank Ltd	I0129593	HABIB OVERSEAS BANK LTD	Bank	Peer 5	0
39	Hbz Bank Ltd	I0165816	HBZ BANK LTD	Bank	Peer 5	0
40	Ubank Ltd	I0341894	UBANK LIMITED	Bank	Peer 5	0

Only two banks, VBS Mutual Bank and African Bank Investment Ltd (ABIL) were labelled with the binary risk flag. The 2017 financial statements of VBS Mutual Bank proved fraudulent and they were placed under curatorship the following year [50]. ABIL was deemed a South African corporate governance failure [44] and was placed under curatorship in August 2014. Given the curatorship, it was assumed that risk was mitigated after the 2015 financial year. To deal with the time-sensitivity of the risk, ABIL was represented as two distinct companies – the old company got suffix (O) and a risk flag (positive) and the new company got suffix (N) and no risk flag (negative).

4.3.3 Document mapping and labelling

Annual report filenames varied across companies and across reports for the same company. For example,

- ICICI Bank annual report-subsiidiaries-ar-2013-14.pdf
- ICICI Bank annual-report-fy2017 (2016).pdf

Table 4.3 is an example of the CSV mapping file used to assign metadata like year and company.

Table 4.3: First 10 rows of the document mapping file.

Doc ID	Document Name	Path Root	Year	Co ID	Company Name	To Do
0	2009-ABSA Group annual-report.pdf	../data/raw/Bank /ABSA	2009	2	Absa Group	1
1	2010- ABSA Group annual-report.pdf	../data/raw/Bank /ABSA	2010	2	Absa Group	1
2	2009-12-31-Absa-Bank-Annual-Report.pdf	../data/raw/Bank /ABSA/Other	2009	3	Absa Bank Ltd	1
3	2010-12-31-Absa-Bank-Annual-Report.pdf	../data/raw/Bank /ABSA/Other	2010	3	Absa Bank Ltd	1
4	1-african-bank-holdings-ir-2017.pdf	../data/raw/Bank/African Bank	2017	34	African Bank Ltd (N)	1
5	1464958532-AfricanBank_IR_FY20131.pdf	../data/raw/Bank/African Bank	2013	35	African Bank Ltd (O)	1
6	1464958705-4144_African-Bank-2014annualreport.pdf	../data/raw/Bank/African Bank	2014	35	African Bank Ltd (O)	1
7	1464958761-4144_African-Bank-2015annualreport.pdf	../data/raw/Bank/African Bank	2015	34	African Bank Ltd (N)	1
8	1517467421-ABILIAR2012 completelowres.pdf	../data/raw/Bank/African Bank	2012	35	African Bank Ltd (O)	1
9	2009 Full report - PDF 7MB - African Bank.pdf	../data/raw/Bank /African Bank	2009	35	African Bank Ltd (O)	1

Not only was it simpler, but the original filename could be retained for traceability to the source. Document identifiers allowed consistent reference through processing iterations, i.e. documents could be appended without duplicating and could also be re-processed. Since ABIL was placed under curatorship in August 2014, it was assumed that the annual report of 2015 may still reflect risk. Annual reports from 2016 were

mapped to the new distinct company sans risk flag. *ToDo=1* flagged 258 documents as appropriate and to be processed. Only 8 of these were mapped to companies tagged with risk resulting in significant imbalance between the two classes.

4.4 Data preparation and analysis

A limitation of data preparation method 1 and 2 is that they assumed one file per annual report. Results were influenced by this limitation in that constituent documents in Table B.4 were analysed and compared to other documents as if they were the full annual report comprising all sections of the annual report.

4.4.1 Booklet identification

Table 4.4 shows that the programme identified 3 files starting with front and back covers, and 12 documents were booklets.

Table 4.4: Number of documents per page ratio.

Page Ratio	Number of Documents	Meaning
3	0	Page0 is cover consisting of front and back pages
347	1	Page0 and Page1 are the same width
12	2	Page1 is twice as wide as Page0 i.e. is booklet

This confirmed that the 2016 report of VBS Mutal Bank was not isolated and that booklets had to be processed automatically. Had there only been the VBS 2016 report (see Figure B.2) then a manual conversion could have been attempted. However, this would not have led to a scalable and repeatable pipeline for creating the corpus.

4.4.2 Text extraction

PyMuPDF's default text extraction did not maintain the natural human reading order in processing multiple columns. The first row of each column was read together instead

of reading within each column. This resulted in segments of sentences being combined that don't belong together. Even though the CFIE–FRSE tool demonstrated the same behaviour, method 2 was an improvement.

Similarly, the treatment of spaces in the extracted text was manually checked to optimise the integrity of individual words (see Appendix B.3.1).

A total of 60 825 pages were saved to a 347MB CSV file. Merging *year*, *coID*, *coName* and *risk* columns from the document and company mapping files had minimal impact on the file size which grew to 349 MB.

4.4.3 Word count

The UK reference report was included with the SA report “VBS-2015-Annual-Report.pdf” in the first iteration for reference control (see Table 4.2).

The new tool counted one additional word compared with the CFIE–FRSE tool for each of the sections comprising a single page of the UK reference report. Results for PDF pages with zero-based numbering were as follows:

1. Page 3 (About Ai Claims Solutions) counted 642 words instead of 641;
2. Page 18 (Finance Director's Review) counted 1102 words instead of 1101;
3. Page 19 (Board of Directors) counted 691 instead of 692.

This word count variance was less than 0.2%. The variance was not seen as material to the study and the word count functionality of the new tool was accepted.

The ‘Performance Overview’ on page 18 of the VBS report returned a word count of 721. This was validated against the length of the `getText('words')` list and splitting on the original `getText('blocks')` text.

The regular expression `re.findall(r'+', text)` over counted due to splitting initials like “P.N.” and contractions like “it's” (see Appendix B.3.8).

4.4.4 Wordlist count

Text formats and word count methods also impacted the computation time. For instance, computation on blocks of text extracted from the page was estimated at 16 days. Instead,

blocks of text were concatenated with a space after which computation took 23h 40m.

Concatenated text and word counts for 60 825 pages was saved to a 211MB CSV file. The file was 30% smaller than the file for text blocks alone. This was due to dropping the block lists and their rectangle coordinates.

4.4.5 Readability score

The `py-readability-metrics` (`r`) package was installed to import the Readability function. Text extracted from the UK reference report was passed to the function for each page. Scores were compared against the CFIE results. Given some discrepancies, the `textstat` (`ts`) package was also installed and compared (see Appendix B.3.2).

As discussed in Section 2.4.2, the largest Gunning Fog values represent the most unreadable annual reports. Table 4.5 lists the most unreadable reports using the Gunning Fog metric from the `py-readability-metrics` package.

Table 4.5: Top 5 hard reading reports using `py-readability-metrics` Gunning Fog.

Company		Document			r_
ID	Name	ID	Name	Year	gf
30	Canara Bank	38	Canara Bank Annual Report 2009 (5324830310).pdf	2009	105074
30	Canara Bank	39	Canara Bank Annual Report 2010 (5324830310).pdf	2010	100414
30	Canara Bank	40	Canara Bank Annual Report 2011 (5324830311).pdf	2011	100414
2	Absa Group	1	2010- ABSA Group annual-report.pdf	2010	106
26	Bank of Baroda	36	Bank of Baroda Annualreport2017-18.pdf	2018	35

Canara Bank and the Bank of Baroda are both Indian banks with branches operating in South Africa. Hence, four of the five annual reports are of banks whose origin is a non-English speaking country. All annual reports from these two banks are written in two languages, English and their native language. Furthermore, the native language does not use the Latin character set (see Figure B.7). It is understandable that annual reports written by banks from non-English speaking countries may not read easily. However, it is not clear why 3 of the 10 reports of Canara Bank have such high Gunning Fog values. ABSA Group, on the other hand, is one of the four large retail banks in South Africa

and its 2010 annual report is even less readable than the 2018 annual report of the Bank of Baroda.

The next table again used the Gunning Fog metric to list the most unreadable reports but this time using the `textstat` package.

Table 4.6: Top 5 hard reading reports using `textstat`'s Gunning Fog.

Company		Document			ts_ gf
ID	Name	ID	Name	Year	
30	Canara Bank	38	Canara Bank Annual Report 2009 (5324830310).pdf	2009	37716
30	Canara Bank	40	Canara Bank Annual Report 2011 (5324830311).pdf	2011	35936
30	Canara Bank	39	Canara Bank Annual Report 2010 (5324830310).pdf	2010	35936
22	Mercantile Bank Ltd	246	NASDAQ_MBWM_2015.pdf	2015	157
22	Mercantile Bank Ltd	243	NASDAQ_MBWM_2012.pdf	2012	142

The three annual reports for Canara Bank for 2009-2011 appear using the `textstat` package as well but the values are a third of what the `py-readability-metrics` package calculated. The 2012 and 2015 annual reports of Mercantile Bank Ltd were highlighted by the `textstat` package even though they did not appear in the top-5 list using the `py-readability-metrics` package.

As discussed in Section 2.4.2, the smallest Flesch Reading Ease values represent the most unreadable annual reports. Table 4.7 lists the most unreadable reports using the Flesch Reading Ease metric from the `py-readability-metrics` package.

Table 4.7: Top 5 hard reading reports using `py-readability-metrics` Flesch Reading Ease.

Company		Document			r_ f
ID	Name	ID	Name	Year	
30	Canara Bank	38	Canara Bank Annual Report 2009 (5324830310).pdf	2009	-266503
30	Canara Bank	40	Canara Bank Annual Report 2011 (5324830311).pdf	2011	-254678
30	Canara Bank	39	Canara Bank Annual Report 2010 (5324830310).pdf	2010	-254678
2	Absa Group	1	2010- ABSA Group annual-report.pdf	2010	-154
23	Sasfin Bank Ltd	273	sasfin-integrated-report-2014.pdf	2014	8

Again the three annual reports for Canara Bank for 2009-2011 as well as the 2010

report of ABSA Group appear using the Flesch Reading Ease metric. Another South African bank makes this top 5 list of most unreadable reports, namely Sasfin Bank Ltd through their 2014 annual report.

The next table again used the Flesch Reading Ease metric to lists the most unreadable reports but this time using the `textstat` package.

Table 4.8: Top 5 hard reading reports using textstat’s Flesch Reading Ease.

Company		Document			ts_
ID	Name	ID	Name	Year	f
30	Canara Bank	38	Canara Bank Annual Report 2009 (5324830310).pdf	2009	-95575
30	Canara Bank	39	Canara Bank Annual Report 2010 (5324830310).pdf	2010	-91057
30	Canara Bank	40	Canara Bank Annual Report 2011 (5324830311).pdf	2011	-91057
22	Mercantile Bank Ltd	246	NASDAQ_MBWM_2015.pdf	2015	-376
22	Mercantile Bank Ltd	243	NASDAQ_MBWM_2012.pdf	2012	-329

The 2009 to 2011 annual reports of Canara Bank appear in all four tables. As with the Gunning Fog index, the `textstat` package again highlights the 2012 and 2015 reports of Mercantile Bank Ltd - this time with the Flesch Reading Ease.

As discussed in Section 3.4.5, the largest Text Standard values represent the most unreadable annual reports. Table 4.9 lists the most unreadable reports using the Text Standard metric from the `textstat` package.

Table 4.9: Top 5 hard reading reports using textstat’s Text Standard.

Company		Document			ts_
ID	Name	ID	Name	Year	ts
30	Canara Bank	38	Canara Bank Annual Report 2009 (5324830310).pdf	2009	36769
30	Canara Bank	40	Canara Bank Annual Report 2011 (5324830311).pdf	2011	35033
30	Canara Bank	39	Canara Bank Annual Report 2010 (5324830310).pdf	2010	35033
22	Mercantile Bank Ltd	246	NASDAQ_MBWM_2015.pdf	2015	163
22	Mercantile Bank Ltd	249	NASDAQ_MBWM_2018.pdf	2018	73

The proposed consensus value across all metrics in `textstat` confirmed the 2009 to 2011 annual reports of Canara Bank as well as the 2015 report of Mercantile Bank Ltd

that appeared using the Gunning Fog and Flesch Reading Ease metrics. However, the 2018 report of Mercantile Bank Ltd was suggested as the fifth most difficult report to read (as opposed to their 2012 report highlighted by the Flesch Reading Ease metric.)

Ignoring scores = 0, Table 4.10 shows the top 5 reports that were easiest to read using the Text Standard metric of the `textstat` package.

Table 4.10: Top 5 easy reading reports using textstat’s Text Standard.

Company		Document			ts
ID	Name	ID	Name	Year	ts
32	ICICI Bank Ltd	188	Subsidiaries_of_ICICI_Bank_Annual_Report.FY2019 2018.pdf	2019	4
23	Sasfin Bank Ltd	277	SASFIN_ar_jun16.pdf	2016	8
22	Mercantile Bank Ltd	240	NASDAQ-MBWM-2009.pdf	2009	12
1	VBS Mutual Bank	357	VBS-2015-Annual-Report.pdf	2015	13
7	Investec Bank Ltd	198	Limited-Annual-Report-FA 2012.pdf	2012	13

The easiest report to read according to the Text Standard metric was the 2019 annual report of ICICI Bank Ltd. The metric suggests that it could be understood by a grade 4 pupil in a US school. Similarly, it suggests that the 2016 annual report of Sasfin Bank Ltd can be read by a grade 8 pupil and the 2009 annual report of Mercantile Bank Ltd can be read by a grade 12 pupil. The results suggest that the annual reports of VBS Mutual Bank and Investec Bank Ltd can be understood by a first year university student. in terms of VBS Mutual Bank, the authors could have made their annual reports difficult to read in an attempt to obfuscate information that could point towards fraudulent business. It was therefore unexpected to observe the 2015 report of VBS Mutual Bank in the list of top-5 easiest reports to read.

As discussed in Section 3.4.12 a wrapper method was used initially to assess these readability scores in combination with the word counts (Section 4.4.4) wordlist counts (Section 4.4.5). Three tests for filter feature selection were also explored in Appendix A.2, namely Chi-Squared dependence, Information Gain and ANOVA.

4.4.6 Header identification

The challenge was to identify sufficient headers in comparison to the CFIE–FRSE tool as well as being able to identify sufficient headers in the few positive class SA reports. As such, the Python notebook was developed incrementally and iteratively as follows:

1. AI Claim Solutions 2007 was selected as the UK reference (1 report);
2. VBS Mutual Bank was selected as the SA bank positively labelled for risk;
 - (a) Annual reports included were for the years 2015 and 2016 (2 reports);
3. JP Morgan Chase and Co was selected as the SA bank negatively labelled for risk;
 - (a) At the time of developing the notebook, JP Morgan Chase and Co was a bank for which a full set of annual reports had been collected;
 - (b) Annual reports included in header identification were for the years 2009 to 2019 (11 reports);
4. The full set of collected annual reports for SA Banks.

A total of 24 257 bookmarks were extracted from 93 annual reports. However, for 51 of these reports the bookmark text did not provide the header text. Only 42 of 93 reports had the word ‘contents’ in a bookmark.

The result was that 167 of 209 annual reports did not have tables of content properly bookmarked in the PDF file.

Further challenges in identifying headers related to the layout and format of the contents page which is not always a table and is sometimes untitled (see Appendix [B.3.3](#)).

Algorithm [3.2](#) was enhanced by adding an early stop condition once the first 20% of the pages had been read. Since text blocks were extracted whilst preserving white space, values like ‘c o n t e n t s’ had to have spaces replaced before doing the table of content wordlist match. A total of 1 744 headers were extracted manually and appended to the 4 084 bookmarks.

4.4.7 Header classification

Only 161 of the 5 828 headers could be classified using fuzzy lookups of types defined in Table 3.7. As an example of where headers could be classified, Table 4.11 shows ten headers with their associated class, the item in the wordlist that matched best and the match score.

Table 4.11: Sample of header text classification.

Document		Header		Matching	
ID	Name	Text	Class	Item	Score (%)
48	BANK OF CHINA LTD JHB BRANCH 2015 Annual Report	Corporate Social Responsibilities	CSR disclosures	corporatesocialresponsibility	81
48	BANK OF CHINA LTD JHB BRANCH 2015 Annual Report	Corporate Governance	Governance statement	corporategovernance	87
48	BANK OF CHINA LTD JHB BRANCH 2015 Annual Report	Independent Auditors' Report	Auditors report	independentauditorsreport	83
76	capitec.integrated annualreport.2011.pdf	Chairman's letter	Chairman's statement	chairmanletter	84
76	capitec.integrated annualreport.2011.pdf	Corporate govenance	Governance statement	corporategovernance	89
355	UBank-Integrated-Report- Full.pdf	Annual financial statements	Financial review	financialstatement	80
357	VBS-2015-Annual-Report.pdf	Chief Executive report	CEO review	chiefexecutivesreport	84
357	VBS-2015-Annual-Report.pdf	Independent auditor's report	Auditors report	independentauditorsreport	91
358	VBS-2016-Annual-Report.pdf	chief executive officer's review	CEO review	chiefexecutiveofficerreview	92
358	VBS-2016-Annual-Report.pdf	Independent auditor's report	Auditors report	independentauditorsreport	91

A score of 75% was used as the threshold for classifying a header. Lower scores led to incorrect classification. Table 4.12 shows the frequency of header when filtering on

keyword combinations.

Table 4.12: Header type frequency.

No	Header Type Description	Filter words	Count
1	Letter from board chair	‘board’ and ‘report’	80
		‘board’ and ‘statement’	
		‘board’ and ‘letter’	
		‘chair’ and ‘report’	
		‘chair’ and ‘statement’	
		‘chair’ and ‘letter’	
2	CEO review	‘exec’ and ‘report’	29
		‘exec’ and ‘statement’	
		‘exec’ and ‘review’	
		‘ceo’ and ‘report’	
		‘ceo’ and ‘statement’	
		‘ceo’ and ‘review’	
3	Governance statement	‘govern’ and ‘report’	53
		‘govern’ and ‘statement’	
		‘govern’ and ‘review’	
4	Remuneration report	‘remuneration’ and ‘report’	30
		‘remuneration’ and ‘statement’	
		‘remuneration’ and ‘review’	
5	Business review	‘business’ and ‘report’	13
		‘business’ and ‘statement’	
		‘business’ and ‘review’	
6	Financial review	‘financial’ and ‘report’	180
		‘financial’ and ‘statement’	
		‘financial’ and ‘review’	

Continued on next page

Table 4.12 – *Continued from previous page*

No	Header Type Description	Filter words	Count
7	Operating review	‘operating’ and ‘report’	4
		‘operating’ and ‘statement’	
		‘operating’ and ‘review’	
8	Highlights	‘highlight’	66
9	Auditors report	‘independent’ and ‘auditor’	66
		‘auditor’ and ‘report’	
10	Risk management	‘risk’ and ‘report’	199
		‘risk’ and ‘manage’	
11	Chairman’s governance	‘chair’ and ‘gov’	0
12	CSR disclosures	‘sustain’	69
		‘social’	
13	Audit committee report	‘audit’ and ‘committee’	37
14	Company secretary declaration	‘secretar’	35

With only four occurrences, the SA corpus of annual reports does not typically contain sections for the ‘Operating review’. Similarly, none of the SA reports had a section for ‘Chairman’s governance’. Conversely, SA annual reports had sections for the ‘Audit committee report’ and ‘Company secretary declaration’ that were not configured with the tool for UK reports. Challenges experienced in classifying headers is also attributed to diverse structures and header text used across the SA reports. Financial reporting and regulation in South Africa will benefit from companies improving the quality of their PDF publishing and driving conformance of the header/section structure of annual reports even if it is not completely standardised.

4.4.8 Section aggregation

There were 113 of the 258 reports for which headers could not be identified. When headers were identified, words were counted between headers for each section.

Table 4.13 shows the tone of each section within the following annual reports:

1. AI Claim Solutions 2007 (the UK reference report);

2. VBS Mutual Bank 2015;
3. VBS Mutual Bank 2016.

Table 4.13: Tone for UK reference report and VBS Mutual Bank reports by section.

	LMtone	H6tone	H8tone
About Ai Claims Solutions	-0.3	1.0	1.0
Why our ethical business model is a recipe for success	0.28	0.33	0.33
Questions and Answers David Sandhu, Chief Executive Officer	0.39	0.76	0.75
How have we performed?	-0.18	0.58	0.58
Chairman's Statement	0.53	0.85	0.85
Chief Executive's Review	0.38	0.85	0.85
Finance Director's Review	0.27	0.54	0.53
Board of Directors	0.41	1.0	1.0
Executive Management	-0.5	1.0	1.0
Remuneration Report	0.56	0.57	0.57
Report of the Directors	-0.48	0.0	0.0
Report of the Independent Auditor	-0.75	-1.0	-1.0
Consolidated Profit and Loss Account, Consolidated Statement of Total Recognised Gains and Losses	-0.67	nan	nan
Consolidated Balance Sheet	-0.71	-0.5	-0.5
Company Balance Sheet	-0.71	-0.5	-0.5
Consolidated Cash Flow Statement, Reconciliation of Net Cash Flow to Movement in Net Debt	-0.43	0.33	0.33
Notes to the Financial Statements	-0.4	0.0	0.0
Five Year Summary	-0.71	-1.0	-1.0
Notice of Annual General Meeting Ai Claims Solutions PLC	-0.32	-0.04	0.0
Officers and Professional Advisers	-0.6	1.0	1.0
Introduction to VBS	0.46	0.47	0.47
Governance	-0.44	-0.59	-0.59
Chairman's report	-0.19	-0.45	-0.45
Audit and Risk Committee report	-0.4	-0.77	-0.77
Chief Executive report	0.33	0.43	0.54
Performance Overview	0.22	0.47	0.47
Directors' responsibility statement	-0.5	-1.0	-1.0
Directors' report	-0.14	0.2	0.2
Independent auditor's report	-0.82	-1.0	-1.0
Statement of financial position	1.0	nan	nan
Statement of comprehensive income	-0.71	nan	nan
Statement of cash flows	0.6	0.6	0.6
Statement of changes in shareholders equity	-1.0	nan	nan
Notes to the annual financial statements	-0.38	-0.35	-0.35
introduction to VBs	0.23	0.13	0.11
Governance	-0.41	-0.67	-0.67
chairman's report	0.22	0.2	0.2
Audit and risk committee report	-0.35	-0.75	-0.75
chief executive officer's review	0.38	0.33	0.33
Performance overview	0.23	0.59	0.59
Directors' responsibility statement	nan	nan	nan
Directors' report	-0.43	0.0	0.0
Independent auditor's report	-0.8	-1.0	-1.0
Statement of financial position, Statement of comprehensive income	-0.2	nan	nan
Statement of cash flows, Statement of changes in shareholders equity	0.33	0.33	0.33
Notes to the annual financial statements	-0.44	-0.48	-0.48

Different measures of tone were used as it was unclear which may be best suited to the language used in UK annual reports. The three measures of tone were:

1. LMtone: Loughran and McDonald [42];
2. H6tone: Henry [30];
3. H8tone: Henry [29].

The pink shading for both of the Henry measures illustrates that initial sections of the UK reference report are positive as well as the 'Officers and Professional Advisors'. By contrast, the SA reports have fewer sections with positive tone and the tone scores

are also lower. The ‘Statement of financial position’ of VBS in 2015 was the only section to score +1. Since this was for LMtone and both Henry tones could not be calculated, this result was ignored. ‘Performance overview’ was then the most positive sections. The ‘Chief Executive report’ was similarly positive for 2015 but was less positive for 2016 on the Henry measures. This could be an indicator of a negative trend.

It should be noted that the two South African annual reports are those of a bank whilst the UK reference report) is for an insurer. Annual reports of companies from different industry sectors may not be directly comparable. Further discussion on the net tone comparison as well as benchmarking institutions against their peers can be found in Sections 6.1.4 and 7.2.1.

Tables 4.14 - 4.16 show scaled counts for forward-looking, uncertain, causal, performance and strategic words for the same three reports.

Table 4.14: Scaled counts by section for UK reference.

headText	forward2	uncert	causalM	perf	strat
About Ai Claims Solutions	0.78%	3.74%	0.31%	1.09%	4.67%
Why our ethical business model is a recipe for success	0.66%	8.28%	1.32%	1.99%	13.58%
Questions and Answers David Sandhu, Chief Executive Officer	1.41%	4.54%	0.43%	0.22%	8.97%
How have we performed?	0.25%	3.68%	0.25%	3.68%	1.47%
Chairman’s Statement	1.13%	4.97%	1.13%	1.72%	7.81%
Chief Executive’s Review	1.32%	4.94%	0.53%	0.71%	7.81%
Finance Director’s Review	0.64%	3.99%	2.45%	2.36%	3.18%
Board of Directors	0.14%	1.74%	0.00%	0.14%	4.49%
Executive Management	0.22%	0.89%	0.22%	0.67%	6.25%
Remuneration Report	1.09%	3.83%	0.18%	0.55%	5.46%
Report of the Directors	0.98%	1.69%	0.21%	0.42%	2.53%
Report of the Independent Auditor	0.28%	1.38%	0.14%	0.55%	0.69%
Consolidated Profit and Loss Account_Consolidated Statement of Total Recognised Gains and Losses	0.00%	1.42%	0.47%	7.58%	1.42%
Consolidated Balance Sheet	0.00%	0.55%	0.00%	1.10%	0.55%
Company Balance Sheet	0.00%	0.57%	0.00%	1.14%	0.57%
Consolidated Cash Flow Statement_Reconciliation of Net Cash Flow to Movement in Net Debt	0.00%	0.78%	0.00%	0.00%	0.78%
Notes to the Financial Statements	0.56%	1.45%	0.27%	1.47%	1.38%
Five Year Summary	0.00%	2.00%	1.00%	6.50%	1.00%
Notice of Annual General Meeting Ai Claims Solutions PLC	2.37%	1.81%	0.33%	0.09%	0.88%
Officers and Professional Advisers	0.00%	1.07%	0.00%	0.00%	2.41%

Table 4.15: Scaled counts by section for VBS Mutual Bank 2015.

headText	forward2	uncert	causalM	perf	strat
Introduction to VBS	0.76%	4.26%	0.15%	0.76%	5.17%
Governance	1.17%	2.99%	0.39%	0.09%	5.33%
Chairman's report	0.63%	2.35%	0.39%	0.55%	5.33%
Audit and Risk Committee report	0.39%	2.96%	0.73%	0.22%	3.69%
Chief Executive report	1.62%	2.58%	0.65%	1.29%	4.85%
Performance Overview	0.83%	7.07%	3.05%	2.91%	3.74%
Directors' responsibility statement	0.81%	2.42%	0.40%	0.40%	2.82%
Directors' report	0.00%	0.46%	0.23%	0.46%	1.16%
Independent auditor's report	0.35%	1.74%	0.35%	0.17%	2.78%
Statement of financial position	0.00%	1.63%	0.00%	0.00%	2.17%
Statement of comprehensive income	0.00%	0.62%	0.00%	9.26%	0.00%
Statement of cash flows	0.00%	2.06%	0.00%	0.41%	7.00%
Statement of changes in shareholders equity	0.00%	0.00%	0.00%	2.55%	1.53%
Notes to the annual financial statements	0.68%	2.38%	0.50%	1.83%	3.07%

Table 4.16: Scaled counts by section for VBS Mutual Bank 2016.

headText	forward2	uncert	causalM	perf	strat
introduction to VBs	0.24%	2.61%	0.04%	0.28%	4.20%
Governance	0.89%	2.63%	0.31%	0.10%	5.46%
chairman's report	1.39%	3.78%	0.80%	0.40%	5.37%
Audit and risk committee report	0.34%	2.40%	0.59%	0.20%	3.09%
chief executive officer's review	1.01%	3.18%	0.43%	1.30%	4.91%
Performance overview	0.63%	5.48%	2.66%	2.35%	3.13%
Directors' responsibility statement	0.56%	1.69%	0.28%	0.85%	2.26%
Directors' report	0.00%	0.63%	0.21%	0.42%	0.84%
Independent auditor's report	0.41%	2.28%	0.41%	0.21%	2.70%
Statement of financial position_Statement of comprehensive income	0.00%	0.85%	0.00%	2.82%	0.85%
Statement of cash flows_Statement of changes in shareholders equity	0.00%	0.61%	0.81%	0.81%	3.64%
Notes to the annual financial statements	0.79%	2.46%	0.56%	1.48%	2.98%

The most forward looking section of the UK reference was the 'Notice of Annual General Meeting' (2.37%). The Chief Executive's report for VBS in 2015 was most forward looking (1.62%) and the Chairman's report for 2016 (1.39%). Note that their relative rank was reversed across these years with the Chairman's report for 2015 (0.63%) and the Chief Executive's report for 2016 (1.01%). Greater uncertainty is evident in VBS's performance overview, 7.07% for 2015 and 5.48% in 2016 in comparison to the UK ref-

erence at 3.68%. The UK reference report had few causal words regarding performance (0.25%) whereas VBS had 3.05% and 2.66% for 2015 and 2016 respectively. This may relate to VBS describing a causal relationship e.g. that something was responsible for causing another thing to happen regarding their performance. The most strategic narrative was the section on ethics in the UK reference report (13.58%). It can also be observed that a number of the sections are more strategic than the VBS reports. Except for the ‘Statement of cashflows’ in 2015, the ‘Governance’ and ‘chairman’s report’ sections for both VBS reports were the most strategic.

4.4.9 Document aggregation

A total of 60 212 non-empty pages were grouped into 261 documents. Three annual reports were removed. Two reports had zero count of keywords in the wordlist and one report was a duplicate (see Appendix B.3.5).

Having aggregated individual pages and word counts by report, it was possible to calculate an overall tone. Table 4.17 contrasts the tones between negatively and positively labelled reports as well as consecutive financial years.

Table 4.17: Tone for selected SA reports.

Report	Risk	LMtone	H6tone	H8tone
JP Morgan Chase and Co - annualreport-2015.pdf	0	-0.20	0.00	-0.01
JP Morgan Chase and Co - annualreport-2016.pdf	0	-0.19	0.01	0.00
VBS-2015-Annual-Report.pdf	1	-0.26	-0.19	-0.19
VBS-2016-Annual-Report.pdf	1	-0.28	-0.26	-0.26

It is evident that the tone of the VBS reports were significantly more negative than the JP Morgan reports. The latter was essentially neutral on both Henry scores. VBS on the other hand, was 19% negative in 2015 and declined to 26% in 2016.

The aggregated scaled word counts for each report from forward-looking, uncertain, causal, performance and strategic wordlists is shown in Table 4.18.

The VBS Mutual Bank annual reports scored lower for every word list. The VBS Mutual Bank annual reports had forward-looking words, performance related words and strategic words all constituting a lower proportion of the overall text compared with

Table 4.18: Scaled counts for selected SA reports.

Report	Risk	forward2	uncert	causalM	perf	strat
JP Morgan Chase and Co - annualreport-2015.pdf	0	1.00%	3.72%	1.04%	1.54%	4.41%
JP Morgan Chase and Co - annualreport-2016.pdf	0	1.03%	3.78%	1.03%	1.51%	4.45%
VBS-2015-Annual-Report.pdf	1	0.70%	2.64%	0.56%	1.31%	3.65%
VBS-2016-Annual-Report.pdf	1	0.67%	2.51%	0.52%	1.04%	3.43%

the UK reference. By not addressing performance and longer term views, the narrative may be an indicator of risk. This is counteracted by VBS Mutual Bank’s fewer words of uncertainty. Had VBS Mutual Bank used many words of uncertainty, then the vagary of the report may be an indicator of risk, however this was not the case.

4.4.10 Validation

The text extraction and word count method developed for SA annual reports proved repeatable and consistent with the UK reference. Variances on word counts per section were relatively small with three sections having no delta (see Appendix B.3.6).

The newly developed code counted the exact number of occurrences of words from the casual wordlist in every section of the AI Claim Solutions 2007 report with the exception of the ‘Chairman’s Statement’ where one less keyword instance was found. Counts for all of these wordlists matched exactly for every section:

- forward1;
- forward2;
- LMneg;
- LMpos;
- H6neg;
- H6pos;
- H8neg;
- H8pos;

- Uncert.

This effectively validated the new text extraction and word counting algorithms and Python code.

4.4.11 Training and test data sets

Years 2018 and 2019 of the data set had 25 and 21 documents respectively with risk=0. African Bank's 2014 annual report and VBS Mutual Bank's 2016 annual report were both labelled with risk=1 and were appended to these to form a test data set of 48 documents, i.e. 18.5% of the data was held out for testing as shown in Appendix B.3.7.

The remaining 210 documents contained 204 documents without risk and 6 documents with risk. The first attempt at training the machine learning models with this original data set could not consistently return F1 scores for the 5-fold cross validation due to the significant imbalance between positive (risk=1) and negative (risk=0) classes.

The second iteration resampled the 6 positive documents with replacement such that there were 204 documents in each class. The third iteration capped the resampled positive documents to 41 resulting in an 80:20 ratio between negative and positive classes.

4.5 Data processing and analysis

4.5.1 Bag of words

Table 4.19 ranks the most common 10 tokens for n-grams in the range 1 to 3.

Table 4.19: Top 10 tokens for n-grams in the range 1 to 3.

Unigram	Count	Bigram	Count	Trigram	Count
financial	171973	fair value	60966	consolidated financial statements	11945
risk	153586	financial statements	44292	annual financial statements	11213
bank	137390	annual report	29061	year ended december	8557
group	124279	risk management	22750	fair value profit	7587
assets	113425	balance sheet	21250	fair value loss	7577
value	101768	financial assets	20640	year ended march	5341
year	99139	profit loss	19555	loans advances customers	5061

Continued on next page

Table 4.19 – *Continued from previous page*

Unigram	Count	Bigram	Count	Trigram	Count
management	94202	credit risk	18659	designated fair value	4659
total	86238	year ended	18579	million million million	4245
credit	85765	loans advances	17100	non executive directors	4019

Given the task of identifying risk from these annual reports, it is interesting to note the frequency with which the unigrams ‘risk’, ‘assets’ and ‘credit’ occur. Similarly for bigrams, ‘risk management’, ‘financial assets’ and ‘credit risk’. The trigram ‘million million million’ is likely to have resulted from extracting text from tabular financial statements within the annual report.

4.5.2 Topic analysis

For all n-gram experiments assuming one topic in the “Audit and risk committee report” section of the VBS Mutual Bank 2016 annual report, wordclouds based on Latent Dirichlet Allocation (LDA) of the TFIDF features closely matched wordclouds for TFIDF without the LDA. Figure 4.1 contrasts the results for both approaches on unigrams.

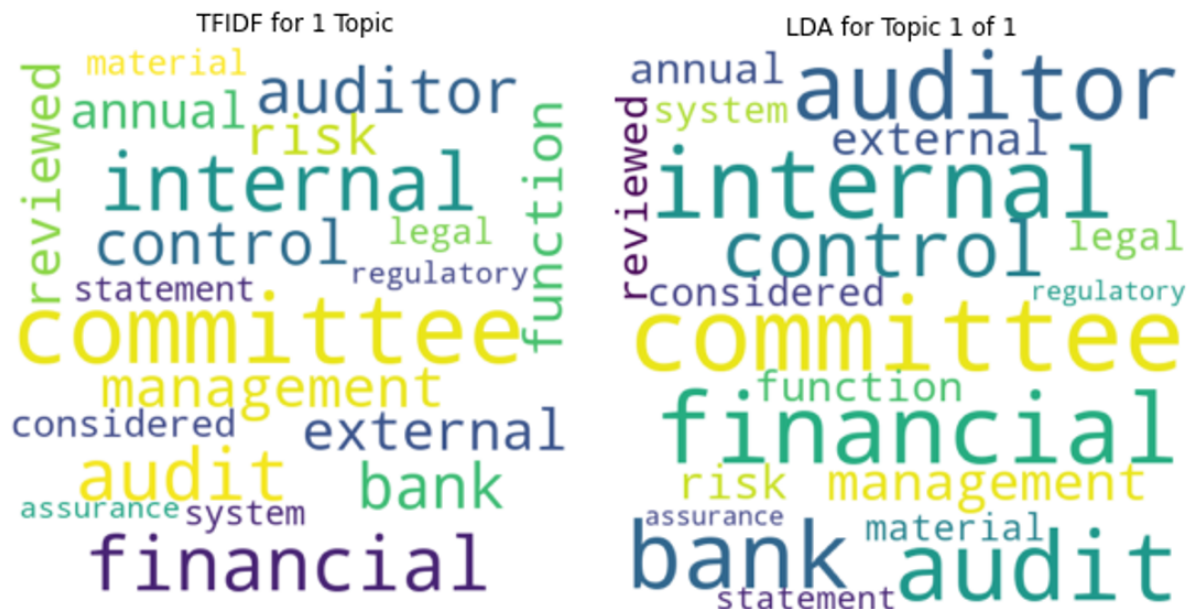


Figure 4.1: Unigram wordclouds for one topic in VBS Bank Audit committee report 2016.

The close match provided a degree of validation for the LDA used for the remainder of the topic analysis. The next figure contrasts bigrams and trigrams using LDA.



Figure 4.2: 2- and 3-gram wordclouds for one topic in VBS Bank Audit committee report 2016.

Noteworthy features of the trigram wordclouds for one topic were:

- ‘breakdown internal control’;
- ‘material breakdown internal’;
- ‘information technology risk’.

See Appendix A for wordclouds assuming a range of topics from 2 to 5 as well as plots using t-distributed Stochastic Neighbour Embedding (t-SNE).

4.6 Summary

Section 4.1 showed that it was not viable to re-use the CFIE–FRSE tool that was developed for the corpus of UK annual reports. Section 4.2 itemised the 258 annual reports

manually sourced and downloaded from bank websites. Challenges like company names, gaps in report availability and inconsistent report numbers were highlighted.

Section 4.3 revealed that the AI Claim Solutions 2007 report from the UK was included with the SA reports for reference control purposes. It also explained why VBS Mutual Bank and African Bank Investment Ltd were annotated with risk leading to 8 annual reports with the risk flag.

The results of data preparation were given in Section 4.4. The Python code was developed to extract text inclusive of 12 reports being in booklet format. Over 60 000 pages of text was extracted. The text block method proved more useful and appropriate than the default text extraction. Overall word counts and wordlist counts correlated well with the CFIE–FRSE reference providing sufficient validation of the text extraction and word count mechanism. Header identification was challenging and limited to 56% of the documents. Even though classifying headers and aggregating pages into sections was not possible across the corpus, counts could be validated against the UK reference and the VBS Mutual Bank annual reports were assessed as having higher uncertainty in its performance overview section. Pages aggregated into documents showed that VBS Mutual Bank’s annual reports had more negative tone than JP Morgan and lower scores for 5 linguistic features.

‘Risk management’ and ‘credit risk’ were among the most common tokens listed in Section 4.5, whilst topic analysis uncovered ‘material breakdown internal’ and ‘information technology risk’.

Chapter 5

Modelling

This chapter shares the outcome of executing the modelling steps from Chapter 3.

- Section 5.1 sets out the modelling goals;
- Section 5.2 compares the performance of the classifiers;
- Section 5.3 applies classifiers to predict documents with highest risk;
- Section 5.4 provides some model interpretation;
- Section 5.5 predicts answers from the corpus to questions using pre-trained models;
- Section 5.6 summarises the end results of modelling experiments.

See Appendix B for more detailed results that complement the tables and figures in this chapter.

5.1 Goals

The goals for modelling started with overcoming the challenge of highly imbalanced classes in a small data set. Training the models assumed the calculation of F1 scores for five-fold cross-validation. The following section on modelling results explains that these calculations were seldom possible on the original data set.

The main modelling goal was to discover and learn the conditions under which classifiers could be trained to predict the few risk-positive documents in the test set. The intention was to minimise the number of features and ensure that the classifier generalises well. The aim was an F1 score of unity to ensure that the labelled documents could be identified without predicting false-positives. This would answer the research question *Can risk be identified in unstructured company annual reports using Natural Language Processing (NLP)?*

Once models were trained on content features, the next goal was to determine which linguistic features to add to improve classifier performance.

5.2 Results

The models were unable to train adequately on the original training data set of 210 annual reports. Positive samples either did not occur within the fold of cross-validation or the models were unable to predict the sample leading to warnings. Ultimately neither of the two positive samples could be identified within the 48 test reports (see Appendix B.4.3). The results quoted in this section are for the 50:50 training set. The six positive training documents were sampled with replacement to result in balanced binary classes, i.e. 204 non-risk documents and 204 risk documents.

Eight models were able to classify the two annual reports from among 48 test documents without predicting false positives. CNN2 predicted one of the two risk documents using either TF, TFIDF or word embedding. CNN1a did the same using either TF or TFIDF. All classifiers generalised better with the min-max normalisation (Equation (3.3)) except for the LSTM models that were more suited to standardisation using the z-score of Equation (3.4).

The simple Logistic Regression (LR) attained a test F1=1 using unigram TF or TFIDF for only 40 features. To sustain this performance, the features limit had to be increased to 60 when n-grams in the range 2 to 6 were allowed. The LR model with word embeddings did not generalise well and could not equal this test performance. The Support Vector Machine with auto scaling (SVMa) achieved the same with 40 unigram TFIDF features, but required 100 unigram TF features. Adding other n-grams reduced

generalisation performance unless 80 or more TFIDF features were allowed. The SVMa also got test $F1=1$ using the mean transform of CBOW embeddings.

The Support Vector Machine with scaling specified (SVMs) did not perform as well with BoW, requiring 100 unigram TF features for a $F1=1$ test result.

Four of the Convolutional Neural Network models (CNN1b, CNN1c, CNN1d and CNN3) also accomplished test $F1=1$. As with the LR and SVMa models, CNN1b, CNN1c and CNN1d did this with TF and TFIDF of 40 unigrams. CNN1b maintained this performance on TFIDF in adding n-grams in the range 3 to 6. Test $F1$ reached unity using the mean transform of CBOW for words occurring at least 4 times and an embedding size of 500. CNN1c with 40 TFIDF features kept test $F1=1$ when adding n-grams in the range 2 to 6. It also used a smaller vector of size 300 for the mean transform of CBOW by counting words from 5 occurrences. CNN1d required 60 TFIDF of n-grams in the range 1 to 6 but could not attain a perfect test $F1$ score with the mean transform of CBOW vectors. Adding n-grams up to 3 or up to 6 allowed CNN3 to reach $F1=1$ on the test data using a maximum of 40 features. An embedding size of 300 and counting words with frequencies of at least 4, 5 or 6 allowed CNN3 to give test $F1=1$ using the CBOW mean.

The MLP only required 40 features with TF and TFIDF of unigrams to get test $F1=1$. N-grams in the range 2 or 5 also worked for 40 TFIDF features. This test performance could not be met with CBOW mean. The LSTM models did not perform as well using unigram TF as the other classifiers, although test $F1=1$ was seen with 40 unigram TFIDF features. The LSTM1 model also achieved this using the CBOW mean, an embedding size of 300 and counting words with frequencies of at least 4, 5, 6 or 7.

Stemming did not improve classifier performance. As an example, the LR model achieved training $F1=0.976$ on the mean transform of CBOW vectors of size 300 and a minimum word count of 5. This dropped to 0.9770 and 0.9778 using the Porter Stemmer and Lancaster Stemmer respectively. Test performance was not affected by the Lancaster stems, but test $F1$ dropped from 0.667 to 0.500 using the Porter stems.

The wrapper method of Recursive Feature Elimination (RFE) was sensitive to the initial group of features and gave different results on 20, 30 and 40 TFIDF features. Without specifying a minimum number of features, RFE selected only 9 content features

from the TFIDF vocabulary. The RFE returned features as ranked in Table B.21 when specifying a minimum of 18 features.

Table 5.1: Recursive Feature Elimination.

Feature	Rank
abil	1
fair	1
management	1
r_sm	1
directors	1
market	1
company	1
net	1
cash	1
capital	1
business	1
ts_f	1
non	1
term	1
ts_lw	1
african	1
investments	1
income	1
liabilities	2
wordCount	3
risk	4
perf_scaled	5
value	6
r_f	7
tax	8
ts_ari	9
LMneg	10
share	11
shares	12
credit	13

RFE ranks the features specified as 1 i.e. the first 18, after which each feature's rank

is incremented by 1. See Appendix B.2 for tests of feature importance using Chi-Squared, Information Gain and ANOVA.

Figure 5.1 shows the effect on classifier training and test performance of adding select linguistic features to the 40 TFIDF content features where

- $L1$ is SMOG, Flesch Reading Ease, Linsear Write;
- $L2$ is SMOG, Flesch Reading Ease, Linsear Write, wordCount, perf_scaled, Automated Readability Index;
- $L3$ is H6neg_scaled, Coleman-Liau, H8tone, wordCount.

$L1$ was the set of features selected by RFE with cross-validation. $L2$ is the same set as $L1$ with the next 3 most significant linguistic features added. $L3$ is the set of linguistic features proposed by ANOVA.

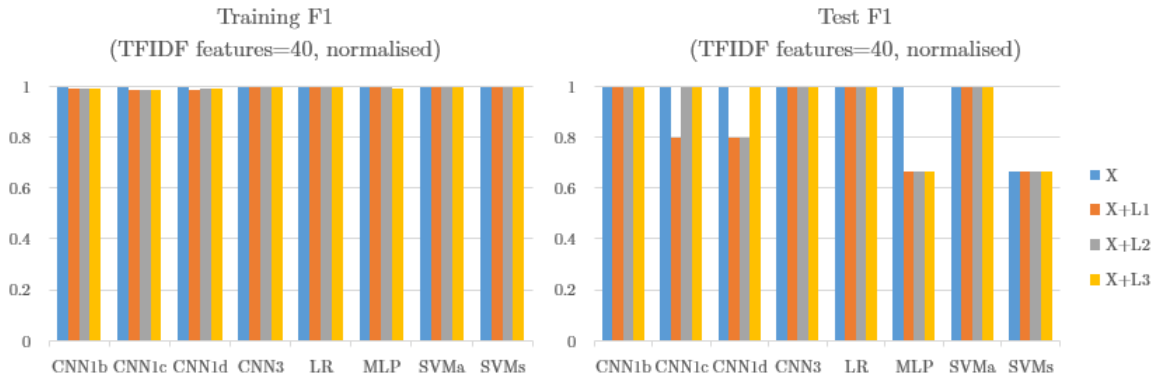


Figure 5.1: Performance on additional linguistic features

F1 training performance on 40 TFIDF features was near unity e.g. 0.9966 for the LR model. Adding the linguistic features reduced training F1 to 0.9948. The difference is marginal and cross-validation variability also needs to be accounted for. The impact on test performance was more noticeable. CNN1c was adversely affected by the addition of $L1$. CNN1d was adversely affected by the addition of $L1$ and $L2$. The test performance of the MLP model declined with adding any of the linguistics features sets $L1 - 3$.

See Appendix B.4 for further analysis of the experiments and how the classifiers performed. Appendix B.4.7 explains the difficulties experienced in using the embedding layers of the LSTM and CNN models.

5.3 Application

The top performing SVMa model was used to predict the probability of reports representing risk. Table 5.2 shows the top 5 reports.

Table 5.2: Top 5 reports by probability of risk using SVMa.

Doc ID	Document Name	Year	Co ID	Company Name	Risk	Prob
6	1464958705-4144_African-Bank-2014annualreport.pdf	2014	35	African Bank Ltd (O)	1	0.985114
358	VBS-2016-Annual-Report.pdf	2016	1	VBS Mutual Bank	1	0.973201
aa1	30 April 2018.pdf	2018	aa	Bank A	0	0.247932
12	african-bank-holdings-limited-annual-integrated-report-2019.pdf	2019	34	African Bank Ltd (N)	0	0.085518
cc1	2019 Annual Report.pdf	2019	cc	Bank C	0	0.034302

The predicted probability of risk in the 2018 annual report of Bank A is still low at 0.24%; however it is notably greater than the probability of next two reports (0.09% and 0.03%). This may prompt a regulator to investigate further.

Similarly, the top performing LR model was used to predict the probability of reports representing risk. Table 5.3 shows the top 5 reports.

Table 5.3: Top 5 reports by probability of risk using LR.

Doc ID	Document Name	Year	Co ID	Company Name	Risk	Prob
6	1464958705-4144_African-Bank-2014annualreport.pdf	2014	35	African Bank Ltd (O)	1	0.906858
358	VBS-2016-Annual-Report.pdf	2016	1	VBS Mutual Bank	1	0.789731

Continued on next page

Table 5.3 – *Continued from previous page*

Doc ID	Document Name	Year	Co ID	Company Name	Risk	Prob
12	african-bank-holdings-limited-annual-integrated-report-2019.pdf	2019	34	African Bank Ltd (N)	0	0.19037
13	african-bank-integrated-report-2018-final-website.pdf	2018	34	African Bank Ltd (N)	0	0.104735
aa1	April 2018.pdf	2018	aa	Bank A	0	0.069963

The 2019 annual report of African Bank Ltd has a relatively large predicted probability. Further investigation will be necessary to understand if this relates to the other African Bank Ltd annual reports labelled as risk or whether there are intrinsic features of this report that elevate the prediction of risk. With the LR, the 2018 annual report of Bank A predicted probability of risk appears negligible.

5.4 Interpretation

In total, 48 experiments were requested of SP-LIME to understand how a SVM model can accurately predict risk across the full test set with only 40 tokens of unigrams complemented by wordlist counts and readability indices.

For the purposes of this report, only two positive predictions and two negative predictions are explained. Blue shading is associated with risk=0 (the negative class) and orange shading is associated with risk=1 (the positive class).

Figure 5.2 shows the 10 most important features in predicting the 2014 report of Africa Bank as positive risk.

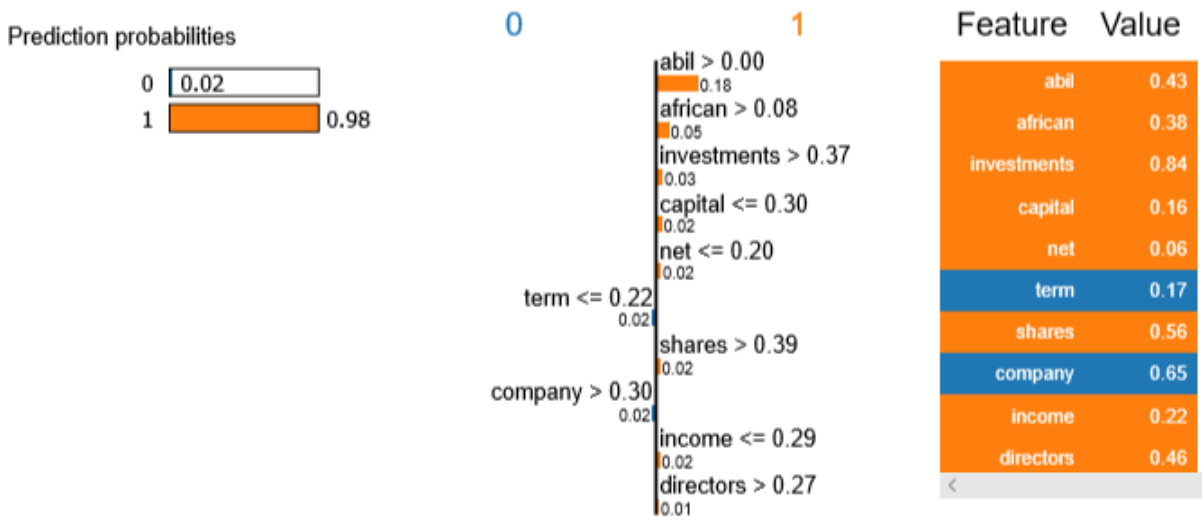


Figure 5.2: SP-LIME explanation of Africa Bank 2014 report

The true positive (TP) prediction on this report being risk positive was 98%. The presence of the ‘abil’ unigram supported the prediction positively, i.e. its frequency of 0.43 was greater than zero. This feature had the greatest relative importance of 0.18. Similarly, the frequency value of ‘african’ at 0.38 exceeded the threshold of 0.08 to contribute positively to the prediction. This feature had the second greatest relative importance of 0.05. This may not come as a surprise understanding that ABIL is the abbreviation for African Bank Investments Limited, however there were two other African Bank reports in the test set that were predicted with negative risk.

Conversely, the frequency value of ‘term’ at 0.17 counteracted the prediction by being below the threshold of 0.22, but its relative importance was a low 0.02. Similarly, the frequency value of ‘company’ at 0.65 exceeded the threshold of 0.30 to counteract the prediction. This feature also had a low relative importance of 0.02.

Figure 5.3 shows the 10 most important features in predicting the 2016 report of VBS Mutual Bank as positive risk.



Figure 5.3: SP-LIME explanation of VBS Mutual Bank 2016 report

The TP prediction on this report being risk positive was 96%. The absence of the ‘abil’ unigram counteracted the prediction and had the greatest relative importance of 0.20. Similarly, the frequency value of ‘investments’ at 0.14 counteracted the prediction by being below the threshold of 0.16, but its relative importance was a low 0.02.

Positive contributions to the prediction came from unigrams ‘cash’, ‘african’, ‘term’, ‘non’ and ‘capital’ in decreasing order of importance.

Figure 5.4 shows the 10 most important features in predicting the 2018 report of Africa Bank as negative risk.

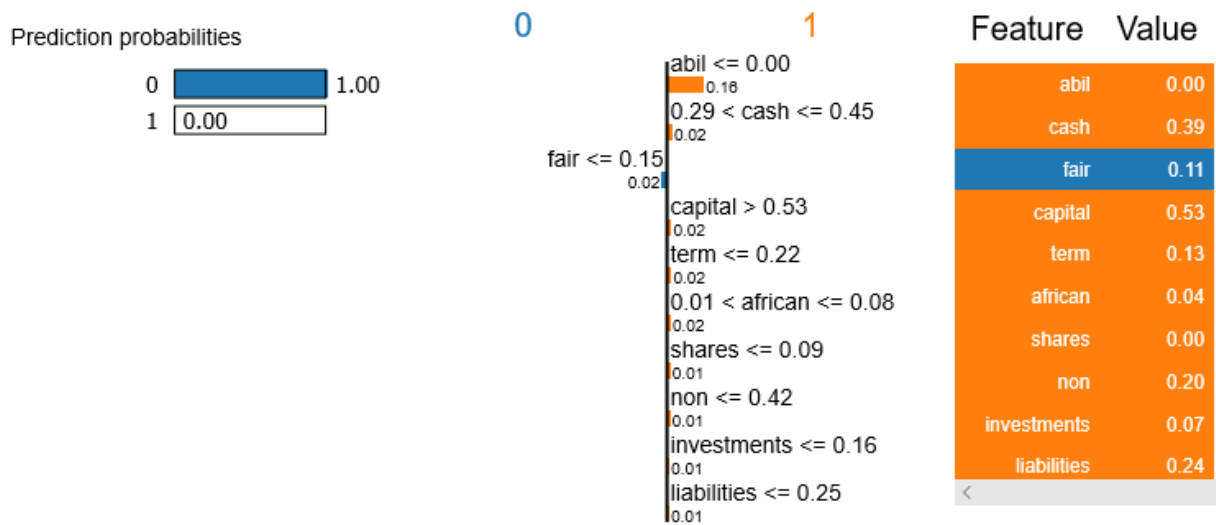


Figure 5.4: SP-LIME explanation of Africa Bank 2018 report

The TP prediction on this report being risk negative was 100%. The absence of the ‘abil’ unigram counteracted the prediction and had the greatest relative importance of 0.16. Conversely, the frequency value of ‘fair’ at 0.11 supported the prediction by being below the threshold of 0.15 despite its relative importance being a low 0.02. Other features contributing to the prediction are not in the top 10 list and would require further investigation.

Figure 5.5 shows the 10 most important features in predicting the 2019 report of Africa Bank as negative risk.

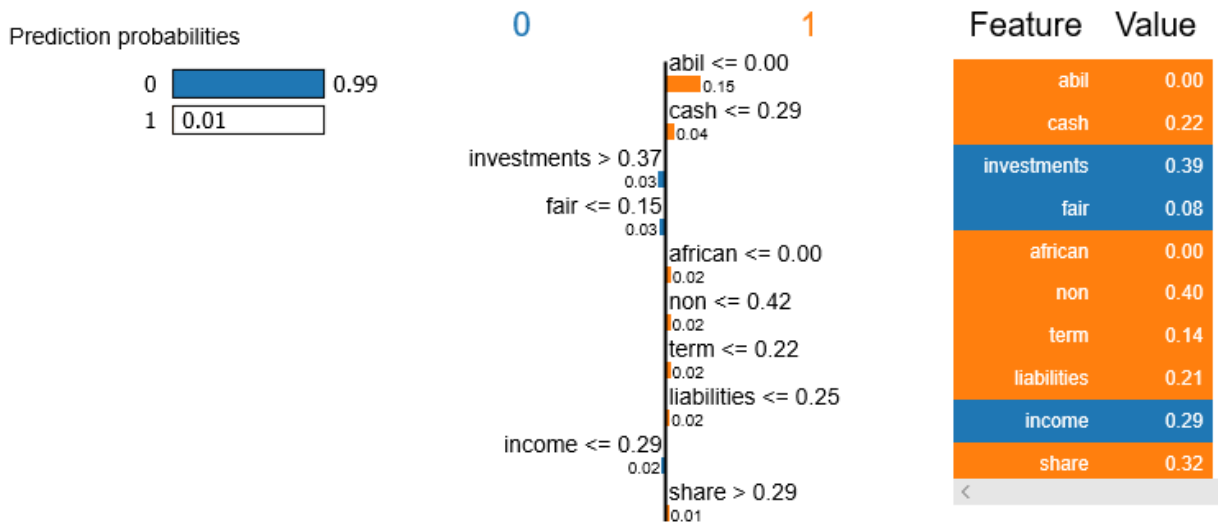


Figure 5.5: SP-LIME explanation of Africa Bank 2019 report

The TP prediction on this report being risk negative was 99%. Again, the absence of the ‘abil’ unigram counteracted the prediction and had the greatest relative importance of 0.15. Unlike the prediction on the 2018 report, another two features appear in the top 10 list that contribute to the prediction, namely ‘investments’ and ‘income’. Even though the relative importance is rounded to the same 0.02 value, ‘investments’ ranked above ‘fair’ with a frequency of 0.39 being above the 0.37 level. Similarly, the frequency value of ‘income’ at 0.29 was below the threshold of 0.29 to contribute to the prediction. This feature also had a relative importance of 0.02.

It was interesting to note that none of the linguistic features were among the 10 most important predictors, neither word counts nor readability indices.

5.5 Question answering

Pre-trained reader models `bert-squad_1.1` and `distilbert-squad_1.1` were downloaded and cdQA pipelines were constructed for them. The Google Colab session crashed after using all 12.72GB RAM attempting to fit paragraphs to the BERT cdQA pipeline. The cdQA `filter_paragraphs` utility was then applied which reduced the number of sentence tokens for the VBS Mutual Bank 2015 annual report from 498 to only 69. See

Appendix B.5 for the example content that was filtered out. Following the filtration, paragraphs were then successfully fitted to both BERT and DistilBERT cdQA pipelines.

Table 5.4 gives the three best predicted answers by both pipelines. Two initial questions were asked as follows:

1. Who are the auditors for VBS Mutual Bank?
2. What is the 2016 asset value of VBS Mutual Bank?

Table 5.4: Example question and answers on Bank annual reports.

Q1a Who are the auditors for VBS Mutual Bank?				
No	Model	Answer	Title	Paragraph
1	BERT	board of directors	VBS-2015-Annual-Report.pdf	The annual financial statements of VBS Mutual Bank, as identified in the first paragraph, were approved by the board of directors. . .
1	Distil BERT	board of directors	VBS-2015-Annual-Report.pdf	The annual financial statements of VBS Mutual Bank, as identified in the first paragraph, were approved by the board of directors. . .
2	BERT	AUDIT AND RISK COMMITTEE	VBS-2015-Annual-Report.pdf	VBS MUTUAL BANK AUDIT AND RISK COMMITTEE REPORT Audit and Risk Committee Report to VBS Mutual Bank. . .
2	Distil BERT	NKS Malaba Chartered Accountant (SA)	VBS-2016-Annual-Report.pdf	Per NKS Malaba Chartered Accountant (SA) Registered Auditor Director Independent Auditor's Report To the Shareholders of VBS Mutual Bank We have audited the financial statements of VBS Mutual Bank
3	BERT	Ramavhunga Chairman: Board of Directors Chief Executive Officer The directors	VBS-2015-Annual-Report.pdf	Ramavhunga Chairman: Board of Directors Chief Executive Officer The directors are responsible for the preparation and fair presentation of the annual financial statements
3	Distil BERT	VBS MUTUAL BANK AUDIT AND RISK COMMITTEE	VBS-2015-Annual-Report.pdf	VBS MUTUAL BANK AUDIT AND RISK COMMITTEE REPORT Audit and Risk Committee Report to VBS Mutual Bank. . .

Continued on next page

Table 5.4 – Continued from previous page

Q1b Who are the <i>independent</i> auditors for VBS Mutual Bank?				
No	Model	Answer	Title	Paragraph
1	BERT	Registered Auditor Director Independent Auditor's Report To the Shareholders	VBS-2016- Annual- Report.pdf	Per NKS Malaba Chartered Accountant (SA) Registered Auditor Director Independent Auditor's Report To the Shareholders of VBS Mutual Bank We have audited the financial statements of VBS Mutual Bank
1	Distil BERT	board of directors	VBS-2015- Annual- Report.pdf	The annual financial statements of VBS Mutual Bank, as identified in the first paragraph, were approved by the board of directors...
2	BERT	board of directors	VBS-2015- Annual- Report.pdf	The annual financial statements of VBS Mutual Bank, as identified in the first paragraph, were approved by the board of directors...
2	Distil BERT	NKS Malaba Chartered Accountant (SA)	VBS-2016- Annual- Report.pdf	Per NKS Malaba Chartered Accountant (SA) Registered Auditor Director Independent Auditor's Report To the Shareholders of VBS Mutual Bank We have audited the financial statements of VBS Mutual Bank
3	BERT	AUDIT AND RISK COMMITTEE	VBS-2015- Annual- Report.pdf	VBS MUTUAL BANK AUDIT AND RISK COMMITTEE REPORT Audit and Risk Committee Report to VBS Mutual Bank...
3	Distil BERT	Shareholders	VBS-2016- Annual- Report.pdf	...Independent Auditor's Report To the Shareholders of VBS Mutual Bank...

Continued on next page

Table 5.4 – *Continued from previous page*

Q2 What is the 2016 asset value of VBS Mutual Bank?				
No	Model	Answer	Title	Paragraph
1	BERT	millions	VBS-2015-Annual-Report.pdf	VBS Mutual Bank Asset growth Value – millions 300 250 200...
1	DistilBERT	300	VBS-2015-Annual-Report.pdf	VBS Mutual Bank Asset growth Value – millions 300 250 200...
2	BERT	statement of financial position at 31 March 2016	VBS-2016-Annual-Report.pdf	Per NKS Malaba Chartered Accountant (SA) Registered Auditor Director Independent Auditor's Report To the Shareholders of VBS Mutual Bank We have audited the financial statements of VBS Mutual Bank
2	DistilBERT	2015	VBS-2015-Annual-Report.pdf	VBS MUTUAL BANK AUDIT AND RISK COMMITTEE REPORT Audit and Risk Committee Report to VBS Mutual Bank...
3	BERT	2015	VBS-2015-Annual-Report.pdf	VBS MUTUAL BANK AUDIT AND RISK COMMITTEE REPORT Audit and Risk Committee Report to VBS Mutual Bank...
3	DistilBERT	31 March 2016	VBS-2016-Annual-Report.pdf	...statement of financial position at 31 March 2016, and the statements of profit or loss and other comprehensive income, changes in equity...

Remarkably by returning ‘NKS Malaba Chartered Accountant (SA),’ DistilBERT answered the first Question Q1(a) correctly with its second most probable prediction. With Question Q1(b), the first question was further qualified by including the word ‘independent’ in the hope that predictions would improve. Even though BERT made a new best prediction and demoted its two best answers from Q1(a), the answer was not correct. It predicted the correct paragraph but did not extract the appropriate substring as DistilBERT did. DistilBERT kept its first two predictions from Q1(a) and did not promote the correct answer to the first option. It also replaced the third best prediction based on a paragraph containing the bigram ‘Independent Auditor’.

Although ‘millions’ and ‘300’ were the best predictions for the second question, both were sourced from the incorrect 2015 annual report. Furthermore, dates formed part of the answers in the remaining predictions instead of a rand value.

5.6 Summary

Section 5.2 explained modelling goals covering discovery, minimising features and adding linguistic features. The results of Section 5.2 showed that eight models could be trained to correctly classify the test set of annual reports. All models performed best on normalised features except for the LSTM models that prefer standardised features. CNN1a and CNN2 models did not perform well relative to the others. Recursive Feature Elimination (RFE) reduced 40 TFIDF features to only 9. Setting the minimum number of features to 18 uncovered some linguistic features. Adding these linguistic features to the 40 TFIDF features did not improve classifier performance nor did stemming.

Section 5.3 applied the traditional classifiers to predict the top 5 annual reports representing risk. In addition to the two positive test documents, the models also returned the 2018 report of Bank A and the 2019 report of African Bank. Finally, a SVM classifier using TFIDF, wordlist counts and readability features was explained in Section 5.4 using Sub modular Pick LIME.

Notwithstanding the relatively small initial data set, the results of the classification and top-5 predictions are encouraging. The objective of predicting financial risk through NLP of the reports was proven based solely on banks placed under curatorship. Future work will undoubtedly improve the data set, processing pipeline and classifiers for the risk prediction as proposed in Section 7.2.

Besides risk prediction, closed domain Question Answering (cdQA) in Section 5.5 demonstrated another use case of the annual report data set. However, predicted answers to questions using paragraph filtered text and pre-trained models had room for improvement. Examples included correct paragraphs but incorrect answers, correct answers not being the first prediction as well as proposed date values instead of financial values.

Chapter 6

Discussion

This chapter discusses the research performed, issues experience, results obtained and what has been learnt through the study. Toward answering the research question and progressing solutions to the risk identification problems,

- Section 6.1 discusses the results of data preparation, processing and analysis prior to modelling;
- Section 6.2 discusses the results of modelling.

6.1 Pre-processing

Results of the experiments in Chapter 4 started with a feasibility test of re-using a software programme developed for UK annual reports.

6.1.1 Feasibility test

The outcome of the feasibility test (Section 4.1) proved that the existing CFIE–FRSE tool was not suitable for re-use within the South African setting. Annual reports in booklet form (Section 4.4.1) could not be processed by CFIE–FRSE. The UK tool could also not identify headers which had not been bookmarked as part of the PDF publication. The developers were approached and unfortunately the source code could not be shared to

allow for these enhancements. As part of this study, a new extraction tool was developed in Python.

6.1.2 Data collection

The initial corpus contained 252 annual reports from 34 banks in South Africa (see Section 2.4.2) which is smaller than the data set of 408 reports used by Minhas and Hussain [48]. Relative to the objective of creating a text corpus of annual reports from 2009 to 2018 of South African companies, the corpus also included 24 annual reports for 2019. Time limits for this research and no access to a single source of company annual reports did not allow for the inclusion of insurers as a second type of financial institution. The SA data set can be extended with future work (Section 7.2.1).

6.1.3 Data labelling

The objective of labelling a subset of reports with risk was partially fulfilled. Legitimate sources of SA companies at risk was a limitation. Companies involved in fraud investigations are available from the US Securities Exchange Commission (SEC) Accounting and Auditing Enforcement Releases (AAERs) [21] and the UK Serious Fraud Office. Similar sources for South African companies could not be found. As a consequence, publicly available reports and well publicised risk events had to be used.

Only one financial issue involving VBS Mutual Bank and one governance issue involving African Bank were reported. The rationale used to label reports from African Bank as both positive and negative dependent on the financial year assumed a time series and recognised the change of management. Reports for the years under which African Bank were under curatorship were not labeled with risk.

It is unlikely that the 8 documents annotated with the risk flag based on this publicly available data are complete or representative of the risk profile across all 259 annual reports. Notwithstanding improvements to be made on labelling (Section 7.2.1), the initial labelling was sufficient to answer in the affirmative the main research question, *Can risk be identified in unstructured company annual reports using NLP?*

6.1.4 Data preparation

To understand the initial research question *"Can text be extracted from annual reports of South African companies?"* Section 4.4 showed that text was successfully extracted despite 12 of the SA bank annual reports being in booklet form (Section 4.4.1). Both the CFIE–FRSE tool and a UK report were used to validate the new Python extraction tool contributed by this work. Validation metrics included overall word counts as well as the frequency with which words in predefined wordlists were found in the text. The natural reading order of text across columns was preserved as well as ensuring traceability to the location of text by page and by text block coordinates.

Furthermore, algorithms were developed to identify 1744 headers within the text of 167 annual reports that lacked bookmarked tables of content (Section 4.4.6). The reports for which headers could not be identified as well as headers that could not be classified leaves room for future work (Section 7.2.1). Some examples were found that require numbers and text to be matched not only on vertical coordinates but also horizontal coordinates (see Appendix B.3.3).

There was no noticeable benefit to extracting larger bundles of words from the text. The best performing models on the smallest number of features relied on unigrams and did not require n-grams in the range of 2 to 6 (Section 5.2). This is aligned with Fissette's [21] experiments with Naive Bayes (NB) and Support Vector Machine (SVM) models where these additions could not improve performance beyond the level achieved with solely unigrams.

Initial analysis of the two annual reports of VBS Mutual Bank in Section 4.4.8 showed that the net tone of the SA company was lower than the reference report of the UK company. The level of uncertainty in the SA bank's performance overview was notably higher whilst the UK annual report contained more strategic commentary. The relative uncertainty and lack of strategic focus may be indicators of elevated risk. Furthermore, year-on-year comparisons can be made to establish trends with features like tone and uncertainty that can warn of imminent or underlying problems facing the company. Section 4.4.8 provided an example where the tone of VBS Mutual Bank annual reports declined to -26% in 2016 from -19% in 2015 using the Henry scores. They also scored lower for forward-looking, uncertain, causal, performance and strategic keywords than

another SA bank (JP Morgan Chase and Co). These linguistic features may complement traditional metrics in benchmarking financial institutions against their peers to influence their risk classification. The use of causal words in particular requires interpretation relative to risk. For example, could numerous references to cause and effect suggest that the company is trying to make excuses for poor performance or matters out of their control. Alternatively, do numerous references to cause and effect reduce the inherent risk since analysts and investigators can evaluate the company’s explicit reasoning.

Each feature selection method in Appendix A.2 gave differing views of feature importance. Furthermore, the wrapper method of Recursive Feature Elimination was sensitive to the initial group of features. Adding selected sets of linguistic features did not improve classifier performance. This was consistent with Fissette’s [21] results.

This study showed that TFIDF alone can answer the research question “*Which features can be extracted from annual reports to distinguish companies at risk?*” However improvements were observed through the addition of linguistic features, specifically keyword counts from wordlists and readability indices. However, additional linguistic features should be extracted and used to train future models as described in Section 7.2.1.

6.2 Modelling

5-fold cross validation F1 scores were often not returned in training the first Logistic Regression (LR) model. A number of the folds did not contain a positive sample. This was due to the class imbalance with the original training data set having only 6 risk documents. To remedy this, random oversampling with replacement was applied to positive documents to balance the classes as discussed in Section 3.4.11. This had the adverse effect of overfitting with the training F1 being over optimistic compared with test performance.

Another observation in training was a noticeable increase in computation time of the pipeline of pre-processing and classification when the number in the n-grams range was increased. Even though the SVMs model showed the best training performance, it was probably overfitting and did not generalise well.

The MLP model was introduced at a late stage in the lifecycle of the study. Pre-built

neural networks available within Python libraries could have been used earlier in the study as an easier alternative to building the neural network.

Having trained Logistic Regression (LR) and Support Vector Machine (SVM) classifiers on annual reports over a 9 year period, parameters were selected that allowed generalisation on the test set of 2018 and 2019 annual reports and two other reports labelled with risk. Test prediction achieved a perfect F1 score of unity with no false positives and no false negatives providing a positive answer to the main research question, *Can risk be identified in unstructured company annual reports using NLP?*

Probabilities of the top 5 predictions were explored to reveal reports and companies that may have higher levels of inherent risk albeit below the threshold for classification. This illustrated one way in which the objective of predicting financial risk through Natural Language Processing of the reports could be met. Even though these models demonstrated adequate predictability, the tools and models developed through this study are far from being ready for real-world applications like regulators choosing to investigate the report or the company further. Rather they provide a baseline toward proving the concept that financial risk can be identified through Natural Language Processing processing of company annual reports in South Africa. Work proposed in Section 7.2.1 will address the stability, reproduceability and scalability needed.

The SVM model with auto-scaling was concluded as the answer to the investigative research question, *“Which classifiers perform best at identifying risk using the text features of annual reports?”* The classifier had test F1=1 for a combination of 40 content features and 17 linguistic features. Through further work proposed in Section 7.2.1 other model type or configurations will certainly improve upon this finding.

An explanation of the model using Sub modular Pick LIME (SP-LIME) highlighted solely TFIDF tokens in the top 10 predictors of the two positive and two negative documents examined. Many of these tokens were generic terms like ‘shares’, ‘income’, ‘directors’, ‘capital’ and ‘cash’ which may not be strong predictors to distinguish between low and high risk annual reports. This may have resulted from the small data set and the simple risk labelling. It was noted that many of the features had the same rounded relative importance value. This suggests that all 60 features should be examined to understand the contribution of the 17 wordlist counts and 3 readability indices.

6.3 Summary

Section 6.1 discussed how the existing text extracting tool could not process booklets and why a new tool had to be developed for this study. The initial data set of bank annual reports was compared to the data sets used in similar studies. Limitations of the labelling performed in this study were highlighted. Comparisons were made between UK and SA documents as well as annual reports over time using the linguistic features. Section 6.2 discussed observations in training and testing the models. Applying the model for top-5 predictions as well as model interpretation were discussed.

Chapter 7

Conclusion

This chapter summarises conclusions drawn in identifying financial risk in annual reports of banks in South Africa.

- Section [7.1](#) summarises the conclusions of this study;
- Section [7.2](#) considers future directions that related research could take.

7.1 Summary of Conclusion

A methodology was proposed in Chapter [3](#) for collecting unstructured PDF annual reports of South African companies and creating a text corpus that can be analysed and used to train machine learning models. This approach addressed opportunities highlighted in the literature review of Chapter [2](#) by extending research beyond quantitative financial values into the use of unstructured text to identify risk. It also initiated a corpus of annual reports that was lacking for South African companies. The new pipeline of risk predictors has the potential to complement existing warning systems and mitigate problems related to the financial collapse of a company as set out in Section [1.1](#).

Annual reports collected in this study was solely for banks making the corpus relatively small. Although extracting text from these unstructured PDF documents remains a challenge, this study hopes to provide techniques and algorithms to improve the completeness and grain of extraction. Annual reports in booklet form were processed when

previously they had been excluded. Algorithms developed in the study may assist others toward identifying headers when not properly bookmarked in the PDF. This will allow for risk identification based on section-specific features rather than features aggregated for the document. Notwithstanding these technical improvements, there is much room for improving the quality with which annual reports of banks in South Africa are published into PDF. Any convergence toward a more uniform structure would improve descriptive analysis as well as machine learning outcomes.

Only two of the banks had their annual reports annotated with risk due to the banks being placed under curatorship. To overcome a highly imbalanced data set, the few positive documents were resampled with replacement allowing for effective cross-validation of classifier training performance. n-grams, stemming and the addition of linguistic features did not improve the classifier performance. Nonetheless, there is benefit in extracting linguistic features. This study used these features to profile and contrast documents in each class. Risk-positive documents showed elevated levels of uncertainty and negative trends in tone that would hopefully serve as an early warning for regulators to investigate.

7.2 Future work

Given the limited time available for the study, the goal was to develop an initial corpus of South African company annual reports and a pipeline for reading, processing, labelling and classifying the data in terms of risk. It is hoped that through future elaboration and refinement this work can lead to real-world application. Extensions to this work as well as areas that could not be focussed on are discussed below.

7.2.1 Pre-processing

Models developed during this study are limited to an analysis of the qualitative data in company annual reports. Future work should add financial numeric data to the models and evaluate whether classification performance can be improved.

The data set of SA companies created from this work can be refined and elaborated upon. Depending on the use case, it may prove prudent to remove integrated annual

reports of the group company from the data set or to combine them with the subsidiary company. With the pipeline established, annual reports of other institution types can be added to the corpus. Adding insurers to the corpus may increase the number of companies to approximately 260 and annual reports to approximately 2800. This would make the SA data set larger than the US data set used by Fissette [21]. Investigation into public and proprietary platforms that can service as a single source of SA company annual reports should be pursued.

Access to risk ratings for each of the South African companies will improve the validity of learning and prediction tasks. Models developed in this study can be retrained if data becomes available that allows more relevant annotation of risk or higher fidelity of the risk rating. Instead of using data confined to the public domain, analysts or regulators can use their internal risk ratings as a means of annotating the annual reports. These ratings may allow formulation of a multi-class problem that can build upon the binary classification used in this work. Risk ratings may be categorical values like low, medium and high which would allow for three classes instead of 2. Furthermore, risk ratings may be continuous (e.g. from 0 to 10) which may allow formulation of a regression problem.

Algorithm 3.2 can be enhanced to improve the identification of table of content pages. Header-specific features can be added as predictors to improve risk classification. For instance, features extracted from the independent auditor report section may be more important than other sections or the report as a whole. Experiments using TF and TFIDF can set the index to n as opposed to allowing a range of 1 to n . This will allow performance to be compared when features are restricted to the n -gram as opposed to setting an upper bound on the n -gram value.

The unique context of a developing economy and South Africa's language and cultural differences provides an opportunity to future research to update wordlists. The effect of these changes on linguistic features like tone and uncertainty as well as their effect on risk identification can be assessed. Further exploratory data analysis will allow data cleaning to be introduced to remove obvious incongruities like 'million million million' observed on tri-grams (Section 4.5.1). Existing content features can be extended to include measures of vocabulary and quantitative content [21].

Existing linguistic features can be further analysed. UK annual reports can be seg-

mented into industry sectors to ring-fence banks, insurers and other financial institution types. This will allow SA annual reports to be appropriately benchmarked against their UK counterparts to answer questions like: *Are the tone scores of banks lower than other institution types? Do annual reports of SA banks use a more negative tone than UK banks?* Time series analysis of these features can establish trends and warning limits for companies at risk. Other feature selections and methods can be used assessed for further understanding of the research question *Which features can be extracted from annual reports to distinguish companies at risk?* Consideration should be given to Correlation-based Feature Selection (CFS) or application of the Fisher Ratio (FiR). Both are suited to numeric features and categorical targets, specifically binary classes for the latter [12].

The Linguistic Inquiry and Word Count (LIWC) tool may be purchased to extract grammatical features and psychological categories. The Stanza Python NLP package (which superseded the Stanford parser used by Fissette [21]) can be used to extract relations between words. Similarly, Named Entity Recognition (NER) may be used to add further features.

Additional scaling techniques can be applied to the features. In particular, scaling the feature to have a range of $[-1,1]$ can be compared to the results achieved with the range $[0,1]$. Each document's row of features can be treated as a vector. The vector can be scaled to unit length e.g. using the L1-norm (sum of absolute values of vector elements) or L2-norm (square root of the sum of squared vector element values). Each feature column can also be treated as a vector and scaled to unit length using these norms.

7.2.2 Modelling

Since the focus of this work was establishing a data set and data set pipeline for future machine learning, many opportunities exist to tune the initial models as well as experiment with other machine learning models. Instead of aiming for an F1 score of unity, false positives could be tolerated. These annual reports would require further investigation to resolve as false or a possible high risk report.

The performance of classifiers using ensemble methods can be compared with these baseline models. SVM models can be tuned by finding improved C and gamma hyper pa-

rameters for instance through applying grid search techniques. Additional experimentation with SVM's using linear, polynomial and sigmoid kernels may improve performance [65].

Although the CFIE–FRSE tool was not used to extract text from SA annual reports, it could be used to recreate the UK corpus so that the output is a set of full texts and header texts for each of the UK annual reports. This would allow training of machine learning models on the larger UK data set for application as a pre-trained model on the SA corpus. Predictors of risk in these baseline models can be studied further and compared with predictors identified in the US and UK markets. Further experimentation with closed domain Question Answering (cdQA) may lead to more meaningful answer text. In particular, using regular expression patterns or omitting the paragraph filter may improve the utility of the cdQA.

The tools and models developed in this study will benefit from refactoring and continuous improvement to ensure stability and reproduceability. Having started with banks within the financial sector, an incremental approach to the scope is recommended. Logical next sub-sectors will be other financial institutions like insurers, pension funds, medical schemes, collective investment schemes and friendly societies. Expanding the scope to industrial and resource sectors will make significant inroads into the scope of companies listed on the JSE. Finally, unlisted private companies can be addressed as collecting annual reports for them may present a greater challenge than those listed on the JSE.

By expanding on this work and making these improvements, it may not be long before Natural Language Processing of annual reports can help regulators and other stakeholders mitigate their risk. If potential risk can be identified and investigated earlier, companies have a better chance of financial recovery and ultimate prosperity for their stakeholders and their country.

Bibliography

- [1] Snowball, 2021. Available online: [<https://snowballstem.org/>].
- [2] J. Anderson. LIX and RIX: Variations on a little-known readability. *Journal of Reading*, 26(6):490–496, 1983.
- [3] B. Ay, M. Talo, I. Hallac, and G. Aydin. Evaluating deep learning models for sentiment classification. *Concurrency and Computation: Practice and Experience*, 07 2018. Available online: [[doi:10.1002/cpe.4783](https://doi.org/10.1002/cpe.4783)].
- [4] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *Journal of Machine Learning Research*, 6(1):20–29, 2004. Electronic journal. Available online: [<https://doi-org/10.1145/1007730.1007735>].
- [5] R. Beale and T. Jackson. *Neural Computing: An Introduction*. CRC Press, 1990.
- [6] D. Biber and F. Barbieri. Lexical bundles in university spoken and written registers. *English for Specific Purposes*, 26(3):263–286, 2007. Electronic journal. Available online: [<https://doi.org/10.1016/j.esp.2006.08.003>].
- [7] P. Branco, L. Torgo, and R. Ribeiro. A Survey of Predictive Modelling under Imbalanced Distributions, 2015.
- [8] J. Brownlee. Sequence Classification with LSTM Recurrent Neural Networks in Python with Keras, 2016. Machine Learning Mastery. Available online: [<https://machinelearningmastery.com/>]

- [sequence-classification-lstm-recurrent-neural-networks-python-keras](#)]
(Accessed: 22 Sept. 2020).
- [9] J. Cai, F. Eidam, A. Saunders, and S. Steffen. Syndication, Interconnectedness, and Systemic Risk. *Journal of Financial Stability*, 34:105–120, 2018.
- [10] S. Chen. Detection of fraudulent financial statements using the hybrid data mining approach. *Springerplus*, 5(1):1–16, 2016. Available online: [[doi:10.1186/s40064-016-1707-6](#)].
- [11] F. Chollet. Using pre-trained word embeddings in a Keras model, 2016. The Keras Blog. Available online: [[https://blog.keras.io/using-pre-trained-word-embeddings-in-a-keras-model.html](#)] (Accessed: 28 Sept. 2020).
- [12] M. Ciortan. Overview of feature selection methods, 2019. Towards Data Science. Available online: [[https://towardsdatascience.com/overview-of-feature-selection-methods-a2d115c7a8f7](#)] (Accessed: 12 Nov. 2020).
- [13] J. J. Dabrowski, C. Beyers, and J. P. de Villiers. Systemic banking crisis early warning systems using dynamic Bayesian networks. *Expert Systems With Applications*, 62:225–242, 2016.
- [14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. Available online: [[https://www.aclweb.org/anthology/N19-1423](#)].
- [15] M. El-Haj, P. Alves, P. Rayson, M. Walker, and S. Young. Retrieving classifying and analysing narrative commentary in unstructured glossy annual reports published as PDF files. *Accounting and Business Research*, 50(1):6–34, 2020. Available online: [[doi:10.1080/00014788.2019.1609346](#)].

- [16] M. El-Haj and P. Rayson. CFIE-FRSE tool, 2020. GitHub repository. Available online: [<https://github.com/drelhaj/CFIE-FRSE/>] (Accessed: 1 Mar. 2020).
- [17] M. El-Haj, P. Rayson, M. Walker, S. Young, and V. Simaki. In Search of Meaning: Lessons, Resources and Next Steps for Computational Analysis of Financial Discourse. *Journal of Business Finance Accounting*, 46(3-4):265–306, 2019. Available online: [[doi:10.1111/jbfa.12378](https://doi.org/10.1111/jbfa.12378)].
- [18] M. El Haj, S. Young, and P. Rayson. UK annual report narratives dataset: CFIE-FRSE May 2019, 2019. GitHub repository. Available online: [[doi:10.17635/lancaster/researchdata/296](https://doi.org/10.17635/lancaster/researchdata/296)].
- [19] I. E. Fisher, M. R. Garnsey, and M. E. Hughes. Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. *Intelligent Systems in Accounting, Finance and Management*, 23(3):157–214, 2016. Available online: [[doi:10.1002/isaf.1386](https://doi.org/10.1002/isaf.1386)].
- [20] R. A. Fisher. *Statistical methods for research workers*. Oliver and Boyd, Edinburgh, United Kingdom, 1934.
- [21] M. V. M. Fissette. *Text mining to detect indicators of fraud in annual reports world wide*. Doctoral thesis, University of Twente, Enschede, The Netherlands, 2017. Available online: [[doi:10.3990/1.9789036544207](https://doi.org/10.3990/1.9789036544207)].
- [22] R. Flesch. *How to Write Plain English : A Book for Lawyers and Consumers*. Harper Row, New York, 1979.
- [23] S. Goel, J. Gangolly, S. Faerman, and O. Uzuner. Can Linguistic Predictors Detect Fraudulent Financial Filings? *Journal of Emerging Technologies in Accounting*, 7(1):25–46, 2010. Available online: [[doi:10.2308/jeta.2010.7.1.25](https://doi.org/10.2308/jeta.2010.7.1.25)].
- [24] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT Press, Cambridge, United States of America, 2016.
- [25] R. Gunning. *The Technique of Clear Writing*. McGraw-Hill, New York, United States of America, 1968.

- [26] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002. Electronic journal. Available online: [<https://doi.org/10.1023/A:1012487302797>].
- [27] P. Hajek and R. Henriques. Mining corporate annual reports for intelligent detection of financial statement fraud - a comparative study of machine learning methods. *Knowledge-Based Systems*, 128:139–152, 2017. Available online: [[doi:10.1016/j.knosys.2017.05.001](https://doi.org/10.1016/j.knosys.2017.05.001)].
- [28] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge Data Engineering*, 21(9):1263–1284, 2009. Electronic journal. Available online: [<https://doi.org/10.1109/TKDE.2008.239>].
- [29] E. Henry. Are investors influenced by how earnings press releases are written? *Journal of Business Communication*, 45(4):363–40, 2006.
- [30] E. Henry. Market Reaction to verbal components of earnings press releases: Event study using a predictive algorithm. *Journal of Emerging Technologies in Accounting*, 3(1):1–19, 2006.
- [31] G. Hinton and S. Roweis. Stochastic neighbor embedding. In *16th Annual Neural Information Processing Systems Conference, NIPS 2002*, Vancouver, Canada, 2002. Neural Information Processing Systems.
- [32] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. Electronic journal. Available online: [<https://doi.org/10.1162/neco.1997.9.8.1735>].
- [33] X. Huang, H. Zhou, and H. Zhu. A framework for assessing the systemic risk of major financial institutions. *Journal of Banking and Finance*, 33(11):2036–2049, 2009.
- [34] J. Hunter, D. Dale, E. Firing, and M. Droettboom. Interpolations for imshow, 2020. Matplotlib Version 3.3.2. Available online: [https://matplotlib.org/gallery/images_contours_and_fields/interpolation_methods.html] (Accessed: 17 Oct. 2020).

- [35] K. Hussainey, T. Schleicher, and M. Walker. Undertaking large-scale disclosure studies when AIMR-FAF ratings are not available: the case of prices leading earnings. *Accounting and Business Research*, 33(4):275–294, 2003.
- [36] M.-C. Kincaid and C. D. Manning. Stanford typed dependencies manual, 2016. The Stanford Natural Language Processing Group. Available online: [https://nlp.stanford.edu/software/dependencies_manual.pdf] (Accessed: 31 Oct. 2020).
- [37] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. Available online: [<http://arxiv.org/abs/1412.6980>].
- [38] S. Kotsiantis, E. Koumanakos, D. Tzelepis, and V. Tampakas. Forecasting Fraudulent Financial Statements using Data Mining. *International Journal of Economics and Management Engineering*, 1(12):844–849, 2007. Electronic journal. Available online: [<https://publications.waset.org/7180/pdf>](Accessed: 7 Sept. 2019).
- [39] C. Lee and G. G. Lee. Information gain and divergence-based feature selection for machine learning-based text categorization. *Information Processing Management*, 42(1):155–165, 2006. Electronic journal. Available online: [<https://doi.org/10.1016/j.ipm.2004.08.006>].
- [40] J. Li, K. Cheng, S. Wang, F. Morstatter, R. Trevino, J. Tang, and H. Liu. Feature selection: a data perspective. In *ACM Computing Surveys (Csur)*, volume 50, pages 1–45, 2018. Available online: [<https://doi.org/10.1145/3136625>].
- [41] H. Liu and R. Setiono. In Chi2: feature selection and discretization of numeric attributes. In *Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence Herndon, VA, USA 1995 Nov. 5 - 1995 Nov. 8. (1995)*, pages 388–391, 1995. Available online: [<https://doi.org/10.1109/TAI.1995.479783>].
- [42] T. Loughran and B. McDonald. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65, 2011.

- [43] V. Marivate, A. Modupe, T. Sefara, V. Chabalala, K. Makhaya, T. Mokgonyane, and R. Mokoena. Investigating an approach for low resource language dataset creation, curation and classification: Setswana and sepedi. In *Proceedings of the First workshop on Resources for African Indigenous Languages (RAIL), 11–16 May 2020*, pages 15–20, Luxemburg:, 2020. European Language Resources Association (ELRA). Available online: [<https://www.aclweb.org/anthology/2020.rail-1.3.pdf>].
- [44] M. Marrian. African Bank Investment Ltd (ABIL): A South African corporate governance failure. In *African Governance Showcase 2017*. Institute of Directors Southern Africa, 2017.
- [45] R. O. Mason. Four Ethical Issues of the Information Age. *MIS Quarterly*, 10(1):5–12, 1986.
- [46] D. Mease, A. J. Wyner, and A. Buja. Boosted classification trees and class probability/quantile estimation. *Journal of Machine Learning Research*, 8(3):409–439, 2007. Electronic journal. Available online: [<https://doi.org/10.5555/1248659.1248675>].
- [47] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space, 2013.
- [48] S. Minhas and A. Hussain. From spin to swindle: Identifying falsification in financial text. *Cognitive Computation*, 8(4):729–745, 2016. Available online: [[doi:10.1007/s12559-016-9413-9](https://doi.org/10.1007/s12559-016-9413-9)].
- [49] K. C. Moffitt. Using Lexical Bundles to Discriminate between Fraudulent and Non-fraudulent Financial Reports. In *Rutgers Accounting Web Spring Seminar*. Rutgers Accounting Web, 2011.
- [50] T. Motau. VBS Mutual Bank - The Great Bank Heist, 2018. Werksmans Attorneys, Sandton, South Africa. Available online: [<https://www.resbank.co.za/Lists/News%20and%20Publications/Attachments/>]

- [8830/VBS%20Mutual%20Bank%20-%20The%20Great%20Bank%20Heist.pdf](#)] (Accessed: 18 May 2019).
- [51] C. D. Paice. Another stemmer. *ACM SIGIR Forum*, 24(3):56–61, 1990. Electronic journal. Available online: [<https://doi.org/10.1145/101306.101310>].
- [52] R. Panai. Wordcloud of bigram using Python, 2019. Available online: [<https://stackoverflow.com/questions/49537474/wordcloud-of-bigram-using-python>] (Accessed: 16 Oct. 2020).
- [53] K. Pearson F.R.S. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900. Electronic journal. Available online: [<http://dx.doi.org/10.1080/14786440009463897>].
- [54] K. Pearson F.R.S. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. Electronic journal. Available online: [<https://doi.org/10.1080/14786440109462720>].
- [55] J. Perols. Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms. *Auditing*, 30(2):19–50, 2011. Available online: [[doi:10.2308/ajpt-50009](https://doi.org/10.2308/ajpt-50009)].
- [56] M. Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137, 1980. Electronic journal. Available online: [<https://doi-org.uplib.idm.oclc.org/10.1108/eb046814>].
- [57] Profile Group (Pty) Ltd. ShareData online, 2020. Available online: [<http://www.sharedata.co.za>] (Accessed: 10 June 2020).
- [58] L. Purda and D. Skillicorn. Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection. *Contemporary Accounting Research*, 32(3):1193–1223, 2015. Available online: [[doi:10.1111/1911-3846.12089](https://doi.org/10.1111/1911-3846.12089)].

- [59] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986. Electronic journal. Available online: [<https://doi.org/10.1007/BF00116251>].
- [60] Reuters. Steinhoff says could issue shares to help settle lawsuits, 2019. Eyewitness News, Johannesburg, South Africa. Available online: [<https://ewn.co.za/2019/11/08/steinhoff-says-could-issue-shares-to-help-settle-lawsuits>] (Accessed: 15 Nov. 2019).
- [61] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-august-2016*, pages 1135–1144, 2016. Available online: [<https://doi.org/10.1145/2939672.2939778>].
- [62] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature : International Weekly Journal of Science*, 323(6088):533–536, 1986. Electronic journal. Available online: [<https://doi.org/10.1038/323533a0>].
- [63] D. L. C. J. . W. T. Sanh, V. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, 2019.
- [64] A. Sawarn. Keras for Multi-label Text Classification, 2019. Towards AI - Multidisciplinary Science Journal. Available online: [<https://medium.com/towards-artificial-intelligence/keras-for-multi-label-text-classification-86d194311d0e>] (Accessed: 7 Sept. 2020).
- [65] Scikit-learn developers. Scikit-learn machine learning in python, 2020. Available online: [<https://scikit-learn.org/stable>] (Accessed: 16 Oct. 2020).
- [66] W. Silva, H. Kimura, and V. A. Sobreiro. An Analysis of the Literature on Systemic Financial Risk: A Survey. *Journal of Financial Stability*, 28, 2017. Available online: [[doi:10.1016/j.jfs.2016.12.004](https://doi.org/10.1016/j.jfs.2016.12.004)].

- [67] M. Stone. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147, 1974. Electronic journal. Available online: [<http://www.jstor.org/stable/2984809>].
- [68] The Financial Times Limited. The Banker Database, 2020. Available online: [<https://www.thebankerdatabase.com>] (Accessed: 10 June 2020).
- [69] V. N. Vapnik. *The nature of statistical learning theory*. Springer, New York, United States of America, 1995.
- [70] Vault. The Vault Exchange, 2020. Available online: [<https://thevault.exchange>] (Accessed: 11 June 2020).
- [71] D. Vu. Generating WordClouds in Python, 2019. Available online: [<https://www.datacamp.com/community/tutorials/wordcloud-python>] (Accessed: 17 Oct. 2020).
- [72] WordNet. About WordNet, 2010. Available online: [<https://wordnet.princeton.edu/>].
- [73] Y. Wu, H. Wang, B. Zhang, and K.-L. Du. Using radial basis function networks for function approximation and classification. *ISRN Applied Mathematics*, pages 1–34, 2012. Electronic journal. Available online: [<https://doi.org/10.5402/2012/324194>].
- [74] Yumpu. Yumpu, 2020. Available online: [<https://www.yumpu.com/en>] (Accessed: 11 June 2020).

Appendix A

Exploratory Data Analysis

Additional figures and tables are provided based on initial Exploratory Data Analysis (EDA).

- Section [A.1](#) profiles the text and features extracted from the annual reports;
- Section [A.2](#) considers the most significant predictors of risk from within the feature pool using tests for Chi-Squared, Information Gain and ANOVA;
- Section [A.3](#) displays additional wordclouds based on topic analysis;
- Section [A.4](#) provides visualisations based on dimension reduction;
- Section [A.5](#) concludes the appendix with a summary.

A.1 Data profiling

The pandas profiling report was used as part of the exploratory data analysis performed on the corpus of SA annual reports. 40 TFIDF features were selected to start the analysis since the SVMa model was able to return a F1 score = 1 using unigrams supplemented with the overall word count and scaled word count features from the wordlists. All available linguistic features were appended, namely all features derived from the wordlists and all readability indices from both the `py-readability-metrics` and `textstat` packages.

Table A.1 shows highly correlated features largely due to the use of multiple wordlists and multiple readability indices.

Table A.1: Correlation between features.

Feature 1	Feature 2	ρ
value	fair	0.94815
forward2	forward1	0.99935
LMneg	forward2	0.97104
LMpos	LMneg	0.92948
H6neg	LMpos	0.92046
H6pos	H6neg	0.9041
H8neg	H6pos	0.90235
H8pos	H8neg	0.90333
uncert	H8pos	0.95986
causal	uncert	0.97015
causalM	causal	0.96908
causalM50	causalM	0.99712
perf	causalM	0.9039
strat	causalM50	0.92046
forward2_scaled	forward1_scaled	0.99065
H6neg_scaled	LMneg_scaled	0.91718
H6pos_scaled	LMpos_scaled	0.97141
H8neg_scaled	H6neg_scaled	0.99967
H8pos_scaled	H6pos_scaled	0.99891
causalM50_scaled	causalM_scaled	0.97945
H6tone	LMtone	0.9443
H8tone	H6tone	0.99917
r_fk	r_gf	0.9947
r_ari	r_fk	0.97113
r_lw	r_ari	0.97654
r_sp	r_lw	0.95026

Continued on next page

Table A.1 – *Continued from previous page*

Feature 1	Feature 2	ρ
ts_fk	ts_gf	0.99075
ts_ari	ts_fk	0.97725
ts_sm	ts_fk	0.94018

Some of the word counts show correlations across different wordlists e.g.

- positive and negative words;
- uncertainty and Henry’s positive words;
- uncertainty and causal words.

A.1.1 Word count and pages

Overall word count measured the length of the narrative within annual reports. Figure A.1 shows the frequency distribution across the range of 2 442 words to 589 568 words.

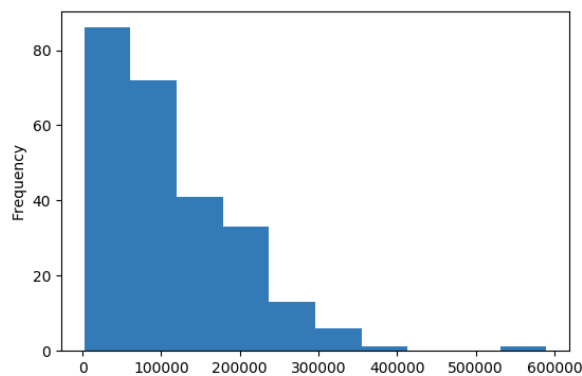


Figure A.1: Histogram of word count.

Table A.2 lists the 5 documents containing the most words.

APPENDIX A. EXPLORATORY DATA ANALYSIS

Table A.2: Top 5 documents with most wordCount.

Company		Document		pages	word Count
ID	Name	ID	Name		
188	Subsidiaries_of_ICICI _{Bank} _Annual_Report_FY2019_2018.pdf	32	ICICI Bank Ltd	748	589568
175	ICICI Bank annual report-subsidiaries-ar-2015-16.pdf	32	ICICI Bank Ltd	492	386033
174	ICICI Bank annual report-subsidiaries-ar-2014-15.pdf	32	ICICI Bank Ltd	444	342960
71	ddr2017_bnp-paribas_gb.pdf	12	Bnp Paribas South Africa Branch	580	326917
70	ddr2016gb.pdf	12	Bnp Paribas South Africa Branch	560	319964

The Indian multinational ICICI Bank was responsible for the 3 longest narratives followed by the French multinational Bnp Paribas for their 2017 and 2016 annual reports. By contrast, 3 of the Bnp Paribas reports are in the list of 5 shortest reports (Table A.2).

Table A.3: Top 5 documents with least wordCount.

Company		Document		pages	word Count
ID	Name	ID	Name		
236	NASDAQ_MBWM_2004.pdf	22	Mercantile Bank Ltd	12	2442
237	NASDAQ_MBWM_2005.pdf	22	Mercantile Bank Ltd	12	2488
66	bnp-ra2015-va_e-accessible.pdf	12	Bnp Paribas South Africa Branch	63	17201
69	bnp-paribas_ra_2014_en.02.pdf	12	Bnp Paribas South Africa Branch	64	18045
63	2018_bnp-paribas_integrated_report.pdf	12	Bnp Paribas South Africa Branch	56	18258

The 2004 and 2005 annual reports of Mercantile Bank Ltd were significantly shorter than the average 113 930 word count.

Table A.4 lists the 5 documents comprising the most pages.

APPENDIX A. EXPLORATORY DATA ANALYSIS

Table A.4: Top 5 documents with most pages.

Company		Document		pages	word Count
ID	Name	ID	Name		
188	Subsidiaries_of_ICICI_Bank_Annual_Report_FY2019_2018.pdf	32	ICICI Bank Ltd	748	589568
109	Deutsche_Bank_Annual_Report_2014.pdf	14	Deutsche Bank AG	610	314868
71	ddr2017-bnp-paribas.gb.pdf	12	Bnp Paribas South Africa Branch	580	326917
108	Deutsche_Bank_Annual_Report_2013.pdf	14	Deutsche Bank AG	572	292890

Only two of the reports with the longest narrative have the most pages. Two reports from Deutsche Bank AG and one report from Firststrand Bank Ltd complete the list.

Counts of words in wordlists were scaled relative to the number of words in the report overall. Table A.5 lists the 5 documents having the highest proportion of words from the uncertainty wordlist.

Table A.5: Top 5 documents with least pages.

Company		Document		pages	word Count
ID	Name	ID	Name		
236	NASDAQ_MBWM_2004.pdf	22	Mercantile Bank Ltd	12	2442
237	NASDAQ_MBWM_2005.pdf	22	Mercantile Bank Ltd	12	2488
68	bnp-paribas_2019_integrated_report_en.pdf	12	Bnp Paribas South Africa Branch	34	20990
298	Annual_Report_2017.pdf	25	S A Bank of Athens Ltd	35	28593
297	Annual_Report_2015.pdf	25	S A Bank of Athens Ltd	36	26154

A.1.2 Uncertain and causal word counts

Counts of words in wordlists were scaled relative to the number of words in the report overall. Table A.6 lists the 5 documents having the highest proportion of words from the uncertainty wordlist.

APPENDIX A. EXPLORATORY DATA ANALYSIS

Table A.6: Top 5 documents with most uncertainty.

Company		Document		uncert_ scaled
ID	Name	ID	Name	
279	SASFIN_ar_jun18.pdf	23	Sasfin Bank Ltd	0.048319
277	SASFIN_ar_jun16.pdf	23	Sasfin Bank Ltd	0.047951
278	SASFIN_ar_jun17.pdf	23	Sasfin Bank Ltd	0.046332
280	sasfin_integrated-report_2019.pdf	23	Sasfin Bank Ltd	0.045309
276	SASFIN_ar_jun15.pdf	23	Sasfin Bank Ltd	0.04496

All 5 documents are from Sasfin Bank Ltd exceeding the mean of 0.03 by 50% to 60%. Reports with the least use of uncertainty words are listed in Table A.7.

Table A.7: Top 5 documents with least uncertainty.

Company		Document		uncert_ scaled
ID	Name	ID	Name	
47	Bank of Canara Annual Report 2018 (5324830319).pdf	30	Bank of Canara	0.009793
1	2010- ABSA Group annual-report.pdf	2	Absa Group	0.010358
44	Bank of Canara Annual Report 2015 (5324830316).pdf	30	Bank of Canara	0.011112
35	Bank of Baroda Annualreport2016-17.pdf	26	Bank of Baroda	0.011219
34	Bank of Baroda Annualreport2015-16.pdf	26	Bank of Baroda	0.011661

Of the five reports using the least uncertainty words, Bank of Baroda has four with ABSA Group's 2010 report making up the balance. The frequency distribution of the scaled uncertainty word counts are graphed in Figure A.2.

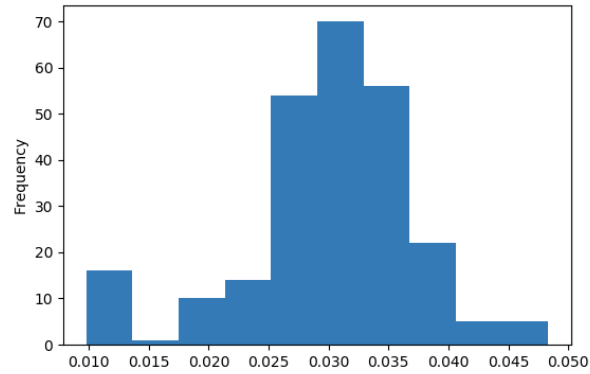


Figure A.2: Histogram of uncertain word count scaled.

Figures A.3 and A.4 show how the histograms for the causal and causal wordlists differ.

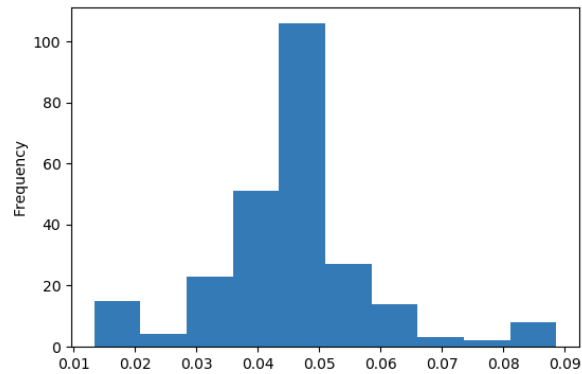


Figure A.3: Histogram of causal word count scaled.

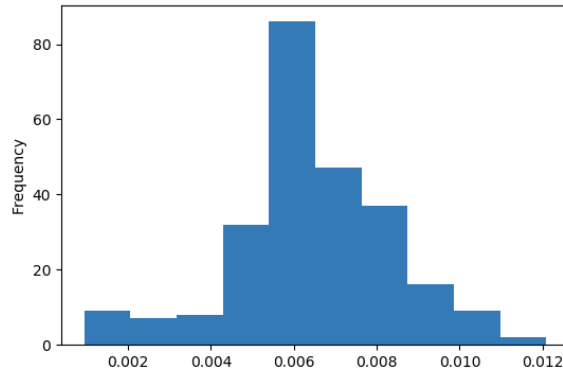


Figure A.4: Histogram of causalM word count scaled.

The reports of Mercantile Bank Ltd have the most causal words and the Bank of Canara the least, as reflected in Tables A.8 and A.9.

Table A.8: Top 5 documents with most causal words.

Company		Document		causal_
ID	Name	ID	Name	scaled
249	NASDAQ_MBWM.2018.pdf	22	Mercantile Bank Ltd	0.025576
250	NASDAQ_MBWM.2019.pdf	22	Mercantile Bank Ltd	0.025494
248	NASDAQ_MBWM.2017.pdf	22	Mercantile Bank Ltd	0.025314
246	NASDAQ_MBWM.2015.pdf	22	Mercantile Bank Ltd	0.0252
247	NASDAQ_MBWM.2016.pdf	22	Mercantile Bank Ltd	0.025152

Table A.9: Top 5 documents with least causal words.

Company		Document		causal_
ID	Name	ID	Name	scaled
44	Bank of Canara Annual Report 2015 (5324830316).pdf	30	Bank of Canara	0.004118
45	Bank of Canara Annual Report 2016 (5324830317).pdf	30	Bank of Canara	0.004124
41	Bank of Canara Annual Report 2012 (5324830314).pdf	30	Bank of Canara	0.004129
42	Bank of Canara Annual Report 2013 (5324830314).pdf	30	Bank of Canara	0.004403
43	Bank of Canara Annual Report 2014 (5324830315).pdf	30	Bank of Canara	0.004419

The causal wordlist had no effect on the bottom 5 Bank of Canara reports (Table A.9)

but returned a different leading document in the 2019 annual report of Sasfin Bank Ltd and four reports from Citibank N.A. (Table A.10)

Table A.10: Top 5 documents with most causalM words.

Company		Document		causalM_
ID	Name	ID	Name	scaled
280	sasfin.integrated-report.2019.pdf	23	Sasfin Bank Ltd	0.013532
101	Citi Bank NA Annual Report 2017.pdf	13	Citibank N.A	0.013231
98	Citi Bank NA Annual Report 2014.pdf	13	Citibank N.A	0.012986
100	Citi Bank NA Annual Report 2016.pdf	13	Citibank N.A	0.012626
102	Citi Bank NA Annual Report 2018.pdf	13	Citibank N.A	0.012615

Table A.11: Top 5 documents with least causalM words.

Company		Document		causalM_
ID	Name	ID	Name	scaled
44	Bank of Canara Annual Report 2015 (5324830316).pdf	30	Bank of Canara	0.001778
47	Bank of Canara Annual Report 2018 (5324830319).pdf	30	Bank of Canara	0.001778
41	Bank of Canara Annual Report 2012 (5324830314).pdf	30	Bank of Canara	0.00185
45	Bank of Canara Annual Report 2016 (5324830317).pdf	30	Bank of Canara	0.001918
42	Bank of Canara Annual Report 2013 (5324830314).pdf	30	Bank of Canara	0.001961

A.1.3 Performance, strategy and forward-looking word counts

Reports using the most words relating to the performance wordlist are given in Table A.12

Table A.12: Top 5 documents with most performance words.

Company		Document		perf_
ID	Name	ID	Name	scaled
290	Societe Generale_Financial Statements 2018.pdf	17	Societe Generale Johannesburg Branch	0.023395
291	Societe Generale_Financial Statements 2019.pdf	17	Societe Generale Johannesburg Branch	0.021181
337	State Bank of India_AR.2012_English.pdf	28	Bank of India - Johannesburg Branch	0.021171
287	Societe Generale_Financial Statements 2015.pdf	17	Societe Generale Johannesburg Branch	0.020694
288	Societe Generale_Financial Statements 2016.pdf	17	Societe Generale Johannesburg Branch	0.02041

The Johannesburg Branch of Societe Generale had 4 of the 5 reports with the most frequent reference to performance relative to the overall narrative. The 2012 report of the Johannesburg Branch of Bank of India also made the list. Mercantile Bank Ltd, Bnp Paribas South Africa Branch and Bank of Canara were responsible for reports with the least reference to performance (Table A.13).

Table A.13: Top 5 documents with least performance words.

Company		Document		perf_ scaled
ID	Name	ID	Name	
240	NASDAQ_MBWM_2009.pdf	22	Mercantile Bank Ltd	0.002207
69	bnp_paribas_ra_2014.en.02.pdf	12	Bnp Paribas South Africa Branch	0.002217
66	bnp-ra2015-va_e-accessible.pdf	12	Bnp Paribas South Africa Branch	0.0025
44	Bank of Canara Annual Report 2015 (5324830316).pdf	30	Bank of Canara	0.002672
42	Bank of Canara Annual Report 2013 (5324830314).pdf	30	Bank of Canara	0.002713

Figure A.5 shows a left skewed distribution of scaled performance word counts

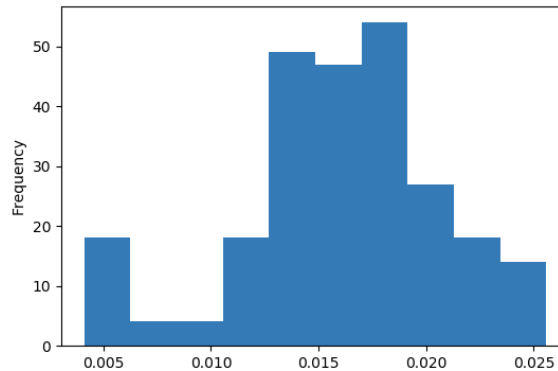


Figure A.5: Histogram of performance word count scaled.

This skewness shows that most reports make fewer references to performance than the median.

Reports with the most strategic words are given in Table A.14.

APPENDIX A. EXPLORATORY DATA ANALYSIS

Table A.14: Top 5 documents with most strategic words.

Company		Document		strat_ scaled
ID	Name	ID	Name	
279	SASFIN_ar_jun18.pdf	23	Sasfin Bank Ltd	0.088623
68	bnp_paribas_2019_integrated_report_en.pdf	12	Bnp Paribas South Africa Branch	0.08728
277	SASFIN_ar_jun16.pdf	23	Sasfin Bank Ltd	0.086836
63	2018_bnp_paribas_integrated_report.pdf	12	Bnp Paribas South Africa Branch	0.085223
12	african-bank-holdings-limited- annual-integrated-report-2019.pdf	34	African Bank Ltd (N)	0.084964

Sasfin Bank Ltd, Bnp Paribas South Africa Branch and the new African Bank Ltd report nearly twice the proportion of strategic words compared with the mean of 0.046. Bank of Canara were responsible for all 5 reports with the least strategic words as reflected in Table A.15.

Table A.15: Top 5 documents with least strategic words.

Company		Document		strat_ scaled
ID	Name	ID	Name	
47	Bank of Canara Annual Report 2018 (5324830319).pdf	30	Bank of Canara	0.01341
44	Bank of Canara Annual Report 2015 (5324830316).pdf	30	Bank of Canara	0.01456
41	Bank of Canara Annual Report 2012 (5324830314).pdf	30	Bank of Canara	0.015002
45	Bank of Canara Annual Report 2016 (5324830317).pdf	30	Bank of Canara	0.015538
46	Bank of Canara Annual Report 2017 (5324830318).pdf	30	Bank of Canara	0.016244

Figure A.6 shows a normal distribution of scaled strategic word counts with increased frequencies at lower and upper values.

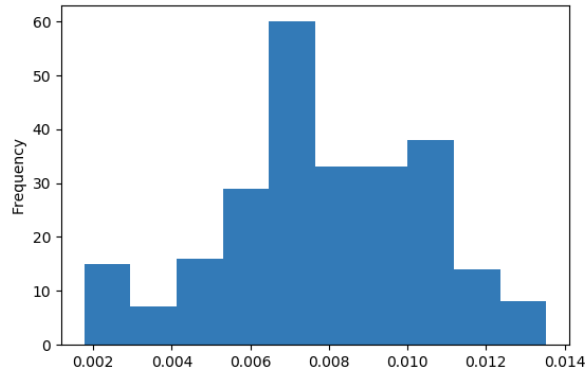


Figure A.6: Histogram of strategic word count scaled.

Table A.16 reveals that the 2016 report of the new African Bank Ltd has highest proportion of forward-looking words with Mercantile Bank Ltd making up the balance.

Table A.16: Top 5 documents with most forward-looking words.

Company		Document		forward1_ scaled
ID	Name	ID	Name	
14	doc-1-african-bank-ir-28-nov-2016.pdf	34	African Bank Ltd (N)	0.012089
240	NASDAQ_MBWM_2009.pdf	22	Mercantile Bank Ltd	0.011918
250	NASDAQ_MBWM_2019.pdf	22	Mercantile Bank Ltd	0.010784
245	NASDAQ_MBWM_2014.pdf	22	Mercantile Bank Ltd	0.010493
248	NASDAQ_MBWM_2017.pdf	22	Mercantile Bank Ltd	0.010468

Similar to the use of strategic words, the Bank of Canara has 4 of 5 reports with the least forward-looking words (Table A.17).

Table A.17: Top 5 documents with least forward-looking words.

Company		Document		forward1_ scaled
ID	Name	ID	Name	
1	2010- ABSA Group annual-report.pdf	2	Absa Group	0.000945
45	Bank of Canara Annual Report 2016 (5324830317).pdf	30	Bank of Canara	0.00163
42	Bank of Canara Annual Report 2013 (5324830314).pdf	30	Bank of Canara	0.001793
43	Bank of Canara Annual Report 2014 (5324830315).pdf	30	Bank of Canara	0.001821

Continued on next page

Table A.17 – Continued from previous page

Company		Document		forward1_
ID	Name	ID	Name	scaled
33	Bank of Baroda Annualreport2014-15.pdf	26	Bank of Baroda	0.001897

Figure A.7 shows a normal distribution of scaled forward-looking word counts.

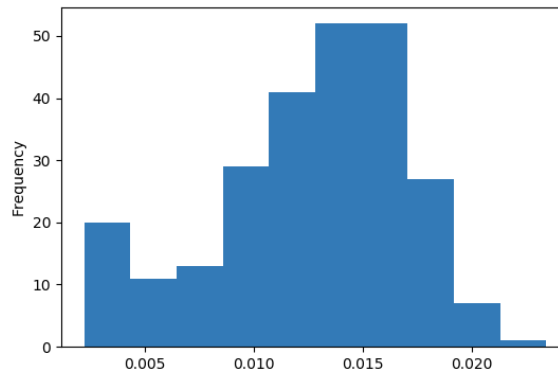


Figure A.7: Histogram of forward-looking word count scaled.

A.2 Feature importance

The literature review on feature extraction (Section 2.4) identified two papers in which the Chi-Squared dependencies between features and output classes were used for feature selection. Since the review also questioned the suitability of this test for numeric features, Information Gain and ANOVA were also performed.

A.2.1 Chi-Squared

The Pearson Chi-Squared statistical test [53] was applied to both content and linguistic features to understand the goodness of fit. Chi-Squared tests the relationship between features (independent input variables) and the target class (dependent output variable). These results can be used to change the features selected to train the classifiers.

The null hypothesis was that the feature and the risk classification are independent of each other. The alternative hypothesis was that the risk classification is dependent

on the feature. Only features with p-values lower than the significance level of 5% were assumed dependent. This provided 95% certainty that the null hypothesis could be rejected.

A total of 40 TFIDF features were selected to start the analysis, since the SVMa model was able to return an F1 score = 1 using features when supplemented with the overall word count and scaled word count features from the wordlists. Features were normalised since Chi-Squared cannot process negative values as was found in the calculations of net tone. Figure A.8 lists the first 40 features in order of significance along the x-axis and their p-value along the y-axis relative to the significance level.

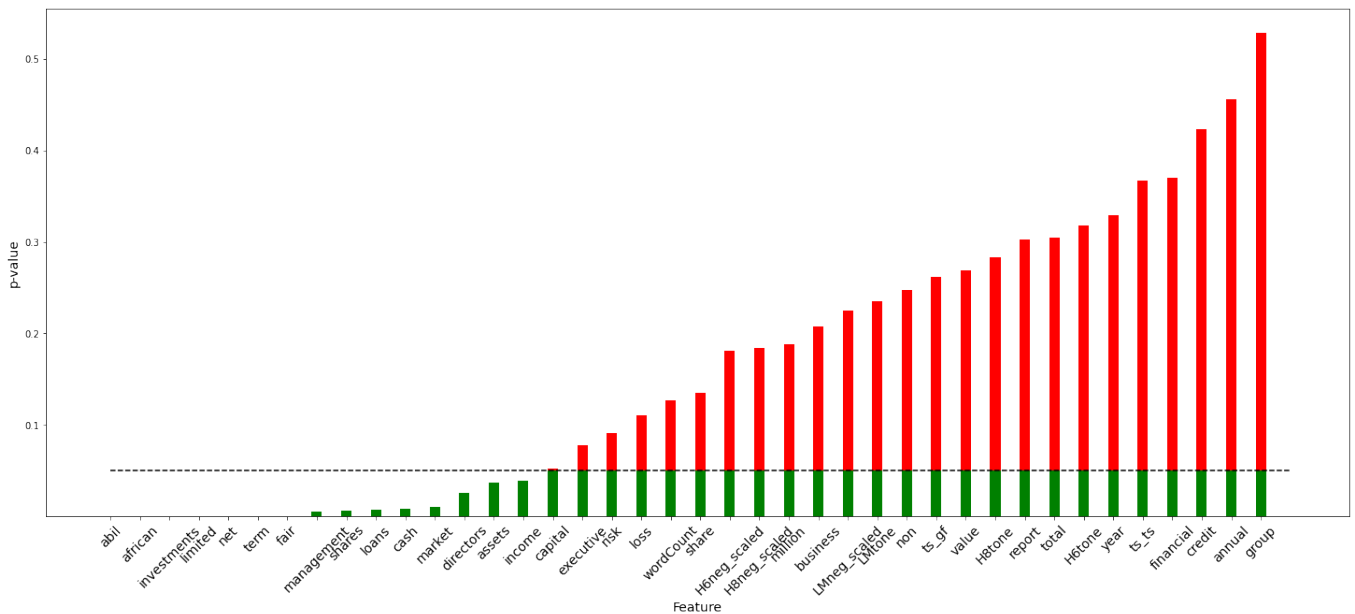


Figure A.8: Chi-Squared on 40 TFIDFs, selected counts and readability indices.

Merely 15 features were deemed significant for having p-values <0.05. All of these were word frequencies. The report length, represented by the overall word count, was the most significant other feature type with p-value=0.13. Features derived from wordlists did not appear beneath the 5% significance level; however the three negative wordlists (LMneg, H6neg and H8neg) had their scaled counts and net tone reflected in the top 40 features. Of the three `textstat` readability indices selected, Gunning Fog (`ts_gf`) was more significant than the consensus Text Standard (`ts_ts`) with p-value=0.26 compared

with 0.37.

It is unsurprising to see ‘abil’, ‘african’, ‘investments’ and ‘limited’ since African Bank was one of two banks for which reports were labelled with risk. On the other hand, it is interesting to note that the frequency of terms like ‘net’, ‘term’, ‘fair’, ‘cash’, ‘market’, ‘directors’, ‘assets’ and ‘income’ are statistically significant.

The second scenario tested all available linguistic features. Absolute word counts were appended to the scaled counts of words from the wordlists. Also, all readability indices from both the `py-readability-metrics` and `textstat` packages were included. Figure A.9 lists the first 40 features in order of significance from the larger feature set.

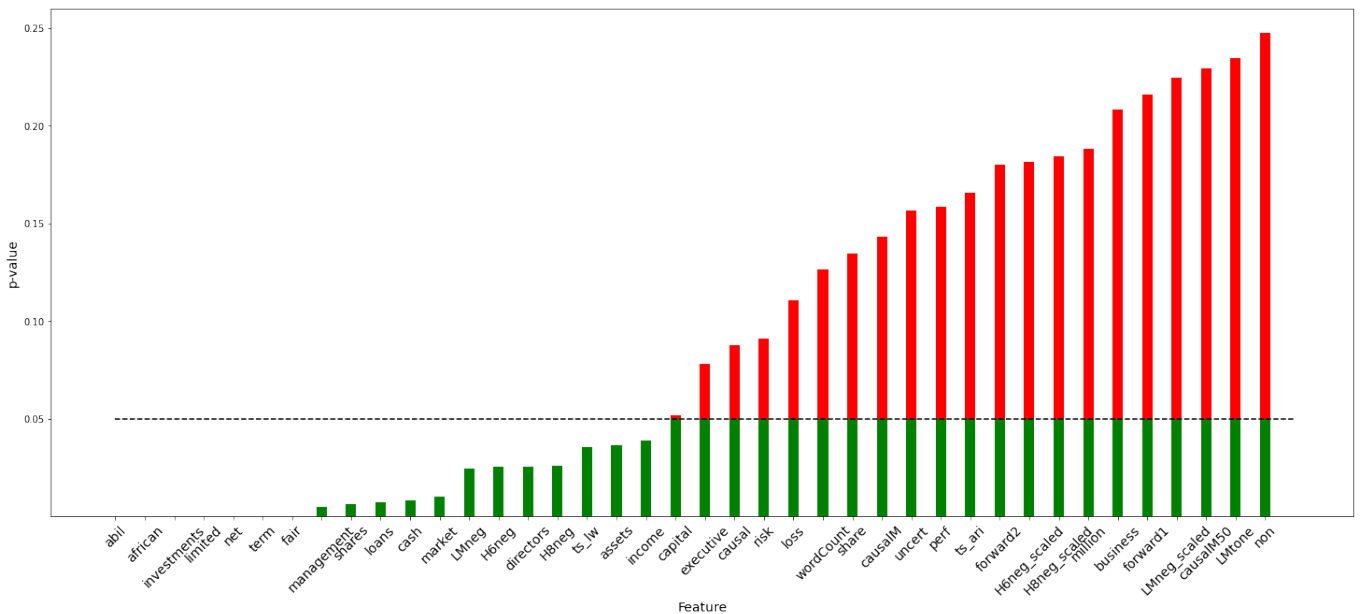


Figure A.9: Chi-Squared on 40 TFIDFs, all wordlist counts and readability indices.

The absolute count of words in the three negative wordlists (LMneg, H6neg and H8neg) had p-values beneath the 5% significance level (0.0244, 0.0253 and 0.0258) and lower than their scaled counterparts. Of the 20 readability indices (Table 3.5) it was only Linsear Write that fell below the threshold and only the Automated Readability Index within the top 40. Absolute counts of words in the causal and causal word lists preceded counts for the uncertainty (uncert), performance (uncert) and forward looking wordlists (forward2 and forward1).

A.2.2 Information Gain

The Information Gain theoretical test [59, 39] was applied to the features from the second Chi-Squared test (Figure A.9) namely 40 TFIDF features and available linguistic features. Figure A.10 shows the top thirty features in terms of importance.

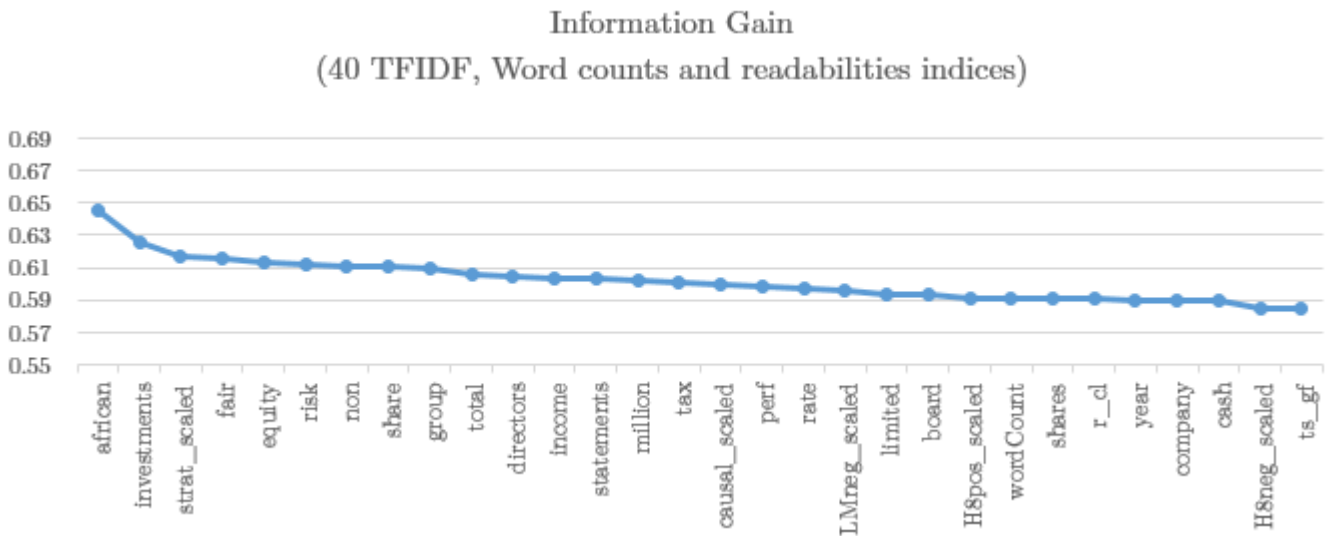


Figure A.10: Information Gain on 40 TFIDFs, all wordlist counts and readability indices.

The scaled count of strategic keywords *strat_scaled* was placed third. This was the highest rank for a linguistic feature across the four methods used (RFE ranked the SMOG readability index forth). The other scaled word counts were from causal, negative (LMneg) and positive (H8pos) wordlists in positions 16, 19 and 22. The absolute count of performance words was ranked 17. It is not recommended however that the absolute counts get used as these are a function of the report length. The overall word count (position 23) can be used as a measure of report length. Readability indices show up in positions 25 and 30, specifically the Coleman-Liau and Gunning Fog metrics respectively.

A.2.3 ANOVA

The Analysis of variance (ANOVA) statistical test [20] was applied to the same 40 TFIDF features and available linguistic features as used with Chi-Squared and Information Gain.

Figure A.11 shows the top thirty features in terms of importance.

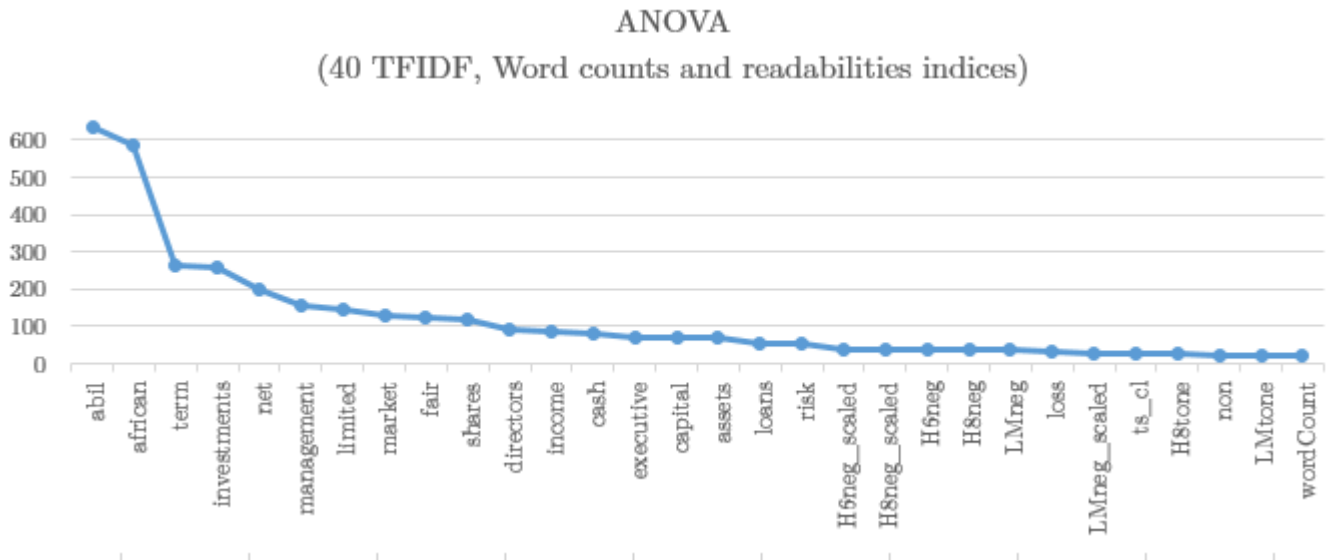


Figure A.11: Information Gain on 40 TFIDFs, all wordlist counts and readability indices.

The token ‘african’ carries more relative weight in the ANOVA score compared with Information Gain. Also, ANOVA ranks the token ‘abil’ first whereas it did not appear in the top-thirty features using Information Gain. Three scaled counts and three absolute counts from negative word list are prioritised. Given similar purpose and correlation, only the scaled count for the *H6neg* wordlist need be used. The absolute word count and Coleman-Liau readability metric makes the list again as they did with the Information Gain. Note that Coleman-Liau was the only readability metric in the top-thirty list. The ANOVA scores also give elevated importance to the net tone using the *H8* word list.

A.3 Topic analysis

A range of topics from 2 to 5 were assumed for LDA. This section contrasts unigrams, bigrams and trigrams using wordclouds to cluster tokens for the assumed topics within sections of VBS Mutual Bank annual reports.

A.3.1 2016 annual report

This subsection analyses topics within the 2016 annual report of VBS Mutual Bank.

Results of the topic analysis within the Audit and risk committee report section of the annual report are shown below.

Figure B.6 places wordclouds for 4- and 5-grams side-by-side assuming one topic for LDA.



Figure A.12: 4- and 5-gram wordclouds for one topic in VBS Bank Audit and risk committee report 2016.

Figure A.13 assumed two topics for the LDA.



Figure A.13: 4- and 5-gram wordclouds for two topics in VBS Bank Audit and risk committee report 2016.

A.4 Visualisation

A few plots using Principal Component Analysis (PCA) [54] and t-distributed Stochastic Neighbour Embedding (t-SNE) [31] give some pictorial insight into the annual report text.

A.4.1 Principal Component Analysis

Tokens across all reports were embedded using Continuous Bag of Words (CBOW) into vectors with 300 dimensions. PCA reduced these dimensions into four components. Two of the components that explained most of the variance (0.09 and 0.07) are plotted in Figure A.14 for the first 1000 tokens of each class.

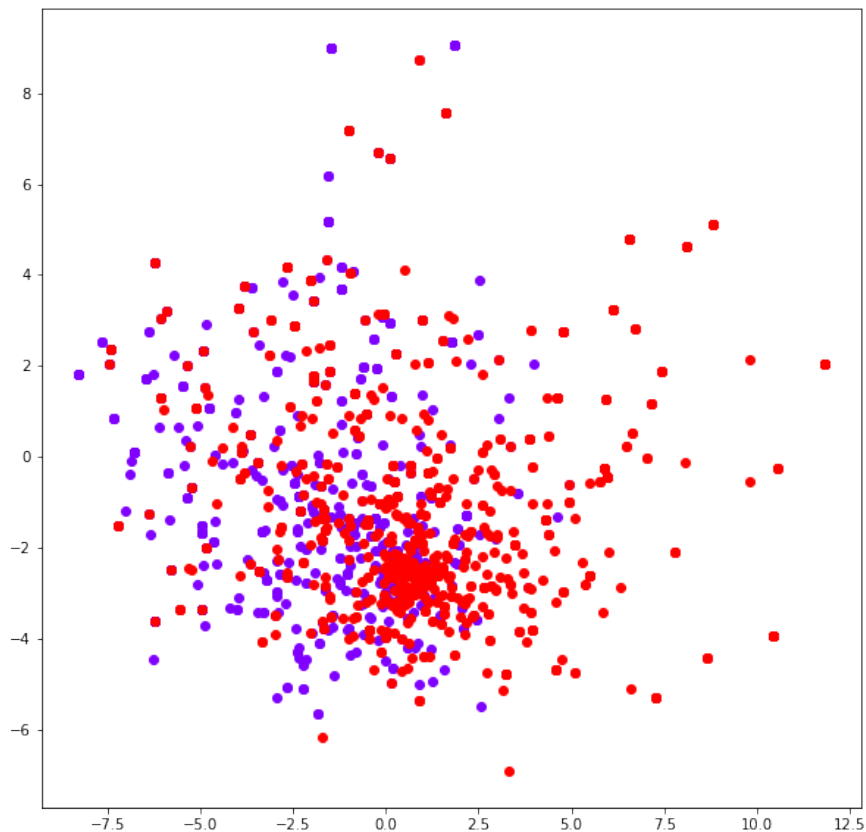


Figure A.14: PCA for the first 1000 words per class.

Features for the low risk documents (purple dots) and the high risk documents (red dots) are highly correlated with much of the scatter plots overlapping. There is a clear distinction of features for high risk documents in the right-hand region of the plot. As it is only the first 1000 tokens of each class, the plot is not representative of all the text nor of the most significant features.

A.4.2 t-distributed Stochastic Neighbour Embedding

t-SNE of the same 1000 tokens using 2 components is plotted in Figure A.15.

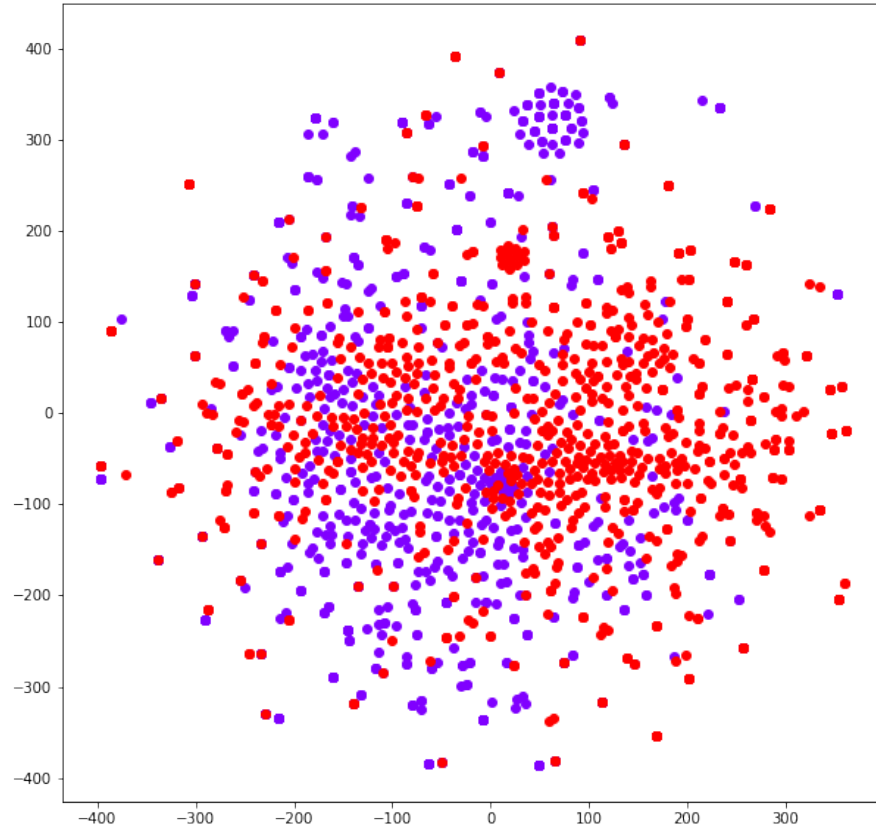


Figure A.15: t-SNE for the first 1000 words per class.

Some of the clusters in this diagram are due to repeated words. For instance, the top purple cluster is the token ‘absa’ repeated for each occurrence.

In Figure A.16, the size of the plot represents the average frequency per token over documents for the class. In total, 80 TFIDF features were embedded using CBOW. Again t-SNE reduced the vectors with 300 dimensions into two components.

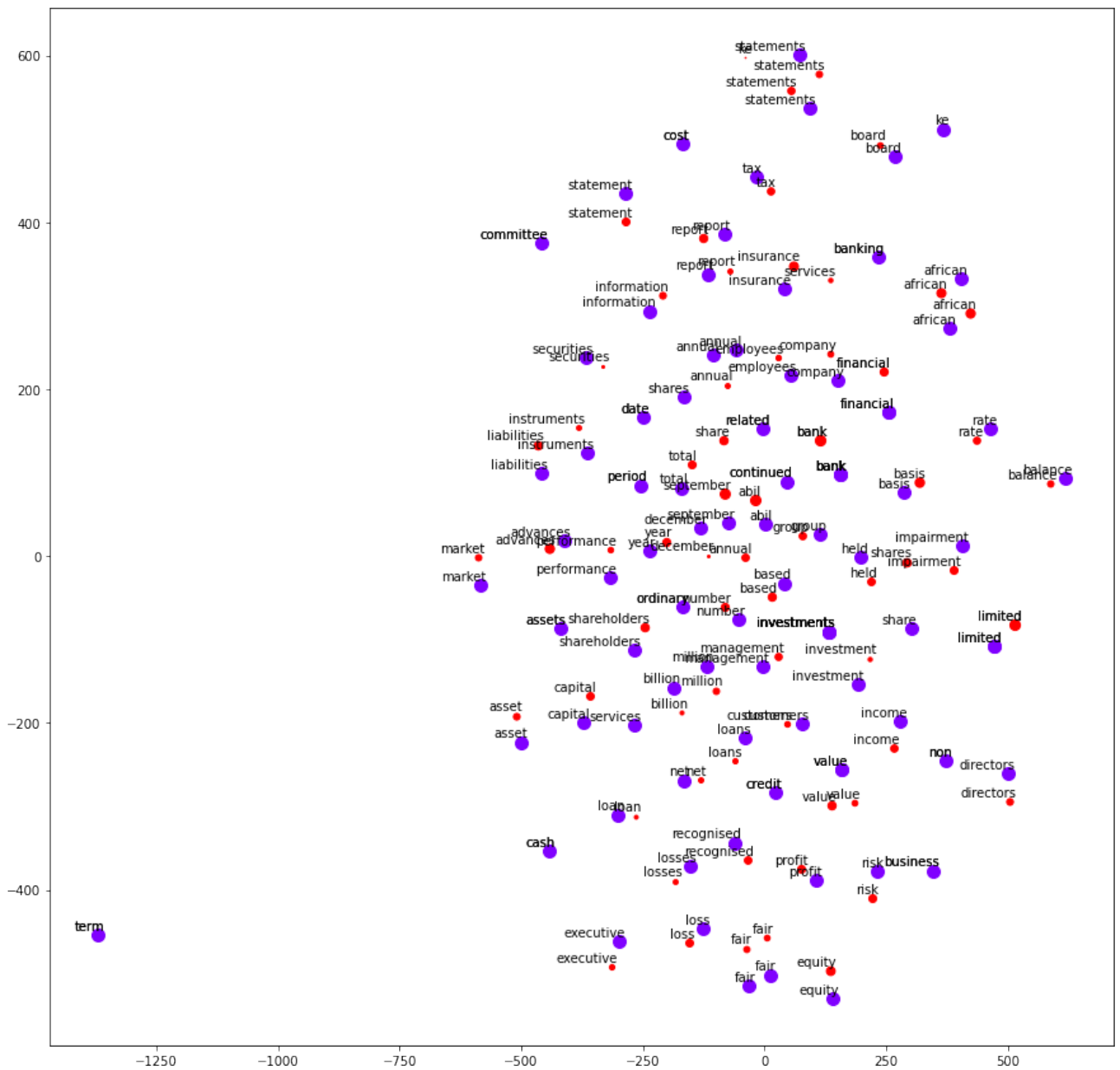


Figure A.16: t-SNE for 80 TFIDF tokens per class.

The distinction between the classes is a combination of the word vector position and the frequency of the word.

A.5 Summary

Although EDA was not a focus within the methodology (Section 3) this appendix recorded some visualisations that prove insightful and inspire further exploration. Section A.1 showed some feature correlations, feature histograms and top 5 documents for some of the feature values. Given its use in the literature (see Section 2), Chi-Squared was tested in Section A.2 to propose important features in predicting risk from the annual reports. Chi-Squared proposed absolute word counts from the three negative word lists probably due to the discrete values resembling categorical values as apposed to scaled counts and tone that had real, continuous values. Nonetheless, linguistic features representing negativity also appeared in the Information Gain and ANOVA tests. Both Information Gain and ANOVA tests highlighted overall word count and the Coleman-Liau readability metric. Only Information Gain highlighted strategic words. Additional word clouds based on 2 to 5 topics were shown in Section A.3. Various scatter plots in Section A.4 visualised words in an effort to distinguish the two classes.

Appendix B

Supporting information

This appendix is a container of artefacts that elaborate on the results reported in Chapter 4. Including this detail in the body of the report may have obscured the story line. Instead, references are made from the body to elements within the appendix to provide further insight depending on the reader's interest.

- Section [B.1](#) provides outputs and findings using the CFIE–FRSE tool on both UK and SA annual reports;
- Section [B.2](#) lists annual reports comprising multiple documents;
- Section [B.3](#) show various implementation details in preparing the data for machine learning including validation and anomalies;
- Section [B.4](#) provides results to compare classifier performance;
- Section [B.5](#) compares sentences before and after the cdQA paragraph filter was applied;
- Section [B.6](#) summarises the detailed results of the appendix.

B.1 Feasibility test

The following annual reports of UK companies existed in the pdfs folder of the CFIE–FRSE repository downloaded from GitHub:

- AI CLAIM SOLUTIONS PLC REPORT AND ACCOUNTS 2007.pdf;
- BIOCOMPATIBLES INTERNATIONAL PLC REPORT AND ACCOUNTS 2004.pdf;
- Consumers_Association.pdf;
- hartford group plc report and accounts 2005.pdf;
- HelpForHeroes.pdf.

The first run of the CFIE–FRSE tool produced an output file which contained results for all five documents. The next three tables show a subset of the output file by focusing on the annual report of AI CLAIM SOLUTIONS.

Table B.1 shows the headers that were identified and how they were mapped to a predefined header class.

Table B.1: UK report header identification and classification.

Header ID	HeaderText	Header Type
49	About Ai Claims Solutions	0
50	Why our ethical business model is a recipe for success	0
51	Questions and answers	0
52	How have we performed?	0
53	Chairman’s Statement	1
54	Chief Executive’s Review	2
55	Finance Director’s Review	6
56	Board of Directors	0
57	Executive Management	0
58	Remuneration Report	4
59	Report of the Directors	0
60	Report of the Independent Auditor to the Members of Ai Claims Solutions PLC	9

Continued on next page

APPENDIX B. SUPPORTING INFORMATION

Table B.1 – *Continued from previous page*

Header ID	HeaderText	Header Type
61	Consolidated Profit and Loss Account_Consolidated Statement of Total Recognised Gains and Losses_CONCAT	0
62	Consolidated Balance Sheet	0
63	Company Balance Sheet	0
64	Consolidated Cash Flow Statement_Reconciliation of Net Cash Flow to Movement in Net Debt_CONCAT	0
65	Notes to the Financial Statements	0
66	Five Year Summary	0
67	Notice of Annual General Meeting	0
68	Officers and Professional Advisers	0

HeaderText for header 64 was “Consolidated Cash Flow Statement_Reconciliation of Net Cash Flow to Movement in Net Debt_CONCAT” - a concatenation of two headers that appeared on the same page, namely

- Consolidated Cash Flow Statement;
- Reconciliation of Net Cash Flow to Movement in Net Debt.

The same applied to header 61. Headers 53, 54, 58 and 60 were mapped to header types 1, 2, 4 and 9 respectively. Section 3.4.7 provides a description of these header types. These types were preconfigured in the CFIE–FRSE tool.

Table B.2 identifies start and end page numbers for each header in the report.

Table B.2: UK report header start and end identifiers.

HeaderText	Header Page	Actual Page	End Page	Section Number Pages
About Ai Claims Solutions	2	4	5	2
Why our ethical business model is a recipe for success	4	6	7	2

Continued on next page

APPENDIX B. SUPPORTING INFORMATION

Table B.2 – *Continued from previous page*

HeaderText	Header Page	Actual Page	End Page	Section Number Pages
Questions and answers	6	8	9	2
How have we performed?	8	10	11	2
Chairman’s Statement	10	12	13	2
Chief Executive’s Review	12	14	18	5
Finance Director’s Review	17	19	19	1
Board of Directors	18	20	20	1
Executive Management	19	21	21	1
Remuneration Report	20	22	22	1
Report of the Directors	21	23	24	2
Report of the Independent Auditor to the Members of Ai Claims Solutions PLC	23	25	25	1
Consolidated Profit and Loss Account_Consolidated Statement of Total Recognised Gains and Losses_CONCAT	24	26	26	1
Consolidated Balance Sheet	25	27	27	1
Company Balance Sheet	26	28	28	1
Consolidated Cash Flow Statement_Reconciliation of Net Cash Flow to Movement in Net Debt_CONCAT	27	29	29	1
Notes to the Financial Statements	28	30	44	15
Five Year Summary	43	45	45	1
Notice of Annual General Meeting	44	46	47	2
Officers and Professional Advisers	46	48	52	5

The ActualPage column name is somewhat misleading as it refers to the actual PDF page number as displayed in an application like Adobe Reader and not the (actual) page number published on the report. Page 1 refers to the first PDF page as there is no page zero. The EndPage uses the same numbering range and refers to the PDF page at the end of the file or prior to the next header. The HeaderPage column name refers to the page numbers published on the report which are not the same as the ActualPage for the AI Claim Solutions 2007 report.

Counts of words overall and words specific to wordlists in Table B.3 are per section of the AI Claim Solutions 2007 report.

All five UK reports had 0 values for *bookletFlag* and had 0 values for *NoTOC* columns

APPENDIX B. SUPPORTING INFORMATION

meaning that the sample documents could be processed as they were not booklets and had tables of content bookmarks.

Table B.3: Word counts and wordlist-specific counts of UK reference report.

Header ID	Word Count	Forward Looking	LM		Henry				uncertainty	causal	causal-Martin	Performance
			neg	pos	Neg 2006	Pos 2006	Neg 2008	Pos 2008				
49	641	5	37	20	0	12	0	12	24	4	2	7
50	302	2	13	22	6	12	6	12	25	4	4	6
51	923	13	14	32	2	15	2	14	42	9	4	2
52	405	1	11	8	2	7	2	7	15	2	1	15
53	1506	17	20	65	3	53	3	52	75	28	17	26
54	2259	30	43	96	6	75	6	73	112	21	12	16
55	1101	7	18	31	8	27	8	26	44	33	27	26
56	690	1	5	12	0	8	0	8	12	3	0	1
57	448	1	15	5	0	5	0	5	4	4	1	3
58	549	6	4	14	3	11	3	11	21	7	1	3
59	1390	14	31	11	8	8	8	8	24	16	3	6
60	725	2	14	2	1	0	1	0	10	3	1	4
61	170	0	5	1	0	0	0	0	3	1	1	16
62	161	0	6	1	3	1	3	1	1	0	0	2
63	156	0	6	1	3	1	3	1	1	0	0	2
64	211	0	5	2	1	2	1	2	2	1	0	0
65	5424	33	120	51	40	40	40	40	86	67	16	87
66	188	0	12	2	2	0	2	0	4	3	2	13
67	2165	51	35	18	14	13	13	13	39	24	7	2
68	367	0	12	3	0	2	0	2	4	0	0	0

The 2015 annual report of VBS Mutual Bank used in the second run of the tool had the bit set for the *NoTOC* column. Figure B.1 shows that the report had two bookmarks but that these were not a table of contents.

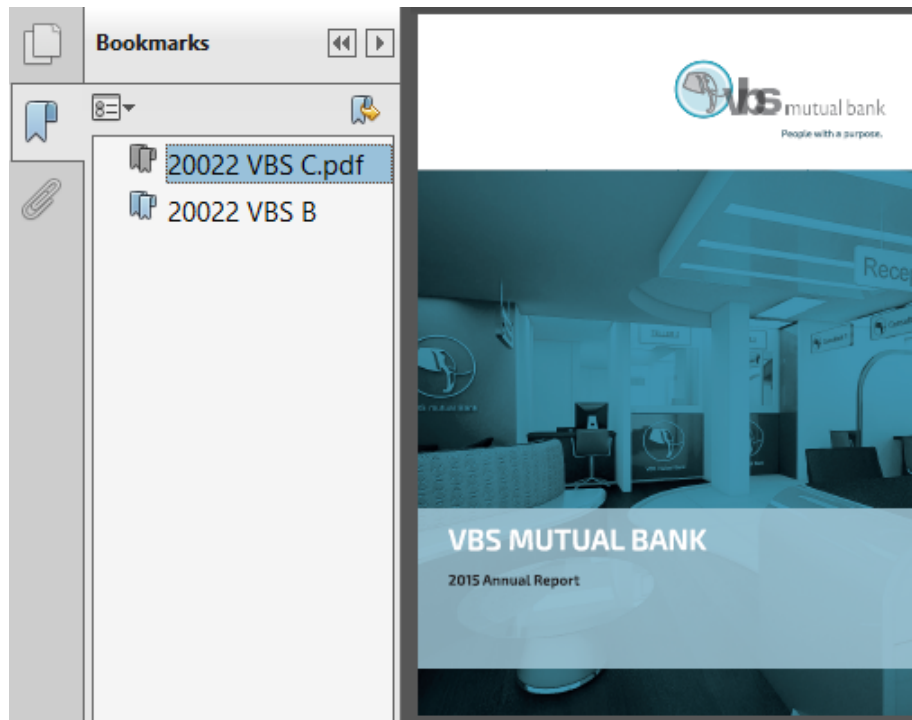


Figure B.1: Bookmarks of VBS Mutual Bank 2015 annual report.

B.2 Data collection

Of the 258 documents downloaded, 238 represent one annual report each. Table B.4 shows the exceptions where annual reports comprise multiple documents.

Table B.4: Annual reports comprising many documents.

Company	Year	Filename
Investec Group	2014	Investec-DLC-Annual-Report-2014.pdf
		Investec-DLC-annual-report-volume-2.pdf
		Investec-DLC-annual-report-volume-3.pdf
	2015	annual-report-2015-volume-1.pdf
		annual-report-2015-volume-2.pdf
		annual-report-2015-volume-3.pdf
	2016	2016-annual-report-volume-1.pdf
		2016-annual-report-volume-2.pdf
		2016-annual-report-volume-3.pdf
	2018	FINAL-Investec-DLC-Vol 1-15-June-2018.pdf
		FINAL-Investec-DLC-Vol2-15-June-2018.pdf
		FINAL-Investec-DLC-Vol3-15-June-2018.pdf
	2019	Investec-DLC-annual-report-volume-1.PDF
		Investec-DLC-annual-report-volume-2.pdf
		Investec-DLC-annual-report-volume-3.pdf
Nedbank Group	2010	2010_Nedbank_Group_Annual_Report_Volume_1.pdf
		2010_Nedbank_Group_Annual_Report_Volume_2.pdf
Standard Bank Group	2011	1428497731-2011-Book-I-Business-Review.pdf
		1428497824-2011-Book-II-Risk-and-Capital-Management-Report-and-AFS.pdf
		1428498029-2011-Book-III-Shareholder-Information.pdf

A limitation of this study is that these 20 documents were treated as 20 separate annual reports.

B.3 Data preparation

B.3.1 Text extraction

Figure B.3 provides an illustration of how PyMuPDF’s default text extraction did not maintain the natural human reading order in processing multiple columns.

Raw PDF document:

WHAT DO WE MEAN BY ETHICAL?	CLAIMS REFERRAL	VEHICLE REPAIR MANAGEMENT
At Ai Claims Solutions we recognise that, in the end, all services are paid for by policyholders.	State-of-the-art Contact Centre Following motoring accidents, insurance companies, brokers, bodyshops and	Guaranteed work, rapid response Our Vehicle Repair Management services actively manage the workflow through our

PyMuPDF’s default text extraction:

WHAT DO WE MEAN BY ETHICAL? CLAIMS REFERRAL VEHICLE REPAIR MANAGEMENT
At Ai Claims Solutions we recognise State-of-the-art Contact Centre Guaranteed work, rapid response
that, in the end, all services are Following motoring accidents, insurance Our Vehicle Repair Management services
paid for by policyholders. companies, brokers, bodyshops and actively manage the workflow through our

Figure B.3: PyMuPDF’s default text extraction across columns.

In method 2, the first block of the 2009 ABSA Group annual report was extracted as follows:

(265.5209045410156, 810.9342041015625, 528.857421875, 820.6837158203125, ‘1\nAbsa Group Limited Annual report for the year ended 31 December 2009’, 0, 0)

Table B.5 shows the effect of using PyMuPDF’s `fitz flag = ‘TEXT_INHIBIT_SPACES’` in `getText(‘blocks’, flags)`.

APPENDIX B. SUPPORTING INFORMATION

Table B.5: Flags for text extraction.

TEXT_PRESERVE_LIGATURES	TEXT_PRESERVE_LIGATURES TEXT_INHIBIT_SPACES
'Lower than anticipated economic growth rates, unemployment trends, rising costs and greater consumer indebtedness all contributed to a tough economic environment.'	'Lowerthananticipatedeconomicgrowth rates,unemploymenttrends,risingcostsand greaterconsumerindebtednessall contributedtoatougheconomicenvironment.'

The flag '*TEXT_INHIBIT_SPACES*' was not used since the fidelity of some individual words was lost. Text extracted per block using Algorithm 3.1 took 31 minutes. Warnings received are reflected in Table B.6.

Table B.6: Warnings on text block extraction.

ID	Document Name	Warning
14	doc-1-african-bank-ir-28-nov-2016.pdf	invalid page object
27	Albaraka_Annual_Report_2016.pdf	expected 'obj' keyword (3084714 0 ?)
28	Bank of Baroda Annual report2009-10.pdf	zlib error: invalid distance too far back
109	Deutsche_Bank_Annual_Report_2014.pdf	invalid page object
224	Ithala Integrated Annual Report 2013 - 14.pdf	FT_New_Memory_Face(LSYZZZ+Avenir-Roman-SC700): unknown file format FT_New_Memory_Face(LSYZZZ+Avenir-Heavy-SC700): unknown file format

B.3.2 Readability score

Scores obtained for the two metrics *GunningFog* and *FleschReadingEase* using the `py-readability-metrics` package on the UK reference report are shown in Table B.7.

Table B.7: Validation of `py-readability-metrics` of UK reference.

Header Text	Gunning Fog			Flesch Reading Ease		
	Reference	r	Variance	Reference	r	Variance
Finance Director's Review	18	16	13%	51	47	8%
Board of Directors	22	14	35%	37	34	9%
Executive Management	20	15	28%	41	40	4%
Remuneration Report	21	12	41%	36	36	-1%
Report of the Independent Auditor to the Members of Ai Claims Solutions PLC	24	19	21%	38	18	53%
Consolidated Profit and Loss Account_ Consolidated Statement of Total Recognised Gains and Losses.CONCAT	57	41	28%	-69	-11	84%
Consolidated Balance Sheet	23	67	-195%	25	-75	399%
Company Balance Sheet	21	65	-207%	30	-67	327%
Consolidated Cash Flow Statement _Reconciliation of Net Cash Flow to Movement in Net Debt.CONCAT	69	92	-32%	-80	-131	-64%
Five Year Summary	26	32	-21%	32	26	20%

The readability of the two sections relating to 'Balance Sheet' was much worse than the CFIE results. The difference can be attributed to the number of sentences calculated.

The Gunning Fog metric was calculated on the UK extracted text for 'Consolidated Balance Sheet' using first principles and a score of 19 was obtained. This assumed three sentences based on the presence of three colon ':' characters. The statistics for `py-readability-metrics` confirmed that it had assumed one sentence. Inspection of the text revealed the following substrings where colons occurred:

- ‘Creditors: Amounts falling due within one year’;
- ‘Creditors: Amounts falling due after one year’;
- ‘The Financial Statements were approved by the Board on 10 October 2007 and signed on its behalf by: David Sandhu’.

The `py-readability-metrics` package gave errors on 5 pages that had less than the 100 words that it required (see Table B.8). Also, it could only calculate the SMOG metric for 7 of the 52 pages that had at least 30 sentences.

The `textstat` package did not report any errors and calculated scores for every page. In particular, Table B.8 shows its results for pages where the `py-readability-metrics` gave errors.

Table B.8: UK report pages scored by `textstat`.

pdf Page	ts_gf	ts_f	ts_fk	ts_dc	ts_ari	ts_cl	ts_lw	ts_sm	ts_ts
0	14	43	12	9	13	12	13	0	13
48	16	35	13	10	15	16	14	0	14
49	12	44	10	9	7	10	6	0	10
50	0	207	-16	0	0	-16	-1	0	0
51	26	19	21	11	27	16	32	0	0

The scores from the `textstat` package were compared with the `py-readability-metrics` in Table B.9.

Table B.9: Py-readability-metrics compared with `textstat`.

Header Text	Gunning Fog			Flesch Reading Ease		
	ts	r	Variance	ts	r	Variance
Finance Director’s Review	17	16	5%	30	47	-56%
Board of Directors	17	14	18%	21	34	-59%
Executive Management	16	15	8%	28	40	-42%
Remuneration Report	13	12	3%	42	36	13%

Continued on next page

APPENDIX B. SUPPORTING INFORMATION

Table B.9 – *Continued from previous page*

Header Text	ts	r	Variance	ts	r	Variance
Report of the Independent Auditor to the Members of Ai Claims Solutions PLC	20	19	1%	16	18	-11%
Consolidated Profit and Loss Account_ Consolidated Statement of						
Total Recognised Gains and Losses.CONCAT	39	41	-5%	-168	-11	94%
Consolidated Balance Sheet	68	67	1%	-227	-75	67%
Company Balance Sheet	66	65	1%	-231	-67	71%
Consolidated Cash Flow Statement						
_Reconciliation of Net Cash Flow to Movement in Net	89	92	-2%	-265	-131	51%
Debt.CONCAT						
Five Year Summary	40	32	21%	-83	26	131%

From the Gunning Fog Equation (2.1) it is clear that the score is very sensitive to the number of sentences derived from the text. Unsurprisingly, variation was greatest on sections displaying numeric data with little commentary. The variation was accepted since this study focuses on the narrative and similar results were obtained using two different packages. Also, developing code to calculate the readability metric more accurately or appropriately was outside of the scope of this study.

B.3.3 Header identification

Table B.10 is one example of a document where the bookmark text did not provide the header text.

Table B.10: First 5 bookmarks extracted as an example of non-header text.

Document		Bookmark	
ID	Name	No	Text
0	2009-ABSA Group annual-report.pdf	1	NEW COVER A4
0	2009-ABSA Group annual-report.pdf	2	1-3
0	2009-ABSA Group annual-report.pdf	3	●4-5
0	2009-ABSA Group annual-report.pdf	4	6
0	2009-ABSA Group annual-report.pdf	4	7-9

Table B.11 shows a breakdown of the 4 084 headers extracted.

Table B.11: Header counts for documents with table of contents extracted.

ID	Document name	Header Count
48	BANK OF CHINA LTD JHB BRANCH 2015 Annual Report.pdf	42
65	bnp-ra-en-accessible.pdf	16
66	bnp-ra2015-va_e-accessible.pdf	26
69	bnp_paribas_ra_2014_en_02.pdf	32
74	capitec_integratedannualreport_2009.pdf	32
76	capitec_integratedannualreport_2011.pdf	29
86	China Construction Bank Corporation Annual Report 2012.pdf	24
87	China Construction Bank Corporation Annual Report 2013.pdf	27
88	China Construction Bank Corporation Annual Report 2014.pdf	28
89	China Construction Bank Corporation Annual Report 2015.pdf	39

Continued on next page

APPENDIX B. SUPPORTING INFORMATION

Table B.11 – *Continued from previous page*

ID	Document name	Header Count
90	China Construction Bank Corporation Annual Report 2016.pdf	37
91	China Construction Bank Corporation Annual Report 2017.pdf	37
92	China Construction Bank Corporation Annual Report 2018.pdf	36
100	Citi Bank NA Annual Report 2016.pdf	107
101	Citi Bank NA Annual Report 2017.pdf	105
103	Citi Bank NA Annual Report 2019.pdf	102
108	Deutsche Bank Annual Report 2013.pdf	176
162	HSBC - -holdings-plc-annual-report-and-accounts-2015.pdf	214
163	HSBC - 170221-annual-report-and-accounts-2016.pdf	505
165	HSBC - 190219-annual-report-and-accounts-2018.pdf	116
168	HSBC - annual-report-and-accounts-2010.pdf	8
169	HSBC - annual-report-and-accounts-2011.pdf	146
170	HSBC - annual-report-and-accounts-2012.pdf	371
225	JP Morgan Chase and Co - annualreport-2009.pdf	29
226	JP Morgan Chase and Co - annualreport-2010.pdf	68
227	JP Morgan Chase and Co - annualreport-2011.pdf	136
228	JP Morgan Chase and Co - annualreport-2012.pdf	128
229	JP Morgan Chase and Co - annualreport-2013.pdf	166
230	JP Morgan Chase and Co - annualreport-2014.pdf	164
231	JP Morgan Chase and Co - annualreport-2015.pdf	164
232	JP Morgan Chase and Co - annualreport-2016.pdf	153
233	JP Morgan Chase and Co - annualreport-2017.pdf	136
234	JP Morgan Chase and Co - annualreport-2018.pdf	143
235	JP Morgan Chase and Co - annualreport-2019.pdf	148
245	NASDAQ_MBWM_2014.pdf	23
247	NASDAQ_MBWM_2016.pdf	75
248	NASDAQ_MBWM_2017.pdf	12

Continued on next page

APPENDIX B. SUPPORTING INFORMATION

Table B.11 – Continued from previous page

ID	Document name	Header Count
261	NedbankIR2013.pdf	50
301	1428495414-SBSA-AR-2012.pdf	42
308	1524203604-SBSAFY17AnnualReport.pdf	39
328	standard-chartered-plc-full-year-2013-report.pdf	59
329	standard-chartered-plc-full-year-2014-report.pdf	94

Figure B.4 is an example where the contents page does not have a header and the contents layout is not a table but rather a journey.

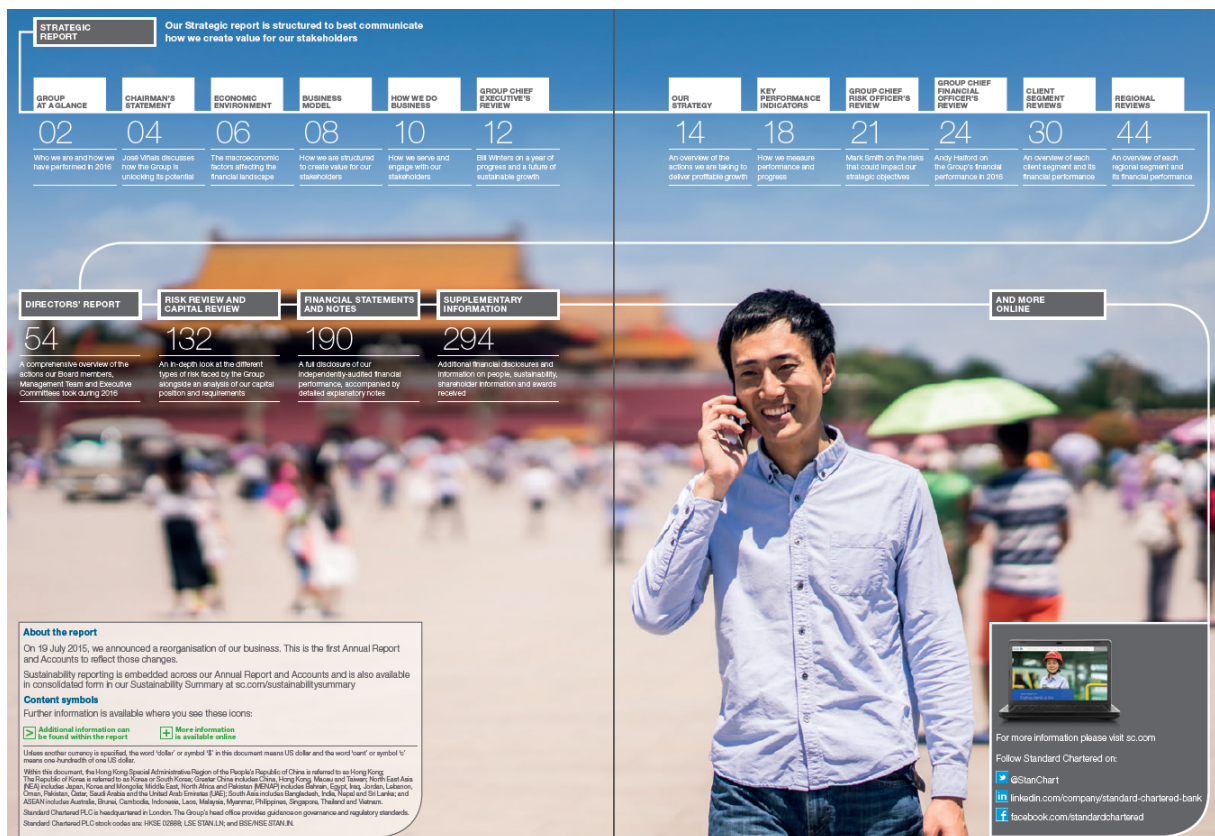


Figure B.4: Journey layout of contents in 2016 annual report of Standard Chartered PLC.

Algorithm 3.2 did not identify this as a page with a table of contents as it does display

the words ‘contents’ or ‘index’. Furthermore, headers have their page numbers displayed beneath them.

Figure B.5 has an example of another contents page from the same report that was detected by Algorithm 3.2.

NOTES TO THE FINANCIAL STATEMENTS		
Contents – Notes to the financial statements		
CONTENTS		
Section	Note	Page
Basis of preparation	1 Accounting policies	207
Performance/return	2 Segmental information	209
	3 Net interest income	210
	4 Net fees and commission	211
	5 Net trading income	212
	6 Other operating income	212
	7 Operating expenses	213
	8 Impairment losses on loans and advances and other credit risk provisions	214
	9 Other impairment	216
	10 Taxation	217
	11 Dividends	220
	12 Earnings per ordinary share	221
Assets and liabilities held at fair value	13 Financial instruments	221
	14 Derivatives financial instruments	237
	15 Investment securities	240
Financial instruments held at amortised cost	16 Loans and advances to banks and customers	241
Other assets and investments	17 Goodwill and intangible assets	244
	18 Property, plant and equipment	246
	19 Operating lease commitments	247
	20 Other assets	247
Funding, accruals, provisions, contingent liabilities and legal proceedings	21 Debt securities in issue	248
	22 Other liabilities	249
	23 Provisions for liabilities and charges	249
	24 Contingent liabilities and commitments	250
	25 Legal and regulatory matters	250
Capital instruments, equity and reserves	26 Subordinated liabilities and other borrowed funds	252
	27 Share capital, other equity instruments and reserves	253
	28 Non-controlling interests	256
Employee benefits	29 Retirement benefit obligations	256
	30 Share based payments	261
Scope of consolidation	31 Investment in subsidiary undertakings, joint ventures and associates	265
	32 Structured entities	268
Cash flow statement	33 Cash flow statement	270
	34 Cash and cash equivalents	271
Other disclosure matters	35 Related party transactions	271
	36 Post balance sheet events	272
	37 Auditor's remuneration	272
	38 Standard Chartered PLC (Company)	273
	39 Related undertakings of the Group	276
	40 IFRS 9 Financial Instruments	291

Figure B.5: Contents for subset of 2016 annual report of Standard Chartered PLC.

B.3.4 Section aggregation

Annual reports were found to have ‘out-of-context’ text interlaced between the text of interest. See Figure B.6 for an example of a value statement within the Chief Executive’s Review section of the UK reference report.



Figure B.6: ‘Out-of-context’ text within section of UK reference report.

Words counted between headers and attributed to the section are inflated by these inserts.

B.3.5 Document aggregation

Two annual reports of Canara Bank had no wordlist counts. From Figure B.7 it can be seen that they have two-column format. The first column is their native language using a non-Latin alphabet and the second column is English.



Figure B.7: Foreign language in Canara Bank Annual Report 2010 (5324830310).pdf.

Despite the dual language and dual column, this was not deemed to be the root cause of the problem. Opening the documents in Adobe Reader, launched a dialog box

warning that the document is trying to open in full screen mode. Figure B.8 shows the dialog.

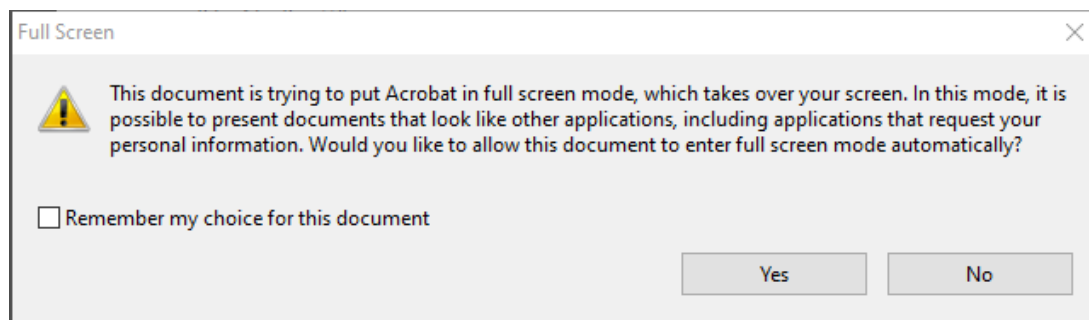


Figure B.8: Warning dialog upon opening Canara Bank reports.

B.3.6 Validation

Table B.12 shows the word counts per section of the UK reference report validated against the CFIE-FRSE tool.

Table B.12: Word count validated against UK reference report by section.

Header Text	CFIE-FRSE	Word Count	Delta
About Ai Claims Solutions	641	642	1
Why our ethical business model is a recipe for success	302	302	0
Questions and answers	923	925	2
How have we performed?	405	408	3
Chairman's Statement	1506	1510	4
Chief Executive's Review	2259	2266	7
Finance Director's Review	1101	1102	1
Board of Directors	690	691	1
Executive Management	448	448	0

Continued on next page

Table B.12 – *Continued from previous page*

Header Text	CFIE- FRSE	Word Count	Delta
Remuneration Report	549	549	0
Report of the Directors	1390	1423	33
Report of the Independent Auditor to the Members of Ai Claims Solutions PLC	725	725	0
Consolidated Profit and Loss			
Account_Consolidated Statement of Total Recognised Gains and Losses_CONCAT	170	211	41
Consolidated Balance Sheet	161	181	20
Company Balance Sheet	156	176	20
Consolidated Cash Flow			
Statement_Reconciliation of Net Cash Flow to Movement in Net Debt_CONCAT	211	257	46
Notes to the Financial Statements	5424	5938	514
Five Year Summary	188	200	12
Notice of Annual General Meeting	2165	2150	-15
Officers and Professional Advisers	367	373	6

Table B.13 shows the counts and variances for the causal wordlist.

Table B.13: Causal wordlist count validated against UK
reference report by section.

Header Text	CFIE- FRSE	Word Count	Delta
About Ai Claims Solutions	4	4	0
Why our ethical business model is a recipe for success	4	4	0
Questions and answers	9	9	0
How have we performed?	2	2	0

Continued on next page

Table B.13 – *Continued from previous page*

Header Text	CFIE- FRSE	Word Count	Delta
Chairman's Statement	28	27	-1
Chief Executive's Review	21	21	0
Finance Director's Review	33	33	0
Board of Directors	3	3	0
Executive Management	4	4	0
Remuneration Report	7	7	0
Report of the Directors	16	16	0
Report of the Independent Auditor to the Members of Ai Claims Solutions PLC	3	3	0
Consolidated Profit and Loss			
Account_Consolidated Statement of Total Recognised Gains and Losses_CONCAT	1	1	0
Consolidated Balance Sheet	0	0	0
Company Balance Sheet	0	0	0
Consolidated Cash Flow			
Statement_Reconciliation of Net Cash Flow to Movement in Net Debt_CONCAT	1	1	0
Notes to the Financial Statements	67	67	0
Five Year Summary	3	3	0
Notice of Annual General Meeting	24	24	0
Officers and Professional Advisers	0	0	0

B.3.7 Training and test data sets

Table [B.14](#) lists 48 documents held out for testing.

APPENDIX B. SUPPORTING INFORMATION

Table B.14: Documents in test data set.

Company		Document			Risk
ID	Name	ID	Name	Year	
35	African Bank Ltd (O)	6	1464958705-4144_African-Bank-2014annualreport.pdf	2014	1
1	VBS Mutual Bank	358	VBS-2016-Annual-Report.pdf	2016	1
34	African Bank Ltd (N)	13	african-bank-integrated-report-2018-final-webs...	2018	0
19	Albaraka Bank Ltd	22	alBaraka 2018 AnnualReport.pdf	2018	0
26	Bank of Baroda	36	Bank of Baroda Annualreport2017-18.pdf	2018	0
30	Canara Bank	47	Canara Bank Annual Report 2018 (5324830319)...	2018	0
20	Bidvest Bank Ltd	62	Bidvest+Bank+Annual+Report,+year+ended+30+June...	2018	0
12	Bnp Paribas South Africa Branch	63	2018_bnp_paribas_integrated_report.pdf	2018	0
36	Capitec Bank	73	capitec.bank.integrated_annual_report_2018.pdf	2018	0
31	China Construction Bank Corporation - Jhb Branch	92	China Construction Bank Corporation Annual Rep...	2018	0
13	Citibank N.A	102	Citi Bank NA Annual Report 2018.pdf	2018	0
14	Deutsche Bank AG	113	Deutsche_Bank_Annual_Report_2018.pdf	2018	0
5	Firststrand Bank Ltd	126	firststrand-annual-report 2018.pdf	2018	0
37	Grindrod Bank	132	2018 Integrated Annual Report Grindrod.IAR.201...	2018	0
15	HSBC Bank Plc - Johannesburg Branch	165	HSBC - 190219-annual-report-and-accounts-2018.pdf	2018	0
32	ICICI Bank Ltd	177	ICICI Bank annual-report-fy2018 (2017).pdf	2018	0
7	Investec Bank Ltd	190	FINAL-Investec-IBP -15-June-2018-22h04.pdf	2018	0
21	Ithala Ltd	222	Ithala Corporate Annual Report (final) AUGUST ...	2018	0
16	JP Morgan Chase and Co	234	JP Morgan Chase and Co - annualreport-2018.pdf	2018	0
22	Mercantile Bank Ltd	249	NASDAQ_MBWM_2018.pdf	2018	0
9	Nedbank Ltd	259	2018 Nedbank Limited Annual Report.pdf	2018	0
23	Sasfin Bank Ltd	279	SASFIN_ar_jun18.pdf	2018	0
17	Societe Generale Johannesburg Branch	290	Societe Generale_Financial Statements 2018.pdf	2018	0
25	S A Bank of Athens Ltd	299	GroCap_Financials_IB_30_April 2018.pdf	2018	0

Continued on next page

APPENDIX B. SUPPORTING INFORMATION

Table B.14 – *Continued from previous page*

Company		Document			Risk
ID	Name	ID	Name	Year	
11	Standard Bank of S A Ltd	309	1555481849-SBSAAAnnualReport2018LORESSingles.pdf	2018	0
18	Standard Chartered Bank	333	standard-chartered-plc-full-year-2018-report.pdf	2018	0
28	Bank of India - Johannesburg Branch	342	State Bank of India_AR_2018_English.pdf	2018	0
34	African Bank Ltd (N)	12	african-bank-holdings-limited-annual-integrate...	2019	0
19	Albaraka Bank Ltd	23	alBaraka 2019 Full Annual Report.pdf	2019	0
26	Bank of Baroda	37	Bank of Baroda Annualreport2018-19.pdf	2019	0
27	Bank of China Ltd Jhb Branch	51	BANK OF CHINA LTD JHB BRANCH 2019 Annual Repor...	2019	0
12	Bnp Paribas South Africa Branch	68	bnp-paribas_2019_integrated_report_en.pdf	2019	0
36	Capitec Bank	82	integrated_annual_report_2019.pdf	2019	0
13	Citibank N.A	103	Citi Bank NA Annual Report 2019.pdf	2019	0
14	Deutsche Bank AG	114	Deutsche_Bank_Annual_Report_2019.pdf	2019	0
5	Firststrand Bank Ltd	127	firststrand-annual-report 2019.pdf	2019	0
37	Grindrod Bank	133	2019 Integrated Annual Report.pdf	2019	0
15	HSBC Bank Plc - Johannesburg Branch	166	HSBC - 200218-annual-report-and-accounts-2019.pdf	2019	0
32	ICICI Bank Ltd	188	Subsidiaries_of_ICICI_Bank_Annual_Report_FY201...	2019	0
7	Investec Bank Ltd	195	investec-limited-annual-report-2019.pdf	2019	0
16	JP Morgan Chase and Co	235	JP Morgan Chase and Co - annualreport-2019.pdf	2019	0
22	Mercantile Bank Ltd	250	NASDAQ_MBWM_2019.pdf	2019	0
9	Nedbank Ltd	260	2019 Nedbank Limited Consolidated Annual Finan...	2019	0
23	Sasfin Bank Ltd	280	sasfin.integrated-report.2019.pdf	2019	0
17	Societe Generale Johannesburg Branch	291	Societe Generale.Financial Statements 2019.pdf	2019	0
11	Standard Bank of S A Ltd	310	1587728295-SBSA2019AnnualReport.pdf	2019	0
18	Standard Chartered Bank	334	standard-chartered-plc-full-year-2019-report.pdf	2019	0
28	Bank of India - Johannesburg Branch	343	State Bank of India_AR_2019_English.pdf	2019	0

The 6 positives samples within the original training set of 210 documents are reflected in Table B.15.

Table B.15: Positive documents in training data set.

Company		Document		
ID	Name	ID	Name	Year
35	African Bank Ltd (O)	5	1464958532-AfricanBank_IR_FY20131.pdf	2013
35	African Bank Ltd (O)	8	1517467421-ABILIAR2012completelowres.pdf	2012
35	African Bank Ltd (O)	9	2009 Full report - PDF 7MB - African Bank.pdf	2009
35	African Bank Ltd (O)	10	2011 Full integrated annual report - African B...	2011
35	African Bank Ltd (O)	11	ABIL_ar_sep10.pdf	2010
1	VBS Mutual Bank	357	VBS-2015-Annual-Report.pdf	2015

B.3.8 Bag of words

A total of 15 370 829 words were extracted from 259 annual reports in the complete data set using regular expression pattern 1. However, the second regular expression pattern (Option 2) extracted 136 429 tokens compared with 60 722 of pattern 1 for a particular report. For other documents, the variance was smaller, e.g. 79 296 compared with 62 968. Out of concern that the first pattern was filtering out too much text, the second pattern was selected for the remainder of the study.

B.4 Classifier performance

B.4.1 Grid search on TFIDF

All classifiers were trained and tested on unigram TFIDF features normalised using the min-max scaler of Equation (3.3). Systematically reducing the number of features from 100 to 20 in decrements of 20 showed that test performance diminished below 40 as illustrated in Figure B.9.

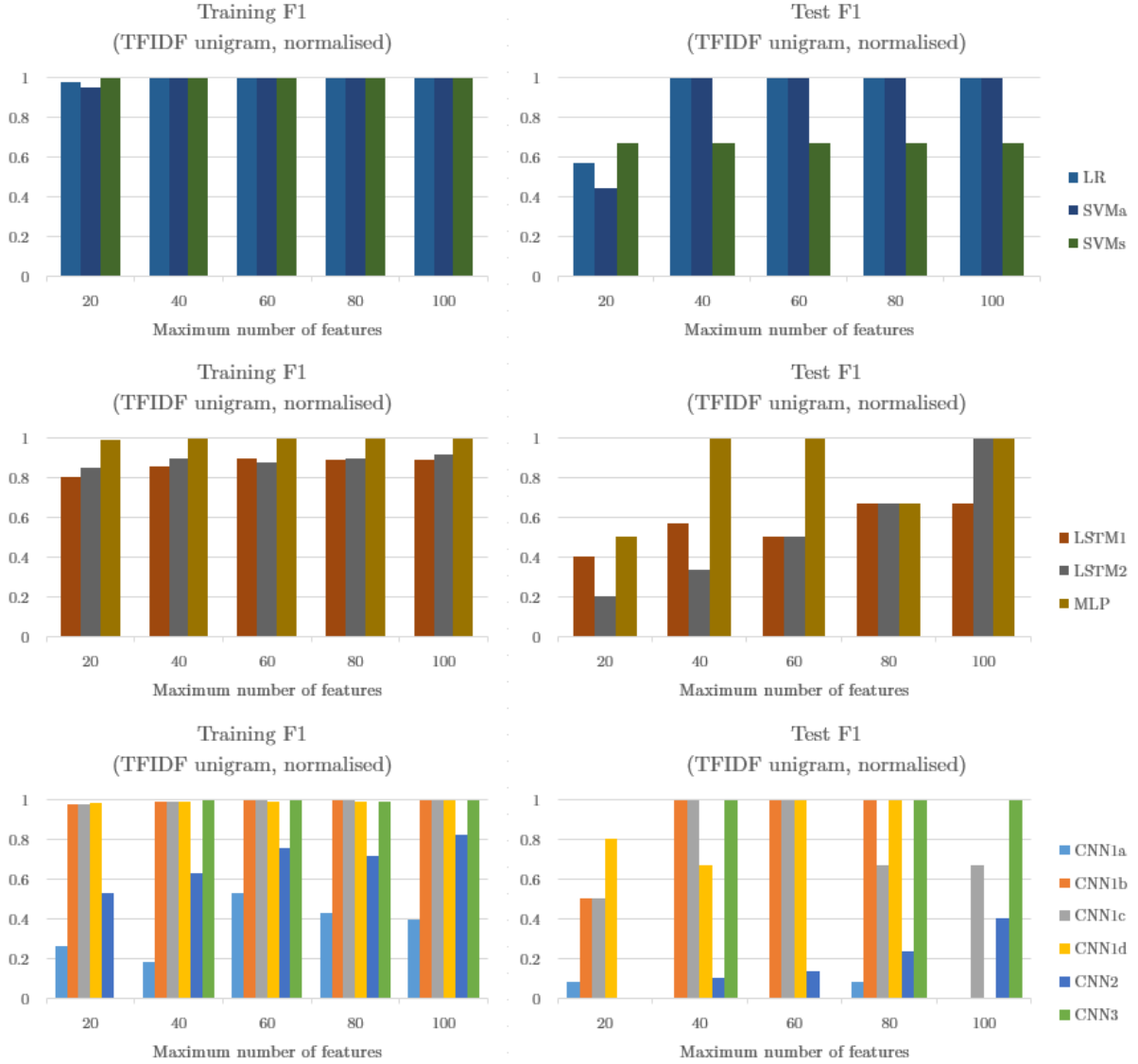


Figure B.9: Performance of classifiers on limited TFIDFs (normalised).

The SVMs model showed the best training performance but did not generalise well. The LR and SVMa models had better test performance than the SVMs with 40 unigrams still returning F1=1. This was also accomplished by four modern classifiers (CNN1b, CNN1c, CNN3 and MLP).

Figure B.10 repeats the analysis for features standardised using the z-score Equation (3.4).

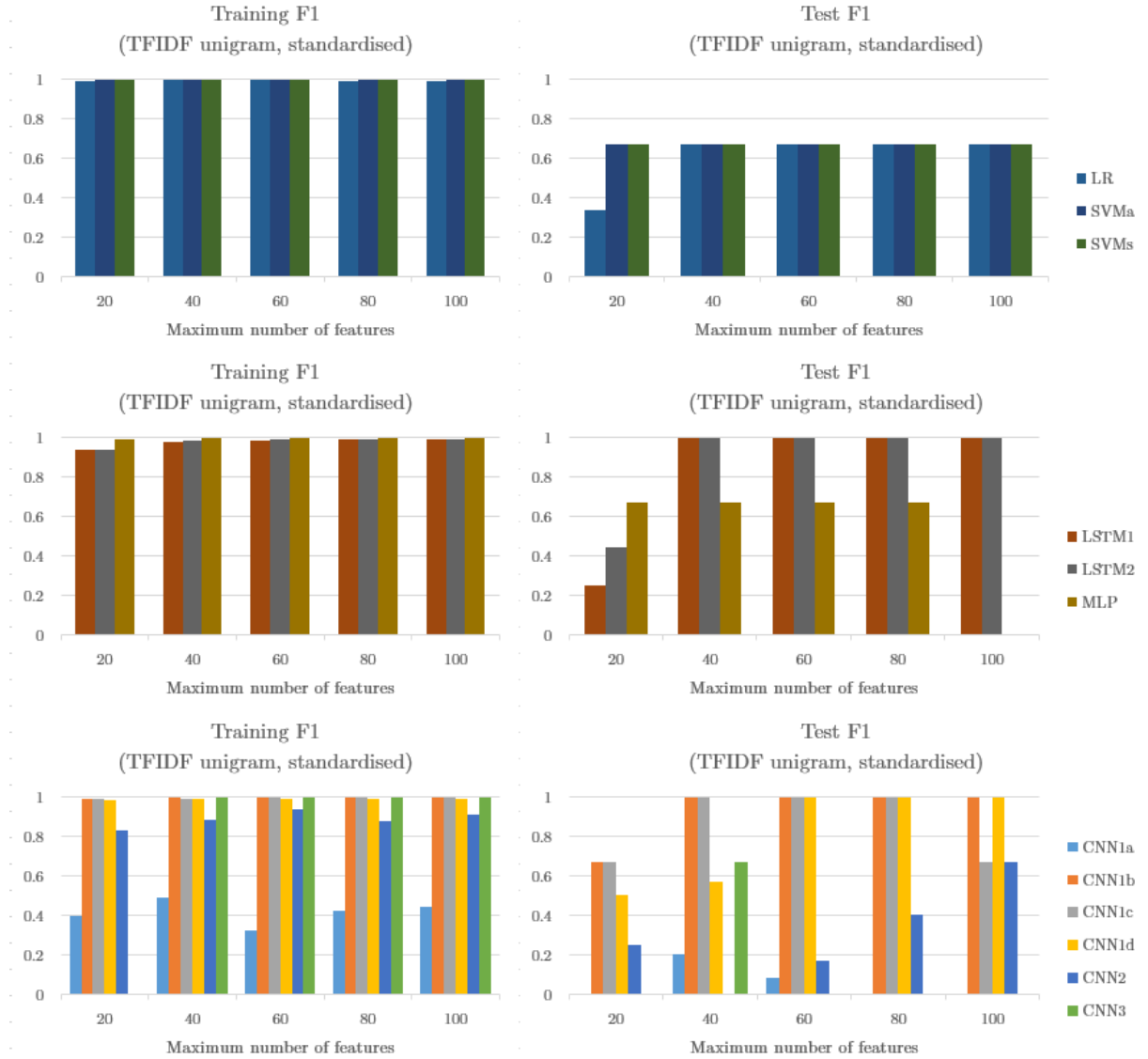


Figure B.10: Performance of classifiers on limited TFIDFs (standardised).

For 40 standardised features, the LR, SVMa, MLP and CNN3 classifiers no longer returned $F1=1$. The LSTM generalised better with test $F1=1$.

The test performance of CNN1a and CNN2 was poor for 40 TFIDF features using both scaling functions. CNN1a and CNN2 are filtered from all other figures to simplify reading and comparison of the results. The remainder of the analysis used standardised inputs for the LSTM models and normalised features for the rest of the classifiers.

B.4.2 TF and TFIDF compared

Classifier performance on TF was compared with TFIDF for 40 unigram features. Figure B.11 shows classifiers using normalised scaling.

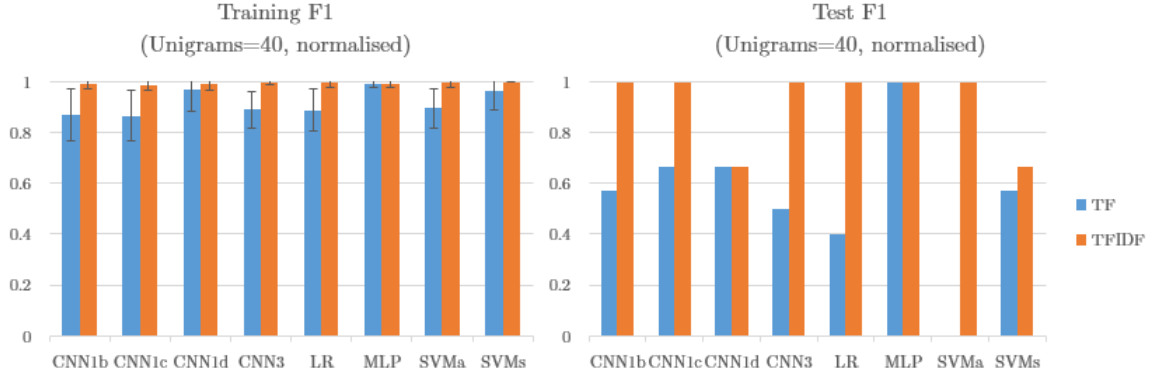


Figure B.11: Performance of classifiers on 40 unigrams (normalised).

All classifiers in Figure B.11 showed improved training performance using TFIDF over TF. The absolute F1 scores improved and the 95% confidence interval reduced for each model. The CNN1d and MLP models delivered the same test performance for both TF and TFIDF. The remaining classifiers generalised better using TFIDF.

LSTM performance on TF and TFIDF was compared using standard scaling in Figure B.12.

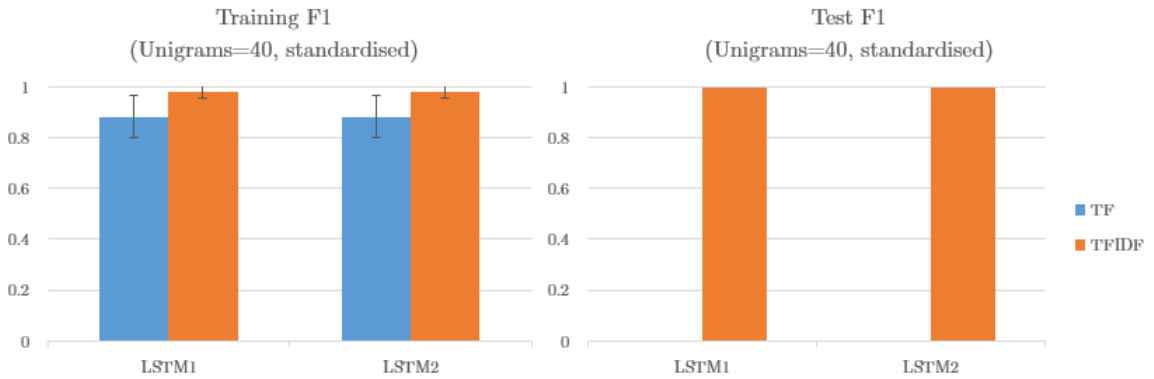


Figure B.12: Performance of classifiers on 40 unigrams (standardised).

Classifier training results for standardised features also proved TFIDF superior to using TF. The LSTMs could not generalise using TF whilst test F1=1 with TFIDF.

B.4.3 Data sets compared

The traditional classifiers using TFIDF for 40 unigram features normalised were used to compare the effect of resampling the original training data. The results are shown in Figure B.13.

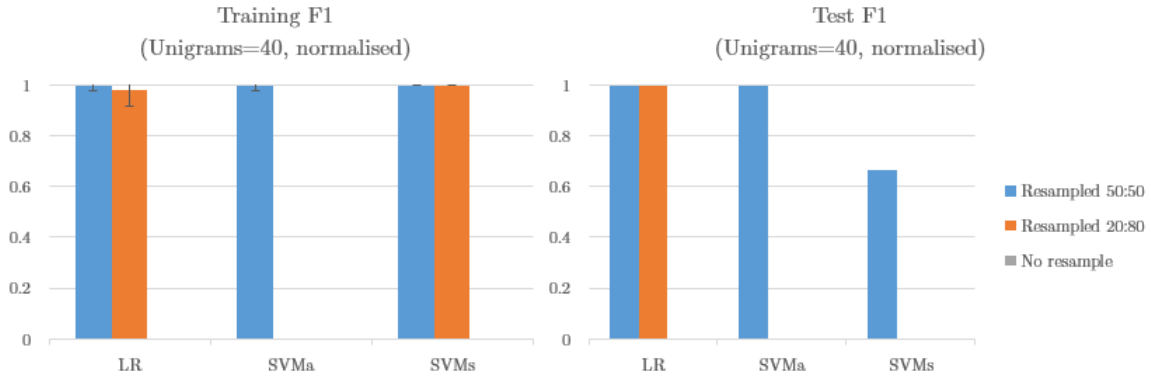


Figure B.13: Performance of classifiers across data sets using 40 TFIDF (normalised).

None of the classifiers were able to predict positive risk documents in the original data set. On the data set resampled to 20:80 (risk:no-risk) the SVMs model was validated with F1=1 in training but it could not predict the test documents. The LR training F1 score of 0.9783 was lower than 0.9951 on the data set resampled to 50:50 and the 95% confidence interval grew from ± 0.0172 to ± 0.0603 . The LR model still returned F1=1 for the 20:80 data set.

Given the improved training and test results, the remainder of the report describes experiments on the data set balanced 50:50.

B.4.4 N-gram comparison

A mix of n-grams in the range [1,6] were incrementally added after which TFIDF features were limited to 20, 40 and 60 respectively. Figure B.14 shows classifier performance on these normalised features.

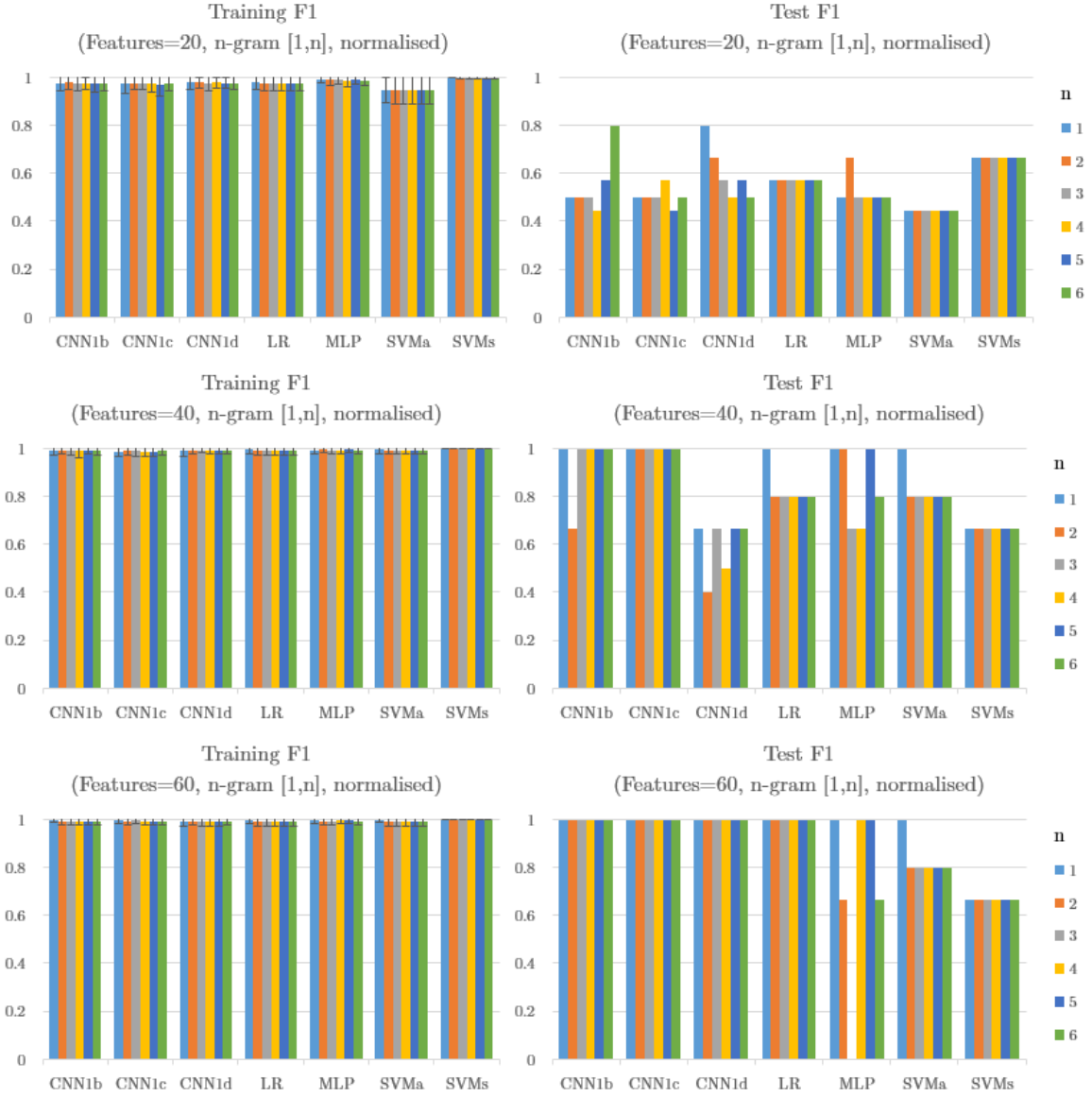


Figure B.14: Performance of classifiers across TFIDF n-grams (normalised).

When features were limited to 20, then CNN1b showed slight improvement in test F1 with up to 5 or 6 n-grams. The MLP classifier also showed slight improvement with the addition of bigrams. Increasing unigram tokens to a range of n-grams made no improvement in the test performance when limiting features to 40 and 60.

B.4.5 Word normalisation compared

Graphs in Figure B.15 compare training and test performance of an SVMa model using the F1 metric across the five word normalisation options (abbreviations given in Section 3.5.1). The model used a minimum word count of 8 and an embedding size of 300 with no feature scaling. Given poor performance of maximum, power of -1 and power of -2 functions in transforming word vectors, these experiments were not shown.

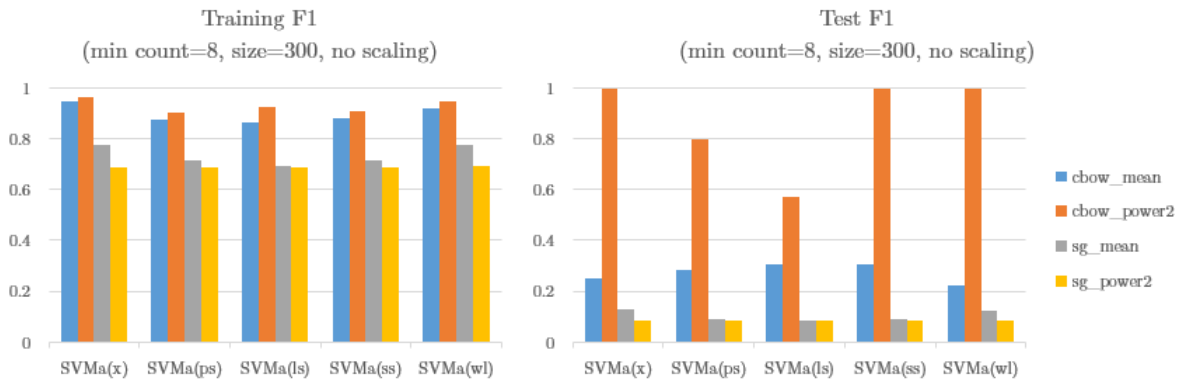


Figure B.15: Performance of an SVMa model for word normalisation and embedding.

For Continuous Bag of Words (CBOW), training performance was better when using the power of 2 function to transform the word vector than using the mean. Conversely for Skip-Gram (SG) word vectors the training performance was better using the mean function.

The model was able to generalise better with CBOW vectors than SG vectors. The power of 2 transform of the CBOW vector had test F1 scores markedly better than the mean transform as well as all SG transforms. Normalising words with the Porter Stemmer (ps) and Lancaster Stemmer (ls) had an adverse effect on performance whereas unit scores were achieved with the Snowball Stemmer, WordNet Lemmatizer and without word normalisation.

Scaling features between 0 and 1 for this model lead to overfitting where all training F1 scores were one and all test F1 scores were zero.

B.4.6 Linguistic features compared

Figure B.16 shows the effect of adding 17 word count features to limited numbers of TFIDF unigrams across the three classifiers. No word normalisation was used but feature values were normalised into the range $[0,1]$.

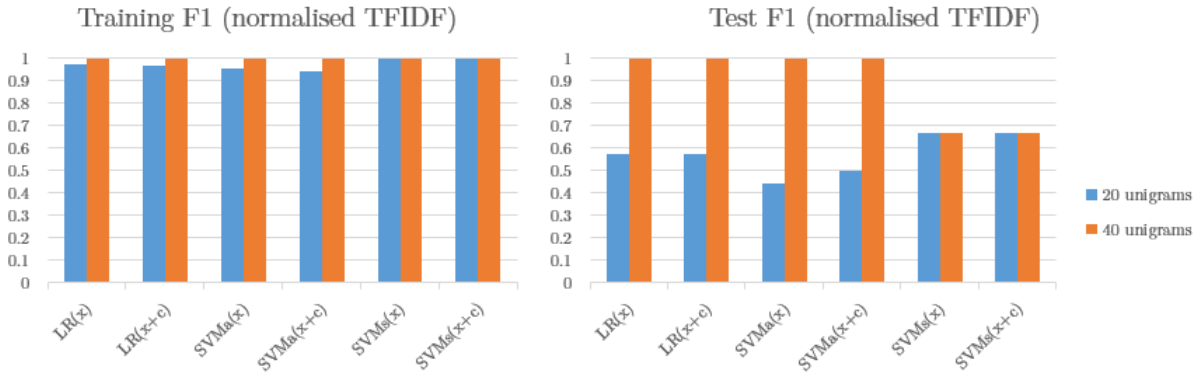


Figure B.16: Performance of classifiers on limited TFIDFs plus word counts (normalised).

With the inclusion of word count features, the SVMa model showed a slight improvement training on 40 unigrams (F1=0.997 as apposed to 0.995) and generalising on 20 unigrams (F1=0.500 as apposed to 0.444). Performance of the LR was not improved.

B.4.7 Using embedding layer of LSTM and CNN models

Less experiments were run on the Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) classifiers given resource challenges in using the embedding layer of these models. Table B.16 provides a summary of experiments performed using the LSTM1 model on different runtime environments. Initially training data was the balance after holding out 10% of the data for testing. The training data changed when 2018 and 2019 annual reports were held out for testing as discussed in Section 4.4.11.

APPENDIX B. SUPPORTING INFORMATION

Table B.16: Training of Long Short Term Memory Recurrent Neural Networks.

LSTM	Runtime	Test Data	Input Size	Epoch	Time	Comment
1	Local CPU	10% holdout	151796	2	8h 42m	50% of maximum report length
1	Colab CPU	10% holdout	151796			Crashed using available RAM
1	Colab TFU	10% holdout	151796			Crashed using available RAM
1	Colab GPU	10% holdout	151796			Error
1	Colab CPU	10% holdout	62933			Crashed using available RAM
1	Colab TFU	10% holdout	62933			Crashed using available RAM
1	Colab GPU	10% holdout	62933			Error
1	Local CPU	10% holdout	1200	5		Did not converge
1	Local CPU	2018/2019	99087	2		50% of maximum report length

The input sequence was attempted for the longest annual report in the training data, namely 196 221 tokens. The local CPU environment ran out of resources and could not train the model. The input size was halved to 99 087 to allow computation. The LSTM1 model was also created and compiled with the () strategy on Google Colab. It was discovered that the tensorflow distribution strategy does not allow class weights to be set. For this reason, it was not used on the original data set with its significant imbalance in classes.

Experiments reverted to the local environment, this time reducing the sequence length by half. Only 8 documents were longer than this and had word sequences truncated.

This also constrained training and test performance results as reflected in Table B.17 and Table B.18.

Table B.17: Training performance of LSTM.

LTSM	Input Size	Features	Epoch	Precision	F1	Comment
1	99087	TFIDF	2	0.4925	0.2988	Did not converge
1	99087	TFIDF + word counts	2	0.4927	0.5407	Did not converge

APPENDIX B. SUPPORTING INFORMATION

Table B.18: Test performance of LSTM.

LSTM	Input Size	Features	Epoch	Threshold	Precision	F1	Comment
1	99087	Sequence	2	0.486	0.0526	0.1000	Can additional epochs be run?
1	99087	TFIDF	2	0.497	0.0417	0.0588	
1	99087	TFIDF+word counts	2	0.497	0.0417	0.0588	

Recall of 1 was easily achieved by predicting many false positives (FP). Hence the models struggled with precision and the F1 score was poor. Table B.19 provides the confusion matrix for the first LSTM model using an input sequence of tokens which makes the 36 false positives explicit.

Table B.19: Confusion matrix of first LSTM model.

		Actual		Prediction	
		P=1	N=0	Total	Precision
Prediction	P=1	2 TP	36 FP	38	0.0526
	N=0	0 FN	10 TN	10	
Actual Total:		2	46	48	
Recall:		1.0000			

In experimenting with the CNN classifiers, one of three aspects were changed: either the CNN models, their batch size or the number of epochs were adjusted. Table B.20 provides a summary of experiments performed using different CNN models on the Google Colab CPU runtime environment. Increasing the batch sizes and epochs required more CPU although increasing the batch size reduced the computation time.

Table B.20: Training of Convolutional Neural Networks.

No	CNN	Runtime	Features	Scaling	Input Size	Epoch	Batch Size	Time	Comment
1	1a	Colab CPU	Sequence	None	98110	10	16	1h47m	
2	1a	Colab CPU	Sequence	[0,1]	98110	10	16	1h46m	

Continued on next page

APPENDIX B. SUPPORTING INFORMATION

Table B.20 – *Continued from previous page*

No	CNN	Runtime	Features	Scaling	Input Size	Epoch	Batch Size	Time	Comment
3	1a	Colab CPU	Sequence	Standard	98110	10	16	1h46m	
4	1b	Colab CPU	Sequence	None	98110	10	16	1h47m	
5	1b	Colab CPU	Sequence	None	98110	2	32	0h27m	
6	1c	Colab CPU	Sequence	None	98110	3	16	0h44m	
7	1d	Colab CPU	Sequence	None	98110	3	32	0h30m	
8	2	Colab CPU	Sequence	None	98110	3	64	N/A	Crashed using available RAM
9	2	Colab CPU	Sequence	None	98110	3	16	6h34m	Much slower than CNN1
10	2	Colab CPU	Sequence	None	98110	3	32	4h37m	
11	2	Colab CPU	Sequence	None	98110	10	32	14h17m	
12	3	Colab CPU	Sequence	None	98110	10	64	N/A	Crashed using available RAM
13	3	Colab CPU	Sequence	None	98110	10	32	1h1m	
14	3	Colab CPU	Sequence	[0,1]	98110	10	32	1h0m	

The CNN2 model was the most resource intensive and computationally expensive probably due to LSTM layer. It failed to run with the original 64 batch size and its computation time was much greater than the others when it did complete.

The CNN classifiers performed better than LSTM classifiers during training as reflected in Table B.21.

Table B.21: Training performance of CNN models.

Exp	CNN	Loss	Precision	Recall	F1	Comment
1	1a	0.0035	0.5110	0.5813	0.3777	Did not converge
2	1a	0.0041	0.4701	0.3938	0.2591	Did not converge. Worse than raw sequence.
3	1a	0.0035	0.5116	0.2750	0.1952	Did not converge. Worse than normal sequence.
4	1b	1.9E-07	1.0000	1.0000	1.0000	Reduce to 2 epochs; Try normalising inputs
5	1b	8.4E-04	0.9801	0.9308	0.9553	Reduce to 1 epoch; Try normalising inputs
6	1c	1.6E-04	0.9816	1.0000	0.9883	Reduce to 2 epochs; Try normalising inputs
7	1d	6.1E-11	1.0000	1.0000	1.0000	Reduce to 2 epochs; Try normalising inputs
9	2	0.0030	0.5682	1.0000	0.7143	Increase batch size and epochs
10	2	0.0030	0.5682	1.0000	0.7193	Insignificant improvement over experiment 9

Continued on next page

APPENDIX B. SUPPORTING INFORMATION

Table B.21 – *Continued from previous page*

Exp	CNN	Loss	Precision	Recall	F1	Comment
11	2	0.0030	0.5682	1.0000	0.7237	Insignificant improvement over experiment 9 and 10
13	3	4.9E-06	1.0000	1.0000	1.0000	Reduce to 3 epochs; Try normalising inputs
14	3	5.5E-06	1.0000	1.0000	1.0000	Reduce to 2 epochs

Four experiments returned training F1=1 and two had F1>0.9 which was significantly better than F1=0.5407 achieved by the LSTM. Despite the improved training performance, Table B.22 shows that test performance remained poor.

Table B.22: Test performance of CNN models.

No	CNN	Loss	Precision	Recall	F1	Comment
1	1a	0.8008	0.5551	0.0417	1.0000	0.1111 Predicts only positive
2	1a	1.4447	0.1144	0.0417	1.0000	0.1111 Predicts only positive
3	1a	0.8712	0.5876	0.0417	1.0000	0.1111 Predicts only positive
4	1b	0.5018	0.0001	0.0000	0.0000	0.0000 Did not generalise
5	1b	0.1245	0.0020	0.0714	1.0000	0.1333 Custom threshold
6	1c	0.5660	0.0001	0.0000	0.0000	0.0000 Did not generalise
7	1d	1.7052	0.0001	0.0000	0.0000	0.0000 Did not generalise
9	2	0.7158	0.5670	0.0500	1.0000	0.1176
10	2	0.7158	0.5670	0.0500	1.0000	0.1176 Same as 9
11	2	0.6943	0.5670	0.0488	1.0000	0.1176 Slightly worse than 9 and 10
13	3	0.4273	0.0200	0.0000	0.0000	0.0000 Did not generalise
14	3	0.4259	0.0200	0.0000	0.0000	0.0000 Did not generalise

Since model CNN1b was overfitting on training, 20% was set for the dropout layer and epochs were reduced to 3 for CNN1c. However, both the CNN1b and CNN1c models failed to generalise. Given difficulties in getting these neural network models to perform adequately as well as time limits on the study, the balance of this study removed the embedding layer.

B.5 Question answering

Figure 17 illustrates the impact of having applied the `cdQA filter_paragraphs` utility to the the VBS Mutual Bank 2015 annual report.

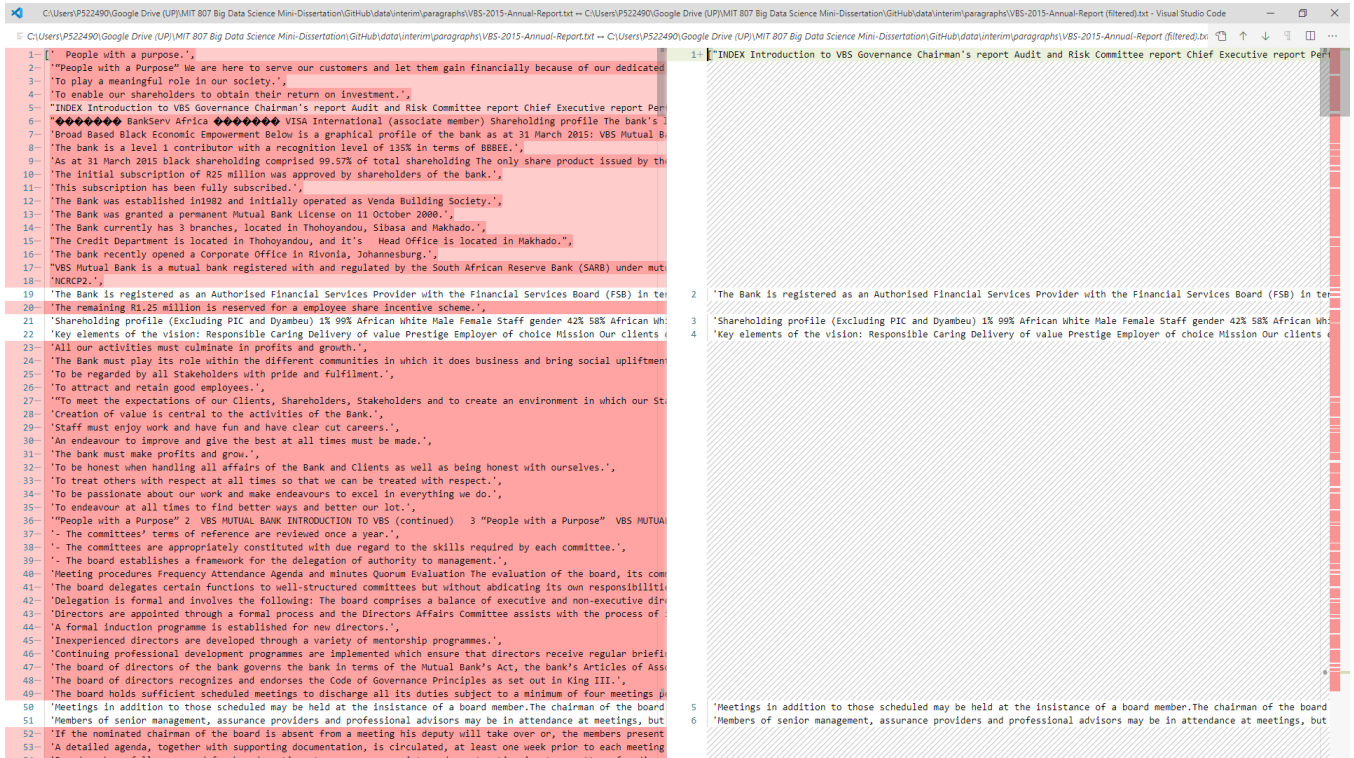


Figure B.17: Impact of `cdQA filter_paragraph` utility on 2015 report of VBS Mutual Bank.

The filtered paragraphs on the right are scant compared with the tokenised sentences on the left. The significant loss of sentences are further illustrated by the red shaded text. For instance, prior to the ‘INDEX’ being recognised as the first paragraph by the filter, four tokens were eliminated, namely:

1. ‘ People with a purpose.’;
2. ““People with a Purpose” We are here to serve our customers and let them gain financially because of our dedicated assistance.’;
3. ‘To play a meaningful role in our society.’;

4. ‘To enable our shareholders to obtain their return on investment.’.

B.6 Summary

This appendix provided detailed information relating to the experimental results in Section 4.

Select outputs of the CFIE–FRSE tool were shared in Section B.1 as part of the feasibility test. Annual reports published into separate documents were listed in Section B.2 as this study treated them as separate annual reports. Section B.3 showed various implementation details in preparing the data for machine learning including a listing of the test data set and positive documents in the training data set. The detailed performance results of the classifiers were compared in Section B.4. Similarly, the impact of the cdQA filter paragraph utility was shown in Section B.5.

Appendix C

Symbols

Symbols are described under the chapter in which they first appear.

Chapter 2: Literature review

<i>gf</i>	Readability metric: Gunning Fog	[Eq. (2.1), pg. 13]
<i>f</i>	Readability metric: Flesch Reading Ease	[Eq. (2.2), pg. 14]

Chapter 3: Methodology

<i>pageRatio</i>	Index for width of <i>page0</i> relative to <i>page1</i>	[Eq. (3.1), pg. 24]
<i>page0</i>	first PDF page	[Eq. (3.1), pg. 24]
<i>page1</i>	second PDF page	[Eq. (3.1), pg. 24]
<i>rect</i>	rectangular coordinates of the page	[Eq. (3.1), pg. 24]
<i>x0, y0, x1, y1</i>	Coordinates of the rectangular text block	[pg. 26]
<i>block_no</i>	Sequential number of block occurrence on the page	[pg. 26]
<i>block_type</i>	Index denoting text or image block	[pg. 26]
<i>r_</i>	<code>py-readability-metrics</code> package readability metric	[pg. 32]
<i>ts_</i>	<code>textstat</code> package readability metric	[pg. 32]
<i>fk</i>	Readability metric: Flesch-Kincaid Grade Level	[pg. 32]

<i>dc</i>	Readability metric: Dale-Chall	[pg. 32]
<i>ari</i>	Readability metric: Automated Readability Index	[pg. 32]
<i>cl</i>	Readability metric: Coleman-Liau	[pg. 32]
<i>lw</i>	Readability metric: Linsear Write	[pg. 32]
<i>sm</i>	Readability metric: SMOG	[pg. 32]
<i>sp</i>	Readability metric: SPACHE	[pg. 32]
<i>ts</i>	Readability metric: Text Standard	[pg. 32]
<i>wordCount</i>	Count of all words	[pg. 45]
<i>forward1_ scaled</i>	Scaled count of words in the forwardLooking wordlist	[pg. 45]
<i>forward2_ scaled</i>	Scaled count of words in the forwardLookingNew wordlist	[pg. 45]
<i>LMneg_scaled</i>	Scaled count of words in the LMnegative wordlist	[pg. 45]
<i>LMpos_scaled</i>	Scaled count of words in the LMpositive wordlist	[pg. 45]
<i>H6neg_scaled</i>	Scaled count of words in the HenryNeg2006 wordlist	[pg. 45]
<i>H6pos_scaled</i>	Scaled count of words in the HenryPos2006 wordlist	[pg. 45]
<i>H8neg_scaled</i>	Scaled count of words in the HenryNeg2008 wordlist	[pg. 45]
<i>H8pos_scaled</i>	Scaled count of words in the HenryPos2008 wordlist	[pg. 45]
<i>uncert_scaled</i>	Scaled count of words in the HenryPos2008 wordlist	[pg. 45]
<i>causal_scaled</i>	Scaled count of words in the causal wordlist	[pg. 45]
<i>causalM_ scaled</i>	Scaled count of words in the causalMartinAll wordlist	[pg. 45]
<i>causalM50_ scaled</i>	Scaled count of words in the causalMartin50 wordlist	[pg. 45]
<i>perf_scaled</i>	Scaled count of words in the performance wordlist	[pg. 45]
<i>strat_scaled</i>	Scaled count of words in the newStrategy wordlist	[pg. 45]
<i>LMtone</i>	Net tone using Loughran and McDonald (2011)	[pg. 45]
<i>H6tone</i>	Net tone using Henry (2006)	[pg. 45]
<i>H8tone</i>	Net tone using Henry (2008)	[pg. 45]

x	raw text without normalisation	[pg. 40]
ps	Porter Stemmer	[pg. 40]
ls	Lancaster Stemmer	[pg. 40]
ss	Snowball Stemmer	[pg. 40]
wl	WordNet Lemmatizer	[pg. 40]
γ	Constant in the Radial Basis Function	[Eq. (3.4), pg. 42]

Chapter 5: Modelling

$L1$	MOG, Flesch-Kincaid Reading Ease, Linsear Write features	[pg. 87]
$L1$	MOG, Flesch-Kincaid Reading Ease, Linsear Write, wordCount, perf_scaled, Automated Readability Index features	[pg. 87]
$L1$	H6neg_scaled, Coleman-Liau, H8tone, wordCount features	[pg. 87]

Appendix A: Exploratory Data Analysis

ρ	Coefficient of linear correlation	[pg. 119]
--------	-----------------------------------	-----------

Index

- B**
booklet [6](#), [8](#), [20](#), [23](#), [24](#), [26](#), [56](#), [57](#), [64](#), [82](#), [98](#), [100](#),
[144](#), [147](#)
bookmark [6](#), [33](#), [34](#), [70](#), [98](#), [100](#), [146](#), [154](#)
- C**
CFIE-FRSE tool [4–6](#), [19](#), [20](#), [30](#), [33](#), [35](#), [37](#), [55](#), [65](#),
[70](#), [81](#)
cross validation [38](#), [45–47](#), [51](#), [79](#), [101](#)
- D**
data set [105](#)
- E**
early warning [1](#), [3](#)
- H**
header [6](#), [20](#), [23–25](#), [30](#),
[33–36](#), [43](#), [56](#), [57](#), [70](#), [71](#), [73](#), [82](#), [98](#), [100](#),
[108](#), [142–144](#), [151–158](#), [160–162](#)
- M**
mapping file [23](#), [30](#), [60](#), [63](#), [65](#)
- R**
regular expression pattern [40](#), [41](#)
research question [2](#), [3](#), [12](#), [14](#), [98–100](#), [102](#)
- T**
table of content [6](#), [20](#), [25](#), [34](#), [57](#), [70](#), [100](#), [106](#), [145](#),
[146](#), [156](#)
text block [100](#)
- U**
UK reference [20](#), [22](#), [32](#), [34](#), [37](#), [60](#), [65](#), [66](#), [70](#),
[73–78](#), [82](#), [143](#), [151](#), [161](#)
- V**
validate [22](#), [37](#), [55](#), [60](#), [65](#), [81](#), [82](#)