Assessing Completeness of Solvency and Financial Condition Reports through the use of Machine Learning and Text Classification

Ruairí Nugent
Technological University Dublin

Follow this and additional works at: https://arrow.dit.ie/scschcomdis

Part of the Computer Sciences Commons

Recommended Citation

This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 License
Assessing Completeness of Solvency and Financial Condition Reports through the use of Machine Learning and Text Classification

Ruairí Nugent

C10335353

A dissertation submitted in partial fulfilment of the requirements of Dublin Institute of Technology for the degree of
M.Sc. in Computing (Data Analytics)

June, 2018
DECLARATION

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the test of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute’s guidelines for ethics in research.

Signed: ________________________________

Date: 14/06/2018
ABSTRACT

Text mining is a method for extracting useful information from unstructured data through the identification and exploration of large amounts of text. It is a valuable support tool for organisations. It enables a greater understanding and identification of relevant business insights from text. Critically it identifies connections between information within texts that would otherwise go unnoticed. Its application is prevalent in areas such as marketing and political science however, until recently it has been largely overlooked within economics. Central banks are beginning to investigate the benefits of machine learning, sentiment analysis and natural language processing in light of the large amount of unstructured data available to them. This includes news articles, financial contracts, social media, supervisory and market intelligence and regulatory reports.

In this research paper a dataset consisting of regulatory required Solvency and Financial Condition Reports (SFCR) is analysed to determine if machine learning and text classification can assist assessing the completeness of SFCRs. The completeness is determined by whether or not the document adheres to nine European guidelines. Natural language processing and supervised machine learning techniques are implemented to classify pages of the report as belonging to one of the guidelines.

**Key words:** natural language processing, machine learning, text classification, solvency and financial condition reports
ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my supervisor Brendan Tierney for his expert support and guidance and for keeping me on track throughout this project.

I would also like to thank all of the lecturers who have taught me so much throughout the course of this MSc and without whose support this task would have been impossible.

Finally, I would like to thank my family and friends for their motivational support throughout.
# TABLE OF CONTENTS

1. **INTRODUCTION** ............................................................................................................. 8  
   1.1 Background .................................................................................................................. 8  
   1.2 Research Problem ....................................................................................................... 10  
   1.3 Aims & Objectives ...................................................................................................... 12  
   1.4 Research Methods ...................................................................................................... 13  
   1.5 Scope & Limitations ................................................................................................... 13  
   1.6 Paper Outline ............................................................................................................ 14  

2. **LITERATURE REVIEW** ............................................................................................... 16  
   2.1 Introduction ................................................................................................................ 16  
   2.2 Narrative Reporting ................................................................................................... 16  
   2.3 Text Classification ...................................................................................................... 17  
   2.4 Pre-processing ........................................................................................................... 19  
   2.5 Classification Algorithms .......................................................................................... 21  
   2.6 Approach’s to similar problems ................................................................................ 24  
   2.7 Conclusion ................................................................................................................ 25  

3. **DESIGN AND METHODOLOGY** ................................................................................. 27  
   3.1 Introduction ................................................................................................................ 27  
   3.2 Data Collection .......................................................................................................... 29  
   3.3 Corpus Annotation ..................................................................................................... 30  
   3.4 Data Pre-processing ................................................................................................... 40  
   3.5 Classifiers .................................................................................................................. 41  
   3.6 Model Evaluation ....................................................................................................... 42  

4. **BUILDING AND EVALUATING PREDICTIVE MODELS** ........................................... 43  
   4.1 Data Preparation ........................................................................................................ 43  
   4.2 Data Investigation ...................................................................................................... 43  
   4.3 K-Nearest Neighbour ............................................................................................... 45  
   4.4 Naïve Bayes .............................................................................................................. 48  
   4.5 Decision Tree ............................................................................................................ 50  
   4.6 Neural Network ......................................................................................................... 53  
   4.7 Support Vector Machine ............................................................................................ 56  
   4.8 Initial Results ............................................................................................................. 58  
   4.9 SVM model tuning ...................................................................................................... 60  
   4.10 Summary of results ................................................................................................. 63
5. ANALYSIS, EVALUATION & DISCUSSION ................................................................. 64
   5.1 Evaluation of results ......................................................................................... 64
   5.2 Observations from final model ......................................................................... 66
   5.3 Model Application to Business Problem .......................................................... 72
   5.4 Strengths of the Results ................................................................................ 76
   5.5 Limitations of the Results ............................................................................... 76
6. CONCLUSION ........................................................................................................... 78
   6.1 Research Overview ......................................................................................... 78
   6.2 Problem Definition ......................................................................................... 79
   6.3 Design/Experimentation, Evaluation & Results ................................................ 80
   6.4 Contributions and Impact ............................................................................. 81
   6.5 Future Work & Recommendations ................................................................ 82

BIBLIOGRAPHY ......................................................................................................... 84
# TABLE OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Previous Business Solution</td>
<td>11</td>
</tr>
<tr>
<td>3.1</td>
<td>Business Solution Design</td>
<td>27</td>
</tr>
<tr>
<td>3.2</td>
<td>Experiment Design</td>
<td>29</td>
</tr>
<tr>
<td>3.3</td>
<td>SFCR High-Level Structure</td>
<td>31</td>
</tr>
<tr>
<td>3.4</td>
<td>Analysis of SFCR Length</td>
<td>31</td>
</tr>
<tr>
<td>4.1</td>
<td>Tree Map of Corpus Categories</td>
<td>44</td>
</tr>
<tr>
<td>4.2</td>
<td>Corpus Word Cloud</td>
<td>45</td>
</tr>
<tr>
<td>4.3</td>
<td>KNN tf &amp; tf-idf Comparison</td>
<td>45</td>
</tr>
<tr>
<td>4.4</td>
<td>KNN feature selection</td>
<td>46</td>
</tr>
<tr>
<td>4.5</td>
<td>K Tuning</td>
<td>47</td>
</tr>
<tr>
<td>4.6</td>
<td>Naive Bayes tf &amp; tf-idf Comparison</td>
<td>48</td>
</tr>
<tr>
<td>4.7</td>
<td>Naive Bayes Feature Selection</td>
<td>49</td>
</tr>
<tr>
<td>4.8</td>
<td>Decision Tree tf &amp; tf-idf Comparison</td>
<td>51</td>
</tr>
<tr>
<td>4.9</td>
<td>Decision Tree Feature Selection</td>
<td>51</td>
</tr>
<tr>
<td>4.10</td>
<td>Decision Tree Complexity Tuning</td>
<td>52</td>
</tr>
<tr>
<td>4.11</td>
<td>Neural Network tf &amp; tf-idf Comparison</td>
<td>53</td>
</tr>
<tr>
<td>4.12</td>
<td>Neural Network Feature selection</td>
<td>54</td>
</tr>
<tr>
<td>4.13</td>
<td>Neural Network Weight Decay</td>
<td>55</td>
</tr>
<tr>
<td>4.14</td>
<td>SVM tf &amp; tf-idf Comparison</td>
<td>56</td>
</tr>
<tr>
<td>4.15</td>
<td>SVM Feature Selection</td>
<td>57</td>
</tr>
<tr>
<td>4.16</td>
<td>SVM Cost Tuning</td>
<td>60</td>
</tr>
<tr>
<td>4.17</td>
<td>Final SVM Model Accuracy</td>
<td>61</td>
</tr>
<tr>
<td>5.1</td>
<td>Example of Page from Asset Category</td>
<td>68</td>
</tr>
<tr>
<td>5.2</td>
<td>SFCR other liabilities example</td>
<td>69</td>
</tr>
<tr>
<td>5.3</td>
<td>Business Report Example 1</td>
<td>73</td>
</tr>
<tr>
<td>5.4</td>
<td>Business Report Example 2</td>
<td>74</td>
</tr>
<tr>
<td>5.5</td>
<td>Business Report Example 3</td>
<td>75</td>
</tr>
<tr>
<td>6.1</td>
<td>Research Process</td>
<td>78</td>
</tr>
</tbody>
</table>
TABLE OF TABLES

Table 3.1 Breakdown of SFCR Categories ................................................................. 39
Table 3.2 Pre-Processing Models ............................................................................. 42
Table 4.1 KNN Pre-Processing .................................................................................. 46
Table 4.2 Naive Bayes Pre-Processing ..................................................................... 49
Table 4.3 Decision Tree Pre-Processing ................................................................. 52
Table 4.4 Neural Network pre-processing ............................................................... 54
Table 4.5 SVM Pre-Processing .................................................................................. 58
Table 4.6 Best Performing Models ........................................................................... 59
Table 4.7 SVM Confusion Matrix ............................................................................ 61
Table 4.8 SVM Statistics ......................................................................................... 62
Table 5.1 Individual Category Accuracy from Model Performance on Test Set .......... 66
1. INTRODUCTION

This section details an introduction to the subject matter and the questions tackled by this project. Following this is a brief review of the methods applied and consideration is given to potential limitations there might be in convincingly answering the research questions. The section concludes with an outline of this paper.

1.1 Background

The new, harmonised EU-wide regulatory regime for insurance companies, known as Solvency II, came into effect from 1 January 2016. Within this new pan European regulation, is the initiation of significant reporting requirements to the Central Bank of Ireland. Reporting by companies has evolved significantly beyond quantitative information, to include a variety of new narrative reports on risk exposures, details of models and information on systems of governance. There are approximately 200 insurance firms required to report to the Central Bank. Regulatory reporting matters because it is a significant input into regulators’ risk assessment of the on-going health of firms, and the protection of those firms’ customers. Historically text mining has not been applied to regulation. However, with such a large amount of textual data available regarding firms’ financial stability, it is becoming more prominent.

The Solvency II Directive (2009, s.51) requires a Solvency and Financial Condition Report (SFCR) to be submitted annually by Insurance firms supervised by the Central Bank of Ireland. The SFCR is a narrative report that supports Quantitative Reporting Templates. Articles 290 – 303 of the Delegated Regulation specify the content that must be included within the SFCR. The Solvency II Directive came in to force in January 2106. As such, the central bank has only received two iterations of SFCRs thus far. The second iteration was only received in May 2018, by which stage the majority of this study had been completed. The first iteration was entirely manually checked to be adhering to legislative requirements. Following this exercise, a number of companies were required to resubmit, as they did not comply with legislation. This was a highly time consuming process. The solution described within this paper was
utilised this year in order to automate some of these previous manual and time-intensive checks. This paper proposes and explores the idea of automating some key checks through supervised Machine Learning and Text Classification to determine if SFCRs comply with legislation.

Text mining is a valuable tool that can be used by central banks in order to help achieve policy objectives (Bholat, Hansen, Santos, & Schonhardt-Bailey, 2015). In recent years central banks are beginning to explore the potential insights available to them from text mining of speeches and unstructured reports they receive. Analysing documents that must adhere to legislative requirements is a mostly manual and time-consuming process. Automated text mining is a potential solution to these challenges (Massey, Eisenstein, Antón, & Swire, 2013).

Automated text classification has many applications including indexing of scientific articles, spam filtering, identification of document genre and survey coding (Dharmadhikari, Ingle, & Kulkarni, 2011). Automated text classification is appealing because it can help relieve the expensive and time-consuming job of organising document bases (Sebastiani, 2002). Automating text classification involves information retrieval, machine learning and natural language processing techniques.

Supervised machine learning algorithms use externally provided instances to create general hypotheses that in turn make predictions about future instances (Kotsiantis, 2007). Labelling and Feature Vectors from a corpus of text data are applied to a Machine Learning algorithm in order to train and build a predictive model. Once built feature vectors from new text data can be applied to the model in order to generate an expected label.

The European Insurance and Occupational Pensions Authority (EIOPA) published guidelines stipulating the content that must be included within an SFCR. These guidelines relate to Articles 35, 5, 53, 54, 55, 254 and 256 of Directive 2009/138/EC of the European Parliament and Articles 290 to 298, 305 to 311, 359 and 365 of
Commission Delegated Regulation (EU) 2015/35. There are thirteen guidelines from EIOPA that an SFCR must adhere to. This project will involve classifying each page of text to determine which guideline it is addressing. For the purpose of this study, which is exploring the suitability of such a solution, a subset comprising of nine of the more relevant guidelines is considered.

Completeness analysis of financial regulatory documents through text classification is yet to be thoroughly explored. At least within published literature it does not appear to have been heavily researched thus far. However, this is likely to change in the near future. Central banks and economists are beginning to take a greater interest in text mining. Areas of interest include sentiment analysis of narrative returns and use of text mining to highlight potentially uncompliant returns. When considering the latter, narrative returns are manually assessed by supervisors within regulatory bodies. Analytics teams are aiming to help these supervisors with early warning indicators that can identify which returns to prioritise for assessment. Narrative returns for financial institutions are complex documents; they are company specific and provide the narrative for the performance of their company. As such, there is a large qualitative element to assessing these returns. It is likely they will always require a level of examination for consistency and compliance by humans. Undoubtedly though there can be gains from automating some parts of the checks and providing indicators and identifying returns that are more likely to be uncompliant. Use of text mining and natural language processing techniques can lead to cost avoidance, namely, a reduced need to hire staff to undertake a large amount of manual checks.

1.2 Research Problem

The idea for the study came about following the need to solve a real life business problem. In 2017, insurance companies published the first ever set of solvency and financial condition reports. For the regulator, assessing these for compliance was a tedious, manual and time-consuming process. It is proposed to explore the idea that machine learning and text classification can be used in order to automate some of this work.
Figure 1.1 displays the business solution used for the assessment of SFCRs in 2017. All companies submitted their SFCR to the central bank. Each SFCR was then checked manually by a colleague responsible for the review project. This was a tedious and time-consuming project. Almost two hundred SFCRs were reviewed individually to ensure they were addressing all of the required legislative guidelines. If a company did not meet all of the required guidelines the reviewee informed the relevant supervision team who then conducted their own in-depth review and liaised with the company for correction and resubmission where necessary. A major flaw in this approach was the iterative nature of going through each report one by one. Uncompliant reports towards the bottom of the pile would not be discovered for a considerable time. The project took over four weeks to complete. As this is a publicly available report encouraging transparency, it is not ideal to have an incorrect report publicly available for large periods of time. A solution was required that could quickly flag and prioritise potentially uncompliant returns.

The primary research question under investigation here is to what extent can machine learning be used to assess the completeness of solvency and financial condition reports? Concerning this question, completeness refers to the legislative guidelines being adhered to. For instance an SFCR with all guidelines present is considered...
complete, an SFCR with only half of the guidelines adhered to is considered half complete and so on.

Two secondary research questions are also considered:

1. Can the model produced accurately classify the guideline differences between standard formula and internal models? Particular attention is paid to this guideline as only about 5% of companies have approval to use internal models. With only one year of SFCRs available it is known before the project begins that there will be a shortage of data relating to this guideline.

2. Can the model achieve the same accuracy on new SFCRs? Any model built will be using training data from the year-end 2016 returns. The performance of such model will be validated against 2017 returns. The model will be trained using a dictionary of words from the 2016 returns, it is unlikely this dictionary will vary a great deal year on year and much of the language will be finance/insurance specific. However, in answering the primary research question it will be important to establish if the model will need to be re-trained each year or if the initial model can be used going forward.

1.3 Aims & Objectives

The aim of this research is to develop on the endeavours of previous efforts of text mining, natural language processing and machine learning to classify text. The research will examine if these techniques are applicable to accurately classify text within narrative regulatory reporting. The research focuses on using natural language processing and machine learning techniques in order to classify pages of SFCRs with the objective of assessing if all legislative guidelines are present within the document.

The final solution should be able to assess a company’s SFCR detailing which guidelines are present within the text and what proportions of the document they account for. This will enable a quick assessment of the approximately two hundred companies that publish SFCRs annually. An early identification of companies that
have failed to address guidelines or guidelines that are only briefly addressed will greatly assist company supervisors in their assessments.

1.4 Research Methods

The work began with secondary research into the principal concepts required to understand natural language processing and text classification. The SFCRs are due annually by each insurance company supervised by the Central Bank of Ireland and are publicly available. This is empirical research as the research focuses on testing the feasibility of a solution using empirical evidence. This research is deductive as it is top down. I am starting with a theory progressing to a hypothesis then observation and confirmation. This is a systematic empirical investigation of quantitative properties and phenomena along with their relationships. The aim is to use automated text mining and text classification to determine if the SFCRs of companies adhere to a subset of legislative guidelines set out by the European Insurance and Occupational Pensions Authority.

The Cross Industry Standard Process for Data Mining (CRISP-DM) (Wirth & Hipp, 2000) is also used as a reference standard for the general methodology. These methodologies are iterative processes, which is suitable for the subject matter of this research.

The data for this study comprises of 80 SFCRs from year-end 2016. The 80 SFCRs totalled 4,274 pages. These had be manually labelled with the most applicable guideline. This is a more extensive dataset than that used by (Cosante, Sun, Petkovic, Hartog, 2012) which used 1,049 annotated paragraphs in their efforts to assess completeness of privacy policies.

1.5 Scope & Limitations
This research can only act as an indicator in assessing completeness of SFCRs. It enables a quick establishment of which guidelines are present in SFCRs. It will also be able to provide a summary of what proportion of each SFCR addresses each guideline present. Whilst this solution will be very helpful in identifying reports that need further consideration when assessing almost 200 companies, it does not provide any qualitative review. The solution highlights the existence of a guideline within the text however, it cannot determine if the guideline is addressed to the expected quality. For instance, a company might reference information related to a required guideline but refrain from divulging information that is required.

As a result, scenarios can exist where an SFCR is considered to have alluded to all guidelines within the text but still not be of the expected quality. The examination of SFCRs will still involve human interaction. This solution will help save on time and finance yet it does have its limitations. A subsequent direction of research could be to apply semantic analysis to the guidelines discovered within text in order to further automate the process of determining if they meet qualitative legislative requirements.

1.6 Paper Outline

- **Chapter 2 Literature Review:** - This chapter contains a review of the research already conducted in this space. It introduces narrative reporting and text classification. It provides the key concepts within text classification pre-processing as described within the literature. A review of literature outlining popular and successful algorithms used for text classification is provided. The chapter concludes with a review of approaches taken to solving similar problems.

- **Chapter 3 Design & Methodology:** - This chapter introduces the project design and methodology. The data source and structure is discussed. The legislative guidelines that apply to the SFCR are summarised. The process that is followed in order to label each SFCR page within the dataset is documented
with examples also provided. The chapter closes with an overview of pre-processing techniques and classification algorithms considered for this study.

- **Chapter 4 Building & Evaluating Predictive Models:** This chapter focuses on the individual model training, tuning and performance. Initial results are documented and briefly discussed.

- **Chapter 5 Analysis, Evaluation & Discussion:** This chapter reviews the results obtained in chapter 4 and the inferences from these. The model providing the best accuracy is evaluated and consideration is given to how it performs with respect to each individual guideline. The approach to which this model was then applied to a real life business problem is detailed. The chapter concludes with an evaluation of the strengths and limitations of the findings.

- **Chapter 6 Conclusion:** The final chapter provides a conclusion and a review of the contribution of this experiment to the literature. Suggestions are also put forward for direction of future work.
2. LITERATURE REVIEW

2.1 Introduction

Financial regulation imposes certain requirements and guidelines on financial institutions with the aim of maintaining the integrity of the financial system. The Central Bank of Ireland’s mission statement is ‘Safeguarding Stability, Protecting Consumers’ and this encapsulates the dual priorities of the Central Bank in delivering on its mandate. In recent years, the level of narrative financial reporting has increased substantially. In addition to this, there is a wealth of financial textual information available from Google Trends, newspaper articles and social networks such as Twitter. Text mining is a valuable tool that central banks can use in order to help achieve policy objectives (Bholat, Hansen, Santos, & Schonhardt-Bailey, 2015). In recent years, central banks are beginning to explore the potential insights available to them from text mining of speeches and unstructured reports they receive. Automated text classification is appealing because it can help relieve the expensive and time-consuming job of organising document bases (Sebastiani, 2002). Automated text classification has been utilised with largely positive results in a range of applications. This paper looks to investigate if automated text classification is a viable solution within regulation to aid compliance assessments of narrative returns.

This chapter presents a review of the research undertaken. Initially there is a review of the role and importance of narrative reporting. Following this is an outline of research regarding the text classification approach including pre-processing techniques and popular text classification algorithms.

2.2 Narrative Reporting

In the early 2000’s it became clear that the business-reporting model in place was outdated. It had to develop in order to serve the changing information requirements of the market and provide the necessary information for improved corporate transparency
and accountability (Beattie, McInnes, & Fearnley, 2004). Globally regulators viewed narrative disclosures as a welcome addition to the business-reporting model in place. The consensus was that narrative reporting would help address gaps in the traditional financial reporting model that emphasised retrospective, enumerated, financial information. Regulators and lawmakers are still striving to improve the business-reporting model today.

With respect to this study, pillar 3 of the Solvency II regulation introduced in 2016 by EIOPA focuses on increased transparency and accountability. The SFCR is a new narrative public document designed with the findings of Beattie et al in mind. Financial regulators are still attempting to find ways to ensure financial reporting models are not just backward looking and numeric. Narrative reports are key in helping to give an idea of what is going on at a financial institution with regards to governance and direction. What is crucial for one company may not be important for another. Narrative reporting sets out an analysis of the business through the eyes of the board of directors (Ambler, & Neely, 2007). They provide a lot of detail that is not obtainable from solely analysing the numbers. The narrative reports can also aid the supervisors in understanding movements within the numbers. As such, numeric financial reporting and narrative reports can complement each other well and give a greater insight to the regulator.

2.3 Text Classification

Humans generally tend to struggle addressing problems that attempt to establish relationships between numerous features. Machine learning performs well when addressing such problems. For machine learning algorithms, every instance in the dataset is represented using the same set of features. These features can be continuous, categorical or binary. Supervised machine learning algorithms use externally provided instances with labels provided to create general hypotheses that in turn make predictions about future instances. In unsupervised learning, the instances are unlabelled. Unsupervised learning algorithms aim to discover unknown but useful
insights from data. This study explores text classification with the use of supervised machine learning algorithms. The first step in a supervised machine learning problem is data collection. Following this is the data preparation and data pre-processing stage. Data is split into training and testing in order to estimate classifier performance. The specific algorithm selection is a crucial stage. Once satisfied with preliminary testing of the classifier, it is available for routine use (Kotsiantis, 2007).

Natural Language Processing (NLP) is an area of research that examines how computers can be used to comprehend and manipulate natural language text. A wide number of disciplines provided the foundation for NLP. These include linguistics, mathematics, computer and information sciences, psychology and artificial intelligence. NLP is widely implemented with applications in speech recognition, cross language information retrieval and artificial intelligence. With the information explosion of the worldwide web and digital libraries, NLP became a prominent and well-researched topic (Chowdhury, 2003). The boom in NLP was due to the large increase in availability of digital documents and the subsequent need to organise them led to researchers examining the potential of text classification.

Text classification originated in the early 1960s. However, it was only in the 1990s that it became a prominent subfield of the information systems field due to the rise of NLP, a substantial increase in digital documents and availability of more powerful hardware. Text classification is central to many applications today. An example of these include automatic indexing for Boolean information retrieval systems, document organisation, text filtering, word sense disambiguation and hierarchical categorisation of web pages (Sebastiani, 2002). The leading text classification approaches utilise machine learning techniques. In an inductive process, a classifier learns the characteristics of the categories from a preclassified set of documents. The resulting classifier is then used to assign categories to instances where the values of the predictor characteristics are known but the value of the category is unknown. Automated text classification is appealing because it can help relieve the expensive and time-consuming job of organising document bases.
2.4 Pre-processing

When approaching a text classification problem, document representation is one of the first ports of call. It is a technique used in order to reduce the complexity of documents and makes them easier to handle. Documents are transformed from full text versions to a document vector. Within this study, each individual SFCR will represent one document. Text representation conveys the mapping of documents into a compact form of its contents. Typically, a text document is represented as a vector of term weights from a set of terms, in which each term occurs at least once in a certain number of documents (Khan, Baharudin, Lee, & Khan, 2010). Text classification problems generally contain a very large dimensionality of text data. This is due to the number of potential features often exceeding the number of training documents. As a result, dimensionality reduction is an important part of pre-processing. It is the exclusion of a large number of words, on a statistical basis to create a lower dimension vector. Dimensionality reduction is necessary as irrelevant and redundant features often lower the performance of classification algorithms. These redundant features negatively influence speed and classification accuracy.

The two main approaches for dimensionality reduction are feature extraction and feature selection. Effective dimension reduction leads to significant savings in storage space and ensures a more efficient learning stage (Wang, & Wang, 2005). These gains are due to the removal of immaterial and noisy features. The most common feature extraction methods are tokenisation, stemming and removal of stop words. Tokenisation sees a document treated as a string before being partitioned into a list of tokens. Stemming applies the stemming algorithm that converts words into their canonical form. For instance, the stemming algorithm would reduce the words “fishing”, “fisher” and “fished” to the canonical form “fish”. Removal of stop words focuses on removing noisy insignificant frequently occurring stop words such as “and”, “a” and “the”. The goal of feature selection is to improve efficiency, scalability and accuracy of a text classifier. A suitable feature selection method should consider domain and algorithm characteristics. Feature selection isolates a subset of features from the original documents. Using a pre-set measure of importance for words, it
analyses each word and only selects those above a certain threshold. As feature selection addresses the issue of high dimensionality within text classification, it is commonly used in order to reduce the dimensionality of feature space and improve the efficiency and accuracy of classifiers. Within machine learning, there are two prominent types of feature selection: wrappers and filters. However, wrappers are largely unsuitable for text classification. They use the classification accuracy of learning algorithms as their evaluation function. This means wrappers have to train a classifier for each feature subset that is to be evaluated. This is generally computationally expensive and time consuming. When considering the number of features is usually quite high within text classification, wrappers are rarely suitable when addressing text classification problems. Filters conversely perform feature selection separately of the algorithm that will use the selected features. Filters use an evaluation metric that measures the ability of the feature to differentiate each class (Ogura, Amano, & Kondo, 2009). Two popular feature extraction methods are term frequency and term frequency – inverse document frequency (tf-idf). Term frequency is a count of the number of times a term occurs in a document.

\[
tf(f_i, d_j) = \frac{freq_{ij}}{\max_k freq_{kj}}
\]

Term frequency however only identifies the most frequently occurring features, meaning meaningful features for some categories within documents may be overlooked. Tf-idf addresses this issue; it does so by diminishing the weight of terms that occur very frequently within the corpus and increases the weight of terms that rarely occur.

\[
|tf - idf| = \log \frac{|D|}{\#(f_1)}
\]

It calculates values for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in (Ramos, 2003). It has been found that tf-idf encoding is uncomplicated making it ideal for establishing the basis for more complex algorithms. The majority of text based recommender systems in digital libraries use tf-idf. Many popular search engines including Google also implement tf-idf in order to score and rank documents from the users search. Literature has shown that feature selection is
important to improve algorithm performance by reducing high dimensionality. When considering text classification wrappers are generally unsuitable however, filters can be easily used in order to perform feature selection.

2.5 Classification Algorithms

This section focuses on reviewing successful text classification algorithms within literature. This study will aim to explore if the success of these text classification algorithms will translate to the problem at hand, classifying text within regulatory narrative returns.

K-nearest neighbour (KNN) is a popular algorithm for text classification. KNN is one of the simplest lazy machine learning algorithms. The algorithm does not require the application of training data to perform classification, the training data can be used in the testing stage. KNN is based on discovering the most similar documents from sample groups about the Euclidean distance (Wang, & Li, 2010). The algorithm classifies documents in the Euclidean space as points. It determines how to categorise a document by only looking at the training documents most similar to it. Euclidean distance is the distance between two points in Euclidean space. The corpus is transformed to a weight matrix such as term frequency or inverse document frequency. The following step in KNN classification is to determine the value of K. K is a factor that indicates a required number of documents from within the collection of documents that is closest to the document being classified (Trstenjak, Mikac, & Donko, 2013). The classification utilises the following equation in order to calculate the vectors distance between documents:

\[
d(x, y) = \sqrt{\sum_{r=1}^{N} (a_{rx} - a_{ry})^2}
\]

where \(d(x, y)\) represents the distance between two documents, \(N\) represents the number of unique words within the collection of documents, \(a_{rx}\) is the weight representative of term \(r\) in document \(x\) whilst \(a_{ry}\) is the weight representative of term \(r\) in document \(y\).
KNN has proven itself to be an algorithm worthy of consideration when approaching text classification problems. Literature does warn that KNN can struggle when feature dimensionality is very high. Trstenjak et al found KNN to perform quite well within their study when applied with tf-idf representation.

Decision trees are constructed with the use of a hierarchical split of the underlying data space with the use of different text features. The purpose of the hierarchical split of the data space is to create class partitions that are more skewed in respect of their class distribution. When assessing a text instance, the decision tree determines which partition it is most likely to belong to and uses this to classify the instance. The split of the data space is performed repeatedly in the decision tree until the leaf nodes have a certain minimum number of records. The dominant class label in the leaf node is used for the objective of classification. Decision trees traverse a path in a top-down manner in order to determine the relevant leaf node. Tree pruning may be utilised in order to reduce the issue of overfitting. When considering decision trees with respect to text classification, the bases for decision tree nodes are usually expressed in terms of words in the underlying text collection. For instance, a node could be partitioned into its children nodes subject to the presence or absence of a particular term within the document. Although partitioning on individual terms is not always possible. When this is the case it is possible to partition using a measure of similarity of documents to correlates set of terms. Decision trees considering text classification have three main methods for splitting: single attribute splits, similarity based multi attribute split and discriminant based multi attribute split (Aggarwal, & Zhai, 2012). Single attribute splits consider the presence or absence of particular words or phrases at tree nodes to perform the split. Similarity based multi attribute split considers word clusters and examines the similarity of the documents to these word clusters to perform the split. For discriminant based multi attribute split, discriminants are used to perform the split. For example when the Fisher discriminant is used, documents are projected on the Fisher vector for rank ordering and then split at a particular coordinate. The split is made where the maximum discrimination between classes occurs. A main advantage of decision trees is their comprehensibility; it is easy to understand why a tree classifies an instance as belonging to a certain class. However large decision trees are
Naïve Bayes is traditionally a popular algorithm within machine learning. Its application to text classification problems has been widely studied. It is a simple probabilistic classifier based on applying Bayes theorem. The algorithm applies an independence assumption; it assumes all features of the examples are independent of each other given the context of the class (Kim, Han, Rim, & Myaeng, 2006). The independence assumption has both advantages and disadvantages. Due to the independence assumptions, the computation is more efficient. However, the independence assumptions also severely restrains its applicability. As independent variables are assumed for naive Bayes, only variances of the variables for each class are required. This means it is possible for naive Bayes classifiers to be trained efficiently using a small amount of training data in order to estimate the parameters required for classification. Naïve Bayes generally produces lower classification performance than other discriminative algorithms. It is easy to implement and works well with textual data although the conditional independence assumption is broken by real world data and it performs poorly when features are highly correlated (Khan, Baharudin, Lee, & Kahn, 2010). It has been a very popular algorithm within machine learning due to its simplicity and reasonable performance.

Neural networks are one of the main algorithms within machine learning. They originated back in the 1940s. There was a lot of study undertaken on single layer neural networks in the 1950s and 1960s. Limitations in the single layer neural networks led to a lull in research in the 1970s. In the 1980s, researchers had breakthrough success with multi-layer neural networks for classification. However, they have only really become prominent in the last couple of decades with the introduction of the backpropagation technique. Neural networks comprise of input layers, output layers and hidden layers. The hidden layers consist of units, which transform the input into something functional for the output layer. There are many popular neural network variations including the traditional multi-layer feed forward network, radial basis function networks, adaptive resonance theory models and self-
organising feature-mapping networks (Zhang, & Zhou, 2006). When applied to text classification problems neural networks have the ability to address high dimensional features and noisy data. Neural networks for text classification generally produce good results in complex domains but training is relatively slow. A disadvantage of neural networks is their high computational cost and requirement for large physical memory. In addition, neural networks are particularly challenging to interpret.

Support Vector Machines (SVMs) are one of the newer machine learning algorithms. They work by looking to maximise the margin between classes on a hyperplane. The margin exists either side of two classes on a hyperplane. Maximising this margin has been proven to lower an upper bound on the expected generalisation error. Once the hyperplane is created, the kernel function maps new points into the feature space for classification. Several kernel functions are available and careful consideration to kernel section must be considered for each problem. SVMs have been found through extensive studies to perform well when applied to text categorisation problems especially when the linear SVM kernel is utilised. Text categorisation problems are generally linearly separable and SVMs deal well with problems containing many features. Another advantage of SVMs is that model complexity is unaffected by the number of features within the training data. This means SVMs can perform well in situations where the number of features is large compared to the number of training instances (Kotsiantis, 2007). SVMs notably robust performance when dealing with sparse and noisy data makes them a very attractive option when performing text classification (Furey, Cristianini, Duffy, Bednarski, Schummer, & Hussler, 2000).

2.6 Approach’s to similar problems

SFCRs are a new report and as such, automated analysis of these exact reports remains a largely unexplored area. In this section, I will discuss various approaches to solving similar machine learning text classification problems.
Menzies and Marcus used text mining and machine learning methods to automatically generate predictors for severity levels from the free text entered in defect reports. They used a data miner to learn rules that predict for severity attributes and a rule-covering algorithm. The authors observe large f-measures when evaluating their classifier against data from NASA’s Project and Issue Tracking System leading them to conclude there is great potential for their research in practice (Menzies, Marcus, 2008).

In Automated Text Mining for Requirements Analysis of Policy Documents, the authors use topic models in order to indicate whether a document contains software requirements expressed as privacy protections or vulnerabilities. Topic models are probabilistic utilising the Latent Dirichlet allocation algorithm and as such, the documents identified are technically more likely to contain the goal statements. The author’s results argue their topic model approach could limit a requirements engineer’s search from 2,061 policy documents to in several cases fewer than 100 policy documents that may contain topics related to a particular goal keyword (Massey, Eisenstein, Antón, Swire, 2008).

The authors of A Machine Learning Solution to Assess Privacy Policy Completeness present a solution that automatically evaluates the completeness of a websites privacy policy for users. Through use of machine learning and text classification techniques, they prove the feasibility of their approach. They determine that Linear Support Vector Machine (LSVM) is the best performing classifier out of the six algorithms tested. The LSVM classifier obtains a high F-measure and an accuracy of 92%. This is a similar level to the accuracy the authors found analysing a human classifier (Cosante, Sun, Petkovic, Hartog, 2012).

2.7 Conclusion

The review of literature has provided the basis for this project design and implementation. Work by Khan et al stressed the importance of astute pre-processing in order to obtain successful text classification solutions. Findings by Ramos suggest
that tf-idf implementation can greatly improve classifier performance compared to normal term frequency weighting. A comprehensive review of text classification algorithms found the most popular and successful to be K-Nearest Neighbour, Naïve Bayes, Decision Trees, Neural Networks and Support Vector Machines. This study will examine these classifiers in order to determine how they perform in relation to classifying SFCRs. Approach’s to similar problems have revealed largely encouraging results. Text classification has proven to be a reliable and successful solution to addressing problems around classifying policy documents and assessing completeness of online privacy policies. This gives strong encouragement that text classification could be a valuable tool within financial regulation and assessing narrative SFCRs in particular.
3. DESIGN AND METHODOLOGY

3.1 Introduction

This chapter contains an overview of the experimental design, specifications of hardware and software used and documentation of the data source and contents. Experimentation was undertaken using a HP EliteBook laptop with an Intel i7-6500 processor, Intel HD 520 graphics card and 32 gigabytes of on-board RAM.

R was used in order to carry out the experimentation within this study. R is a free software environment for statistical computing and graphics. The study utilises a number of packages within R including dplyr, tm and caret. The dplyr package provides additional functionality for data manipulation. The tm (text mining) package provides functionality for data import, corpus handling, pre-processing, metadata management and creation of term document matrices. The caret (Classification And REgression Training) package provides functions to simplify the model training process for both complex regression and classification problems.

![Figure 3.1 Business Solution Design](image-url)

Reports have been classified as having addressed all required legislative guidelines.

No immediate action required.

A review of the qualitative content will be undertaken by the company’s supervisor.

Reports which the classifier identifies as having not met all legislative guidelines are flagged with the company’s supervision team for immediate review.

They will investigate further and liaise with the company to request correction and resubmission if necessary.

Close to 200 companies submit their Solvency & Financial Condition Reports

Text classification algorithm in R classifies each page of every SFCR in minutes.

It acts as an early warning indicator & identifies any reports that require prioritisation.
Figure 3.1 displays the proposed business solution design for this project. The text classification model built in R will classify each page of every company’s SFCR. A report is then produced for each company specifying if they have sufficiently addressed each required guideline. Companies that have failed to do so will immediately be flagged with that company’s supervision team. They will then conduct a further investigation, liaise with the company and request correction and resubmission if necessary. Companies that are considered to have met all legislative guidelines within the report require no immediate action. Their SFCR will undergo a basic qualitative review of the content by that company’s supervision team in due course. The benefits of this solution would be a considerable reduction in tedious manual work seen in figure 1.1 where an employee had to manually review every report in order to determine which required further investigation. The time taken to identify all reports requiring immediate follow up is reduced from weeks to a day. This results in a sizeable time saving but also a large financial saving.

Figure 3.2 displays the experiment design in detail and is summarised as follows:

**Data Collection:** The Central Bank of Ireland maintains a repository of SFCRs for regulated companies on their website: [https://www.centralbank.ie/registration/industry-market-sectors/insurance-reinsurance/solvency-ii/solvency-and-financial-condition-report-repository](https://www.centralbank.ie/registration/industry-market-sectors/insurance-reinsurance/solvency-ii/solvency-and-financial-condition-report-repository) The SFCRs were downloaded from the repository and read as a corpus in to R.

**Corpus Annotation:** The corpus is written to csv from R and each document of the corpus (i.e. page of an SFCR) is manually labelled with the appropriate corresponding legislative guideline.

**Data pre-processing:** At this stage, the annotated corpus is read back in to R. All words are converted to lowercase. Stop words, digits and punctuation are removed. Additional whitespace is stripped and term document matrices are created.
**Classifiers:** The classifiers examined within literature in section 2.5 are built. K-fold cross validation is utilised during training in order to check for overfitting and to estimate how accurate the models predictions will be in practice.

**Model Evaluation:** The classifiers performance against the test set will be evaluated. The models will be assessed through examination of their sensitivity, specificity, precision, recall and accuracy metrics.

---

3.2 **Data Collection**

The SFCRs must be made publically available on the website of every insurance company regulated by the Central Bank of Ireland. The Central Bank of Ireland maintains a repository of SFCRs for regulated companies on their website: [https://www.centralbank.ie/regulation/industry-market-sectors/insurance-reinsurance/solvency-ii/solvency-and-financial-condition-report-repository](https://www.centralbank.ie/regulation/industry-market-sectors/insurance-reinsurance/solvency-ii/solvency-and-financial-condition-report-repository)
The repository holds 187 SFCRs for year-end 2016. At the time of writing, the 2017 returns have yet to be published in the repository. A subset of 80 SFCRs were selected for this study. The 80 SFCRs contain 4,274 pages. The SFCRs were downloaded from the repository and read as a corpus in to R.

3.3 Corpus Annotation

The corpus is written to csv from R and each document of the corpus (i.e. page of an SFCR) is manually labelled. Articles 290-303 of the Solvency II Delegated Regulation specify the content that must be included within the SFCR. Figure 3.3 displays a high-level structure of an SFCR, this highlights the key areas that must be addressed. For the purpose of this study, the legislative guidelines have been condensed in to nine categories that which must be addressed within an SFCR:

1. Business & Performance
2. Governance Structure
3. Risk Profile
4. Assets
7. Own Funds
8. Difference between Standard Formula and Internal Model
9. Appendix
Figure 3.4 shows a summary of the number of pages within the 80 SFCRs used in this study. The longest SFCR consists of 107 pages, the shortest is just 25 pages and the median page length is 47. The length of each SFCR generally coincides with the size and complexity of the insurance company in question.
The following is a brief description of the information that should be included within each guideline. Each guideline is accompanied with a real example from the labelled corpus. The examples contain only the most relevant subset of the text as to include the full page of text would be prolonged and unsightly.

1. **Business & Performance**

   Insurance and reinsurance undertakings should disclose the following information regarding their business:

   - The name and location of the persons that are direct and indirect holders of qualifying holdings in the undertaking, the proportion of ownership interest held and the proportion of their voting rights.
   - A list of significant related undertakings that includes their name, legal form, location, proportion of ownership held and their voting rights.
   - An overview of the group structure.
   - Information on any leasing arrangements.
   - An executive summary of the company's performance.

Example: *Allianz plc is a non-life insurance company located at Allianz House, Elmpark, Merrion Road, Dublin 4, Republic of Ireland. The Company has a branch in Belfast located at 3 Cromac Quay, The Gasworks, Ormeau Road, Belfast, Northern Ireland. Ownership structure as at 31 December 2016 - the Company is a subsidiary of Allianz Irish Life Holdings plc (AILH), mainly owned by Allianz SE. Allianz Europe BV (Dutch co, Amsterdam) owns 66.49% of AILH who is owned by Allianz SE. The directors regard Allianz SE (registered in Germany) as the ultimate parent Company, its headquarters in Koeniginstrasse 28, 80802 Munich, Germany and holds the legal form of a European company (Societas Europaea). 33.51% of AILH is owned by minority shareholders. Irish Life Irish Holdings own 30.43% of the shares who in turn are owned by Canada Life Ltd (UK co), a member of the Great West LifeCo Group based in Canada. The remaining 3.08% of AILH is owned by other minorities made up of both individual and corporate shareholders. The ownership percentages are equal to the voting shares.*
2. Governance Structure

Insurance and reinsurance companies should disclose how their key functions have the required authority, resources and operational independence to carry out their jobs and the manner in which they report to and guide management.

Example: The Board of Directors has mandated a basis for effective risk management within the Company dictated by a clear system of governance that covers all significant aspects of the business, provides an open forum for challenge, and allocates clear responsibilities for both collective management committees and individuals. In addition, the Board has established the four key control functions required under the Corporate Governance Requirements for Insurance Undertakings 2015, risk management, actuarial, compliance and internal audit. These functions are responsible for providing oversight of, and challenge to, the business and for providing assurance to the Board in relation to the Company’s control framework. The Board has delegated the day to day running of the Company to the CEO who has been instructed to appoint a management team to assist him in these duties. The CEO reports on these activities at each quarterly board meeting and presents a business update for its approval.

3. Risk Profile

Insurance and reinsurance undertakings should disclose the following information regarding their business:

- A comprehensive breakdown of the company’s risks including market risk, liquidity risk, governance risk, credit risk, underwriting risk and counterparty default risk.

Example: Liquidity risk is defined as the risk that the Company either does not have sufficient financial resources to meet obligations as they fall due or can only secure such financial resources at excessive cost. At 31 December 2016, the company held assets of €705m on its Solvency II Balance Sheet in order to meet the Company’s liabilities and Solvency Capital Requirement (SCR). €700m of these assets were held in liquid investments. The HoAF has also performed an assessment of the liquidity
position of assets representing Own Funds in particular. At 31 December 2016, the Company held Own Funds of €12.9m over and above the Solvency Capital Requirement. The entire €12.9m was held in highly liquid investments. The HoAF is very satisfied with the Company’s current liquidity position. The Company is exposed to liquidity risk as a result of its business operations including cash flow timing mismatches between policyholder obligations and claims and re-insurance recoveries as well as cash flow obligations arising on operating expenses, taxation, dividends and other liabilities.

4. Assets

Insurance and reinsurance undertakings should disclose the following information regarding their business:

- Valuation basis applied to aggregated asset classes with consideration to the nature, function, risk and materiality of these assets.
- Quantitative and qualitative information for each material class of asset including the recognition and valuation basis applied, the methods used and any judgements used other than estimations, which may materially influence the amounts recognised.

Example: Assets held for index-linked and unit-linked funds mainly consists of policyholder financials assets (debt securities, equity shares, unit trusts, trackers, investment properties and derivatives) that are valued at fair value through profit and loss ("FVTPL") in the IFRS financial statements determined in accordance with IFRS 13 Fair value measurement. The overarching valuation principle under Article 75 (Article 75 (1) of Directive 2009/138/EC)) is that assets are required to be valued at the amount for which they could be exchanged between knowledgeable willing parties in an arm’s length transaction. Solvency II requires that the valuation methods used should be compatible with International Accounting Standards ("IAS’s") provided that such valuation methods are consistent with Article 75. The accounting standard for determining the fair value of financial assets is IFRS 13. The fair value of financial assets as determined by IFRS 13 is consistent with the Solvency II framework under Article 75. Reinsurance recoverables For Solvency II and IFRS financial statements
valuation, reinsurance recoverable relate to the share of Technical Provisions for ceded business that is determined with reference to the contractual agreement and the underlying gross liability.


Insurance and reinsurance undertakings should disclose the significant methods used to calculate the technical provisions with specific detail provided for methods used to calculate the risk margin.

Example: The methodology and assumptions used in calculating the technical provisions are in accordance with articles 75 to 86 of Directive 2009/138/EC of the European Parliament, articles 17 to 42 of Commission Delegated Regulation (EU) 2015/35 of 10 October 2014 supplementing Directive 2009/138/EC of the European Parliament and the Guidelines on valuation of technical provisions, EIOPA-BoS-14/166. The methodology for calculating the best estimate liability (“BEL”) is consistent with the concept of representing the amount that another insurer would need to be paid to assume these policies. The technical provisions typically consist of a liability equivalent to a best estimate of the future cash flows, along with a risk margin that EIOPA intended would reflect the compensation another insurer would be expected to seek for assuming the associated potential uncertainty. The technical provisions do not include any allowance for transitional measures, matching adjustment or volatility adjustment. The key sources of uncertainty associated with the technical provisions are the number and size of claim payments (in respect of claims incurred prior to the valuation date but not yet paid).

6. Valuation of Other Liabilities

Insurance and reinsurance undertakings should disclose the following information regarding their business:
• Valuation basis applied to aggregated liabilities other than technical provisions with consideration to the nature, function, risk and materiality of these liabilities.

• Quantitative and qualitative information for each material class of liabilities other than technical provisions including the recognition and valuation basis applied, the methods used and any judgements used other than estimations, which may materially influence the amounts recognised

Example: Aside from Technical provisions, the valuations of which are detailed above, Carraig has 2 other principal classes of liabilities;

• Deferred taxation

• Creditors arising out of direct insurance operations

Deferred taxation is provided on all timing differences that have originated but not reversed at the balance sheet date where transactions or events that result in an obligation to pay more tax in the future or a right to pay less tax in the future have occurred at the balance sheet date, except to the extent that deferred tax assets are recognised only when it is more probable than not that there will be future taxable income streams against which such assets can be offset. Timing differences are temporary differences between profits as computed for tax purposes and profits as stated in the financial statements, which arise because certain items of income and expenditure in the financial statements are dealt with in different years for tax purposes. Deferred tax is measured at the tax rates that are expected to apply in the years in which the timing differences are expected to reverse based on tax rates and laws that have been enacted or substantively enacted by the balance sheet date. Deferred tax is not discounted.

7. Own Funds

Insurance and reinsurance undertakings should disclose the following information regarding their business:

• Differentiate between basic and ancillary own fund items
• The availability, duration and other pertinent features for assessing quality for each own fund item
• An analysis of significant changes in own funds throughout the reporting period
• An explanation of any changes to subordinated debt
• Notice of any restrictions to available own funds
• Breakdown of the key elements of the reconciliation reserve

Example: The Company’s approach to capital management and how it manages available own funds (being the excess of assets over liabilities) is outlined in the Company’s Capital Management Policy. Key objectives of the policy are to be compliant with all applicable laws, rules and regulations governing the management of capital and to maintain, at all times, sufficient own funds to cover both the Solvency Capital Requirement and Minimum Capital Requirement. The policy and associated processes help to protect the financial strength of the Company, by identifying various capital levels, and requiring appropriate actions depending on the current level of capital. There were no material changes to the objectives, policies and processes for managing own funds over the reporting period. The Asset & Liability Committee and the Board regularly consider capital assessments and projections for the Company to ensure that capital is managed with continuous adherence to Aegon Ireland’s principles around capital adequacy, financial flexibility and the efficient use of capital. The Own Risk and Solvency Assessment process includes an assessment of the sufficiency of capital available to meet the commitments in light of the risks faced by the business, both now and into the future. Aegon Ireland determines its solvency needs by performing capital projections over the business planning period, allowing for the current and expected business strategy, risk profile and capital management activities.

8. Differences between the Standard Formula and Internal Model
Insurance and reinsurance undertakings should disclose the following information regarding differences between the standard formula and internal model:
• The structure of the internal model
• The aggregation methodologies and diversification effects
• Outline of any risks not covered by the standard formula but considered by the internal model

Example: The Solvency II Regulations introduces a risk based capital requirement which can be assessed either using an internal model or a standard formula. The AXA economic capital model (AXA’s Internal Model) aims to cover all the material and quantifiable risks to which the entity is exposed. AXA’s Internal Model offers a concrete and powerful tool to control and measure exposure to most risks, in line with the Solvency II framework. The economic capital model is based on a common definition of risks used consistently throughout the AXA Group. It aims to ensure that the Company risk mapping is comprehensive and is followed in a consistent way across the Company and that efficient procedures and reporting are in place so that roles and responsibilities are allocated to identify, measure, monitor, manage and report key risks. The Group risk grid identifies all material risks applicable for the Company insurance businesses. AXA’s Economic capital model captures all material risks applicable for the Company insurance businesses in order to assess the different sub risks and the overall aggregation of risks. The underlying methodologies used in the economic capital model are regularly reviewed to ensure that they accurately reflect the Company’s risk profile and new methods are developed and integrated if necessary (in accordance with the internal model change policy). AXA’s Internal Model is calibrated to represent the value at risk of Group Economic Value at the 99.5th percentile over a one year horizon. In other words, the Solvency Capital requirement (SCR) is the capital needed to sustain a 1 in 200 years shock. It strives to include all measurable risks (market, credit, insurance and operational) and reflects AXA’s unique diversified profile.

9. Appendix
• Insurance and reinsurance undertakings must disclose the relevant quantitative reporting templates within their appendix
Example: APPENDICES

The following quantitative templates fall under scope of Solvency II Audit (with the exception of S.05.01.01 and S.05.02.01) which comes into effect for periods from 31st December 2016 and have been added to Appendix 1.

S.02.01.02 - Balance sheet
S.05.01.01 - Premiums, Claims and Expenses
S.05.02.01 - Premiums, claims and expenses by country;
S.17.01.02 - Non Life technical provisions
S.19.01.21 - Claims Developments
S.23.01.01 - Own Funds
S.25.01.21 - SCR using standard formula

<table>
<thead>
<tr>
<th>Category</th>
<th>Training</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business &amp; Performance</td>
<td>522</td>
<td>189</td>
</tr>
<tr>
<td>Governance Structure</td>
<td>783</td>
<td>276</td>
</tr>
<tr>
<td>Risk Profile</td>
<td>447</td>
<td>129</td>
</tr>
<tr>
<td>Assets</td>
<td>116</td>
<td>53</td>
</tr>
<tr>
<td>Valuation of Technical Provisions</td>
<td>159</td>
<td>60</td>
</tr>
<tr>
<td>Liabilities other than Technical Provisions</td>
<td>73</td>
<td>32</td>
</tr>
<tr>
<td>Own Funds</td>
<td>252</td>
<td>100</td>
</tr>
<tr>
<td>Differences between the Standard Formula and Internal Model</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>Appendix</td>
<td>607</td>
<td>295</td>
</tr>
</tbody>
</table>

Table 3.1 Breakdown of SFCR Categories

Table 3.1 displays the split of categories within the training and test data. It is clear that companies tend to focus their SFCRs predominantly on Business & Performance, Governance Structure and Risk profile. The Appendix represents a large amount of pages, as companies must publish various quantitative reporting templates here. Differences between the Standard Formula and Internal Model are seldom present.
however; this is not surprising as only a handful of firms have approval to use an internal model.

3.4 Data Pre-processing

At this stage, the annotated corpus is read back in to R and the following pre-processing techniques are applied:

- Conversion of all characters to lowercase. This is done so words with and without capitals are considered the same word for analysis. For example, “Capital” and “capital” would be considered two different words in the absence of this conversion.
- Removal of all digits will be trialled. In some instances, it is beneficial to leave digits present within the text. SFCRs will reference quantitative reporting templates, which are numbered. So leaving digits in may be beneficial for determining which category a page of the SFCR is addressing.
- Removal of punctuation marks. These are noisy which will not add any value to the performance of our classifiers.
- Stripping of extra whitespace. This removes extra unnecessary whitespace from documents.
- Removal of stop words. Removal of stop words focuses on removing noisy insignificant frequently occurring stop words such as “and”, “a” and “the”.

Following this, a document term matrix is created. This matrix describes the frequency of terms that occur in the corpus. Term frequency however only identifies the most frequently occurring features; because of this, meaningful features for some categories within documents may be overlooked. Tf-idf addresses this issue; it does so by diminishing the weight of terms that occur very frequently within the corpus and increases the weight of terms that rarely occur. Therefore, a tf-idf document term matrix is also created in order to conduct analysis between the performance of term frequency and tf-idf within this study. The data is then split in to training and test data.
3.5 Classifiers

K-fold cross validation is utilised during training in order to check for overfitting and to estimate how accurate the models predictions will be in practice. The value of K within this study is chosen to be 10. 10-fold cross validation is the most popular within research and is capable of detecting overfitting and a decent guide for model performance in practice. The following machine learning classifiers are built from the caret package in R:

- K-Nearest Neighbour
- Naïve Bayes
- Decision Tree
- Neural Network
- Support Vector Machine

Model training within this study is an iterative process. Initial experiments evaluating performance will are undertaken in order to determine the impacts of:

- Term frequency versus term frequency inverse document frequency
- Impact of feature selection
- Various pre-processing decisions

Firstly the classifier’s performance is tested when term frequency representation is used compared to term frequency inverse document frequency. Following this is an exploration of the impact of feature selection. The features are selected using filtering based on the tf or tf-idf scores. The experiment will examine how the model performs with the top 20% of features, top 40%, top 60%, top 80% and all features selected. Finally an examination the effects of pre-processing techniques will be undertaken. Following findings within the literature review, the base model for all initial examination includes conversion to lowercase, removal of stop words, removal of numbers, removal of punctuation and stripping of whitespace. This base model is compared against a second model where numbers are not removed. Although in general it is considered numbers will be noisy, the quantitative return templates are referenced by numbers, so there is a chance that leaving numbers in could positively
influence performance. A third model includes stemming in order to determine if it can improve performance. These models are summarised in table 3.2.

<table>
<thead>
<tr>
<th></th>
<th>Model 1 (base model)</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conversion to lowercase</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Removal of stop words</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Removal of digits</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Removal of punctuation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Stripping of whitespace</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Stemming</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
</tbody>
</table>

*Table 3.2 Pre-Processing Models*

3.6 Model Evaluation

The classifiers accuracy from 10-fold cross validation is initially assessed. Ultimately, we are aiming to obtain the highest accuracy score we can. F-score is very popular within text classification however; this is a legacy from the Information Retrieval domain. This study is interested in how many pages of an SFCR can be correctly classified. Precision makes sense within Information Retrieval making F-score an attractive measure but this is not the case in text classification. The experiments outlined section 3.5 will be used in order to carry out an initial examination of the different classification algorithms. The best performing model will then be analysed in more detail.
4. BUILDING AND EVALUATING PREDICTIVE MODELS

This chapter details the experiment execution accompanied with an evaluation of the methodology. There will be a brief outline of the data preparation and data investigation. Following this, the results for each text classification algorithms are detailed.

4.1 Data Preparation

A large amount of data preparation was required in order to undertake this study. The most time consuming element of this process was the labelling of the corpus. This involved manually labelling over four thousand pages from SFCR documents with their applicable legislative guideline. This process and the requirements of these guidelines were discussed in detail in section 3.3. Following this process the annotated corpus in csv format was read in to R. The following steps outline the process in order to prepare this data for the algorithm building stage:

- Corpus created in R
- Pre-processing techniques applied to the corpus (removal of stop words etc.)
- Document term frequency matrix created
- Term frequency inverse document frequency matrix created
- Both matrices split in to training and test sets
- Individual classifier considerations

4.2 Data Investigation

The corpus consists of 80 SFCRs. The tree map in figure 4.1 demonstrates the 4,274 pages within this corpus consist largely of Governance Structure, Appendix and Business and Performance. These three categories contribute to 65% of the entire corpus. The remaining six categories make up the other 35%. It is notable that Differences between Standard Formula and Internal Model makes up less than 1% of
the corpus with only 24 occurrences. This is due to only a handful of companies having approval to implement an internal model therefore the legislative guideline to disclose this information is not applicable to the majority of firms.

<table>
<thead>
<tr>
<th>SFCR CORPUS COMPOSITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Governance Structure</td>
</tr>
<tr>
<td>Business &amp; Performance</td>
</tr>
<tr>
<td>Risk Profile</td>
</tr>
<tr>
<td>Appendix</td>
</tr>
<tr>
<td>Own Funds</td>
</tr>
<tr>
<td>Valuati... of Techni...</td>
</tr>
<tr>
<td>Provisi...</td>
</tr>
<tr>
<td>Assets</td>
</tr>
<tr>
<td>Liabilities</td>
</tr>
</tbody>
</table>

Figure 4.1 Tree Map of Corpus Categories

Figure 4.2 displays a word cloud where size is indicative of the most frequently occurring words within the corpus. For the creation of this word cloud numbers, punctuation and English stop words were removed. Unsurprisingly business domain words such as insurance, reinsurance, capital, management, assets, risks and board are amongst the most frequently occurring within the corpus of SFCRs.
4.3 K-Nearest Neighbour

K-nearest neighbour classifies documents in the Euclidean space as points. It determines how to categorise a document by only looking at the training documents most similar to it.

![Figure 4.3 KNN tf & tf-idf Comparison](image)

Figure 4.3 shows the accuracy scores calculated for a K-nearest neighbour classifier with 10-fold cross validation with three repeats. Using tf-idf document term matrix as opposed to standard term frequency gives the classifier a very marginal improvement in accuracy. The tf-idf implementation results in an accuracy of 83.34% compared with 82.65% when using the standard term frequency. The tf version took 58.34 minutes to train and the tf-idf model was slightly longer with a training time of 62.51 minutes.
In order to examine the effect of dimensionality reduction with respect to the k-nearest neighbour classifier, feature selection is utilised. This experiment compares how the model performance fluctuates when a feature selection filter is used. Using tf-idf, the top ‘x’ percentage of features are selected. The classifier is trained using the top 20%, 40%, 60% and 80% of features to compare with the first model, which was trained using all features. Figure 4.4 shows the k-nearest neighbour model performance improved to 83.52% accuracy when only the top 20% of tf-idf features were selected. The performance decreases as more features are utilised however, it does increase again when trained with all features. It is important to note however, the variation in accuracy is not that large. With the lowest K-nearest neighbour model obtaining an accuracy of 82.64% and the highest 83.52%.

<table>
<thead>
<tr>
<th>K-nearest Neighbour tf-idf</th>
<th>Model 1 (base model)</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conversion to lowercase</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Removal of stop words</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Removal of digits</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Removal of punctuation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Stripping of whitespace</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Stemming</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Accuracy</td>
<td>83.52</td>
<td>85.77</td>
<td>85.53</td>
</tr>
</tbody>
</table>

*Figure 4.4 KNN feature selection*

*Table 4.1 KNN Pre-Processing*
Examining the effects of pre-processing decisions highlights that performance of the classifier improves if numbers are retained in the corpus. Table 4.1 provides a summary of k-nearest neighbour pre-processing decision variations. Although the SFCRs contain many numbers that will have little meaning, an individual table code represents each quantitative reporting template. For example, the balance sheet is SE.02.01.17.01. The presence of numbers is potentially helping the classifier in identifying categories where certain numbers are usually present and categories that tend to have an absence of numbers. Model 2 shows an improvement of accuracy to 85.77% when numbers are retained in the corpus. Model 3 examines the impact of stemming; it results in a decrease in accuracy to 85.53%.

![Figure 4.5 K Tuning](image)

**Figure 4.5 K Tuning**

K is the number of neighbours that are chosen to vote on a new examples class. Figure 4.5 displays the tuning for K. It is found that the K-nearest neighbour model performs best with a value of five selected for K. The best performing k-nearest neighbour model utilises tf-idf representation and filters such that only the top 20% of features are considered. It uses the pre-processing techniques from model 2 seen in table 4.1. Achieving an accuracy of 85.77%, the k-nearest neighbour model performs well when addressing the problem of classifying the contents of SFCRs.
4.4 Naïve Bayes

Naïve Bayes is a simple probabilistic classifier based on applying Bayes theorem. The algorithm applies an independence assumption; it assumes all features of the examples are independent of each other given the context of the class.

Figure 4.6 shows the accuracy scores calculated for a Naïve Bayes classifier with 10-fold cross validation with three repeats. Using tf-idf document term matrix as opposed to standard term frequency gives the classifier a very marginal improvement in accuracy. The tf-idf implementation results in an accuracy of 86.86% compared with 86.58% when using the standard term frequency. Both classifiers had reasonably quick training times. The tf version took just 13.77 minutes to train and the tf-idf model had a training time of 13.75 minutes.

Figure 4.6 Naïve Bayes tf & tf-idf Comparison
In order to examine the effect of dimensionality reduction with respect to the Naïve Bayes classifier, feature selection is utilised. This experiment compares how the model performance fluctuates when a feature selection filter is used. Using tf-idf, the top ‘x’ percentage of features are selected. The classifier is trained using the top 20%, 40%, 60% and 80% of features to compare with the first model, which was trained using all features. Figure 4.7 shows the Naïve Bayes model does not perform as well with only the top 20% of tf-idf features selected. However when the top 40% of features are selected the model is performing essentially on par with the model containing all features.

<table>
<thead>
<tr>
<th>Naïve Bayes tf-idf</th>
<th>Model 1 (base model)</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conversion to lowercase</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Removal of stop words</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Removal of digits</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Removal of punctuation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Stripping of whitespace</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Stemming</td>
<td>×</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>Accuracy</td>
<td>86.86</td>
<td>87.62</td>
<td>85.77</td>
</tr>
</tbody>
</table>

Table 4.2 Naïve Bayes Pre-Processing

Examining the effects of pre-processing decisions highlights that performance of the classifier improves if numbers are retained in the corpus. Table 4.2 provides a
summary of naïve bayes pre-processing decision variations. Although the SFCRs contain many numbers that will have little meaning, an individual table code represents each quantitative reporting template. For example, the balance sheet is SE.02.01.17.01. The presence of numbers is potentially helping the classifier in identifying categories where certain numbers are usually present and categories that tend to have an absence of numbers. Leaving numbers present within the corpus, results in an increase in classifier accuracy to 87.62%. Model 3 examines the impact of stemming; it results in a decrease in accuracy to 85.77%.

The best performing naïve bayes model utilises tf-idf representation and filters such that only the top 40% of features are considered. It uses the pre-processing techniques from model 2 seen in table 4.2. Thus proving that it is beneficial when considering classifying pages of SFCRs to keep numbers in the corpus. Achieving an accuracy of 87.62%, the naïve bayes model preforms well when addressing the problem of classifying the contents of SFCRs.

4.5 Decision Tree

Decision trees are constructed with the use of a hierarchical split of the underlying data space with the use of different text features.
Figure 4.8 Decision Tree tf & tf-idf Comparison

Figure 4.8 shows the accuracy scores calculated for a decision tree classifier with 10-fold cross validation with three repeats. Using tf-idf document term matrix as opposed to standard term frequency gives the classifier a very marginal improvement in accuracy. Standard term frequency resulted in an accuracy of just 54.16%. The tf-idf model had even worse performance with an accuracy of 53.96%. The tf version took 24.16 minutes to train and the tf-idf model had a training time of 24.86 minutes.

Figure 4.9 Decision Tree Feature Selection

In order to examine the effect of dimensionality reduction with respect to the decision tree classifier, feature selection is utilised. This experiment compares how the model performance fluctuates when a feature selection filter is used. Using term frequency, the top ‘x’ percentage of features are selected. The classifier is trained using the top 20%, 40%, 60% and 80% of features to compare with the first model, which was trained using all features. Figure 4.9 shows that feature selection does not influence the performance of the decision tree. Regardless of the feature filtering, accuracy remains constant at around 54.16%.
Examining the effects of pre-processing decisions highlights that performance of the classifier remains the same if numbers are retained in the corpus. Table 4.3 provides a summary of decision tree pre-processing decision variations. Although the SFCRs contain many numbers that will have little meaning, an individual table code represents each quantitative reporting template. For example, the balance sheet is SE.02.01.17.01. This shows the presence of numbers does not affect the decision tree performance. Model 3 examines the impact of stemming; it results in a decrease in accuracy to 51.79%.

<table>
<thead>
<tr>
<th>Decision Tree Transformation</th>
<th>Model 1 (base model)</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conversion to lowercase</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Removal of stop words</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Removal of digits</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Removal of punctuation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Stripping of whitespace</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Stemming</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Accuracy</td>
<td>54.16</td>
<td>54.16</td>
<td>51.79</td>
</tr>
</tbody>
</table>

*Table 4.3 Decision Tree Pre-Processing*
Figure 4.10 displays the effects of complexity tuning on the decision trees accuracy performance. The complexity parameter dictates how the number of terminal nodes regulates the cost of a tree. A lower value for the complexity parameter leads to larger trees with a potential for overfitting. Conversely, a large value for the complexity parameter produces smaller trees with the possibility of underfitting.

The best performing decision tree model utilises tf representation and filters such that only the top 20% of features are considered. It uses the pre-processing techniques from model 1 seen in table 4.3. Achieving an accuracy of 54.16%, the decision tree model preforms very poorly when addressing the problem of classifying the contents of SFCRs.

### 4.6 Neural Network

![Neural Network accuracy chart](image)

Figure 4.11 shows the accuracy scores calculated for a decision tree classifier with 10-fold cross validation with three repeats. Using tf-idf document term matrix as opposed to standard term frequency gives the classifier a very marginal improvement in accuracy. Standard term frequency resulted in an accuracy of just 86.21%. The tf-idf model had better performance with an accuracy of 88.40%. The tf version took
63.27 minutes to train and the tf-idf model had a slightly longer training time of 64.67 minutes.

In order to examine the effect of dimensionality reduction with respect to the neural network classifier, feature selection is utilised. This experiment compares how the model performance fluctuates when a feature selection filter is used. Using term frequency, the top 20% of features are selected. The classifier is trained using the top 20%, 40%, 60% and 80% of features to compare with the first model, which was trained using all features. Figure 4.12 shows that the neural networks accuracy increases as the number of features it uses increases. With all features used for prediction, the neural network has an accuracy of 88.40%.

<table>
<thead>
<tr>
<th>Neural Network tf-idf</th>
<th>Model 1 (base model)</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conversion to lowercase</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Removal of stop words</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Removal of digits</td>
<td>✔️</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Removal of punctuation</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Stripping of whitespace</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Stemming</td>
<td>✗</td>
<td>✗</td>
<td>✔️</td>
</tr>
<tr>
<td>Accuracy</td>
<td>88.4</td>
<td>87.21</td>
<td>86.36</td>
</tr>
</tbody>
</table>

*Table 4.4 Neural Network pre-processing*
Examining the effects of pre-processing decisions highlights that performance of the classifier decreases to 87.21% if numbers are retained in the corpus. Table 4.4 provides a summary of neural network pre-processing decision variations. Although the SFCRs contain many numbers that will have little meaning, an individual table code represents each quantitative reporting template. For example, the balance sheet is SE.02.01.17.01. This shows the presence of numbers negatively affects the neural network performance. Model 3 examines the impact of stemming; it results in a decrease in accuracy to 86.36%. Thus determining stemming is not a useful pre-processing implementation for the neural network when considering this problem.

Figure 4.13 Neural Network Weight Decay

Figure 4.13 explores weight decay values effect on accuracy for the neural network. Weight decay stipulates regularisation in the neural network. The regularisation term is
used to calculate the backpropagation gradient. Therefore, the weight decay value controls the power the regularisation term has in the gradient computation. The best performing neural network model utilises tf-idf representation and all features are considered for training. It uses the pre-processing techniques from model 1 seen in table 4.4. Achieving an accuracy of 88.40%, the neural network model preforms well when addressing the problem of classifying the contents of SFCRs.

### 4.7 Support Vector Machine

Support Vector Machines (SVMs) are one of the newer machine learning algorithms. They work by looking to maximise the margin between classes on hyperplane. The margin exists either side of two classes on a hyperplane. Maximising this margin has been proven to lower an upper bound on the expected generalisation error. Once the hyperplane is created, the kernel function maps new points into the feature space for classification.

![SVM Accuracy](image)

*Figure 4.14 SVM tf & tf-idf Comparison*

Figure 4.14 shows the accuracy scores calculated for a linear SVM classifier with 10-fold cross validation with three repeats. Using tf-idf document term matrix as opposed
to standard term frequency improves the accuracy of the classifier. Standard term frequency resulted in an accuracy of just 89.20%. The tf-idf model had better performance with an accuracy of 91.78%. The tf version took 30.59 minutes to train and the tf-idf model had a slightly longer training time of 31.58 minutes.

![SVM Feature Selection](image)

**Figure 4.15 SVM Feature Selection**

In order to examine the effect of dimensionality reduction with respect to the SVM classifier, feature selection is utilised. This experiment compares how the model performance fluctuates when a feature selection filter is used. Using term frequency, the top 20% of features are selected. The classifier is trained using the top 20%, 40%, 60% and 80% of features to compare with the first model, which was trained using all features. Figure 4.15 shows that the SVM accuracy is best when all features are used. With all features used for prediction, the SVM has an accuracy of 91.78%. It is notable however, that 20%, 40%, 60% and 80% the classifier performs well with accuracy always around 89.2%. Three is a possibility that the model with all features has been over fit. Nevertheless, the linear SVM model has performed very well.
Examine the effects of pre-processing decisions highlights that performance of the classifier decreases to 89.90% if numbers are retained in the corpus. Table 4.5 provides a summary of SVM pre-processing decision variations. Although the SFCRs contain many numbers that will have little meaning, an individual table code represents each quantitative reporting template. For example, the balance sheet is SE.02.01.17.01. This shows the presence of numbers negatively affects the SVM performance. Model 3 examines the impact of stemming; it results in a decrease in accuracy to 89.72%. Thus determining stemming is not a useful pre-processing implementation for the SVM when considering this problem.

The best performing linear SVM model utilizes tf-idf representation and utilizes all available features. It uses the pre-processing techniques from model 1 seen in table 4.1. Thus highlighting for SVM that it is beneficial when considering classifying pages of SFCRs to remove numbers in the corpus. Achieving an accuracy of 91.78%, the SVM model preforms very well when addressing the problem of classifying the contents of SFCRs.

### 4.8 Initial Results

With consideration to both the research question and business problem, the initial results are encouraging. Average accuracy scores computed from 10-fold cross validation with three repeats have been used to assess model performance. Table 4.6 provides a summary of the best performing model for each classification algorithm.
Table 4.6 Best Performing Models

<table>
<thead>
<tr>
<th></th>
<th>K-Nearest Neighbour</th>
<th>Naïve Bayes</th>
<th>Decision Tree</th>
<th>Neural Network</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bag of words transformation</td>
<td>tf-idf</td>
<td>tf-idf</td>
<td>tf</td>
<td>tf-idf</td>
<td>tf-idf</td>
</tr>
<tr>
<td>Pre-processing model</td>
<td>Model 2</td>
<td>Model 2</td>
<td>Model 1</td>
<td>Model 2</td>
<td>Model 1</td>
</tr>
<tr>
<td>Feature selection %</td>
<td>20</td>
<td>40</td>
<td>20</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Training time (mins)</td>
<td>22.47</td>
<td>4.12</td>
<td>4.77</td>
<td>64.67</td>
<td>31.58</td>
</tr>
<tr>
<td>Accuracy</td>
<td>85.77</td>
<td>87.62</td>
<td>54.16</td>
<td>87.21</td>
<td>91.78</td>
</tr>
</tbody>
</table>

Table 4.6 highlights only the decision tree model as performing particularly poorly. The decision tree appears to struggle with the large number of features associated with this problem. K-nearest neighbour is the second poorest model in terms of accuracy however it manages to classify the SFCR pages to a decent standard at 85.77%. Naïve bayes and the neural network perform well with similar accuracies of 87.62% and 87.21% respectively. The linear support vector machine however was by far the best performing classifier with an accuracy of 91.78%. For k-nearest neighbour, naïve bayes and neural network it was discovered that leaving the numbers in the corpus during the pre-processing stage had a positive impact on the classifier performance. However, the SVM’s accuracy diminished when leaving the numbers in the corpus. Therefore, it cannot be said to be beneficial for all models. This solidifies popular opinions in literature that there is no set pre-processing template for success. A trial and error explorative approach is necessary to tackle unique classification problems.

The investigation in to the impact of feature selection highlighted most algorithms did not need all features in order to train the model successfully. Accuracy of k-nearest neighbour improved when only the top 20% of features were selected. SVM achieved its greatest accuracy with all features considered however; its performance with only 20% of features selected was still as high as 89.24%. Only the neural network required all features to achieve a considerably improved accuracy. The aim of this study is to classify SFCR documents as accurately as possible. As a result training times, feature selection and computational costs are not key concerns. The SVM model in table 4.6 is
Initially the most promising classifier for answering the research question and business problem at hand.

4.9 SVM model tuning

The initial findings showed the linear SVM model to be the best performing classifier with an accuracy of 91.78%. The cost parameter for that model was one. The cost parameter states the level at which the SVM wants to avoid misclassification. The larger the cost value, the smaller the margin is for the hyperplane. A smaller margin in the hyperplane leads to less chance of misclassification. Conversely, a smaller cost value enlarges the margin and can lead to a greater risk of misclassification. Figure 4.16 examines the accuracy performance for the following cost parameter values: 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2.5.

![Figure 4.16 SVM Cost Tuning](image)

Figure 4.16 shows that the large margin at low cost, results in poor accuracy scores. The accuracy improves to an ok level of 77.51% at a cost of 0.25. At 0.5 it has risen to 88.74%. From the cost of 1 onwards the improvement in accuracy slows down.
dramatically as cost increases. However, cost of 2.5 obtains an accuracy of the 93.26%. As a result, a cost value of 2.5 is selected for the final model.

![SVM Accuracy](image)

**Figure 4.17 Final SVM Model Accuracy**

It is visible in figure 4.17 that the linear SVM model obtains an accuracy of 90.46% when examined against the test set. This is a drop off from the 93.26% achieved in the 10-fold cross validation but a good result nonetheless.

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Appendix</th>
<th>Assets</th>
<th>Business</th>
<th>Diff SF &amp; IM</th>
<th>Governance</th>
<th>Liabilities</th>
<th>Own Funds</th>
<th>Risk Profile</th>
<th>Technical Provisions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appendix</td>
<td>275</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Assets</td>
<td>0</td>
<td>45</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Business</td>
<td>11</td>
<td>1</td>
<td>169</td>
<td>1</td>
<td>9</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Diff SF &amp; IM</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Governance</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>262</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Liabilities</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>24</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Own Funds</td>
<td>3</td>
<td>1</td>
<td>7</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>84</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Risk Profile</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>6</td>
<td>120</td>
<td>0</td>
</tr>
<tr>
<td>Technical Provisions</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>54</td>
<td></td>
</tr>
</tbody>
</table>

*Table 4.7 SVM Confusion Matrix*
Table 4.7 displays the confusion matrix obtained from the models predictions against the test set. For the majority of the nine categories the results are good. However, the model did not correctly identify any of the 10 pages belonging to ‘Differences between Standard Formula and Internal Model’. Table 4.8 shows the statistics associated with the confusion matrix. The table highlights the poor performance with relation to the guideline addressing differences between standard formula and internal model, a sensitivity score of 0 is obtained. The only other category with a potentially concerning accuracy for the model is ‘Liabilities other than Technical Provisions’. The model has only correctly identified 24 out of 32 of these instances in the test set, 75%.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Appendix</th>
<th>Assets</th>
<th>Business</th>
<th>Diff SF &amp; IM</th>
<th>Governance</th>
<th>Liabilities</th>
<th>Own Funds</th>
<th>Risk Profile</th>
<th>Technical Provisions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.9322</td>
<td>0.8490</td>
<td>0.8942</td>
<td>0.0000</td>
<td>0.9493</td>
<td>0.7500</td>
<td>0.8400</td>
<td>0.9449</td>
<td>0.9000</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.9882</td>
<td>0.9917</td>
<td>0.9717</td>
<td>0.9991</td>
<td>0.9873</td>
<td>0.9982</td>
<td>0.9827</td>
<td>0.9803</td>
<td>0.9898</td>
</tr>
<tr>
<td>Pos Pred Value</td>
<td>0.9649</td>
<td>0.8333</td>
<td>0.8622</td>
<td>0.0000</td>
<td>0.9597</td>
<td>0.9231</td>
<td>0.8235</td>
<td>0.8571</td>
<td>0.8307</td>
</tr>
<tr>
<td>Neg Pred Value</td>
<td>0.9767</td>
<td>0.9926</td>
<td>0.9789</td>
<td>0.9912</td>
<td>0.9839</td>
<td>0.9928</td>
<td>0.9846</td>
<td>0.9930</td>
<td>0.9944</td>
</tr>
<tr>
<td>Prevalence</td>
<td>0.2583</td>
<td>0.0464</td>
<td>0.1655</td>
<td>0.0080</td>
<td>0.2417</td>
<td>0.0280</td>
<td>0.0876</td>
<td>0.1112</td>
<td>0.0525</td>
</tr>
<tr>
<td>Detection Rate</td>
<td>0.2408</td>
<td>0.0394</td>
<td>0.1480</td>
<td>0.0000</td>
<td>0.2294</td>
<td>0.0210</td>
<td>0.0735</td>
<td>0.1051</td>
<td>0.0473</td>
</tr>
<tr>
<td>Detection Prevalence</td>
<td>0.2496</td>
<td>0.0473</td>
<td>0.1716</td>
<td>0.0000</td>
<td>0.2391</td>
<td>0.0227</td>
<td>0.0893</td>
<td>0.1226</td>
<td>0.0569</td>
</tr>
<tr>
<td>Balanced Accuracy</td>
<td>0.9602</td>
<td>0.9204</td>
<td>0.9329</td>
<td>0.4995</td>
<td>0.9683</td>
<td>0.8741</td>
<td>0.9113</td>
<td>0.9626</td>
<td>0.9449</td>
</tr>
</tbody>
</table>

*Table 4.8 SVM Statistics*
4.10 Summary of results

- Term frequency inverse document frequency (tf-idf) bag of words representation significantly improves classifier performance when compared with the standard term frequency representation.
- The decision tree is the only classification algorithm that appears totally unsuitable for this problem.
- Linear support vector machine proved to be the best classifier for this problem. It has shown that text classification can be successfully applied to the narrative SFCR returns. It correctly classified 90.46% of SFCR pages from the 1,142 pages within the test set.
5. ANALYSIS, EVALUATION & DISCUSSION

The penultimate section focuses on a review of the strength of research undertaken and the results.

- Results are evaluated and their inferences are expanded upon.
- An overview of how the final model was applied to the business problem is detailed.
- The closing two sections focus on the strengths of the experiment and the findings and weaknesses within.

5.1 Evaluation of results

The primary finding from this study is that text classification can be used in order to classify financial narrative documents. Early results have shown that the new narrative Solvency and Financial Condition Report can be classified with a 90% accuracy. This is a promising initial result and gives rise to numerous potential works within the regulatory space.

The decision tree struggled classifying SFCR pages. Achieving a classification accuracy of only 54.16%. The large number of features within the problem has caused issues and the decision tree has struggled with the complexity of the problem. It was the only classifier, which did not benefit from tf-idf representation. However as performance was poor with both tf and tf-idf representation this is a negligible finding. The decision tree is likely only suitable for smaller scale text problems where it’s main appeal is comprehensibility.

K-nearest neighbour performs well with an accuracy of 85.77%. It is a popular algorithm within text classification and has been extensively studied. A very simple lazy machine learning algorithm, it classifies documents as points in euclidean space.
Findings in literature suggested K-nearest neighbour performs well when tf-idf is representation is utilised and that the algorithm can struggle with a large number of features. (Trstenjak, Mikac, & Donko, 2013). This study found both points to be correct. The tf-idf representation outperformed the standard term frequency representation. In addition, the algorithm performed best with only the top 20% of tf-idf features selected. Performance deteriorated as the number of features used to train the model grows.

Known for producing good results even when faced with complex problems, the neural network performed well with an accuracy of 87.21%. Comparatively to the other algorithms tested, the neural network was computationally very expensive and required a large amount of physical memory. The model training time was over an hour. It also only reached its highest accuracy score when utilising all the features for training.

Section 2.5 within the literature review revealed that naïve bayes performed well when applied to text data. It is a popular algorithm due its simplicity and satisfactory performance. It is a simple probabilistic classifier based on applying Bayes theorem. It performed well when attempting to classify SFCR pages, achieving an accuracy of 87.62%. This was the second best performance within the study.

SVMs are one of the newer machine learning algorithms. They work by looking to maximise the margin between classes on hyperplane. Extensive research has shown the linear kernel SVMs perform particularly well with text classification. Text categorisation problems are generally linearly separable and SVMs deal well with problems containing many features. These findings held up with the application the linear SVM to this problem. The SVM achieved a very high accuracy of 93.62% through 10-fold cross validation with three repeats. There was potentially a small element of overfitting as the model accuracy dropped to 90.46% when applied to the test set. However, this is a strong initial result. It has proved machine learning and text classification can be applied to SFCRs successfully. There is definite potential in this
specific report and within the narrative regulatory reporting space in general for considerable time and financial savings through use of text classification.

5.2 Observations from final model

As has been detailed throughout the previous sections, the linear SVM obtained the best accuracy. The model was improved with cost parameter tuning. An increase in the cost parameter to 2.5 led to the model obtaining an accuracy score of 90.46% against the test set. In this section, the performance of the model against the test set will be analysed in detail. The performance of the model with respect to each individual guideline is reviewed and explanations provided for the models performance with respect to each legislative guideline. Table 5.1 displays the accuracy achieved by the model against the test set with respect to each individual guideline. The table elaborates on the confusion matrix illustrated in table 4.7.

<table>
<thead>
<tr>
<th>Guideline</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business &amp; Performance</td>
<td>89.42</td>
</tr>
<tr>
<td>Governance Structure</td>
<td>94.93</td>
</tr>
<tr>
<td>Risk Profile</td>
<td>94.49</td>
</tr>
<tr>
<td>Assets</td>
<td>84.91</td>
</tr>
<tr>
<td>Valuation of Technical Provisions</td>
<td>90.00</td>
</tr>
<tr>
<td>Liabilities other than Technical Provisions</td>
<td>75.00</td>
</tr>
<tr>
<td>Own Funds</td>
<td>84.00</td>
</tr>
<tr>
<td>Difference between Standard Formula &amp; Internal Model</td>
<td>0.00</td>
</tr>
<tr>
<td>Appendix</td>
<td>93.22</td>
</tr>
</tbody>
</table>

*Table 5.1 Individual Category Accuracy from Model Performance on Test Set*

Business & performance is well identified by the model with an accuracy of 89.42% against the test set. It correctly identified 169/189 pages relating to this guideline. The misclassified pages were largely identified as belonging to ‘Own Funds’ (7), ‘Risk
Profile’ (5) and ‘Governance Structure’ (3) by the model. These incorrect classifications are likely because within the Business & Performance category these topics have been referenced. As outlined in section 3.3 regarding corpus annotation the guideline regarding business & performance stipulates an executive summary of the business must be detailed. This executive summary regularly contains references to governance structure, risk profile and own funds. It could be argued that these pages of the SFCR are applicable to more than guideline. The misclassified pages have been labelled as business & performance however; the model is identifying them as belonging to the direct topic they reference. With an accuracy of 89.42% the model performs well at identifying pages of the text relating the business & performance guideline.

Governance structure achieves the highest individual accuracy of any category against the test set. The model correctly identifies pages relating to governance structure in 262/276 pages, 94.93%. The misclassified pages belonging to this guideline were predominantly classified as addressing ‘Business & Performance’ (9). As mentioned previously, there is overlap between some categories as the business and performance guideline contains an executive summary in which all aspects of the business may be addressed at length. The remainder of misclassified governance structure pages were determined to be either ‘Own Funds’ (2) or ‘Risk Profile’ (3). These errors are more difficult to interpret. These topics generally should not contain many similarities or overlap with governance structure. However, they account for a low amount of misclassifications. The model proves to be very good at identifying pages within SFCRs that are relating to governance structure.

The model also performed very well with respect to classifying risk profile, achieving an accuracy of 94.49%. It correctly classified 120/127 pages in the test set labelled as risk profile. The seven misclassified pages were split between ‘Business & Performance’ (2), ‘Assets’ (1), ‘Governance Structure’ (3) and ‘Valuation of Technical Provisions’ (1). The pages addressing governance structure often reference where, the ownership of various risks within the business lie and as such, this could be a reason for some of the risk profile pages being incorrectly classified as governance structure.
The model obtains impressive results when identifying pages within SFCRs that are relating to risk profile.

Pages addressing the assets guideline are correctly addressed with an accuracy of 84.91%. This represents 45/53 pages within the test set. The misclassified pages have been incorrectly identified as ‘Appendix’ (3), Business & Performance (1), ‘Own Funds’ (1) and ‘Valuation of Technical Provisions’ (3). These misclassifications are likely due to the referencing of the financial quantitative reporting templates. These templates are required within the appendix. However, they are also regularly present within pages addressing both assets and technical provisions. For instance, the pages addressing assets are generally accompanied by a subset of the balance sheet, which displays the balance sheet assets. An example of this is present below in figure 5.1. The entire balance sheet is present within the appendix and this is likely leading to confusion within the model. Despite this issue, the model performs to a decent level when identifying pages within the SFCR that are addressing the assets guideline.

Figure 5.1 Example of Page from Asset Category
Valuation of technical provisions is correctly classified with an accuracy of 90% in the test set. In total 54/60 pages relating the guideline are classified correctly. The misclassified pages are split between ‘Appendix’ (1), ‘Assets’ (3) and ‘Liabilities other than Technical Provisions’ (2). There is some overlap in reports where technical provisions and liabilities other than technical provisions are discussed together and as such, this may be causing difficulty for the SVM in mapping the hyperplane. The model obtains impressive results when identifying pages within SFCRs that are relating to the valuation of technical provisions.

The Company does not apply the transitional risk-free interest rate-term structure referred to Article 308c of Directive 2009/138/EC.

The Company does not apply the transitional deduction referred to in Article 308d of Directive 2009/138/EC.

Recoverables from reinsurance and special purpose vehicles
There were no recoverables or special purpose vehicles as at 31 December 2017 (2016 EUR Nil).

Material changes in relevant assumptions made in the calculation of technical provisions
There are no material changes in the relevant assumptions made in the calculation of the technical provisions compared to the previous reporting period.

Other liabilities

Other liabilities at 31 December 2017 were EUR 4,411k (2016: EUR 78k) and are composed of accrued expenses, amounts due to group undertakings and creditors arising out of reinsurance operations.

Alternative Methods for Valuation for other liabilities

The Company does not use any alternative methods for valuation.

Any Other Information

There are no other material matters in respect of the valuation of assets and liabilities.

Figure 5.2 SFCR other liabilities example
Liabilities other than technical provisions achieved an accuracy of only 75% against the test set with 24/32 pages classified correctly. The misclassified pages were labelled as ‘Assets’ (3), ‘Valuation of Technical Provisions’ (3) and ‘Business & Performance’ (1) by the model. For some companies liabilities other than technical provisions are very briefly addressed within a section regarding technical provisions. An example of this is available above in figure 5.2. This is likely to be causing the SVM difficulty in separating these classes on the hyperplane. When considering how occasionally companies do not provide a lot of text in relation to addressing this guideline, an accuracy of 75% may not be too bad.

Own funds is correctly classified by the model with an accuracy of 84% against the test set. Instances of own funds have been classified in to the widest variation of guidelines. The model incorrectly predicted pages labelled own funds as belonging to ‘Appendix’ (3), ‘Assets’ (1), ‘Business & Performance’ (2), ‘Difference between Standard Formula & Internal Model’ (1), Governance Structure’ (1), ‘Risk Profile’ (6) and ‘Valuation of Technical Provisions’ (2). The misclassification with appendix is again likely to be due to the referenced quantitative reporting template for basic own funds S.23.01.01.01. It will be present within the appendix and regularly referenced in sections addressing own funds. Explaining the frequent misclassification of own funds as risk profile is not as elementary. There is no clear-cut reason to explain this misclassification. Nevertheless, the model shows adequate performance with accuracy of 84% in classifying SFCR pages referencing the own funds guideline.

The guideline differences between standard formula and internal model is not dealt with well by the model. It does not correctly classify any of the ten pages in the test set. This is likely because of a combination of factors. Firstly, only fourteen instances of this guideline were in the training set. Secondly, only twelve out of two hundred odd companies use an internal model. As a result, the majority of companies just include one line within their SFCR addressing this guideline. Usually this is along the lines of “We use the standard formula to calculate the SCR, so there are no differences between the standard formula and internal model”. Actual entries from companies
using internal models are generally included within own funds information and refer to both technical provisions, assets and liabilities. As such, they are difficult to map clearly on the hyperplane. Clearly, model performance for guideline is not acceptable. There are a number of potential options to remedy this. The favoured option is to retrain the model with more training data from this category. Unfortunately, at the time of training only one iteration of these reports has been published. As such, access only existed to 12 companies SFCRs using an internal model. This equated to 24 pages of SFCRs addressing this guideline. More data is required to train the classifier for this guideline. This will be included in future work.

The appendix guideline is well classified by the model, it correctly classifies 275/295 pages in the test set, an accuracy of 93.22%. The model incorrectly predicted pages labelled appendix as belonging to ‘Business & Performance’ (11), ‘Governance Structure’ (2), ‘Own Funds’ (3), ‘Risk Profile’ (2) and ‘Valuation of Technical Provisions’ (3). The model incorrectly labels appendix as belonging to business and performance with a small degree regularity. Nevertheless, the model performs very well in achieving an accuracy of 93.22% from the classification of 295 pages.

The model performs well overall achieving an accuracy of 90.46% on the test set. Seven of the nine guidelines are well classified by the model. It encounters some difficulty identifying liabilities other than technical provisions, obtaining an accuracy of 75%. The model fails to identify differences between standard formula and internal model in any instance within the test set. Investigation of misclassifications have resulted in the following findings:

- The model classifies pages of SFCRs. There are instances where one page an SFCR addresses more than one guideline. Consideration could be given to classifying paragraphs instead however; in this case, some paragraphs do not address any guideline. A separate solution could be to introduce multi-labelling for pages.
The guideline difference between standard formula and internal model requires a lot more training data. At the time of training, this data was not yet available. However, the 2017 year-end returns will be available publicly in the near future. Instances from these SFCRs that address this guideline can be used as training data in order to improve the model performance with relation to this guideline.

The quantitative reporting templates that are included in the appendix are also presented in shortened versions elsewhere in the SFCRs. This is leading to difficulty for the SVM when looking to separate some classes on the hyperplane.

5.3 Model Application to Business Problem

The research question for this study evolved from a real life business problem. In 2016, the new Solvency II directive came into effect. This brought about a whole host of new reporting requirements including quantitative reporting templates and narrative reports such as the Own Risk and Solvency Assessment and Solvency and Financial Condition Report. The first iteration of SFCRs were published in early 2017 with a reporting date of year-end 2016. These reports were subject to a large project in order to ensure all requirements were met. This was a large manual and highly time-consuming project. For the 2017 returns, the data analytics team began examining ways of automating some of these checks. Findings showed full automation of some checks is possible whilst others will still require a degree of human intervention.

For example, full automation is possible for a consistency check between the values the companies have disclosed within their financial returns and the values of the quantitative reporting templates appended to the SFCRs. R is used in order to extract the tables from the appendices of the SFCRs. These tables are the quantitative reporting templates such as the balance sheet which companies are obliged to publish. A SQL query is run against the data warehouse to extract the relevant balance sheet entries from company’s submissions to the Central Bank. The balance sheet values from the SFCRs and from the returns for each company can be easily compared then to
ensure there are no discrepancies. This validates that no company has publically reported any value differently to what they reported to the Central Bank.

In order to assess SFCR completeness, human intervention is still required. The model developed throughout this study is used in order to provide early warning indicators for any SFCR that may not have addressed all required guidelines. The company’s supervisors must still review the SFCRs in order to ensure they meet qualitative standards. This solution will provide a quick and effective way of identifying SFCRs for 200 companies that require immediate review, as they may not have addressed all required guidelines. The final report includes a variety of checks. The first draft iteration of the report has been produced in Excel using R and macros. As it is the first version of this report, it is likely to go through a number of iterations and improvements before being finalised. An extract of this report that corresponds to use of the model developed in this study is visible in figure 5.3. This is for a real SFCR with a reporting date of year-end 2017. However, the company is not referenced by name within this paper.

![Figure 5.3 Business Report Example 1](image)

The report allows a user to select the company they are interested in reviewing. A summary section outlines the company’s name, code, impact rating and reporting date referenced. The main section then lists the nine guidelines, states how many pages of the SFCR address this particular guideline. The percentage of the report addressing
each guideline is displayed and the final column states if the guideline has been sufficiently addressed within the report. The guideline being addressed result depends on the percentage of the SFCR, which addresses each guideline. The limits vary for each guideline. With lower limits applying to the guidelines that generally have less text designated to them such as liabilities other than technical provisions.

A traffic light colouring scheme is applied to the guidelines with green indicating a pass, red a fail and orange a warning. A guideline is considered to have failed if it does not reach the minimum percentage of the SFCR that it is expected to. The orange warning indicator is applied to any instance where the difference between standard formula and internal model is zero. This warning is due to the fact for nearly all companies this guideline does not apply however; supervisors of companies who are approved to use internal models will need to investigate further if this is zero for their company. A graphic representation of the amount of pages representing each individual guideline is also included.

A colleague, not involved with the original corpus annotation in this project, labelled the SFCR used for example from figure 5.3. When the model predicted the guidelines for each page an agreement was found with 84/92 pages. This represents an accuracy score of 91.3% by the model. An improvement on the 90.46% obtained against the test set. Importantly it validates the models performance against new data with good performance.

<table>
<thead>
<tr>
<th>Company Code</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>C12345</td>
<td></td>
</tr>
<tr>
<td>Impact Rating</td>
<td>Medium Low</td>
</tr>
<tr>
<td>Reporting Date</td>
<td>31/12/17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Guideline</th>
<th>Pages</th>
<th>%</th>
<th>Guideline Addressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business &amp; Performance</td>
<td>16</td>
<td>17%</td>
<td>✓</td>
</tr>
<tr>
<td>Governance Structure</td>
<td>13</td>
<td>14%</td>
<td>✓</td>
</tr>
<tr>
<td>Risk Profile</td>
<td>14</td>
<td>15%</td>
<td>✓</td>
</tr>
<tr>
<td>Assets</td>
<td>2</td>
<td>2%</td>
<td>✓</td>
</tr>
<tr>
<td>Valuation of Technical Provisions</td>
<td>0</td>
<td>0%</td>
<td>✓</td>
</tr>
<tr>
<td>Liabilities other than Technical Provisions</td>
<td>6</td>
<td>7%</td>
<td>✓</td>
</tr>
<tr>
<td>Own Funds</td>
<td>5</td>
<td>5%</td>
<td>✓</td>
</tr>
<tr>
<td>Difference between Standard Formula &amp; Internal Model</td>
<td>0</td>
<td>0%</td>
<td>✓</td>
</tr>
<tr>
<td>Appendix</td>
<td>12</td>
<td>13%</td>
<td>✓</td>
</tr>
</tbody>
</table>

Figure 5.4 Business Report Example 2
Figure 5.4 shows a second example for a smaller company. The model has not classified any pages as addressing the guideline concerning valuation of technical provisions. This flag indicates a follow up check must be carried out in a timely manner on this company’s SFCR in order to determine if they have or have not addressed this guideline in a satisfactory manner. The same colleague labelled this SFCR. The model achieved an accuracy score of 91.18%. This is again slightly above the models score against the test set.

Figure 5.5 displays a third example for a low impact company. Here it is noticeable that the model has classified a page as addressing the governance structure guideline. However amounting to only 1% of the total report pages it falls below predefined limits and is deemed to have failed to address this guideline sufficiently. The guideline liabilities other than technical provisions has not been classified for any page within the report. As such this document is flagged for review by the relevant supervisor to determine if these guidelines are addressed to a satisfactory manner or not. If they have not been addressed to an acceptable level, a request for correction and resubmission is a likely outcome. The same colleague again labelled this SFCR. The model has performed worse for this report achieving an accuracy of 84.21%. It correctly classified 32/38 pages. A likely struggle for the model here is that the smaller sized companies do not structure or format their SFCRs as coherently. Pages often consist of text addressing more than one guideline.

![Figure 5.5 Business Report Example 3](image-url)
The text classification model developed throughout this study has proven to work when applied to a real world business problem. It has enabled a quick early warning check to be carried out on the SFCRs of almost two hundred companies. This solution has saved a considerable amount of time that would have been lost to tedious manual work. There is also a substantial cost saving associated with this.

5.4 Strengths of the Results

One strength of the results is that they mirror other findings. This study has outlined another successful application of text classification. Highlighting its potential value within narrative regulatory reporting. Strong performance of the linear SVM was demonstrated in classifying online privacy policy categories (Cosante, Sun, Petkovic, Hartog, 2012). That study produced a classifier with an accuracy of 92%. This study has demonstrated the linear SVM is capable of achieving similar results when applied to financial narrative reports.

A second strength of the results is that it successfully provided a solution to a real life business problem. The implementation of the model constructed throughout the course of this study resulted in considerable time and financial savings. It has enabled the quick assessment of SFCRs for almost two hundred companies. Early warning indicators now dictate which SFCRs require further checks and can direct the prioritisation of these reports. The model achieved an accuracy of 90.46% against the test set. When the model was used to classify new individual SFCRs with a reporting date of year-end 2017, it performed well with accuracies in the upper 80s to low 90s.

5.5 Limitations of the Results

The first limitation of the results is that the model cannot classify one of the guidelines to an acceptable standard. Companies using an internal model must address the differences between the standard formula and their internal model. The linear SVM correctly classified 0% of these instances in the test set. The issue is likely to be the
lack of training instances for this guideline. Only a handful of companies have approval to use an internal model. As a result, this guideline is not relevant for the majority of companies. However, it is a limitation within the final model.

A second limitation is that the model only identifies if the guidelines are present or absent within the report. It cannot assess the quality of the content. For instance the report may address own funds in a sufficient number of pages but not actually disclose the required information. The solution can assess completeness of SFCRs in a manner that identifies if references have been made or not to the required guidelines. However, it cannot give any indication of compliance within the content, for guidelines that are present. It cannot detect any uncompliant reports that have referred to each guideline. As such, the solution acts as an early warning indicator and directs which reports require prioritisation.
6. CONCLUSION

The final chapter contains a review of the structure and findings of the project. Contributions to the body of knowledge are discussed and recommendations for further research are addressed.

![Research Process Diagram]

*Figure 6.1 Research Process*

6.1 Research Overview

This research was undertaken to determine if machine learning and text classification could be used in order to assess completeness of SFCRs. The idea for the study came about following the need to solve a real life business problem. In 2017, insurance companies published the first ever set of solvency and financial condition reports. For the regulator, assessing these for compliance was a tedious, manual and time-consuming process. The research question was to what extent machine learning could be used to assess the completeness of solvency and financial condition reports.
Literature provided examples of similar studies in slightly different fields. In Automated Text Mining for Requirements Analysis of Policy Documents, the authors use topic models in order to indicate whether a document contains software requirements expressed as privacy protections or vulnerabilities (Massey, Eisenstein, Antón, Swire, 2008). The authors of A Machine Learning Solution to Assess Privacy Policy Completeness presented a solution that automatically evaluates the completeness of a website’s privacy policy for users. Through use of machine learning and text classification techniques, they prove the feasibility of their approach (Cosante, Sun, Petkovic, Hartog, 2012).

The success of Cosante et al. inspired this study. The research aimed to explore if text classification could be as successful when applied to narrative financial regulatory returns. The SFCR classification problem was chosen, as it was a real life issue that a solution was being sought for. It could be worked on within the time constraints of the project. In addition to this, the data is publically available on the central bank of Ireland’s website.

6.2 Problem Definition

In the last few years, central banks are starting to explore how the vast amount of narrative text data available can be manipulated in order to help achieve policy objectives (Bholat, Hansen, Santos, & Schonhardt-Bailey, 2015). This study aimed to classify pages of SFCRs with the relevant legislative guideline the text within is addressing. Articles 290-303 of the Solvency II Delegated Regulation specify the content that must be included within the SFCR. For the purpose of this study, the legislative guidelines were condensed to nine categories that must be addressed within an SFCR:

1. Business & Performance
2. Governance Structure
3. Risk Profile
4. Assets
7. Own Funds
8. Difference between Standard Formula and Internal Model
9. Appendix

The research focused on exploring the application of five different classification algorithms with the goal of classifying each page of an SFCR with the guideline most relevant to the text within.

6.3 Design/Experimentation, Evaluation & Results

The primary research question under investigation within this study was to what extent can machine learning be used to assess the completeness of solvency and financial condition reports? The study has shown that machine learning can be used effectively to assess completeness of SFCRs. The linear SVM model produced can classify pages of SFCRs with the guideline considered most relevant to that text with an accuracy of 90%. This lead to a business solution where overviews of individual SFCRs can highlight if each guideline has been met or not. The solution acts as an early warning indicator. It highlights reports that it believes require further examination. It cannot assess the report for compliance; this requires human intervention to review the qualitative element within the text. So machine learning can be used as a screening tool when assessing completeness of SFCRs. It cannot however assess if the content is up to the required qualitative standard and the company has disclosed all information they are required to.

One of the secondary research questions asked could the model produced accurately classify the guideline difference between standard formula and internal models. The model is very poor at classifying with respect to this guideline. It does not correctly classify any of the ten pages in the test set. Only twelve companies have approval to
use an internal model as a result there were only twenty-four pages relating to this guideline available for this study. Companies not using an internal model usually included a section along the lines of “We use the standard formula to calculate the SCR, so there are no differences between the standard formula and internal model”. Actual entries from companies using internal models are generally included within own funds information and refer to both technical provisions, assets and liabilities. As such, they are difficult to map clearly on the hyperplane. The secondary research question saw the concerns before the experiment crystallised. However, with this year’s SFCRs available material from companies using an internal model will approximately double. Retraining the model with additional training instances should help improve model performance with relation to this guideline.

The final secondary research question asked could the model achieve the same accuracy on new SFCRs. Any model built will be using training data from the year-end 2016 returns. The performance of such model will be validated against 2017 returns. The model will be trained using a dictionary of words from the 2016 returns, it is unlikely this dictionary will vary a great deal year on year and much of the language will be finance/insurance specific. However, in answering the primary research question it will be important to establish if the model will need to be re-trained each year or if the initial model can be used going forward. This question was answered in section 6.3 where the model built from 2016 year-end reports accurately classified new 2017 year-end reports with high accuracy. This validated the original models suitability for the classifying new SFCRs.

6.4 Contributions and Impact

This research has shown text classification can be a useful tool when applied to narrative financial regulatory returns. It has enabled a quick and effective review of completeness of almost two hundred companies SFCRs. A review of classification models strengthened views in literature that the linear support vector machine is regularly the best option for text classification problems. The model achieved an accuracy of 90.46% against the test set. When the model was used within the business...
solution, it classified pages for individual SFCRs. The model performed well generally achieving accuracy’s of around 90% when compared with human classification of those pages. As such, it has proved a successful business solution. The solution cannot identify uncompliant content within SFCRs. However, it can quickly identify SFCRs that have not addressed guidelines sufficiently. It flags these reports and prioritises them for further checks. The automated solution can assess the completeness of an SFCR with reference to the nine required guidelines. It has resulted in substantial time and financial savings. The solution is in an early stage of its life cycle and will be iteratively improved. It is contributing to compliance and data quality checks.

6.5 Future Work & Recommendations

The future work from this study lends from the limitations. The model struggled classifying the guideline differences between standard formula and internal model. This was due to a limitation of training data for this guideline. Only the first year of SFCRs was available at the time of training. However now the second year of reports have been submitted, there are more instances belonging to this guideline available. Retraining the model with an increased number of pages belonging to this guideline in the training data would be likely to increase model performance.

An exploration of ensemble methods could increase model performance. A 2007 paper revealed that if you are in search of the best possible classification accuracy, use of ensemble methods should be strongly considered (Kotsiantis, 2007). The models in this paper were built using the caret package in R. At the time of training, unfortunately caret did not support multi-class ensemble classification.

The study revealed many instances where an SFCR page addressed more than one guideline. Future work could look at multi-labelling. In such work, one page of an SFCR could be classified as having addressed more than one guideline. Another interesting approach could be to interesting to label each SFCR page as compliant or
not and build a model to see if a likelihood/probability of compliance could be identified.
BIBLIOGRAPHY


Azam, N., Jahiruddin, Abulaish, M., & Halder, NAH. (2015). Twitter Data Mining for Events Classification and Analysis. Second International Conference on Soft Computing and Machine Intelligence (ISCMI), 79-83. DOI: 10.1109/ISCMI.2015.33


Kim, S., Han, K., Rim, H., & Myaeng, S. (2006). Some Effective Techniques for Naive Bayes Text Classification. IEEE Transactions on Knowledge and Data Engineering, 18(11), 1457-1466. DOI: 10.1109/TKDE.2006.180


Zhang, M., & Zhou, Z. (2006). Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization. IEEE Transactions on Knowledge and Data Engineering, 18(10), 1338-1351. DOI: 10.1109/TKDE.2006.162