Norwegian School of Economics
Bergen, Spring 2022

Do Fraudulent Companies Employ Different Linguistic Features in Their Annual Reports?

An Empirical Study Using Logistic Regression and Random Forest Methodologies

Gözde GÖKTÜRK

Supervisor: Christian Langerfeld

Master thesis, Economics and Business Administration

Major: Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

This thesis was written as a part of the Master of Science in Economics and Business Administration at NHH. Please note that neither the institution nor the examiners are responsible - through the approval of this thesis - for the theories and methods used, or results and conclusions drawn in this work.

Acknowledgements

This thesis is the culmination of my master's degree programme at the Norwegian School of Economics, with a focus on Business Analytics. The thesis has been instructive, challenging, and fascinating to write. I would like to thank Christian Langerfeld, my supervisor, for his unwavering support, patience, enthusiasm, and extensive knowledge. His constant advice and recommendations assisted me in the development of my thesis. I would like to express my gratitude to every member of the faculty for their invaluable guidance throughout my time at the Norwegian School of Economics. It has been a fantastic academic experience, and I have acquired valuable knowledge in the business analytics field. I am indebted to my family for their unconditional support, love, and encouragement to pursue my goals.

Thank you!

Gözde Göktürk

Abstract

The use of textual analysis to uncover fraudulent actions in 10-K filings is widespread. The previous studies have looked at the Management Disclosure and Analysis (MD&A) section of annual reports to predict illicit behaviour by analysing the tone of executives, with the majority of those studies dating back 10 years or more. The primary goal of this research is to find patterns in linguistic features of entire annual reports of convicted public businesses, which were found using the Corporate Prosecution Registry database, and compare them to non-fraudulent equivalents in the same industry. The algorithms of logistic regression and random forest are implemented to discover important factors and make accurate predictions. The accuracy rate, ROC-AUC value, and 10-fold cross-validation tools are performed to validate the success of each method. The results of the logistic regression revealed that corrupt organisations utilise a more negative, uncertain, and litigious tone. Furthermore, these businesses employ more words with a high lexical diversity and minimal complexity. Based on the Random Forest machine learning technique, the *litigious* variable is the most important variable in the prediction of untruthful corporations. Moreover, each of the validation methods demonstrates that the Random Forest methodology outperforms logistic regression.

Keywords – sentiment analysis, corporate crime, 10-K filings, logistic regression, Random Forest

Table of Contents

Acknowledgements i
Abstract ii
List of Figures iv
List of Tables v
1. Introduction 1
2. Literature Review 2
2.1 Corporate Crime 2
2.2 Previous Literature on Sentiment Analysis on Fraud Detection 4
2.3 Hypotheses 6
3. Data 8
3.1 Corporate Prosecution Registry 8
3.2 United States Securities and Exchange Commission 10-K Fillings 9
3.3 Data Extraction 11
3.4 Textual Data Analysis 11
4. Methodology 12
4.1 Sentiment Analysis 12
4.2 Parameters After Sentiment Analysis 13
4.3 Logistic Regression Analysis 15
4.4 Random Forest 16
4.5 Validation 18
5. Results 22
6. Discussions 27
6.1 Regression Outcome and Hypotheses 28
6.2 Advantages and Disadvantages of Both Analysis 30
6.3 Limitations 32
7. Conclusion 33
References 34
Appendix 38
A1 Types of Crime 38
A2 SIC Titles of Fraudulent Companies 39
A3 Correlation Table 41
A4 Regression Outcome and Odds Ratio 42
A5 10-Fold Cross-Validation: Average Accuracy and ROC-AUC Values 44

List of Figures

Figure 3.1: Number of occurrences of each category of offence 9
Figure 4.1: Classification phases of Random Forest classifier 17
Figure 4.2: An example of ROC-AUC Curve 20
Figure 5.1: Logistic regression analysis outcome 23
Figure 5.2: Odds Ratio of the Significant Variables 24
Figure 5.3: Variable importance in predicting the crime 25
Figure 5.4: ROC-AUC curve of logistic regression model 26
Figure 5.5: ROC-AUC curve of Random Forest 26
Figure 5.6: ROC-AUC values for each fold 27
Figure A4.1: The regression outcome of overall model 43
Figure A4.2: The odds ratio of significant industry classifications 43

List of Tables

Table 3.1: The name and description of crime types 4
Table 3.2: Number of corrupted companies with their industry titles 10
Table 5.1: Variables used in the analysis and their description 14
Table 5.2: Confusion Matrix for actual and predicted crime, and no crime 19
Table 5.3: Comparison of confusion matrix and accuracy rate for logistic regression and Random Forest 26
Table A1.1: The types and the relevant descriptions 39
Table A2.1: Standard industry classification titles of convicted corporations 40
Table A3.1: Correlation table of independent variables excluding sic 41
Table A5.1: The mean and standard error of both models after 10-Fold Cross-Validation..... 44

1. Introduction

Economic or financial crime is a major offence that has a significant influence on the financial wealth of a country or the entire world. The main purpose of the offenders is to generate financial or professional advantages by using various sources or methods. Such illicit activities result in unemployment, stagnant labour market conditions, unfair competitiveness, reputational damage, and even mental or physical impairment. There are three types of perpetrators in economic crime: external perpetrators, internal perpetrators, and complicity between external and internal offenders (Rivera et al., 2022). Customers, hackers, and suppliers are examples of external culprits who do not work or have no connection to the company. Employees that work for corporations or white-collar contractors are the key actors of internal perpetrators. The secret agreement between exterior criminals and inside offenders is referred to as complicity between external and internal perpetrators. In a recent survey conducted by Rivera et al. (2022), which gathered data from 1296 organisations in 53 different countries, external perpetrators are committed 43% of all economic crimes, while internal criminals are responsible for 31% and collaboration between internal and external offenders accounts for 26%. Nonetheless, illicit activities that involve internal offenders have a much greater impact on financial markets and wealth.

Textual data analysis is a method of extracting data from textual sources, translating it into information, and making it valuable for various types of decision-making. According to Statista Research Department (2021), unstructured text data accounts for 68% of all data types and is the most commonly applied data type in machine learning, artificial intelligence, and data analysis. In today's big data-centred business environment, text mining is a crucial source for spotting irregularities in financial transactions and detecting crimes within organisations. According to Rivera et al. (2022), 46% of respondents experienced fraud within the last two years. The fraudulent activity on financial reports causes not only financial losses for the shareholders but also has a substantial impact on the capital market. Fraudulent actions cost 18% of higher-income companies more than 50 million dollars, while 22% of lower-income businesses lost more than one million dollars (Rivera et al., 2022).

The Securities and Exchange Commission (SEC) requires the majority of publicly traded corporations in the United States to file 10-K reports to notify officials about risk factors,

legal procedures, management perspectives and many other aspects that affect organisations throughout the course of a year (U.S. Securities and Exchange Commission, 2021). Making false or misleading claims in these files is prohibited by law and regulation. The SEC issues comments if statements appear to be in violation of registration guidelines or missing information. Each publicly traded corporation's disclosures must be audited by the SEC at least once every three years. Therefore, SEC is in charge of detecting those crimes proactively and observing irregularities or unlawful acts, in order to prevent such illegal activities.

In this thesis, the 10-K filings and Corporate Prosecution Registry data will be analysed to understand how the written language that is used by the company identifies whether the companies are more likely to commit a crime than non-fraudulent companies. There are seven chapters in this paper. The introduction section provides an overview of economic crime, text analysis and fraud. Chapter 2 includes the description of corporate crime, as well as an understanding of existing studies and hypotheses that are evaluated in this paper. The extraction of data and pre-processing techniques are covered in chapter 3. In chapter 4, the methods employed will be presented in detail and the results of the analysis will be reported in chapter 5. Chapter 6 provides a discussion of the outcome and presents the shortcomings before the remarks of the thesis are concluded in chapter 7.

2. Literature Review

This section will offer pertinent literature on corporate crime as well as previously used analyses to detect fraudulent actions. To begin, a basic explanation of corporate crime will be provided, as well as the types of fraudulent actions that these major corporations engage in. Following that, the methods for detecting patterns of illegal behaviour will be described. Finally, the hypotheses that will be examined in this study will be presented.

2.1 Corporate Crime

Corporate crime, often known as white-collar crime, refers to illegal activities carried out by respectable corporate professionals who use their authority to break the law (Shover & Simpson, 2003). Since corporate crime contributes to the global financial system's imbalance, it's critical to identify and uncover any malicious activity to create equal opportunities for all

the organisations (Said et al., 2014). According to Zahra et al. (2007), the causes for high-profile professionals engaging in criminal wrongdoing, are related to societal, industry, and organisational pressures. The term "societal pressure" alludes to the social aspect of crime, implying that people can use deviant means to fulfil their desires. As for the industry-level pressures, fraud is enticing to business experts due to a number of challenging industrial level conditions such as industry concentration, payback periods and financial returns, environmental antagonism, etc.

Large firms are owned by millions of people around the world therefore, stockholders hire executives to transfer decision-making authority (Zahra et al., 2007). These executives acknowledge that their income, career, and employment are all dependent on the company's short-term achievements. If the firm's historical performance jeopardises the professionals' job security, crime may be substituted for a rigorous work ethic (Alexander & Cohen, 1996). As a result, senior managers may engage in deceptive practices in order to increase the stock value of the company and trigger the growth potential of the organisation, which would also reflect on their wealth (Zahra et al., 2007). This outlines the pressures on managers at the organisational level. The ubiquitous and far-reaching nature of corporate crime influences the shareholders, employees, and society as a whole. Managers' reputations can be harmed through fraud, which can lead to their dismissal or even incarceration. In addition, fines may be imposed on the corporation as a result of employee behaviour.

Disclosures of white-collar crime around the world have ignited a heated debate over the roles of auditing companies, boards of directors, and government agencies in detecting and preventing such crimes (Zahra et al., 2007). Stakeholders began to seek transparent and accountable business operations, as well as prioritise high standards of corporate governance and the development of ethical business partnerships (Said et al., 2014). As a result, corporations are required to exercise ethical behaviour outside the special code of practice of their respective firms and industries. However, according to Paliwal (2006), making ethical decisions on behalf of a corporation is a complex process that cannot be limited to a set of rules that all levels of an organisation can adopt. Ethical evaluation is founded on the workforce's collective decisions, which are subsequently followed by specific codes of conduct, rules, and acts. In order to reach organisational goals, the management team must plan, organise, lead, and control their decision-making process by employing ethical guidelines.

Table 3.1 shows the most prevalent criminal activities that a corporation has committed and has been reported to the Corporate Prosecution Registry. The remaining crime types that are included in the registry data can be found in Appendix A1.

Type	Description
FCPA (Foreign Corrupt Practices Act)	Prevention of U.S corporations from bribing foreign authorities in order to advance their economic interests by checking internal financial recordings (Biegelman & Biegelman D. R., 2010).
Environmental	Pollution of the air, water, and sea, as well as the unauthorized disposal of hazardous waste (Croall, 2001).
Fraud	Unethical or deceptive representation through a statement or conduct with the intent to profit financially or personally (Croall, 2001).
Antitrust	Anti-competitive agreements or abusive behaviour by enterprises with a dominant position in a market (European Commission, 2022).
Pharmaceutical	Anti-kickback and other related allegations involving pharmaceutical sales and branding, as well as charges conducted under The Federal Food, Drug, and Cosmetic Act (FDCA) (Corporate Prosecution Registry, 2022).
Import / Export	Breach of customs regulations, as well as sanctions violations in foreign commerce and financial operations (Corporate Prosecution Registry, 2022).
Bank Secrecy Act	Obligation for financial institutions to collaborate with the federal government to track major money transactions (Lloyd, 2020).

Table 3.1: The name and description of crime types

A company's illegal acts could have a range of consequences, including harming the environment and wildlife, endangering people's health, creating an unbalanced competitive climate, and providing financial opportunities to criminals. Thus, detecting and preventing any form of business crime is indispensable.

2.2 Previous Literature on Sentiment Analysis on Fraud Detection

For analysing the information content of corporate statements, textual analysis has become increasingly widespread. Corporate conference calls, profit statements, media articles, and corporate disclosure have all been evaluated by employing linguistic features (Purda & Skillicorn, 2015). Moreover, Bach et al. (2019) reviewed 123 papers in the text mining literature in order to identify current trends in the field. The findings revealed that a large amount of unstructured data extracted from external sources such as websites, social media,

and news is used to investigate stock price prediction, financial fraud detection, and market forecast. However, many studies have evaluated the language used in annual reports by focusing on the Management Discussion and Analysis (MD&A) section of 10-K filings, which reveals the company's perspective on the preceding fiscal year's business outcomes as well as actions taken in response to industry challenges and threats (U.S. Securities and Exchange Commission, 2021). The language characteristics of annual reports are employed for a variety of purposes, including projecting anomalous stock prices (Hajek, 2017), future financial distress (Hájek & Olej, 2013), earnings drift and accruals (Feldman et al., 2010), and detecting financial statement fraud (Craja et al., 2020; Dong et al., 2016; Goel & Uzuner, 2016; Hajek & Henriques, 2017; Humpherys et al., 2011; Loughran & McDonald, 2011; Skillicorn & Purda, 2012).

In the literature, two main methodologies are proposed for analysing the characteristics of the language used in annual reports: dictionary-based and machine learning. The dictionary-based approach is based on terms linked with a specific emotion, such as optimism, pessimism, dishonesty, or uncertainty, as determined by a financial expert in order to comprehend the document's emotions and tone (Craja et al., 2020; Hajek, 2017). Using a list of negative, uncertain, and litigious financial vocabulary, along with strong and weak modal words, Loughran & McDonald (2011) proposed a financial lexicon and used it to investigate annual reports to predict 10-K filing returns, trading volume, stock return volatility, material weakness, fraud, and unexpected earnings. The dictionary has been widely utilised especially in fraud-detection studies since it was designed for screening annual reports. In addition to the LM glossary, Goel & Uzuner (2016) used two additional lexicons: Linguistic Inquiry and Word Count Categories, which is the categorisation of distinct semantic groups to delve into the emotional, cognitive, and structural aspects of the text to detect the differences in sentiment polarity between prosecuted and non-prosecuted firms, and Multi-Perspective Question Answering Subjectivity Lexicon, which is the measurement subjective clues such as the presence of adjectival and adverbial modifiers. In order to apply a dictionary-based approach, unstructured data from yearly reports can be converted into numerical variables using standard pre-processing and statistical procedures, making classification algorithms easier to apply afterwards (Hajek & Henriques, 2017).

The second strategy relies on machine learning algorithms such as Naive Bayes (Goel et al., 2010; Humpherys et al., 2011) and supports vector machines (SVMs) (Goel et al., 2010;

Purda & Skillicorn, 2015) to produce lexical items and weights autonomously textual categorisation between fraudulent and non-fraudulent texts (Craja et al., 2020; Hajek, 2017). In comparison to prepared lists of terms and cues, Li (2010) claims that this system has several advantages, including the fact that it does not require any adaption to the business setting. Hajek & Henriques (2017) investigated if a better financial fraud detection system could be constructed by merging specific features generated from financial data and managerial remarks in business annual reports using a variety of machine learning approaches including logistic regression, and Random Forest. The results revealed that Random Forest predicts considerably better than logistic regression.

In this paper, the logistic regression and Random Forest machine learning techniques are used to detect corrupt companies based on their 10-K filings. The outcome of the two models will be compared based on three validation techniques, which are accuracy rate, the area under the receiver operating characteristics (ROC-AUC) curve and k-fold cross-validation.

2.3 Hypotheses

The linguistic characteristics of MD&A sections of annual filings have been studied in order to uncover untruthful companies by comparing them to their counterparts. This section will construct the hypotheses that will be analysed in this study based on the findings from the literature.

Loughran & McDonald (2011) used a textual study of the MD&A section of 10-K forms to discerning the language traits used by organisations engaged in criminal operations, focusing on 585 fraudulent firms reported between 1994 and 2004. Loughran & McDonald (2011) discovered that negative, ambiguous, and litigious word lists are all strongly connected to fraud lawsuits after using logistic regression. Using natural language processing technologies, Goel et al. (2010) analyse the oral content and presentation technique of the qualitative component of the annual reports of 126 prosecuted organisations and 622 control groups and inspect the changes in lexical features. According to the findings of the study, untruthful businesses include more ambiguous language in their reports. Furthermore, the study conducted by Humpherys et al. (2011) using MD&A sections of 101 corrupt and 101 truthful companies revealed that executives may promote a stronger picture of the company by

concealing negative developments in order to fulfil Wall Street's expectations and secure future opportunities in a far more positive manner (Dong et al., 2016). Jaeschke et al. (2018) studied a sample of organisations that had been prosecuted for FCPA breaches and discovered that untruthful companies use fewer litigious vocabulary. They described the result as demonstrating that convicted companies use litigious terms to mask existing court proceedings. The following hypotheses are evaluated based on the findings:

> *H_1: Companies that employ fewer negative words are more likely to engage in deceptive practices.*
> *H_2: Companies that employ more uncertain words are more likely to engage in deceptive practices.*
> *H_3: Companies that employ less litigious words are more likely to engage in deceptive practices.*

The same study conducted by Humpherys et al. (2011) also discovered that fraudulent firms create significant amounts of irrelevant content, which increases the number of words, verbs, adjectives, adverbs, and sentences while diminishing the linguistic diversity used in MD&As. Humpherys et al. (2011) concluded that managers who commit fraud are expected to persuade readers of the truth of their remarks while diverting their attention away from potentially detrimental material. Therefore, the hypotheses below will be tested:

> *H_4: Companies that employ more words with fewer unique words are more likely to engage in deceptive practices.*

In addition, Humpherys et al. (2011) found that untruthful companies employ more complex words to render damaging information more challenging to obtain in order to minimise or prolong adverse market reactions. Based on the study, the following hypothesis is assessed:

> *H_5: Companies that employ more complex words are more likely to engage in deceptive practices.*

The burgeoning literature on fraud detection will be expanded by testing these hypotheses. The majority of the literature only investigates the MD&A section of annual reports. Every section of the annual report will be included and tested based on the findings in the literature

to check whether there are any differences in the outcome. Furthermore, because these studies were published 10 years or more ago, it will be determined whether corrupt companies' language has changed over time. Additionally, logistic regression and Random Forest will be performed to obtain a deeper knowledge of the situation, produce more thorough evidence, improve the robustness of the results, and establish a reliable prediction model in order to detect fraudulent activities accurately.

3. Data

This section focuses on how the data was chosen as well as the procedures employed during the data pre-processing stage. Data about corporations convicted of criminal actions in the United States is collated in order to obtain the annual report of these fraudulent firms. As a result, prosecution registry data is utilised to identify corrupt businesses, and the names and/or tickers of those corporations are then used to retrieve their 10-K SEC filings.

3.1 Corporate Prosecution Registry

The data from the Corporate Prosecution Registry, a joint effort of the University of Virginia School of Law and Duke University School of Law that provides the comprehensive and latest information on federal organisational prosecutions in the United States (Corporate Prosecution Registry, 2022), is used to identify companies that have been involved with illegal activities. The registry gathers data on federal corporate prosecutions, including the names of the firms, the date of the verdict, the penalties, the type of offence, the length of probation, and other variables.

The registry had 4,390 criminal offences reported from 1992 to 2021 at the time the data was acquired. However, the data covers cases with the following disposition types: acquittal, declination, dismissal, and no prosecution. Therefore, the observations with these dispositions have been excluded. Furthermore, organisations that do not fall under the criteria of a publicly traded company in the United States have been removed from the data. Following the removal of observations without a ticker and those with an invalid ticker, the remaining data contains 232 prosecuted cases with 189 unique tickers.

There are 21 different categories of violations in the data set. The most common offences committed by the firms, according to Figure 3.1, are violations of the Foreign Corrupt Practices Act (FCPA), unlawful environmental actions, and general fraudulent activities. Other encompasses all remaining offence types, such as tax fraud, money laundering, workplace safety, and so on.

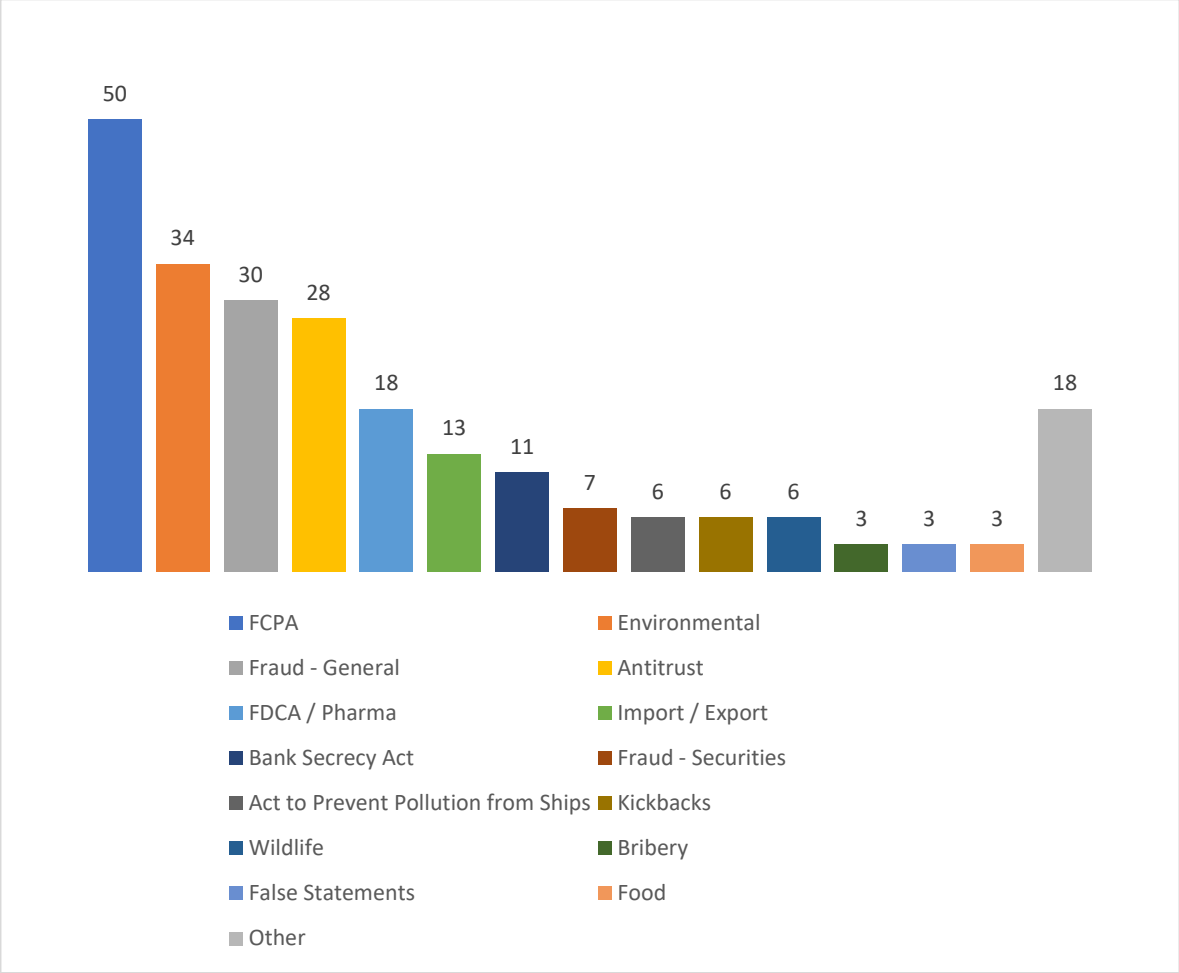

Figure 3.1: Number of occurrences of each category of offence

3.2 United States Securities and Exchange Commission 10-K Fillings

Once the companies implicated in illegal conduct are disclosed, the ticker in the registry data is used to manually retrieve the central index keys (CIK), a registration number granted by the SEC, in order to obtain the annual report fillings of the said corporations. Since the fraudulent action period is unknown, annual reports starting from five years prior to one year prior to the verdict year are taken into account, as a default. For example, if a firm committed a crime in

2005, the company's yearly reports from 2000 to 2004 are retrieved, totalling up to five annual reports. The SEC platform indicated that annual reports for 106 of the 189 corporations are available in its system, therefore, those companies' filings were downloaded. Despite the fact that several companies' fillings were missing, a total of 504 yearly reports of these fraudulent firms were acquired.

66 distinct industry classifications were included in the data. Table 3.2 shows the top ten standard industry classifications that are frequently observed in data. Pharmaceutical preparations companies are responsible for the majority of illegal activities, followed by drilling oil and gas wells, and orthopaedic, prosthetic, and surgical appliances and supplies corporations. The remaining data can be found in Appendix A2.

SIC	Number of Companies	Industry Title
2834	11	PHARMACEUTICAL PREPARATIONS
1381	5	DRILLING OIL & GAS WELLS
3842	5	ORTHOPEDIC, PROSTHETIC & SURGICAL APPLIANCES & SUPPLIES
6021	5	NATIONAL COMMERCIAL BANKS
1311	4	CRUDE PETROLEUM & NATURAL GAS
3714	3	MOTOR VEHICLE PARTS & ACCESSORIES
3841	3	SURGICAL & MEDICAL INSTRUMENTS & APPARATUS
4911	3	ELECTRIC SERVICES
7389	3	SERVICES-BUSINESS SERVICES, NEC
1389	2	OIL & GAS FIELD SERVICES, NEC

Table 3.2: Number of corrupted companies with their industry titles

Following the collection of the prosecuted organisations' filings, the Standard Industry Classification (SIC) numbers are revealed through those reports, allowing non-fraudulent competitors to be identified. On the SEC platform, 158 non-prosecuted competitors are manually detected, by implementing those SIC codes. The annual reports of non-convicted companies are collected depending on the exact years obtained for fraudulent organisations for the accuracy of the analysis. Therefore, 742 forms for firms that had not committed a crime were compiled. As a result, the total number of observations in the data set reached 1246.

3.3 Data Extraction

Extraction allows a variety of data types to be merged and mined for business development. The extraction process detects and discovers relevant unstructured data before processing or transforming it into a structured data format (Flesca et al., 2004). A wrapper is a set of extraction criteria for extracting information from a website. The creation of sophisticated languages for expressing extraction rules and the capacity to generate these rules with the least amount of human input are the two key difficulties facing information extraction systems.

A programming language is a specialised platform that allows users to extract the necessary data for a study by utilising various packages or codes. The R programming language was utilised in this thesis. R is an open-source programming language that anybody can analyse, edit, and improve (Pathak, 2014). R was built by statisticians; thus, statistical analysis is the focus of many of its key language parts. In comparison to other programming languages, the quantity of code required is fairly small. The usage of packages in R makes data manipulation easier. As a result, R will leverage the study by creating extremely detailed, in-depth analyses.

The *edgar* package in R was used in this thesis to extract all of the aforementioned annual reports from 264 fraudulent and non-fraudulent companies.

3.4 Textual Data Analysis

Text documents are unstructured, making it challenging to swiftly analyse the information contained inside them (Bach et al., 2019). The linguistic information gleaned from annual reports is inherently unstructured. As a result, these inputs must be transformed into structured data before using preferred data mining techniques, such as classification and supervised machine learning in this instance. Thereby, the computerised approach to extracting relevant structured data from unstructured text is known as text mining (Bach et al., 2019; Gupta & Gill, 2012).

According to Iezzi & Celardo (2020), there are six steps in textual analysis. The procedure begins with a precise, well-focused, and adaptable definition of the research problem. The

corpus is created in the second step, which includes the collection of a set of texts in order to clarify several elements of the objectives and sample collection. In the third phase, pre-processing stage, the data is stripped of stop words, multiple whitespaces, punctuations, and tags. In this thesis, the 10-K fillings of both the fraudulent and control groups were retrieved in .txt format. The .html tags, punctuation, and digits must be deleted and converted to UTF-8 encoding to create a plain text format. Later, in the same process, each filling is tokenised, which entails breaking down the complete text document into smaller bits such as individual words in order to define the characteristics of text language

In the fourth phase, the tokenised words are converted into a document-term matrix (DTM) (Iezzi & Celardo, 2020). The rows and columns in a DTM correspond to documents and terms, respectively. The following phase involves goal-based methods, approaches, and models which incorporate the logistic regression and Random Forest, in this research. The insights from the model outcome are then used to build a strategy in the final step.

4. Methodology

This chapter describes the approach and techniques utilised to attain the research's objectives. The various forms of linguistic features were employed to decipher the languages used in the annual reports and see whether these were linked to any illicit practices. Sentiment analysis was undertaken to probe into this, and the prosecuted companies were compared to a set of control groups in their respective industrial divisions. The analysis methods employed in this thesis will be provided once a brief definition of sentiment analysis is given. In addition, the validation procedures for testing the model's outcome were identified.

4.1 Sentiment Analysis

The computer task of obtaining author opinions about certain entities such as products, services, organisations, individuals, issues, events, and themes is known as sentiment analysis or opinion mining (Feldman, 2013; Zhang et al., 2018). Many businesses and individuals benefit significantly from sentiment analysis by tracking their brand image and receiving real-time feedback on their products and deeds through various social media platforms, allowing them to react promptly. Not only corporations but local and federal governments also use

textual analysis to reflect public opinions on the policies (Liu, 2015). Due to the volume of data available in numerous sources, accessing and tracking opinions on the web, as well as distilling the information included in statements, remains a difficult endeavour. The average visitor has trouble spotting relevant credible sources, retrieving, and analysing the information contained therein. This is where automated sentiment analysis software tools lend a helping hand to organisations, public bodies, and individuals. According to Feldman (2013), sentiment analysis technology assists marketing managers, public affairs agencies, strategists, legislators, and even stock traders and internet consumers.

In order to perform sentiment analysis, a financial glossary produced by Loughran & McDonald (2011) was applied to gauge negative, positive, uncertainty, litigious, strong modal, moderate, and weak modal words in these fillings. Despite the fact that Harvard General Inquirer (GI) is the most widely used vocabulary, the 10-K filings are based on financial words, hence the Loughran & McDonald (LM) lexicon is employed to conduct analysis in this thesis. Nearly 75% of the Harvard GI's negative words, according to Loughran & McDonald (2011) are improper for capturing a negative tone in commercial applications.

4.2 Parameters After Sentiment Analysis

The thesis compares 504 10-K filings from 106 corporations that have committed wrongful acts against 742 yearly reports from 158 non-fraudulent companies over a 24-year period from 1996 to 2020. The *crime* dummy variable is generated to distinguish the companies and used as a dependent variable. Since this dummy variable indicates illegal activity, the indicator equals 1 for fraudulent companies and 0 otherwise.

As for the independent variables, thirteen explanatory variables are chosen. The total number of words, the number of unique and complex words, the occurrence of words in the LM dictionary, number of negative and positive LM words, number of strong, moderate, and weak modal words in the LM financial dictionary, number of LM uncertainty words, number of LM litigious words, number of Harvard GI negative words defined by LM, and the standard industry classification are all variables that have been considered. The names of all the variables and as well as their descriptions are revealed in Table 5.1.

Variable Name	Description
crime	Dummy variable, crime = 1 if company involved with an illicit behaviour
total.words	Total number of words in a filling text
unique	Number of different words in a filling text
complex	Number of complex words in a filing text (words with three or more vowels)
total.lm.words	Number of words found in the LM reference dictionary in the filing text.
negative	Number of negative words
positive	Number of positive words
strong	Number of strong modal words
moderate	Number of moderate modal words
weak	Number of weak modal words
uncertainty	Number of uncertainty words
litigious	Number of litigious words
hv.negative	Number of negative words listed in Harvard GI, described by LM.
sic	Standard Industry Classification, as a factor

Table 5.1: Variables used in the analysis and their description

In order to have a good comprehension of the concepts, some classification of the words on the LM lexicon must be described. There are 2,337 negative terms in the LM financial lexicon, 353 positive ones, 285 uncertainty words, and 731 litigious terms (Loughran & McDonald, 2011). Some of the frequently used negative words in this dictionary are *restated, litigation, termination, discontinued, penalties, unpaid, investigation, misstatement, misconduct, forfeiture, serious, allegedly, noncompliance, deterioration,* and *felony*. The positive LM bag-of-words features terms with a unidirectional tone, such as *achieve, attain, efficient, improve, profitable,* or *upturn*. According to Loughran & McDonald (2011), the words *approximate, contingency, depend, fluctuate, indefinite, uncertain,* and *variability* in reference to uncertainty do not allude to risk but rather ambiguity. *Claimant, deposition, interlocutory, testimony,* and *tort* are among the words in the litigious list that imply a penchant for legal confrontation. Furthermore, Loughran & McDonald (2011) stated that many of these words fall into more than one classification.

Additionally, the strong, moderate, and weak indicate the degree of likelihood. Words like *always, highest, must* and *will* are examples of strong modal words. *Can, generally,* and *usually* are examples of moderate modal words. *Could, depending, might,* and *possibly* are examples of weak modal words (Loughran & McDonald, 2011).

Furthermore, the correlation table in Appendix A3 suggests that explanatory variables are highly correlated. A high correlation suggests changing the value of one variable, i.e., *total.words*, change the proportion of another variable, i.e., *complex*, nearly in the same fashion. The correlation between variables is unsurprising given that Loughran & McDonald (2011) used the same words in multiple categories.

4.3 Logistic Regression Analysis

Since the response variable in this thesis contains a classification procedure, the ideal method for such variables is logistic regression analysis. This model will predict the likelihood of a corporation belonging to a criminal class that may be classified as either 0 or 1. According to James et al. (2021), logistic regression employs the maximum likelihood method, which is a common methodology for fitting numerous non-linear models. Maximum likelihood estimation (MLE) is a far more rigorous analytics technique for assessing model parameters using only a sample of the data (Nwanganga & Chapple, 2020). The logistic regression for probability was modelled using the equation below (James et al., 2021):

$$p(x) = \Pr(Y = 1|X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_{13} X_{13}}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_{13} X_{13}}}$$

The probability of companies committing a crime is denoted as p(X) in this logistic model and the output can range from 0 to 1. The logistic function will always yield an S-shaped curve therefore, a reasonable prediction will be obtained regardless of the value of X. The intercept is represented by β_0 while the coefficients of the thirteen independent variables, which are described in the data section, are expressed by β_1 to β_{13}. The following equation was obtained by performing a logistic transformation of probability (James et al., 2021; Nwanganga & Chapple, 2020):

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{13} X_{13}$$

On the left-hand side of the equation, the logit or log-odds is shown, indicating the log-odds of P(Y=1|X) versus P(Y = 0|X). The logit of the logistic regression model is linear in X, as seen on the right-hand side (James et al., 2021). This mathematical function converts the log-odds of p(X) to a probability to describe how a unit increase in X increases the log-odds of p(X) by βs (Nwanganga & Chapple, 2020).

The formulation was implemented in R by utilising the *parsnip* package from CRAN.

4.4 Random Forest

Random Forest is a supervised machine learning algorithm for classification and regression analysis that leverages ensemble learning. The random forest approach generates a series of decision trees and a class that is the classification of individual trees during training (Ghavami, 2019; James et al., 2021). A decision tree is a map that depicts a series of recursive splits in a smaller number of stages, with each step identifying the local region. (Alpaydin, 2014). This implies that the same sample may be chosen repeatedly, while other samples may be omitted completely (Belgiu & Drăgu, 2016). The structure of decision trees is depicted in Figure 4.1. The top-down method utilises the predictor input, branches, to provide insight into the target variable, leaves (James et al., 2021). The decision tree is partitioned into two distinct nodes at each division (recursive binary splitting). The probability of class assignment is calculated by arithmetically averaging all constructed trees before arriving at a final classification decision. The generated data is compared to all of the ensemble's decisions in order to find the majority of voters from each tree and assign the outcome to the final class (Belgiu & Drăgu, 2016).

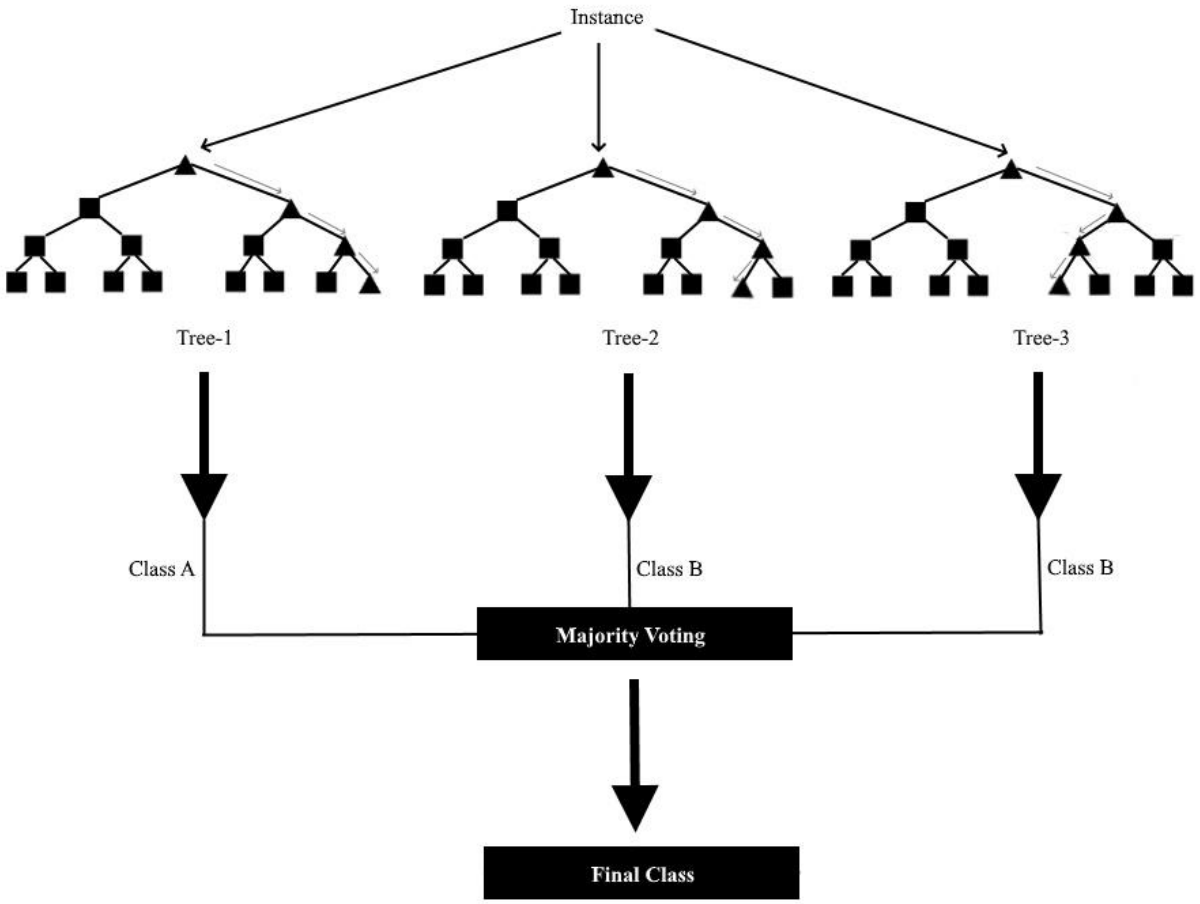

Figure 4.1: Classification phases of Random Forest classifier

The majority of the samples are assigned to the training set in random forest approaches, while the test dataset is utilised to evaluate the performance of the model. Random Forest develops trees with high variation and low bias by growing the forest up to a user-specified number of trees (Belgiu & Drăgu, 2016). Furthermore, each node is split into multiple decision trees based on a user-defined number of predictors.

The most critical hyperparameters must be defined before the Random Forest model needs to be defined:

Number of trees: Increasing the number of decision trees in the model can generally enhance its efficiency because more trees mean more predictions. The model becomes more complex as the number of trees in the model expands, and the R-command requires longer to run. Additionally, performance plateaus at a certain point, at which point adding more trees to the model adds no extra benefit. As a result, the number 600 was selected since it is large enough

to alleviate the overfitting problem while also being small enough to take less computational time. Similarly, the model's performance did not considerably improve after 600.

Number of predictors at each split (mtry): The number of predictors can be estimated by taking the square root of the number of variables, according to James et al. (2021). Therefore, the number of predictors is estimated to be $\sqrt{13} \approx 4$.

The random forest technique is applied in R by utilising the *parsnip* and *randomForest* package from CRAN.

4.5 Validation

In order to understand how effectively the model works, it is vital to investigate the performance of the applied methodologies. Therefore, the validation methods utilised in both logistic regression and random forest analysis will be explained in this section. The data is partitioned into training and test data sets using the *crime* variable as a stratum in order to estimate the predictions. As a consequence, 75% of the total observations (934) were randomly assigned to training data, while the remaining 312 observations were put to test data. The performance metric tools provide feedback on the proposed model, allowing data analysts to select the most trustworthy models for future predictions. The classification metrics and performance measurements have been implemented.

4.5.1 Classification Metrics

The confusion matrices are computed first to assess the accuracy of the prediction, and then the results are visually displayed using the ROC-AUC curve.

4.5.1.1 Confusion Matrix and Accuracy

A confusion matrix is a N x N matrix that depicts the brief outcome of the predictions as a table, where N is the number of target classes (Navlani et al., 2021; Velayutham, 2020). Table 5.2 illustrates a two-dimensional matrix that is applied in this research to display the real class of a criminal and non-criminal firm, as well as the classes of the said companies based on the

prediction outcome. True Positive (TP) and True Negative (TN) denote that the applied technique anticipated the same result as the actual data. If the classifier correctly predicts that illegal practices occurred, it will be assigned to the TP category, and if it accurately estimates no misconduct, it will be placed in the TN category. In contrast, if the model erroneously classifies a company with no fraudulent act, it will be listed in False Positive (FP), or a firm with illicit activities placed into the non-fraudulent class, it will be shown in False Negative (FN).

		Actual	
		Crime	No Crime
Predicted	Crime	True Positive	False Positive **(Type I Error)**
	No Crime	False Negative **(Type II Error)**	True Negative

Table 5.2: Confusion Matrix for actual and predicted crime, and no crime

Accuracy is a metric that evaluates the proportion of samples that are erroneously classified (Raschka & Mirjalili, 2019). The accuracy can be calculated using the formula below:

$$Accuracy = \frac{TP + TN}{TP + FN + TP + TN}$$

As the number of TP and TN rises, the model's accuracy improves. If the accuracy is 0.5, for instance, the likelihood of the model correctly predicting the actual outcome is 50%.

4.5.1.2 ROC Curve and AUC

The Receiver Operating Characteristics (ROC) curve is a graphical representation of the model's performance based on the sensitivity (True Positive Rate or recall) and specificity (1 - False Positive Rate) (Raschka & Mirjalili, 2019). Using the formula below, one may estimate these positive rates:

$$FPR = \frac{FP}{FP + TN}$$

$$TPR = \frac{TP}{FN + TP}$$

The sensitivity and specificity corresponding to a specific decision threshold is represented by each point on the ROC curve. The model's ability to correctly differentiate the positive and negative classes in valuation data is measured by the ROC curves. The best performing model can be validated by calculating the ROC curves of two classification models (Raschka & Mirjalili, 2019).

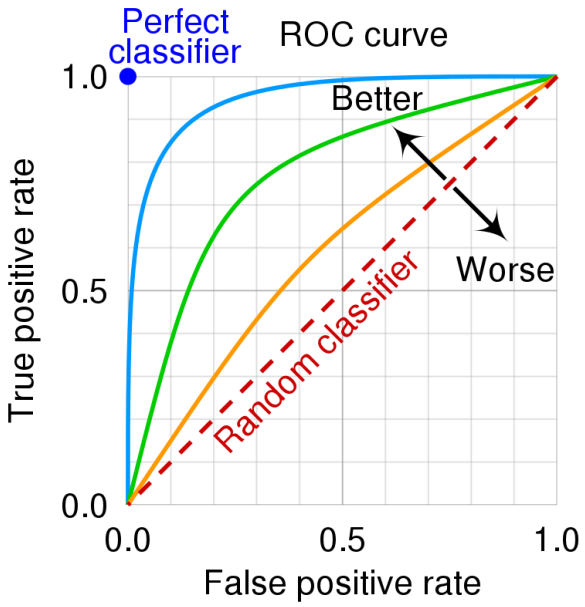

Figure 4.2: An example of the ROC-AUC Curve

Figure 4.2 visualises the ROC curve. The performance of a "no information" classifier is represented by the diagonal red dotted line (James et al., 2021). The ideal ROC curve is arched in the top left corner, with a high true positive rate and a low false-positive rate. The model will be better at forecasting as the curve approaches the sensitivity corner. The AUC (Area Under the ROC Curve) is a measure that ranks the accuracy of the model. Therefore, the greater the ROC-AUC value, the more accurate the classifier is at predicting corrupt companies.

The ROC-AUC is generated in R by using the *yardstick* package from CRAN.

4.5.2 Performance Measurement: k-fold Cross-Validation

In terms of performance evaluation, k-fold cross-validation has been implemented. K-fold cross-validation estimates uncertainty better than the validation set approach, and it takes less time to compute in comparison to the leave-on-out cross-validation method.

The k-fold cross-validation method divides training data into *k* folds with no replacement (Raschka & Mirjalili, 2019). The model employs *k – 1* fold for training and the remaining folds for testing. This process is rehashed *k* times, yielding *k* performance estimation models. In this thesis, the 10-fold cross-validation is applied. The validation data for testing the model is kept one-fold, while the other nine samples are used for training the model and set to be repeated 10 times. Then, the mean accuracy rate and ROC-AUC value for each model will be determined.

The primary advantage of k-fold cross-validation is that each example is used only once for teaching and testing which eliminates bias and produces low variance in model performance estimates (Raschka & Mirjalili, 2019). Since each observation must occur *k – 1* time in the training data and 1-fold in the testing data, the model's performance evaluation is not reliant on a single initial split into training and testing data and the accompanying bias. K-fold cross-validation is commonly used for model tuning to discover the ideal hyperparameter values that give a good classification performance, which is calculated by comparing the model performance on test folds (Raschka & Mirjalili, 2019). Therefore, another important advantage of k-fold cross-validation is that it can be extremely useful when modifying model parameters. Using mean performance numbers from 10-fold cross-validation is more reliable than using values from a basic two-fold validation while comparing the combination of particular parameters and the performance outcome.

The 10-fold cross-validation method is applied in R by using the *workflow* package from CRAN.

5. Results

The empirical findings of the analysis, as well as the accuracy of the applied methodologies, will be presented in this section. After tokenisation of each annual report, the logistic regression method is applied due to the binary nature of the dependent variable *crime* which stands for crime (1), no crime (0). The results reveal the likelihood of a corporation engaging in illegal activities based on the explanatory variables taken from the LM financial lexicon. Furthermore, the Random Forest technique is used since this approach entails the selection of proper hyperparameters, which have a major impact on the model's success. The validation measures are designed to directly test the approaches' accuracy in order to shed light on the language differences found in both fraudulent and control firms' annual reports. Finally, the validation outputs of each technique will be compared to see which model is more effective at detecting the crime.

The coefficients of the logistic model are shown in Figure 5.1, which reflect the degree and direction of that variable's relation to the probability of crime. In addition, the odds ratio in Figure 5.2 shows the probability of a crime occurring versus not occurring if one unit of change occurs in the explanatory variable, holding all other independent variables fixed. According to the findings, the significance level ranges from 1% to 10%. The regression outcome overall model including the industry classification as well as the odds ratio of significant *sic* variables can be found in Appendix A4.

```
Regression Results
============================================
                    Dependent variable:
                ---------------------------
                          crime
--------------------------------------------
total.words                0.0001*
                          (0.0001)
unique                    0.0003**
                          (0.0002)
complex                   -0.0002**
                          (0.0001)
total.lm.words           -0.0001***
                          (0.00005)
negative                   0.001**
                          (0.0004)
positive                    0.001
                           (0.001)
strong                      0.001
                           (0.002)
moderate                   0.0004
                           (0.003)
weak                      -0.004***
                           (0.001)
uncertainty                0.001*
                           (0.001)
litigious                 0.002***
                          (0.0004)
hv.negative                -0.0001
                          (0.0003)
Constant                  -0.936***
                           (0.307)
--------------------------------------------
Observations                 1,246
Log Likelihood             -806.489
Akaike Inf. Crit.          1,638.979
McFadden R-squared           0.041
============================================
Note:            *p<0.1; **p<0.05; ***p<0.01
```

Figure 5.1: Logistic regression analysis outcome

The coefficients with a significance level of 1% have a very significant impact on detecting criminal conduct. The method reveals that the prosecuted organisation uses more litigious terms (*litigious*) in their annual fillings while avoiding using weak model verbs (*weak*), in comparison to their non-convicted counterpart. When detecting the likelihood of crime, the odd ratio visualises that all of the variables will result in a more or less 1% probability increase or decrease, as seen in Figure 5.2. For example, when the number of litigious phrases rises by 1%, the relative probability of committing crime increases by 1% in comparison to not committing a crime, holding everything fixed, whereas the same result will be obtained if the number of weak modal verbs decreases by 1%. Furthermore, the model reveals that if the

number of words in these fillers does not also appear as frequently in the LM-lexicon (*total.lm.words*), the likelihood of capturing crime improves at a 1% significant level.

Variables	Odds Ratio
total.words	1.000
unique	1.000
complex	1.000
total.lm.words	1.000
negative	1.001
weak	0.996
uncertainty	1.002
litigious	1.002

Figure 5.2: Odds Ratio of the Significant Variables

Moreover, the outcome indicates that *complex, unique* and *negative* variables are significant at a 5% significance level. This suggests that fraudulent firms tend to employ more negative terms, less complex phrases, and more unique words in their yearly reports. Furthermore, based on this logistic model, fraudulent firms which 1600 (Heavy Construction other than building construction – contractors) and 6311 (Life Insurance) industry classification are more likely to be discovered than their non-prosecuted rivals (Figure A4.1). The odds ratio reveals that *sic6311* is 9.18% more likely to be captured in our model, implying that organisations in this industry use more distinct words while committing a crime, in comparison to control groups. On the other hand, the odds ratio for *sic1600* is 6.09%, see Figure A4.2.

Finally, the total number of words (*total.words*), as well as the number of uncertainty (uncertainty) phrases, appear to be significant at 10%. In their annual reports, the corrupt corporations include more words and add more uncertain terminology. Additionally, enterprises in the "Aircraft" (3721) and "Search, detection, navigation, guiding, and aeronautical systems" (3812) industries are more likely to be recognised if they are involved in any form of criminal activity, as seen in Figure A4.1. The odds ratio for *sic3812* is calculated as 8.11% whereas, it is 7.55% for *sic3721* (Figure A4.2).

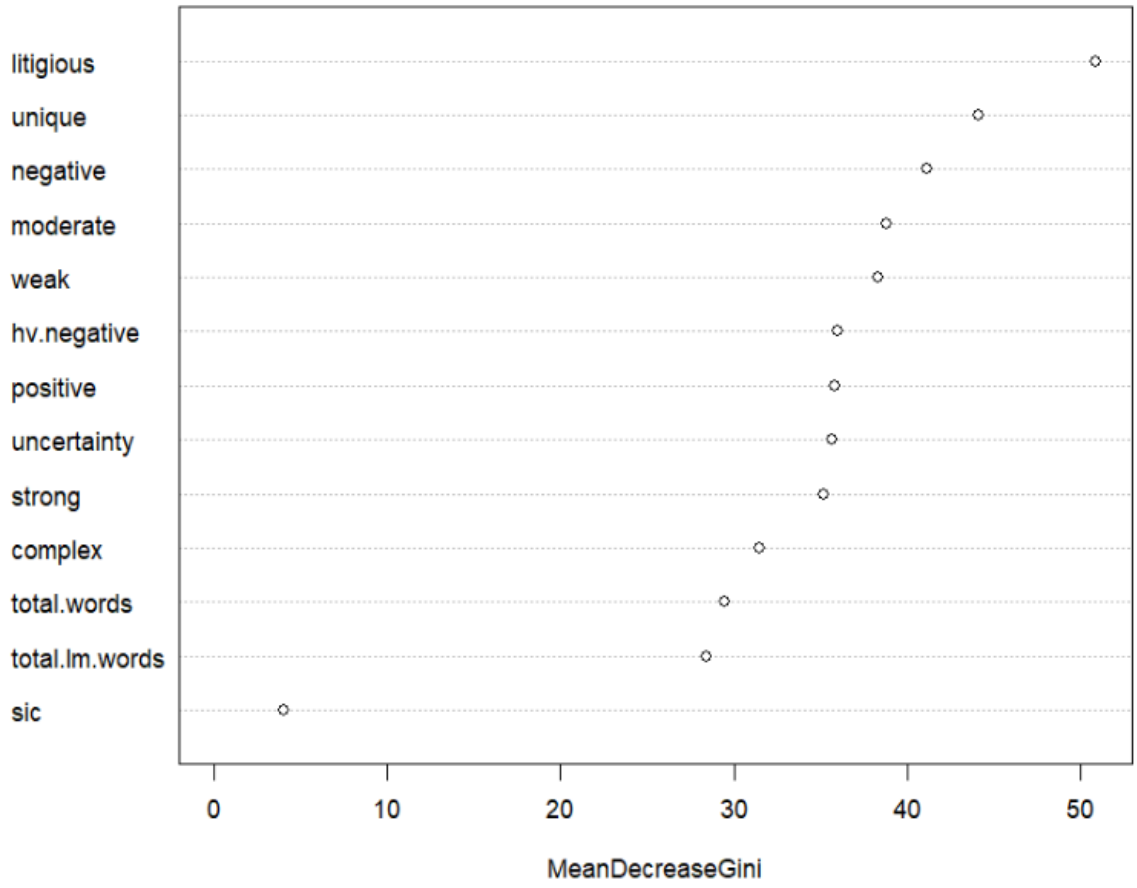

Figure 5.3: Variable importance in predicting the crime

The Random Forest analysis was employed after the logistic regression. The importance of variables in the regression can be ranked using random forests, as shown in Figure 5.3. These rankings are based on the Gini Impurity Index, which assesses the quality of a split and determines whether a variable has been incorrectly categorised (James et al., 2021). Therefore, the larger the magnitude of the mean decreases Gini score, the more important the variable in the model is. Therefore, the most important variable in the Random Forest regression model was identified as *litigious*, followed by *unique* and *negative* variables. The mean decrease Gini of the remaining variables have roughly the same, ranging between 30 to 40. However, the variable *sic* is classified as the least significant variable for the Random Forest analysis.

It is vital to assess the accuracy and reliability of the analysis to see whether the regression is good enough for forecasting the model. The confusion matrix and accuracy rate for the actual and predicted outcomes of the logistic regression and Random Forest are shown in Table 5.3. When it comes to spotting fraudulent enterprises, logistic regression analysis is slightly more

effective than identifying non-fraudulent corporations. Random Forest, on the other hand, is better at spotting both non-criminal firms. Only 65.1% of cases are accurately predicted by the logistic regression model, whereas 78.8% are correctly identified using Random Forest. So, in comparison to logistic regression, the Random Forest model is marginally better at predicting the classification of a random company.

		Logistic Regression		**Random Forest**	
		Actual		Actual	
		Crime	No Crime	Crime	No Crime
Predicted	Crime	172	14	171	15
	No Crime	95	31	51	75
Accuracy		0.651		0.788	

Table 5.3: Comparison of confusion matrix and accuracy rate for logistic regression and Random Forest

Figure 5.4 shows the ROC-AUC curve for logistic regression. The ROC-AUC score of 0.638 indicates that the model recognises more true positives and true negatives than false positives and false negatives. The Random Forest technique in Figure 5.5, on the other hand, has a ROC-AUC of 0.833, which is greater than the previous analysis, which indicates that the model has an 83.3% likelihood of being able to distinguish between positive and negative classes. When compared to Figure 5.4, the curve in Figure 5.5 is much closer to the sensitivity than specificity.

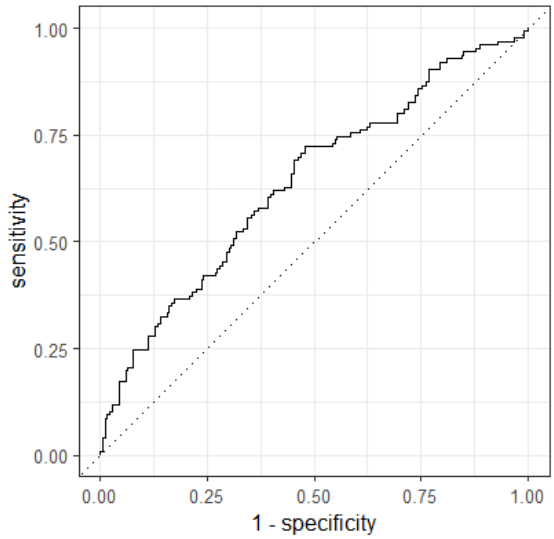

Figure 5.4: ROC-AUC curve of logistic regression model

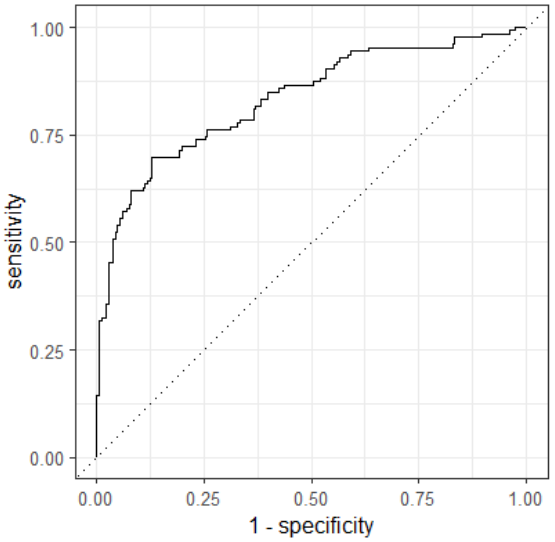

Figure 5.5: ROC-AUC curve of Random Forest

The average accuracy rate for logistic regression was computed as 0.64, using 10-fold cross-validation, whereas the mean accuracy rate for Random Forest was 0.748 (Appendix A5). The estimations of the ROC-AUC value for each fold in the cross-validation for logistic regression and Random Forest models are shown in Figure 5.6. For logistic regression, the ROC-AUC curve values vary from 0.55 to 0.70, with an average of 0.618. However, the mean ROC-AUC score of Random Forest analysis, suggests that the model correctly predicts 83.3% of true and false classifications.

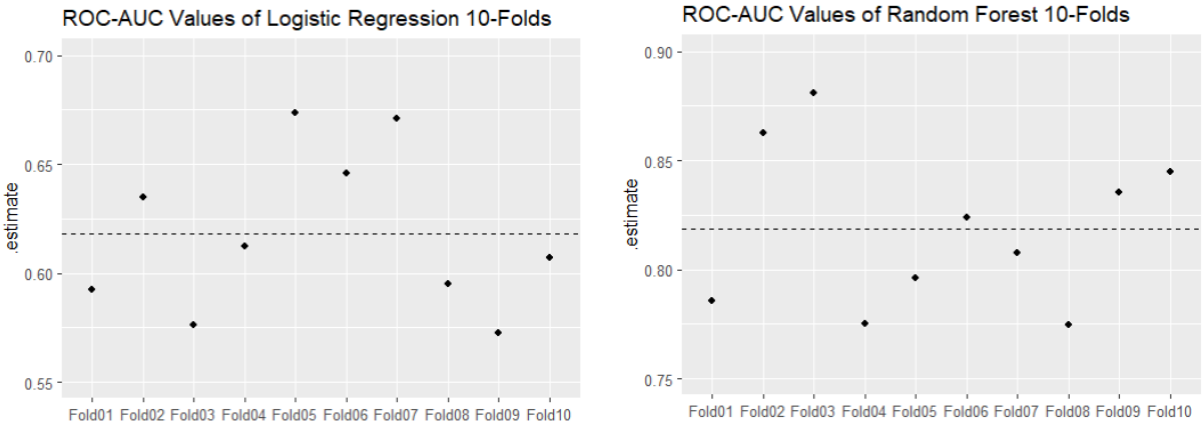

Figure 5.6: ROC-AUC values for each fold

To summarise, the logistic regression model reveals that the variables *total.words*, *unique*, *complex*, *total.lm.words*, *negative*, *weak*, *uncertainty*, *litigious*, *sic1600*, *sic6311*, *sic3812*, and *sic3721* are significant when detecting illegal behaviours based on linguistic aspects in annual reports. According to the Random Forest analysis, the most important variable for this method was *litigious*, while the least important variable was *sic*. The results of validation methods disclose that, overall, the Random Forest classifier outperforms the logistic regression classifier, when it comes to distinguishing the prosecuted firms from their counterfactual parties.

6. Discussions

This section will explain and evaluate the findings of the analysis. The outcomes of the hypothesis will first be presented and compared to those of previous studies. The advantages and disadvantages of logistic regression and Random Forest analysis will next be addressed. Finally, some of the thesis' shortcomings will be outlined.

6. Discussions

6.1 Regression Outcome and Hypotheses

H_1: Companies that employ fewer negative words are more likely to engage in deceptive practices.

According to the results of the analysis, negative financial words are a significant variable for detecting corporate crime and untruthful organisations tend to use more negative words. Previous research, on the other hand, established a negative correlation between crime and the use of negative words.

Two factors might account for the disparity in results concerning the usage of negative words. First, rather than focusing on the MD&A part, as is common in the literature, the analysis focuses on all sections within the annual reports. The other sections of the annual reports may increase the overall negative terms if the company fails to generate sufficient income or is unable to pay its financial obligations. Second, the corrupt executive's behaviour may have changed after the analysis in the literature was published 10 years or more ago. Because earlier research has revealed a pattern in executive behaviour, they may alter their writing tone to fit into the category of non-fraudulent firms. The reputational damage caused by uncovering the fraud could be far more costly to the corporation than the short-term stock price declines following the unfavourable management announcements.

Studies analysing the impact of using a negative managerial tone reveal that a pessimistic voice has shown to increase the capital cost (Feldman et al., 2010) and the volatility in stock returns of the firm (Hájek et al., 2013). Furthermore, market returns are adversely correlated with elevated pessimism and reflect on the future performance of the firms once negative news is released. According to (Hájek & Olej, 2013), financially distressed organisations employ more negative and less positive terms.

H_2: Companies that employ more uncertain words are more likely to engage in deceptive practices.

Based on the findings, prosecuted organisations utilise more uncertain words than control groups, as suggested by the literature. To satisfy predetermined performance objectives and

satisfy the shareholders, corporate executives have a direct motive to offer a favourable image of the company's financial performance. Executives profit from uncertain language that does not convey good or negative aspects of ongoing challenges and risk factors in order to carry out healthy financial growth and secure their position within the company. Another consideration is the explanation of events beyond the company's control, such as the impact of the pandemic on profitability. In which case, communicating in uncertain words could reduce professional guilt or make the fraudulent activities appear more sympathetic. Although managers may gain from uncertain language by diverting attention away from the real picture, Hájek et al. (2013) point out that text uncertainty is correlated with the peculiar volatility in the stock price return. High stock market volatility breeds anxiety and distress in the market, which may erode returns on investment.

H_3: Companies that employ less litigious words are more likely to engage in deceptive practices.

The outcome shows the polar opposite of what the literature predicts, suggesting that convicted businesses employ more litigious language. According to Malik et al. (2022), a litigious tone is correlated with a weaker return and increases the likelihood of shareholders having a poor perception of the internal controls and competence to survive in the current financial environment. Furthermore, a strong litigious tone signals a growth in current and projected defence expenditures, which could have a detrimental effect on the firm's economic situation. The goal of employing more litigious words might be to provide insight into the juridical status of the company in order to minimize any reputational damage and give the company’s perspective on any allegations. This way, the corporation can reduce the risk of a prospective profit drop. Additionally, firms must publish details regarding substantial pending lawsuits or other legal actions, other than regular litigation, in section 3 of annual reports.

H_4: Companies that employ more words with fewer unique words are more likely to engage in deceptive practices.

In terms of employing more words, the analysis' findings are consistent with the literature. Because corrupt firms do not want to divulge illegal behaviour in order to avoid a stock market slump, they will adopt more phrases to conceal the reality. Giving superfluous detail or utilising a description relating to a specific term instead of a single word could be a tactic

for deflecting attention. Although (Humpherys et al., 2011) found that organisations that use more words are unable to produce more unique words, the findings of this paper suggest that fraudulent companies not only use more words but also have a greater lexical variety in their annual filings. Words having specific and obscure meanings can sometimes detract from rather than enrich a shareholder's experience. Executives may create the illusion of professionalism and the professional eloquence could act as a potential smokescreen to criminal acts.

H_5: Companies that employ more complex words are more likely to engage in deceptive practices.

Corrupt companies, according to the regression results, employ fewer complicated phrases than their competitors. Previous research has indicated that in order to avoid or postpone negative market reactions, fraudulent companies utilise more complex terms to prevent damaging news to be disclosed (Humpherys et al., 2011). However, using sophisticated terms can be laborious and time-consuming for investors. CEOs may find that using simpler language makes it easier to communicate with shareholders' regarding the company's current state and attract investment.

Additionally, the prosecuted companies utilise fewer words from the LM lexicon, according to the findings. The deliberate avoidance of adopting pre-determined LM words to conceal their motivations can be one obvious explanation, given that a dictionary is a well-used tool by auditors and investigators for spotting fraudulent organisations.

6.2 Advantages and Disadvantages of Both Analysis

This thesis employs two forms of analysis: The logistic regression model, a type of linear model with independent variables that define a connection to a dependent response variable, and Random Forest, a node-based tree-like structure made up of multiple independent decision trees.

The Random Forest model outperformed logistic regression in terms of accuracy and ROC-AUC score in forecasting convicted firms. Nonetheless, the effectiveness of both models is

influenced by a variety of factors such as the selection of independent variables, the division of data into training and testing, multicollinearity between explanatory variables, or the preference for hyperparameters. Moreover, because the data set contains a small number of total observations, the performance of both models fluctuates dramatically each time they are run. This impact is reduced to some extent by employing a 10-fold cross-validation method.

It may be simpler to use the logistic regression model since the model does not require any selection of hyperparameters. On the contrary, Random Forest necessitates the selection of proper hyperparameters, which has a major influence on how well the process is carried out. Although choosing appropriate parameters might be complex and time-consuming, the adjustment of these hyperparameters is relatively simple when compared to other machine learning algorithms.

Moreover, in logistic regression, the magnitude and direction of the explanatory variable's coefficients, as well as the importance of all variables, are provided transparently. Random Forest, on the other hand, has an interpretability problem and is unable to identify the significance of each variable because of the combination of decision trees. Due to the intricacy of Random Forest, it is very difficult to explain why a given observation has been determined as fraudulent or non-fraudulent.

Another significant distinction is that logistic regression techniques can only detect linear correlations between explanatory variables and logarithmic odds. Random Forest models, on the other hand, can be more successful when dealing with big databases with more complicated, non-linear relationships, which results in highly precise estimates. However, it is critical to select the appropriate hyperparameters to avoid overfitting and poor predictions.

Although both models have advantages and limitations when it comes to spotting fraudulent organisations, the results of this thesis show that the Random Forest model has better at predicting untruthful companies with a higher accuracy rate and ROC-AUC score, which is consistent with the study conducted by Hajek & Henriques (2017).

6.3 Limitations

It is important to highlight the limitations of this study as they have an impact on the quality of the analysis.

First, only publicly traded companies are included in the data. Inclusion of all the available companies in the data might change the degree and significance of the explanatory variables as well as alter the variable importance in Random Forest. There are pressures on publicly traded companies to satisfy shareholders and Wall Street's expectations, which may lead to these companies engaging in fraudulent behaviour that has a significant impact on the financial markets. However, in order to ensure fair competition in the market, it might be beneficial to investigate all fraudulent acts. In addition to this, in terms of the data retrieval process, there were certain challenges in locating publicly traded non-fraudulent corporations within the specified time frame as well as industry classification. Some of the non-fraudulent companies' intermittent documents were attempted to be substituted by another non-convicted equivalent, which may result in a different consequence due to the inconsistency of the documents.

Second, there is a strong correlation between predictor variables. Some variables appear to have a correlation closer to 1, indicating that adjusting one variable without changing another is difficult. This makes it challenging to estimate the connection between each explanatory variable and the response variable separately. Given a little change in the data or model, the model outcomes may be volatile and fluctuate considerably.

Finally, without accounting for the *sic* variable, McFadden's R-squared value is 0.041, as shown in Figure 5.1. This indicates that the 4.1% variation in the dependent variable can be explained by the model. In Figure A4.1, after accounting for industry classifications, the percentage increased to 8.63%. This number is relatively low, implying that there might be other variables that can provide a better explanation for fraudulent company predictions.

7. Conclusion

The main purpose of the thesis was to compare the linguistic features of convicted corporations to those of their rivals in related industries. The focus was to develop empirical evidence using logistic regression and Random Forest algorithms to predict the fraudulent activities by analysing 10-K filings from five years prior to the conviction date.

Since the response variable, crime, is a binary variable, the logistic regression model was employed to predict the likelihood of a firm being fraudulent or not. Additionally, the Random Forest machine learning algorithm was implemented because the algorithm handles both regression and classification tasks and compared the outcome with logistic regression using accuracy rate, ROC-AUC score and 10-fold cross-validation techniques. All the validation tools indicate that the Random Forest model outperforms the logistic regression model, by 13.05% on average.

Several findings emerged from the logistic regression model. Firstly, corrupt companies tend to use more negative words in comparison to non-fraudulent companies. Since the previous study has established a pattern of employing more positive language, the professionals may change their writing tone to place it into the non-fraudulent firm category. Second, the companies that use more uncertain words are more likely to engage in illegal activities. To achieve sustainable profit growth and retain their wealth, executives may benefit from uncertain language that does not indicate positive or negative elements of ongoing problems and risk concerns. Third, corrupt companies use more litigious words. Using more litigious phrases could shed light on the status of the company’s legal proceedings and present the company's viewpoint on any allegations in order to prevent reputational damage. Forth, untruthful companies use more words with high lexical diversity. This strategy could assist managers in creating barriers to extracting damaging information in order to reduce negative market reactions. Finally, corporations that use simpler language tend to engage in deceptive activities. Too many complex words create impediments to comprehending the company's actual financial status, limiting shareholder investment.

References

Alexander, C. R., & Cohen, M. A. (1996). New Evidence on the Origins of Corporate Crime. In *Managerial and Decision Economics*(Vol. 17).

Alpaydin, E. (2014). Decision Trees. In *Introduction to Machine Learning* (pp. 213–238). MIT Press.

Bach, M. P., Krstič, Ž., Seljan, S., & Turulja, L. (2019). Text mining for big data analysis in financial sector: A literature review. *Sustainability (Switzerland)*, *11*(5). https://doi.org/10.3390/su11051277

Belgiu, M., & Drăgu, L. (2016). Random forest in remote sensing: A review of applications and future directions. In *ISPRS Journal of Photogrammetry and Remote Sensing* (Vol. 114, pp. 24–31). Elsevier B.V. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Busch, R. S. (2012). Introduction to Healthcare Fraud. In *Healthcare Fraud: Auditing and Detection Guide* (pp. 1–17).

Corporate Prosecution Registry. (2022). *Data & Documents*. https://corporate-prosecution-registry.com/browse/

Craja, P., Kim, A., & Lessmann, S. (2020). Deep learning for detecting financial statement fraud. *Decision Support Systems*, *139*. https://doi.org/10.1016/j.dss.2020.113421

Croall, H. (2001). *Understanding White Collar Crime*. McGraw-Hill Education.

Dong, W., Liao, S., & Liang, L. (2016). *Financial Statement Fraud Detection Using Text Mining: A Systematic Functional Linguistics Theory Perspective* (Vol. 188). http://aisel.aisnet.org/pacis2016/188

EPA (United States Environmental Protection Agency). (2021). *MARPOL Annex VI and the Act To Prevent Pollution From Ships (APPS)*. https://www.epa.gov/enforcement/marpol-annex-vi-and-act-prevent-pollution-ships-apps

European Commission. (2022). *Competition Policy: Antitrust Overview*. https://ec.europa.eu/competition-policy/antitrust/antitrust-overview_en

Favre, D., Cane, M., Hockel, J., & Konczos, L. (2020). False Statements and False Claims. *American Criminal Law Review 57*, *3*, 727–758.

Feldman, R. (2013). Techniques and applications for sentiment analysis: The main applications and challenges of one of the hottest research areas in computer science. In *Communications of the ACM* (Vol. 56, Issue 4, pp. 82–89). https://doi.org/10.1145/2436256.2436274

Feldman, R., Govindaraj, S., Livnat, J., & Segal, B. (2010). Management's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies*, *15*(4), 915–953. https://doi.org/10.1007/s11142-009-9111-x

Flesca, S., Manco, G., Masciari, E., Rende, E., & Tagarelli, A. (2004). Web wrapper induction: a brief survey. In *AI Communications* (Vol. 17). IOS Press.

Ghavami, P. (2019). Random Forests Techniques. In *Big Data Analytics Methods: Analytics Techniques in Data Mining, Deep Learning and Natural Language Processing* (2nd Edition, pp. 137–141). De Gruyter.

Giroux, G. (2017). Accounting Scandals, A Historical Perspective. In *Accounting Fraud, Second Edition : Maneuvering and Manipulation, Past and Present* (2nd Edition, pp. 1–31). Business Expert Press.

Goel, S., Gangolly, J., Faerman, S. R., & Uzuner, O. (2010). Can linguistic predictors detect fraudulent financial filings? *Journal of Emerging Technologies in Accounting*, *7*(1), 25–46. https://doi.org/10.2308/jeta.2010.7.1.25

Goel, S., & Uzuner, O. (2016). Do Sentiments Matter in Fraud Detection? Estimating Semantic Orientation of Annual Reports. *Intelligent Systems in Accounting, Finance and Management*, *23*(3), 215–239. https://doi.org/10.1002/isaf.1392

Gupta, R., & Gill, N. S. (2012). Financial Statement Fraud Detection using Text Mining. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 3, Issue 12). www.ijacsa.thesai.org

Hajek, P. (2017). Combining bag-of-words and sentiment features of annual reports to predict abnormal stock returns. *The Natural Computing Applications Forum*.

Hajek, P., & Henriques, R. (2017). Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods. *Knowledge-Based Systems*, *128*, 139–152. https://doi.org/10.1016/j.knosys.2017.05.001

Hájek, P., & Olej, V. (2013). Evaluating Sentiment in Annual Reports for Financial Distress Prediction Using Neural Networks and Support Vector Machines. In *CCIS* (Vol. 384).

Hájek, P., Olej, V., & Myšková, R. (2013). *Forecasting Stock Prices using Sentiment Information in Annual Reports-A Neural Network and Support Vector Regression Approach*.

Hasen, E., Alagia, M., Jenets, C., & Miliotes, L. (2021). Obstruction of Justice. *American Criminal Law Review*, *58*, 1251–1292.

Hope, K. R. (2020). Channels of corruption in Africa: analytical review of trends in financial crimes. *Journal of Financial Crime*, *27*(1), 294–306. https://doi.org/10.1108/JFC-05-2019-0053

Humpherys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K., & Felix, W. F. (2011). Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems*, *50*(3), 585–594. https://doi.org/10.1016/j.dss.2010.08.009

Iezzi, D. F., & Celardo, L. (2020). Text Analytics: Present, Past and Future. In D. F. Iezzi, D. Mayaffire, & M. Misuraca (Eds.), *Text Anlytics: Advances and Challenges*. Springer, Cham.

Jaeschke, R., Lopatta, K., & Yi, C. (2018). Managers' use of language in corrupt firms' financial disclosures: Evidence from FCPA violators. *Scandinavian Journal of Management*, *34*(2), 170–192. https://doi.org/10.1016/j.scaman.2018.01.004

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R (Second Edition)*. http://www.springer.com/series/417

Li. (2010). Textual Analysis of Corporate Disclosures: A Survey of the Literature. In *Journal of Accounting Literature* (Vol. 29).

Liu, B. (2015). Introduction. In *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions* (pp. 1–15). Cambridge University Press. https://doi.org/10.1017/CBO9781139084789.002

Lloyd, C. (2020). *The Privacy Revolution Begins: Did Carpenter Just Give Bitcoin Users a Chance to Strike down the Bank Secrecy Act?* (Vol. 88, Issue 1). https://perma.cc/F8E3-8CHF].

Loughran, T., & Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *Journal of Finance*, *66*(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Malik, M. F., Shan, Y. G., & Tong, J. Y. (2022). Does Litigious Tone affect Audit Pricing? *Accounting & Finance*, *62*(S1), 1715–1760.

Mayer, C. H. (2019). *Combating Wildlife Crime in South Africa*. Springer.

Ministry of Justice. (2010). *The Bribery Act 2010*. www.justice.gov.uk/guidance/bribery.htm

Navlani, A., Fandago, A., & Idris, I. (2021). Supervised Learning - Classification Techniques. In *Python Data Analysis* (pp. 291–317). Packt Publishing.

Nwanganga, F., & Chapple, M. (2020). Logistic Regression. In *Practical Machine Learning in R* (pp. 165–219). John Wiley & Sons.

Paliwal, M. (2006). Ethical Choices in Business. In *Business Ethics* (pp. 47–60). New Age International Ltd.

Pathak, M. A. (2014). Introduction. In *Data Science with R*. Springer, Cham.

Purda, L., & Skillicorn, D. (2015). Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection. *Contemporary Accounting Research*, *32*(3), 1193–1223. https://doi.org/10.1111/1911-3846.12089

Raschka, S., & Mirjalili, V. (2019). Learning Best Practices for Model Evaluation and Hyperparameter Tuning. In *Python Machine Learning* (3rd Edition, pp. 191–222). Packt Publishing.

Rivera, K., Murphy, R., Reid, C., Nestler, C., Rigby, M., Qureshi, S., & Heissner, S. (2022). *PwC's Global Economic Crime and Fraud Survey 2022*.

Said, R., Crowther, D., & Amran, A. (2014). Introduction: Corporate Crime and Its Constraint. In *Ethics, Governance and Corporate Crime: Challenges and Consequences* (1st Edition, pp. 1–17). Emerald Group Publishing Limited.

Shover, N., & Simpson, S. S. (2003). Corporate Crime, Law, and Social Control. *Contemporary Sociology*, *32*(4), 500. https://doi.org/10.2307/1556590

Skillicorn, D. B., & Purda, L. (2012). Detecting fraud in financial reports. *Proceedings - 2012 European Intelligence and Security Informatics Conference, EISIC 2012*, 7–13. https://doi.org/10.1109/EISIC.2012.8

Statista Research Department. (2021). *Types of data used by ML, DS, and AI developers worldwide 2021*. Statista. https://www-statista-com.manchester.idm.oclc.org/statistics/1241924/worldwide-software-developer-data-uses/

U.S. Securities and Exchange Commission. (2021). *How to Read a 10-K/10-Q*. https://www.sec.gov/fast-answers/answersreada10khtm.html

Velayutham, S. (2020). Disease Identification in Plant Leaf Using Deep Convolutional Neural Networks. In *Handbook of Research on Applications and Implementations of Machine Learning Techniques* (pp. 47–62). IGI Global.

Wang, K. (2010). Defining Securities Fraud. In *Securities Fraud, 1996-2001 : Incentive Pay, Governance, and Class Action Lawsuits* (pp. 5–14). LFB Scholarly Publishing LLC.

Zahra, S. A., Priem, R. L., & Rasheed, A. A. (2007). Understanding the Causes and Effects of Top Management Fraud. *Organizational Dynamics*, *36*(2), 122–139. https://doi.org/10.1016/j.orgdyn.2007.03.002

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4). https://doi.org/10.1002/widm.1253

Appendix

A1 Types of Crime

Type	**Description**
Accounting Fraud	Misrepresentation of financial data for the purpose of personal enrichment. (Giroux, 2017)
Act to Prevent Pollution from Ships	Violation of the International Convention for the Prevention of Pollution from Ships (MARPOL) that addresses the prevention of marine pollution caused by ships (EPA, 2021).
Antitrust	Anti-competitive agreements or abusive behaviour by enterprises with a dominant position in a market (European Commission, 2022).
Bribery	Illicit transaction in which the offering party gains an advantage through performing improperly while the receiving party is compensated (Ministry of Justice, 2010).
Controlled Substances	Violation of Controlled Substances Act, such as procession or sale of illegal drugs (Corporate Prosecution Registry, 2022).
False Statements	Making false declarations in order to defraud or mislead the government (Favre et al., 2020).
Firearms	Breach of federal criminal weapons licensing and sales legislation (Corporate Prosecution Registry, 2022).
Food	FDCA prohibits violations of federal food safety rules, such as adulteration or misbranding (Corporate Prosecution Registry, 2022).
Gambling	Unlicensed gambling operations or other federal gambling legislation offences (Corporate Prosecution Registry, 2022).
Health Care Fraud	Using false or fraudulent pretences, representations, or pledges to defraud any health-care benefits programme (Busch, 2012).
Immigration	Bringing in and harbouring illegal aliens, as well as breaking immigration restrictions governing non-citizen labour and illegal hiring practices (Corporate Prosecution Registry, 2022).
Kickbacks	Negotiated payments, typically obtained through procurement contracts or project development (Hope, 2020).

Money Laundering	Attempting to conceal the source, ownership, control, or true nature of money obtained through illicit practices (Hope, 2020).
Obstruction of Justice	Any impediment to the efficient administration of justice, such as prosecutors, investigators, or government officials (Hasen et al., 2021).
OSHA (Occupational Safety and Health Act)	Deliberate violations of worker safety regulations are also referred to as workplace safety offences (Corporate Prosecution Registry, 2022).
Securities Fraud	Unlawful conduct by companies in conjunction with the buying or selling of any security in order to mislead public investors (Wang, 2010).
Tax Fraud	Intentional federal tax avoidance and fraud or unsubstantiated statements to tax authorities (Corporate Prosecution Registry, 2022).
Wildlife	Trading plants, animals, or animal products illegally (Mayer, 2019).

Table A1.1: The types and the relevant descriptions

A2 SIC Titles of Fraudulent Companies

SIC	Number of Companies	Industry Title
2015	2	POULTRY SLAUGHTERING AND PROCESSING
2070	2	FATS & OILS
2870	2	AGRICULTURAL CHEMICALS
2911	2	PETROLEUM REFINING
3533	2	OIL & GAS FIELD MACHINERY & EQUIPMENT
3561	2	PUMPS & PUMPING EQUIPMENT
3711	2	MOTOR VEHICLES & PASSENGER CAR BODIES
4931	2	ELECTRIC & OTHER SERVICES COMBINED
5122	2	WHOLESALE-DRUGS, PROPRIETARIES & DRUGGISTS' SUNDRIES
7372	2	SERVICES-PREPACKAGED SOFTWARE
100	1	AGRICULTURAL PRODUCTION-CROPS
1531	1	OPERATIVE BUILDERS
1600	1	HEAVY CONSTRUCTION OTHER THAN BLDG CONST - CONTRACTORS
2800	1	CHEMICALS & ALLIED PRODUCTS
2836	1	BIOLOGICAL PRODUCTS, (NO DISGNOSTIC SUBSTANCES)
2844	1	PERFUMES, COSMETICS & OTHER TOILET PREPARATIONS
2890	1	MISCELLANEOUS CHEMICAL PRODUCTS
3523	1	FARM MACHINERY & EQUIPMENT
3571	1	ELECTRONIC COMPUTERS
3578	1	CALCULATING & ACCOUNTING MACHINES (NO ELECTRONIC COMPUTERS)
3690	1	MISCELLANEOUS ELECTRICAL MACHINERY, EQUIPMENT & SUPPLIES
3713	1	TRUCK & BUS BODIES
3721	1	AIRCRAFT
3724	1	AIRCRAFT ENGINES & ENGINE PARTS
3812	1	SEARCH, DETECTION, NAVAGATION, GUIDANCE, AERONAUTICAL SYS
3829	1	MEASURING & CONTROLLING DEVICES, NEC
3845	1	ELECTROMEDICAL & ELECTROTHERAPEUTIC APPARATUS

4011	1	RAILROADS, LINE-HAUL OPERATING
4213	1	TRUCKING (NO LOCAL)
4400	1	WATER TRANSPORTATION
4412	1	DEEP SEA FOREIGN TRANSPORTATION OF FREIGHT
4512	1	AIR COURIER SERVICES
4813	1	TELEPHONE COMMUNICATIONS (NO RADIOTELEPHONE)
4924	1	NATURAL GAS DISTRIBUTION
5080	1	WHOLESALE-MACHINERY, EQUIPMENT & SUPPLIES
5090	1	WHOLESALE-MISC DURABLE GOODS
5140	1	WHOLESALE-GROCERIES & RELATED PRODUCTS
5150	1	WHOLESALE-FARM PRODUCT RAW MATERIALS
5160	1	WHOLESALE-CHEMICALS & ALLIED PRODUCTS
5211	1	RETAIL-LUMBER & OTHER BUILDING MATERIALS DEALERS
5331	1	RETAIL-VARIETY STORES
5411	1	RETAIL-GROCERY STORES
5812	1	RETAIL-EATING PLACES
6022	1	STATE COMMERCIAL BANKS
6035	1	SAVINGS INSTITUTION, FEDERALLY CHARTERED
6036	1	SAVINGS INSTITUTIONS, NOT FEDERALLY CHARTERED
6189	1	ASSET-BACKED SECURITIES
6199	1	FINANCE SERVICES
6211	1	SECURITY BROKERS, DEALERS & FLOTATION COMPANIES
6282	1	INVESTMENT ADVICE
6311	1	LIFE INSURANCE
6324	1	HOSPITAL & MEDICAL SERVICE PLANS
7370	1	SERVICES-COMPUTER PROGRAMMING, DATA PROCESSING, ETC.
7373	1	SERVICES-COMPUTER INTEGRATED SYSTEMS DESIGN
8742	1	SERVICES-MANAGEMENT CONSULTING SERVICES

Table A2.1: Standard industry classification titles of convicted corporations

Although the total number of companies in this table is 109, there are three firms that appear in two separate industry classifications. Three of the companies shift industries within a five-year span. As a result, the data contains 106 unique firms.

A3 Correlation Table

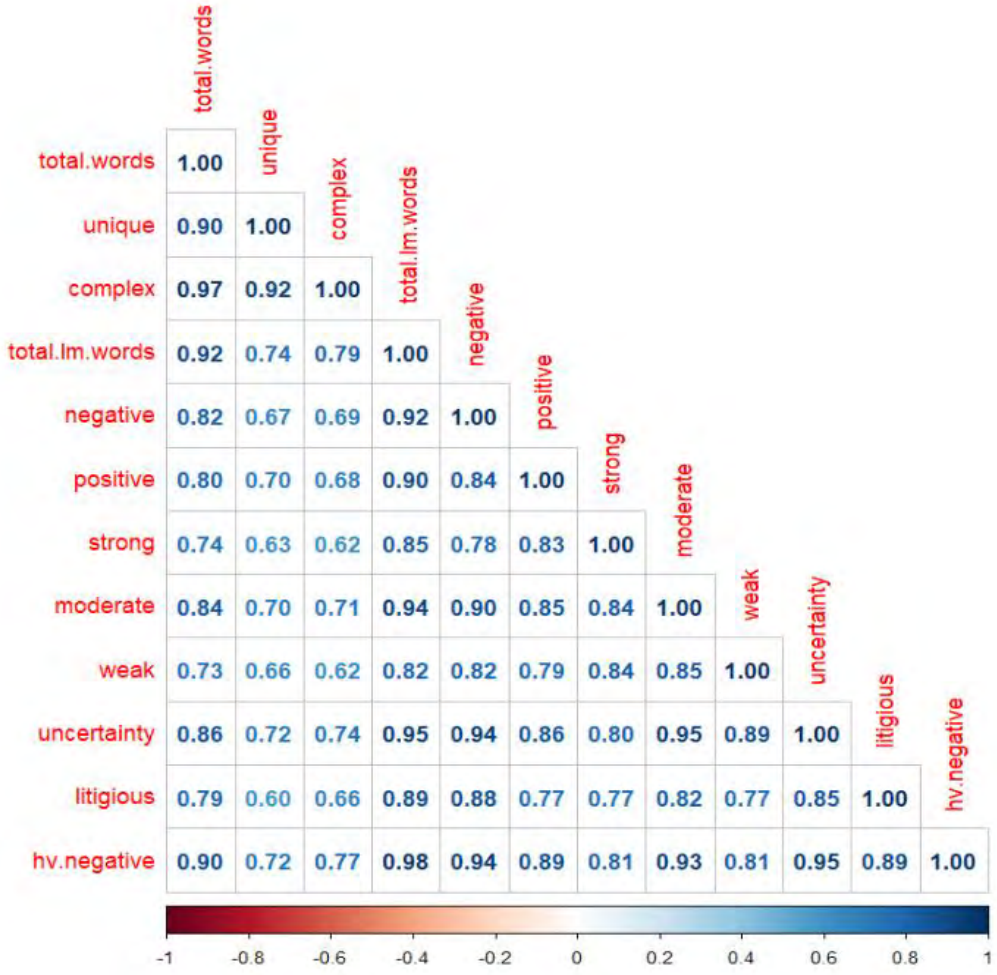

Table A3.1: Correlation table of independent variables excluding sic

A4 Regression Outcome and Odds Ratio

Regression Results

	Dependent variable:		Dependent variable:
	crime		crime
total.words	0.0002**	sic2834	0.515
	(0.0001)		(0.713)
unique	0.0004**	sic2836	0.776
	(0.0002)		(0.858)
complex	-0.0002**	sic2844	1.456
	(0.0001)		(0.950)
total.lm.words	-0.0001***	sic2870	0.624
	(0.00005)		(0.802)
negative	0.001**	sic2890	-1.079
	(0.0004)		(1.039)
positive	0.001	sic2911	0.828
	(0.001)		(0.838)
strong	0.001	sic3523	-0.069
	(0.002)		(1.310)
moderate	0.001	sic3533	0.644
	(0.003)		(0.811)
weak	-0.004***	sic3561	-0.803
	(0.001)		(1.034)
uncertainty	0.002*	sic3571	-0.543
	(0.001)		(1.289)
litigious	0.002***	sic3578	0.644
	(0.0004)		(0.957)
hv.negative	-0.0002	sic3690	1.423
	(0.0003)		(0.980)
sic1311	0.746	sic3711	-0.231
	(0.740)		(0.863)
sic1381	0.389	sic3713	0.369
	(0.760)		(1.010)
sic1389	-0.671	sic3714	0.507
	(0.928)		(0.782)
sic1531	0.421	sic3721	2.021*
	(0.873)		(1.081)
sic1600	1.807**	sic3724	1.133
	(0.918)		(0.908)
sic2015	0.969	sic3812	2.094*
	(0.828)		(1.077)
sic2070	0.867	sic3829	0.702
	(0.797)		(0.950)
sic2200	-12.076	sic3841	1.138
	(535.412)		(0.758)
sic2800	-0.413	sic3842	0.496
	(0.942)		(0.734)

	Dependent variable:
	crime
sic3845	-0.146
	(0.914)
sic4011	0.291
	(0.983)
sic4213	-0.378
	(1.088)
sic4400	-0.466
	(1.115)
sic4412	0.744
	(0.886)
sic4512	1.221
	(0.859)
sic4813	1.378
	(0.900)
sic4911	0.643
	(0.748)
sic4924	0.282
	(0.879)
sic4931	1.283
	(0.799)
sic5080	0.712
	(0.890)
sic5090	1.209
	(1.070)
sic5122	0.059
	(0.898)
sic5140	0.915
	(0.806)
sic5150	0.639
	(0.789)
sic5160	0.262
	(0.981)
sic5211	0.893
	(0.779)
sic5331	0.619
	(0.868)
sic5411	0.874
	(0.858)
sic5812	1.093
	(0.854)

	Dependent variable:
	crime
sic6021	0.747
	(0.725)
sic6022	-0.064
	(0.820)
sic6035	0.563
	(0.903)
sic6036	1.190
	(0.866)
sic6189	1.071
	(1.337)
sic6199	0.343
	(0.929)
sic6211	0.886
	(0.938)
sic6282	0.627
	(0.864)
sic6311	2.217**
	(0.957)
sic6324	-0.424
	(0.965)
sic6770	-12.317
	(377.066)
sic7370	1.279
	(0.944)
sic7372	0.891
	(0.784)
sic7373	0.498
	(0.854)
sic7389	0.357
	(0.764)
sic8742	1.251
	(0.817)
Constant	-1.662**
	(0.724)
Observations	1,246
Log Likelihood	-768.222
Akaike Inf. Crit.	1,694.445
McFadden R-squared	0.086
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure A4.1: The regression outcome of the overall model

Variables	Odds Ratio
sic1600	6.092
sic3721	7.545
sic3812	8.114
sic6311	9.182

Figure A4.2: The odds ratio of significant industry classifications

A5 10-Fold Cross-Validation: Average Accuracy and ROC-AUC Values

	Metric	Mean	Standard Error
Logistic Reg	*Accuracy*	0.640	0.0147
	ROC-AUC	0.618	0.0116
Random Forest	*Accuracy*	0.748	0.0111
	ROC-AUC	0.819	0.0117

Table A5.1: The mean and standard error of both models after 10-Fold Cross-Validation