



Master in Business Analytics and Data Science

Department of Business Administration

Thesis Title

Financial Fraud Detection

of

Evmorfia Maria Konstantinos Iosifidou

**Submitted as required for the acquisition of the postgraduate diploma in
Business Analytics and Data Science**

Supervisor: Dr. Eleftheriadis Iordanis

January 2023

Dedications

This thesis is dedicated to my loving mother.

Acknowledgments

I would like to especially thank my supervisor Dr. Iordanis Eleftheriadis for his invaluable guidance and feedback.

I am extremely grateful to my family for their endless love, care, and sacrifices throughout all these years. I would also like to thank my friends and my partner for their continuing encouragement and warm support to complete my research work.

Abstract

Data science plays a significant role in the decision-making process of businesses in today's era. The vast amount of data being generated and collected, and the ever-evolving mechanisms and software, are transforming the way the accounting and auditing operations are performed, as professionals in these fields are now utilizing data analytics techniques to analyze financial data and identify trends and hidden patterns. A critical application of data science that has attracted the attention of both the academic and the business community is the detection of Financial Statement Fraud. The integration of advanced analytics, machine learning platforms, and automated models in this area allows organizations to more efficiently determine potentially fraudulent activity and thus, be more proactive and attentive during the financial reporting practices.

To this end, this study proposes two approaches for detecting fraudulent financial statements, both of which are based on the Management Discussion and Analysis section of the annual SEC company filings. The first methodology utilizes linguistic variables, related to the context, the structure, and the sentiment of the document, whereas the second one uses the full textual information of the MD&A section in the form of words and phrases (N-grams). Natural Language Processing (NLP) tools and the Random Forest classification algorithm are employed in both models. With regards to the metrics, Accuracy, Sensitivity, Specificity, Precision and F1-score are calculated to evaluate the performance of the models. In addition to achieving the best possible predictive results, this research aims to provide specific "red-flag" indicators at a word and phrase level, which could assist the auditing decision-making procedures.

In conclusion, this dissertation forms a complete, competent and interpretable solution to the Financial Statement Fraud Detection problem. It can serve as a foundation for internal or external financial reporting audits, as well as a thorough tool for detecting fraud with the appropriate adjustments to suit the specific needs of a business based on its size, industry, and operating environment.

Keywords: financial statement fraud detection, MD&A, accounting, auditing, NLP, Machine Learning, text analytics

Declaration

I, Evmorfia Maria Iosifidou, hereby declare that all data that I acquired and analyzed in this research are presented according to the rules and principles of academic deontology and that the work presented herein is my own and has not been published or submitted elsewhere. I also affirm that I refer to all the sources of information which I have cited within this thesis and do not constitute an original creation of mine.

Table of Contents

Dedications.....	2
Acknowledgments	3
Abstract	4
Declaration	5
1. CHAPTER 1: INTRODUCTION.....	10
1.1. Background of the study.....	10
1.2. Purpose of the study	11
1.3. Research Questions	12
1.4. Structure of the study.....	12
2. CHAPTER 2: LITERATURE REVIEW.....	13
2.1. Conceptual Approach of Financial Fraud.....	13
2.1.1. Definition of Fraud.....	13
2.1.2. Financial Fraud Types	13
2.1.2.1. Financial Statement Fraud.....	13
2.1.2.2. Other types of Financial Fraud.....	14
2.2. The Auditor’s Role.....	17
2.2.1. Auditing and Fraud: Conceptual approach and audit types.....	17
2.2.2. The Auditor’s Profile: Characteristics and Commitments.....	18
2.2.3. Audit Process.....	19
2.3. Creative Accounting and Fraud.....	22
2.3.1. Creative Accounting Definitions	22
2.3.2. Creative Accounting and Financial Fraud	23
2.3.3. Creative Accounting Techniques.....	24
2.4. Techniques and characteristics associated with Fraudulent Financial Statements	27
2.4.1. Revenue-based Schemes	27
2.4.2. Asset-based Schemes.....	30
2.4.2.1. Improper capitalization of costs	30
2.4.2.2. Improper asset valuation.....	31
2.4.3. Expenses and Liabilities Schemes.....	33
2.4.4. Other Financial Statement Fraud Schemes.....	34
2.5. Financial Statement Fraud Detection	37
2.5.1. Financial Statement Fraud Detection Models.....	37
2.5.1.1. Benford’s Law	37
2.5.1.2. Beneish M-Score	39
2.5.2. The role of Data Science in Financial Statement Fraud Detection.....	41
2.5.2.1. Related Research Approaches	41

2.5.2.2.	Machine Learning Algorithms and Data Mining Techniques	49
2.6.	Financial Fraud Cases and Scandals.....	51
2.6.1.	Enron Corporation	52
2.6.2.	Parmalat.....	53
2.6.3.	Bank of Crete.....	54
2.6.4.	Folli Follie	55
2.7.	Financial Statement Fraud and COVID 19.....	57
3.	CHAPTER 3: RESEARCH METHODOLOGY	59
3.1.	Identifying the problem and setting the target.....	59
3.2.	Specifying research questions and hypotheses	60
3.3.	Choosing the Research Design.....	61
4.	CHAPTER 4: DATA ANALYSIS	64
4.1.	Data Collection.....	65
4.2.	Data Preparation	69
4.3.	Feature engineering	72
4.4.	Data Exploration.....	75
4.4.1.	Word Clouds and Bar Charts.....	76
4.4.2.	Polarity Histogram.....	79
4.4.3.	Correlation Matrix	80
4.5.	Modelling	81
4.6.	Results	86
4.6.1.	Classification Model: Linguistic Features	86
4.6.2.	Classification Model: MD&A Text.....	89
4.6.2.1.	Classification Model: Unigrams	89
4.6.2.2.	Classification Model: Unigrams, Bigrams and Trigrams	91
4.6.2.3.	Classification Model: Bigrams, Trigrams and 4-grams.....	93
5.	CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS.....	97
5.1.	Discussion	97
5.2.	Limitations of the study.....	99
5.3.	Recommendations for future research.....	100
	References	101
	Appendix	110

List of Tables

Table 1:Full Labelled Master Dataset of SEC Filings.....	68
Table 2:Dataset with raw text data from API	69
Table 3:Flesch Reading Ease Score Levels (Wikipedia).....	73
Table 4:Summary of the features created in the Feature Engineering step.....	75
Table 5:Hyperparameter tuning values.....	84
Table 6:Classification metrics	85
Table 7:Optimal hyperparameters of the first model.....	87
Table 8:Prediction metrics of the first model	87
Table 9:Optimal hyperparameters of the Unigrams model.....	90
Table 10:Prediction metrics of the Unigrams model	90
Table 11:Optimal hyperparameters of the Unigrams/Bigrams/Trigrams model	92
Table 12:Prediction metrics of the Unigrams/Bigrams/Trigrams model.....	92
Table 13:Optimal hyperparameters of the Bigrams/Trigrams/Fourgrams model.....	94
Table 14:Prediction metrics of the Bigrams/Trigrams/Fourgrams model	94

List of Figures

Figure 1:Flexibility in accounting. (Jones, 2011)	23
Figure 2:First, Second, Third, and Fourth Digit Proportions of Benford's Law (Nigrini,1996)	38
Figure 3:FFG Sourcing FS 2017 Differences (FF Group, 2018).....	56
Figure 4:Flow diagram of proposed research design	63
Figure 5:AAER Dataset.....	66
Figure 6:Master Data TSV File for SEC Filings of Q1 2012	67
Figure 7:Full Master Dataset of SEC Filings	67
Figure 8:Raw text data sample before preprocessing	70
Figure 9:Text data sample after preprocessing	71
Figure 10:Flesch Reading Ease Score Formula (www.readable.com)	73
Figure 11:Gunning Fog Score Formula (www.readable.com)	73
Figure 12:Word Cloud of 100 most common phrases of Fraudulent MD&A	76
Figure 13:Bar chart of 20 most common phrases of Fraudulent MD&A	77
Figure 14:Word Cloud of 100 most common phrases of non-Fraudulent MD&A.....	77
Figure 15:Bar chart of 20 most common phrases of non-Fraudulent MD&A.....	78
Figure 16:Bar chart of 20 most common phrases: Comparison of the two classes	79
Figure 17:Histogram of Polarity scores: Comparison of the two classes	79
Figure 18:Histogram of Subjectivity scores: Comparison of the two classes.....	80
Figure 19:Correlation Matrix of the Linguistic Features.....	81
Figure 20:Feature importances of the first model.....	88
Figure 21:Feature importances of the Unigrams model	91
Figure 22:Feature importances of the Unigrams/Bigrams/Trigrams model	93
Figure 23:Feature importances of the Bigrams/Trigrams/Fourgrams model.....	95

CHAPTER 1: INTRODUCTION

1.1. Background of the study

The business world of the modern era faces a number of challenges as a result of the immense technological advancements of the last twenty years (i.e. IoT, Means of communications and information sharing, Remote working, Cloud computing and digital transformation tools, the use of AI etc). In particular, the intense competition, the need for continuous development, the aim to extend business operations, hit market targets and increase profitability are the key drivers of today's enterprises. In this path towards accomplishing efficiency and growth, companies may often proceed to illegal or unethical actions, contrary to regulatory framework, standards and code of corporate ethics, which can broadly be described under the term 'Fraud'.

From a financial perspective, Fraud can take several forms, including Bribery, Embezzlement, Tax Fraud, Financial Statement Fraud etc. According to ACFE, although asset misappropriation and corruption activities are carried out with a greater frequency, the losses associated with the Financial Statement Fraud are significantly more severe (ACFE Report to the Nations, 2020). Indeed, the Financial Statements, being the main reports to depict a company's performance and position, constitute a critical source of information and decision-making mechanism for the investors, creditors, shareholders and other authorities. In the event of financial instability, reduced profitability or deteriorated operating capability and position, management is often motivated or forced to resort to fraudulent financial reporting practices, in order to beautify the financial results and distort the company's true picture.

Following the reveal of multiple Financial Statements Fraud scandals, stricter enforcement actions, regulations and monitoring practices have been applied with regards to the corporate financial reporting activities. Despite these impositions, as well as the internal control mechanisms and external audit procedures carried throughout the accounting cycle, fraudsters still find clever ways to falsify the financial figures. To this direction, there have been major developments in the techniques and methods to support both the prevention and the detection of fraudulent financial statements.

The academic community has expressed great interest in the use of data mining models and mechanisms for the financial statement fraud detection. Data mining includes the application of data-driven approaches, such as statistical, machine learning, artificial intelligence techniques etc, in order to extract meaningful insights and conclusions to a problem. Being able to process different types and a large number of records, data-mining models manage to discover knowledge which can be utilized to predict the most probable outcome of an event, e.g., whether a company's financial statements are fraudulent or not.

The subject of financial statement fraud detection has been examined by the research community from multiple aspects and using different kinds of data and algorithms. While initially the academics focused on the analysis of numerical data and financial ratios, their attention was soon shifted towards the textual information derived from the Management Discussion and Analysis (MD&A) section of the financial statements, which provides information regarding management's view of the reporting company's risks, future plans and prospects.

Given that in most cases senior management is actually aware and involved in the fraudulent financial reporting activities of a business and considering that MD&A contains managers' opinions, predictions and presumptions, it has been observed that this particular section can be highly informative and provide linguistic cues that facilitate the identification of fraud.

1.2. Purpose of the study

On a theoretical basis, the purpose of this study is to describe and present in detail the conceptual framework of Financial Statement Fraud and the role of Data Science in the Detection of Financial Statement Fraud. Specifically, this research covers the definition of Financial Fraud and the determination of the categories associated with it, the identification of the Auditor's role, the description of Creative Accounting, the analysis of the techniques and features of Fraudulent Financial Statements, the examination of the detection models and methodologies applied by the research community, as well as the presentation of important Financial Fraud scandals at both national and global level.

In addition, this study aims to apply diverse data mining models and examine their efficiency in detecting fraudulent companies, using the textual information included in the MD&A section of their published Financial Statements. The proposed methods, not only demonstrate a promising predictive performance but also provide specific word and phrase elements, which could be perceived as indicators or “red flags” of fraudulent financial reports.

1.3. Research Questions

In line with the above points, the following research questions are posed:

- Research Question 1: Is the analysis of the textual data reported in the MD&A section of the Financial Statements effective at detecting Fraudulent Financial Statements?
- Research Question 2: Do the linguistic features extracted from the MD&A text (sentiment, fog index, Flesch-ease score, the proportion of compound words, the average length of sentence, the proportion of positive/negative/litigious/uncertainty words) present significant predictive power in Financial Statement Fraud Detection?
- Research Questions 3: Can the proposed models provide “red flag” indicators on word and phrase level to assist with the decision-making related to the Financial Statement Fraud Detection?

1.4. Structure of the study

This study is structured as follows: This section (Chapter 1) introduces the background, the subject, and the research questions of the thesis. Chapter 2 includes a review of the relevant literature, describing Financial Fraud and Types, the role of the Auditor, Creative Accounting, Methods and Characteristics of Fraudulent Financial Statements, Models and Techniques for Financial Statement Fraud Detection, and notable Financial Fraud Cases. Chapter 3 provides a detailed discussion of the research methodology used in this study. Chapter 4 describes in detail the empirical model applied, by exploring the four main stages of a Data Science problem (Data Collection, Data Preparation, Data Exploration, and Data Modelling). The model results are then presented. Finally, in Chapter 5 the conclusions, limitations of this study, and recommendations for future research are highlighted.

CHAPTER 2: LITERATURE REVIEW

2.1. Conceptual Approach of Financial Fraud

2.1.1. Definition of Fraud

Fraud is a complex and broad concept, which encompasses a multidimensional range of forms, actions, and techniques, thus, it can be explained through several definitions. Corporate Finance Institute (CFI) defines fraud as the act of deceiving, with the use of illegal practices by an individual with the ultimate aim of obtaining a benefit or value. According to the Association of Certified Fraud Examiners (ACFE), fraud “includes any intentional or deliberate act to deprive another of property or money by guile, deception, or other unfair means”. Wang et al. (2006) incorporate into the aforementioned definition the intention of the fraudster to gain unauthorized financial advantage, whereas Rubasundram (2015) adopts the view of AICPA that fraud leads to the loss of the victim and the gain of the perpetrator.

2.1.2. Financial Fraud Types

In the context of financial activities, fraud can be recognized in different instances, which all fall under the scope of financial fraud. Ngai et al. (2010) classify financial fraud into four main categories, including Bank Fraud, Insurance Fraud, Securities and Commodities Fraud, and Other Related Financial Fraud, which comprises corporate and mass marketing fraud. Similarly, according to Sadgali et al. (2019), financial fraud includes Insurance Fraud, Securities and Commodities Fraud, Money Laundering, Financial Statement Fraud, Credit Card Fraud, and Mortgage Fraud. From a more crime-based perspective, Cerullo et al. (1999) recognize Misrepresentation of Material Facts, Failure to Disclose Material Facts, Embezzlement, Larceny, and Bribery as classes of financial fraud, whereas ACFE distinguishes Internal Fraud, consisting of Corruption, Asset Misappropriation, and Financial Statement Fraud, External Fraud, including Tax or Loan Fraud and Fraud Against Individuals, e.g. Ponzi schemes.

2.1.2.1. Financial Statement Fraud

Financial Statement Fraud has attracted increased interest in the research community over the years, since it has been associated with remarkable accounting scandals

worldwide and generally affects many business parties, resulting in significant corporate losses. “Management Fraud” or “Corporate Fraud” have been widely used in the literature as alternative concepts of Financial Statement Fraud because management executives are usually aware of or amenable to it (Goel et al., 2010). One can summarize Financial Statement Fraud as the intentional misrepresentation of a company’s financial position by the deliberate misstatement or omission of material facts in financial reports, in hopes of deceiving the interested parties (ACFE). Nguyen (2010) suggests that fraudulent financial reporting arises from the premeditated failure to prepare financial statements conforming to the generally accepted accounting principles, whereas Spathis (2002) considers financial reports as falsified when their elements distort their real picture. According to Rezaee (2005), financial statement fraud is an intentional attempt by a clever team of knowledgeable perpetrators to deceive the users of financial reports, especially investors and creditors, with the use of carefully designed practices. Likewise, Nieschwietz et al. (2000), state that “typical frauds involve scheming by highly motivated, clever teams of knowledgeable managers with the capacity for considerable persuasion and intimidation of both their employees and their auditors”.

It is therefore clear that the falsification of financial statements is a fraud deliberately committed by the top-level management of an organization in order to improve its financial positions and results, by altering or concealing specific financial elements. Furthermore, the “fraud triangle” components (Cressey, 1950) can be applied to financial statement fraud as follows: First, motives or pressures to misstate or omit material financial statements information are imperative, second, there must be opportunities for the execution of these actions and third, there should exist beliefs and perspectives, which lead one to consciously perpetrate and rationalize such an act (Montgomery et al., 2002).

2.1.2.2. Other types of Financial Fraud

Apart from Financial Statement Fraud, we can distinguish additional subcategories of Financial Fraud, following the example of the aforementioned researchers:

- A. Bank Fraud: According to the Title 18 of the U.S. Code (18 U.S.C. § 1344), “someone commits Bank or Financial Institution Fraud if they “knowingly execute, or attempt to execute, a scheme or artifice to:”

- a. defraud a financial institution; or
- b. obtain any of the money, funds, credits, assets, securities, or other property owned by, or under the custody or control of, a financial institution, by means of false or fraudulent pretenses, representations, or promises.”

The broad category of Bank Fraud involves Credit Card Fraud, which is associated with the unauthorized use of a payment card against the interest of its beneficiary, and Identity Theft, which occurs when a person uses another person’s data, such as their name, date or place of birth, to make bank applications, e.g. to open an account or obtain a credit card. An additional subclass of Financial Institution Fraud is Money Laundering, which generally refers to the process of concealing any illegal income and subsequently passing it through the financial system, so that it appears legitimate (Zdanowicz, 2009). Furthermore, Mortgage Fraud is one of the most common bank crimes and includes the use of false statements or omissions to obtain a mortgage loan relied upon a lender (FBI). Last but not least, the major rise of digital banking has provoked Online Banking Fraud, which is often carried out employing phishing, cyber-attacks, and malware infection practices, leading to significant financial costs (Wei et al., 2013).

B. Insurance Fraud: Insurance Fraud can be defined as any deceptive action committed against or by an insurance organization or agent with the aim of financial benefit (Insurance Information Institute). As Ngai et al. (2010) state, this type of fraud can happen in every phase of the insurance process and by various individuals, including consumers, brokers, and healthcare providers. Typical examples of insurance fraud are life insurance fraud, including the misstatement of a person’s health, income, or other personal information as a means to get a lower premium, and health care insurance fraud, which is related to health care benefits claimed by deceivers and automobile insurance fraud, including staged accidents and collisions and fake traffic deaths or injuries (Wikipedia, Ngai et al. (2010)).

- C. Tax Fraud/Evasion: One can describe tax evasion as the fraud perpetrated by individuals or organizations, with a view to the reduction of their tax obligations, by understating incomes, sales, wealth or overstating deductions, exemptions, credits (Alm & Torlger, 2011). Tax evasion is faced by many countries worldwide, and due to its impact on the economic and social infrastructure, fiscal policy, and other macroeconomic aggregates and practices, tax authorities have been putting global efforts to strengthen transparency and combat tax fraud (Benkraiem et al., 2021).
- D. Securities and Commodities Fraud: Securities and Commodities Fraud, often referred to as stock or investment fraud, includes the falsification or the improper use of confidential information by individuals, in order to make trading and investment decisions and consequently earn profit in the financial market (Criminal Lawyer Group). Some of the most common techniques used by trading fraudsters today involve “The Ponzi Scheme”, “The Pyramid Scheme”, “Foreign Currency Fraud”, “Broker Embezzlement”, “Late Day Trading”, “Advance Fee Fraud” and “High Yield Investment Fraud” (Ngai et al., 2010, Criminal Lawyer Group).
- E. Corruption Offenses: Bribery and embezzlement are noteworthy forms of financial corruption, which are equal to dishonest or illegal behavior by an entity, usually entrusted with power and authority, to acquire personal gain (World Bank, 2005 as cited in Wikipedia). More specifically, bribery is the act of offering, promising, giving, accepting, or soliciting an advantage as an inducement for influence or action in return (Black’s Law Dictionary, Deloitte). Embezzlement, on the other hand, is committed when somebody withholds illegally assets and properties entrusted to them, with the aim of the conversion of these belongings (Wikipedia). Fan et al. (2010) emphasize that bribery increases corporate costs more directly, whereas embezzlement is more easily detectable
- F. Cryptocurrency Fraud: The ever-evolving cryptocurrency market has given rise to new kinds of fraudulent activities, such as high-yield investment programs/online Ponzi schemes, mining-investments

scams, scam wallets, and exchanges (Vasek et al., 2015), in which the perpetrators exploit the cryptocurrencies' popularity, decentralized nature and lack of sufficient regulatory framework (Al-Hashedi et al., 2021), in order to gain illegally millions of dollars. Through their comprehensive analysis, Vasek et al. (2015) have found that nearly \$11 million worth of Bitcoin has been generated from 13.000 distinct Cryptocurrency investors through scams.

2.2. The Auditor's Role

2.2.1. Auditing and Fraud: Conceptual approach and audit types

The American Accounting Association generally defines auditing as the systematic process of objectively obtaining and evaluating audit evidence regarding assertions about economic actions and events, with a view to ascertain the degree of correspondence between those assertions and established criteria, for the communication of the results to interested users. Depending on the subject matter under evaluation, there are several types of audits, including the audit of financial statements, the audit of internal control over financial reporting, or the compliance audit (PwC, 2013). Additionally, an audit may be divided into an internal and an external process, with the difference between them being the identity of the people carrying out the analysis. More specifically, internal auditors are trained professionals, who are hired by the audited company and are responsible for monitoring the business processes, providing advice and consulting assistance to the employees, and preparing audit reports in accordance with the senior management instructions. On the contrary, an external audit is carried out by certified public accountants (CPA), who are independent of the audited firm and provide auditing and assurance services in compliance with the applicable accounting standards of the entity, such as the Generally Accepted Accounting Principles (GAAP) or the International Financial Reporting Standards (IFRS). In many cases, the aforementioned types are combined, so that better organized and coordinated audit services are provided to the involved entity.

In the context of an audit of financial statements, the auditor is required to conduct the audit, following the International Standards on Auditing (ISAs), the current tax legislation, as well as the rules of ethical behavior. (Negakis and

Tachynakis, 2013). According to the International Standard on Auditing 200, the main purpose is to provide interested parties and intended users with an opinion by the auditor on whether the financial statements are prepared and presented fairly, in all material aspects, in accordance with an applicable financial reporting framework, so as to enhance their degree of confidence. The standards require the auditor to obtain a high, yet not an absolute level of assurance about whether the financial statements as a whole are free from material misstatement, whether due to fraud or error. This “reasonable” degree of assurance is justified by the nature of the audit evidence, which is persuasive rather than conclusive, as well as the inevitable use of personal judgment during the audit process: thus, it should be accompanied by an acceptable level of risk taken by the auditor. In fact, the auditor provides assurance, by reducing the audit risk to an acceptable low level. Audit risk can be defined as “the risk that the auditor expresses an inappropriate audit opinion when the financial statements are materially misstated” (ISA 200) and is mitigated through sufficient and appropriate audit evidence. Sufficiency and appropriateness measure the quantity and quality of audit proof respectively and therefore are of great importance for supporting the conclusions on which the auditor’s opinion relies. To conclude, the auditor should design and conduct the audit in such a way that it meets the applicable standards and pay attention to the occurrences that may imply deceptive financial statements, however, as Toit (2008) suggests, they are not solely in charge of detecting and identifying such events. Consequently, the auditor does neither legally determine whether fraud has occurred nor is given particular legal powers, that may be necessary for such an investigation.

2.2.2. The Auditor’s Profile: Characteristics and Commitments

As defined above, the objectives of the auditor include the recognition and evaluation of the risks associated with the material financial reporting misstatements, the support of the proper audit evidence on these speculations, and the suitable responses to fraud or suspected fraud identifications. Apart from the required knowledge and skill adequacy, certifications, and due diligence, the auditing job depends on specific qualities and operations. More specifically, a skilled auditor should first and foremost be independent, in the sense that is as objective as possible and is free of any undue influence. Furthermore, an auditor should demonstrate integrity and commitment to quality and continuous development, based on an understanding of the entity and its

environment, including internal control activities. It is also critical for an auditor to be insightful and act proactively, as well as to communicate efficiently the audit findings to the interested parties, such as the financial statement users, the management, or those charged with governance.

Concerning the plan and conduct of the audit, an auditor should proceed to judgments over the concept of materiality, which determines the responsibility of observing and reporting the perceived misstatements. Of course, the accepted materiality levels depend on the surrounding circumstances and are influenced by the auditor's perception of the financial information needs of the users of the financial statements, and by the size or nature of a misstatement. It is hence clear that professional judgment, encompassing relevant expertise and experience, is a key component of the audit occupation.

One of the most significant and at the same time challenging mentalities an auditor should adopt is professional skepticism, namely an ongoing questioning mindset and critical evaluation of the audit evidence, acknowledging that a material misstatement due to fraud is constantly possible, despite any past experience with the company's management honesty and integrity. A qualified auditor has the ability to think in a critical manner about how changes in risks and opportunities of the business environment can affect an entity's financial statements or may indicate possible fraud, develop proactive audit techniques, and reach well-informed professional decisions.

2.2.3. Audit Process

In general, a complete audit process involves 5 fundamental stages, which according to PwC (2013) can be summarized as follows:

- a. Planning:** This step typically includes the formal signing of the contract between the audit firm and the client-audited company, the assessment of compliance with the independence and ethics standards, the appointment of the audit team, and the determination of the nature, timing, and extent of audit operations, so that the audit is conducted in an effective manner. All of the above points are mainly described by the ISA 210 and ISA 300, which specify the set of actions to be followed during this stage.
- b. Risk assessment:** In this phase, auditors use a plethora of information, and follow various procedures, in order to identify and assess the risks that

could lead to material misstatements in the financial statements. This step requires a good understanding of the business, the industry, and the wider environment in which the audited firm operates and is significantly affected by the knowledge and experience of the auditors. At the beginning of this stage a discussion among the audit team members is crucial, so that issues regarding the susceptibility of the financial statements to material misstatement, the appropriate response to such events as well as the efficient communication of the results of audit procedures are agreed and predefined (ISA 315, ISA 240). Common risk assessment processes include:

- i. Inquiries of management regarding the financial statements risk evaluation, identification, and response actions, as well as the communication of these practices to the directors and employees of the company.
 - ii. Understanding of the supervision activities of management's processes by those charged with governance.
 - iii. Evaluation of unusual or unexpected relationships that may indicate risks due to fraud.
 - iv. Consideration of other available information concerning the entity and its environment.
 - v. Assessment of Fraud Risk Factors, which are often present in circumstances indicating incentives, pressures, or opportunities for committing fraud.
- c. **Audit Strategy and Plan:** This stage begins with the development of an overall audit strategy for handling the observed risks of financial reporting misstatements. In particular, the audit team is called upon to identify and specify the number of several issues such as the selection of the proper audit evidence and analytical procedures, the design of testing actions over the various financial statement items, the set of a detailed timetable, the decision upon the usage of internal controls during the audit process as well as the task allocation to the team members. It is worth noting that the audit strategy is continually reevaluated and adjusted, depending on the new conditions and information arising from the process (ISA 300).

- d. Gathering audit evidence:** During this phase, the audit team gathers and evaluates all-sufficient and appropriate evidence, by validating the figures and disclosures of the financial statements to the entity's accounting books and records, testing the available internal control mechanisms, and evaluating management's representations and assumptions used in financial reporting. As mentioned above, throughout this process, auditors apply their professional skepticism and judgment, so as to carry out the audit efficiently. According to ISA 240, if the auditor doubts the authenticity of a document, the auditor shall investigate further these cases, e.g., by confirming directly with the third party.
- e. Finalization:** Finally, the auditors, based on the aforementioned operations, reach an overall conclusion, which supports the formation of the audit opinion.

According to the Association of Chartered Certified Accountants (ACCA), when an auditor assesses that the financial statements are presented, in all material aspects, according to the applicable financial reporting framework and are free from material misstatement, then they issue an "unmodified opinion". Otherwise, the auditor expresses a "modified opinion", which can be one of the following types (ISA 705):

- i. The "qualified opinion", which states that the observed misstatements or the possible effects of undetected misstatements are material, but not pervasive to the financial statements.
- ii. The "adverse opinion", with which the auditor concludes that misstatements are both material and pervasive to the financial statements.
- iii. The "disclaimer of opinion": In this case, the auditor is unable to form an opinion on financial statements, due to either lack of sufficient and appropriate audit evidence or multiple uncertainties in extraordinary cases.

As technology evolves and available analysis tools increase, audit firms are advancing their auditing process, by using new and innovative techniques and operations, namely (EY,2020):

- a.** The use of data analytics for fraud detection
- b.** The utilization of news and social media information

- c. The crosscheck of audit evidence with electronic confirmation documents
- d. Forensic analysis for the detection of potential fraud opportunities
- e. Mandatory fraud training for audit professionals

2.3. Creative Accounting and Fraud

Di Lullo (2006) has likened financial reporting to art, suggesting that although two entities may be similar in terms of financial performance, one could present profits and the other losses, with both approaches being acceptable at the same time. How could this be possible? Creative accounting is the answer.

2.3.1. Creative Accounting Definitions

According to Kamal Naser (1993), creative accounting appears, when accounting figures are manipulated, by taking advantage of the omissions of the accounting principles or even ignoring them, with the intention of presenting financial statements according to management's interests. Similarly, Jones (2011) believes that companies exploit the flexibility of accounting in order to fulfill the financial statements preparers' needs regarding financial reporting and presentation, without stepping outside the regulatory system. On the contrary, Mulford & Comiskey (2002) explain creative accounting as including fraud, along with aggressive choices within and beyond the generally accepted accounting standards. Baralexis (2004), recognizes two kinds of creative accounting: a) Legitimate, which exploits the weaknesses of accounting rules and regulations, and b) Illegitimate, which violates the accounting legal framework. Both types aim to modify the financial picture of a company in its favor and can be applied at the same time, with the former being the most common. It is hence understandable, that the opinions of researchers regarding the matter of distinction between creative accounting and financial fraud differ. In the American accounting system, creative accounting is considered illegal and is generally described under the terms of "income smoothing", "cosmetic accounting" or "financial engineering" (Ciocan, 2018), whereas in European Countries it is not perceived as fraud.

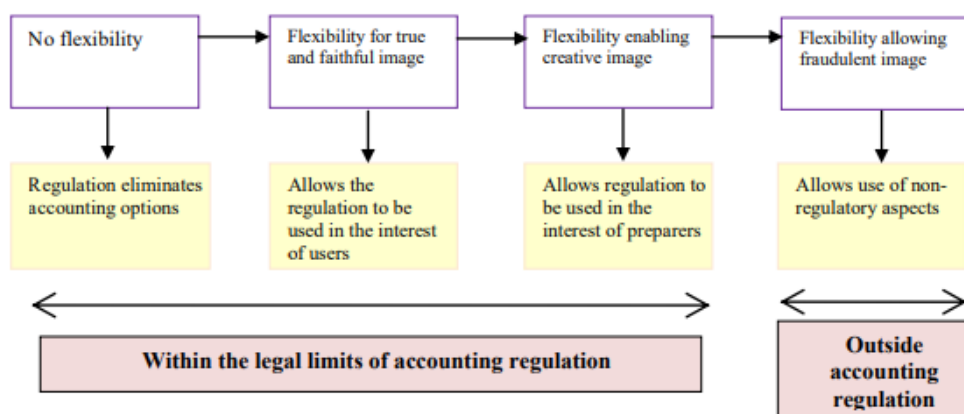


Figure 1: Flexibility in accounting. (Jones, 2011)

2.3.2. Creative Accounting and Financial Fraud

Following the aforementioned views of the scientific community, there is clearly a fine line between “Creative Accounting” and “Fraud”. As Jesus et al. (2020) suggest, it is pretty difficult to determine the exact barriers between the two practices, concluding that creative accounting is placed at the intersection between financial crime and ethical business behavior.

Ciocan (2018) reasonably observes that both concepts require from the involved parties a decent level of creativity and inventiveness, in order to be properly applied. Business professionals need to demonstrate an innovative mindset and apply crafty techniques, either fraudulent or not. To that end, both creative accounting and financial fraud denote dishonest practices and form an explicit violation of ethics in the accounting and financial reporting industries. As Sabau et al. (2020) mention, they are considered intentional actions, which usually appear in periods of financial distress. Furthermore, the two practices undoubtedly distort the true and fair financial picture of the company, by manipulating figures with a view to satisfying management’s benefits. In case the financial statements falsification is revealed, there are similar consequences between the two methods, including financial costs and losses, reputational damage, and business operational crisis.

The main difference between fraud and creative accounting, according to Jones (2011), is that the latter “is operating under the regulatory system”. Indeed, the two practices may both disregard the spirit of the law, but typically creative accounting does not infringe the letter of the law. Furthermore, while fraud is perpetrated exclusively in bad faith, creative accounting may also be applied in good

faith, e.g., if the “creative” use of accounting standards’ flexibility results in appropriate financial values.

2.3.3. Creative Accounting Techniques

The literature describes a wide variety of methods, that managers and accounting professionals can apply, so as to manipulate financial statements without violating the applied accounting framework and regulations. Breton and Stolowy (2008) suggest that creative accounting is accomplished through four main techniques:

- a. Big Bath, which is used when management records a large one-time write-off, often in a poor-performance reporting period or in a period with unusual events (e.g., a merger process or changes in ownership interests), in order to artificially inflate earnings in future periods.
- b. Earnings management, a strategy aiming to manipulate earnings to the desired level
- c. Income smoothing, with which managers wish to present stable earnings and reduce fluctuations of income from one period to another
- d. Window dressing, which includes a set of adjustments applied, so that the financial statements are presented in a more favorable way

Many researchers use the above concepts interchangeably or regard them as components of one another.

A different taxonomy of creative accounting methods recognizes four wider categories (Kevin, 2003; Ashok, 2015):

- a. Accounting policies-based techniques, in which creative accounting is based on the flexibility, the lack or the complexity of the various accounting regulations
- b. Management’s abuse of judgment, mainly in areas where estimations and assumptions are acceptable
- c. Artificial transactions, which are created so that balance sheet figures are modified or earnings are transferred between accounting periods
- d. The choice of timing for genuine transactions, which can result in a more favorable picture of specific accounts

Finally, Jones (2011) specifies five strategies of creative accounting, each of which incorporates several techniques, as described below:

1. **Income Increase:** Increasing sales is one of the most common practices of creative accounting, which is basically used to inflate profit in the statement of comprehensive income. It is usually carried out, using one of the following tactics:
 - i. Recognition of “uninvoiced” sales
 - ii. Increase of interest receivable
 - iii. Recognition of non-operating profits
 - iv. Treatment of loans as sales
 - v. Utilization of Swaps

2. **Expenses Decrease:** The second strategy used to increase profit is to reduce expenses, by applying one of the methods below:
 - i. Use of provision accounting
 - ii. Tax reduction schemes
 - iii. Big bath
 - iv. Increase of inventory closing balance
 - v. Expense capitalization
 - vi. Increase of estimated useful life of asset/ Reduction of depreciation expense
 - vii. Avoidance of recognizing provision for bad debts

3. **Assets Increase:** Management proceeds to the increase of assets, so as to strengthen the entity’s net worth. This can be achieved by following the approaches further down:
 - i. Boost of Goodwill
 - ii. Increase of Intangible Assets (e.g., Brand value)
 - iii. Revaluation of Fixed Assets
 - iv. Mark to market/ Fair Value accounting

4. **Liabilities Decrease:** The second used to enhance a company’s net worth is to decrease liabilities, often by:

- i. Off-balance sheet financing
 - ii. Reclassifying Debt as Equity
5. Operating Cash Flows Increase: This strategy aims to enhance the cash flow statement of a company and consequently boost its short-term liquidity. As cash flows from investing or financing activities are generally more difficult to manipulate, managers focus on operating cash flows, either by maximizing operating cash inflows or by minimizing operating cash outflows.

Although the aforementioned techniques refer to creative accounting, they could also be considered as methods of financial statement fraud, in case they were applied in ways that infringe on the accounting framework and regulations.

2.4. Techniques and characteristics associated with Fraudulent Financial Statements

In line with the aforementioned creative accounting techniques, financial statement fraud is typically committed through specific methods, which can be classified into four main categories (Joseph T. Wells, 2017; Gerard M. Zack, 2012):

- i. Revenue-based Schemes
- ii. Asset-based Schemes
- iii. Expenses and Liabilities Schemes
- iv. Other financial statement fraud Schemes, including Consolidation practices and Disclosure fraud

These techniques are usually combined with the falsification or intentional omission of material financial records, accounts, transactions, and supporting documents, the deliberate misapplication of accounting standards and recognition, measurement and reporting principles as well as the inadequate financial reporting disclosures (Zhou and Kapoor, 2011).

2.4.1. Revenue-based Schemes

The Anti-Fraud Collaboration, analyzing data of SEC enforcement actions on fraudulent financial statements (2021), found that the most common financial statement fraud scheme involves misstatements of revenue (43%). Considering the wide variety of methods associated with improper revenue recognition, it has been the most significant fraudulent practice consistently over the years, as several reports suggest.

One of the most typical revenue manipulation schemes includes the improper timing of sales recording, namely the shifting of revenues between different accounting periods, either prematurely or belatedly. Companies are often tempted to boost their net profits during a reporting period at expense of previous or next ones, failing to meet the revenue recognition criteria strictly defined by U.S. GAAP and IFRS. In most cases, they engage in premature timing schemes, described as follows (Deloitte, 2009; Joseph T. Wells, 2017; Gerard M. Zack, 2012):

1. Backdating, namely recording sales with the use of a falsely modified document date, which is prior to the actual period when revenue is realized and earned.
2. Improperly holding accounting periods open, so that subsequent revenue transactions are recognized in the current reporting period.
3. Inflating the percentage of completion in long-term contracts revenue recognition.
4. Utilizing various shipping schemes, either by creating phony shipping documents or by proceeding to the early shipment of goods, usually at the end of the reporting period and before the actual finalization of the transaction event and typically ensuring an extended transit or a delayed delivery period, often with the use of in-between depositories.
5. In a similar fashion, manipulating “bill and hold” transactions, in which companies are tempted to book sales orders prior to the delivery and the transfer of ownership of the products held to the customers.
6. Making side agreements, namely in the form of sales incentives, providing rights of return, extended credits, and refunds or engaging in consignment sales, and failing to apply the proper accounting.
7. Channel stuffing, which refers to the sale of an unusually large amount of a product to distributors, often accompanied by large discounts and lengthy payment terms. Although this practice may not always indicate fraud, it should be viewed skeptically, as in many cases it does not meet the criteria for revenue recognition until certain terms are completed.
8. The recognition of the total amount of up-front or one-time initiation fees in a single accounting period, before the corresponding services are performed.
9. Falsely identifying a multiple element transaction or improperly allocating revenues into segments, in favor of those recognized first and at expense of those that can be deferred in subsequent periods.
10. Failing to record sales reductions or liabilities of future obligations, arising from customer loyalty products, future sales returns, or exercise of customer incentives, e.g., gift cards, discount coupons, etc.

In cases of an extremely profitable current reporting year, businesses are often motivated to push current revenues to future accounting periods, to maintain their success and secure themselves from potential future losses. These schemes often involve the recording of income in a phony reserve or a deferred revenue account, the falsification of invoice documents, or the deliberate delay of invoicing of trade receivables. Furthermore, the use of estimates and provisions through the reporting process enables managers to create excess reserves as liabilities, i.e. “cookie jar” reserves, which are easily converted to income in subsequent unprofitable periods.

Sales can be also manipulated using fabrication techniques, either by creating fictitious revenue or by inflating income. More specifically, in order to improve profitability, conceal losses or simply appear larger, a company may create artificial invoices and transactions, with real, fictional, or even without corresponding clients. The falsified transactions are often booked in the reporting period and reversed in the subsequent. Many times, this type of fraud includes also the recording of fictitious cost of goods sold so that gross margin remains analogous and increased revenue does not raise any suspicions. In addition, inflated sales can be part of authentic transactions, by altering amounts, namely line items, product quantities, and sales prices in the original documents. This type of scheme is typically more difficult to detect, as it includes legitimate customers, as opposed to fictitious transactions, in which there are more conspicuous signs of fabrication. In most cases though, businesses act artfully and methodically, so that everything in the accounting system seems legit. (Joseph T. Wells, 2017; Gerard M. Zack, 2012)

Misclassification schemes form another type of revenue manipulation techniques, and typically include the financial presentation of amounts in misplaced line items in the Financial Statements. Compared to the aforementioned methods, they are associated with a lower cost (Sarah Elizabeth Mcvay, 2006), since they do not change bottom-line profits. They are usually operated by misclassifying extraordinary or one-time income into operating sales, which are supposed to include revenue exclusively from core business activities. This mainly impacts gross profits and margins as well as the expectations of the investors, since a false picture of business operations is presented (Gerard M. Zack, 2012). Additionally, misclassification of revenue may occur, when a company reports a received purchase incentive (e.g., cash received from a vendor) as income, instead of decreasing the cost of sales.

Correspondingly, a vendor may recognize such an incentive as a cost, instead of decreasing revenues, as supposed.

Many times, companies aim to appear larger and present a higher volume of business activity, in order to meet market expectations and investors' needs (Gerard M. Zack, 2012). To achieve this, they often result in gross-up schemes, in which a greater number of transactions is recorded, involving both sides of income and expenses. Although net profits are not affected, the size and business potential of the involved company are distorted. Gross-up schemes are generally perpetrated in terms of recycling products, services, or cash through barter or round-trip transactions, often utilizing subsidiaries or affiliated companies. Additionally, when an "agent" company records income and costs as a transaction principal, gross-up fraud has been carried out. In many cases, the easiest way to inflate Income Statement is simply to record artificial, equal amounts of revenue and expenses simultaneously.

2.4.2. Asset-based Schemes

A significant area of financial reporting fabrication includes the overstatement of asset accounts. Especially working capital balances, namely trade receivables and inventories, as well as tangible and intangible assets valuations, which are dependent on a variety of management's estimations, provisions, and assumptions, are highly prone to manipulation (Marilena & Corina, 2012). In general, asset-based schemes can be divided into two wide categories:

1. Schemes related to improper capitalization of costs
2. Schemes related to improper asset valuation

2.4.2.1. Improper capitalization of costs

Improper cost capitalization occurs when a company decides to convert certain expenses to assets, without them being eligible for capitalization. This accounting practice enables businesses to amortize such expenditures over more than one accounting period, instead of recording their full amount in one single period, with a clear impact on net profits (Deloitte, 2009). Although expense capitalization is acceptable under some circumstances, there are specific cost categories, which are extremely susceptible to manipulation (Gerard M. Zack, 2012):

- i. Start-up costs, related to one-off expenses during the start-up phase of a new entity, are supposed to be expensed as incurred, not capitalized.
- ii. Research and Development Costs, associated with the discovery of knowledge, and the planning and development of new products, are supposed to be expensed when incurred, except if they concern activities with an alternative future use.
- iii. Some expenses relating to tangible assets, e.g., repair or maintenance costs, which do not expand the asset's useful life or increase its performance and thus, are not suitable for capitalization.
- iv. Software development and acquisition costs, which should be recorded as expenses, if they include indirect expenditures, which are not related to software advancements or when they precede the technological feasibility phase.
- v. Website costs, which are only allowed to be recognized as assets during the application and content development stages.
- vi. Advertising costs, which ought to be expensed either when occurred or at the first time of advertising, unless they are for direct-response purposes.
- vii. Prepaid expenses, for goods or services already delivered in the current period.
- viii. Capitalization of labor or other costs as inventories, in cases of manufacturing entities.

2.4.2.2. Improper asset valuation

Even though accounting standards define explicitly the initial recognition, valuation, and subsequent measurement of assets, there are still many opportunities for financial reporting fraud, mainly concerning inventory, trade receivables, property and equipment, intangible assets, cash, and investments accounts. Fraudulent activities typically involve either the modification of the value or the failure to record an impairment associated with an asset. The most common asset-based schemes, according to Deloitte (2009), Joseph T. Wells (2017), and Gerard M. Zack (2012) can be described as follows:

- i. Inventory Valuation Schemes: Companies usually inflate inventory balances, especially year-end amounts, in order to reduce the cost of sales

and thus, report higher annual earnings. One common method for inventory overstatement is the fabricated increase of inventory quantities, which can be accomplished in various ways, including the manipulation of count sheets or receiving reports, the deliberate multiple physical counting of inventory items, often combined with their transfer to multiple locations or “consignment” warehouses, as well as the alteration of vendors’ delivery notes and corresponding records. The total cost of inventory is an additional area of falsification, including the manipulation of unit prices of inventory items, the improper application of generally accepted valuation methods, intentional stock mislabeling, and the modification of suppliers’ invoices and other receiving reports. Sometimes, a company may even create fictitious inventory, supported by phony documents, fictional vendors, and imaginary line items. Finally, the postponement, the under-reporting, or complete avoidance of recording impairment losses, in case of inventory obsolescence, is another commonplace action of inventory-related fraud.

- ii. Accounts Receivables Schemes: Trade and other receivables are susceptible to two kinds of misstatement. First, they are frequently presented increased, along with fictitious client data and false balance confirmations and documentation, so as to accompany inflated revenues. Secondly, many times businesses modify document dates or proceed to specific “exchange” transactions with their customers, so that accounts receivables remain “current” as per their aging, and thus a high expected credit losses provision is avoided. In some cases, managers may completely omit to write down bad debt accounts, due to their negative effect on net profits.
- iii. Tangible and Intangible Assets Schemes: Fixed asset schemes usually encompass the measurement in a value higher than the approved acquisition cost of the asset, especially in cases of non-cash transactions, where fair value accounting is typically involved. There are circumstances, under which companies record fake assets with purchase notes, or misclassify property to false accounts, in order to meet budget expectations, skew financial ratios or satisfy debt covenants. Another usual practice is the understatement of the depreciation or amortization expense,

which negatively impacts the Income Statement. This is mainly achieved by determining an excessive asset's useful life, delaying the starting date of depreciation, or by assigning an improper salvage value. Finally, failure to record impairment losses is a common fraud technique, especially with regard to intangible assets with "indefinite" useful lives.

- iv. **Cash and Cash Equivalents Schemes:** Since liquidity is of primary importance for businesses, managers often proceed to cash manipulations, by adding back outstanding checks or "adjusting" amounts to cash balances. These activities usually engage one "reconciling" journal entry, supported with forged bank certificates or the inappropriate alteration of check registers.
- v. **Investments Accounts Schemes:** Investments and other relevant financial assets are overstated, either by modifying the existing or creating fake investment deals. Since the valuation of investments is based on their category, many entities intentionally misclassify financial instruments as "held for sale", so that their corresponding unrealized gains or losses are recorded in Other Comprehensive Income (OCI) and do not affect Income Statement. Furthermore, there are cases where losses from the impairment of investments are not recorded at all.

2.4.3. Expenses and Liabilities Schemes

Financial reporting frauds concerning expenses and liabilities are generally carried out through omissions, understatements of amounts, or manipulation of specific accounting treatments. It is typically more complex for auditors to detect liabilities-related schemes, since they usually leave no audit trail, in comparison with revenue-based fraudulent activities. (Gerard M. Zack, 2012)

Omissions of amounts are commonly perpetrated in the form of timing schemes when a company conveniently avoids recording liabilities for expenses incurred in the current accounting period. As a result, both Balance Sheet and Income Statement are beautified, with the latter reporting higher net earnings than the actual ones. Using a cash instead of an accrual basis of accounting, businesses delay the recognition of payables, until they pay off their vendors. This improper practice may involve the hiding of expense documents and may even be aided by the suppliers

under a mutual deal of invoicing postponement or with the use of non-cash arrangements. Some categories of liabilities are particularly prone to deliberate “missing” in the accounting records, e.g., the liabilities arising from the right of return or the warranty on product sales, especially in the retail industry. Another case of easily “forgotten” items is provisions enclosing contingencies, e.g., pending litigations. Due to the uncertain nature of such elements, companies frequently manipulate their recognition criteria, claiming a low probability of their occurrence or an inadequate basis for the estimation of their corresponding loss and therefore, they postpone their recording until the period of settlement. Debt obligations, guarantees, compensations relating to employee bonuses and benefits, as well as asset retirement obligations are also highly susceptible to reporting omissions. (Joseph T. Wells (2017) and Gerard M. Zack (2012))

Understating amounts is another way of falsifying liabilities, especially when estimations or fair value accounting are involved. More specifically, many times entities measure incorrectly or reduce improperly liabilities concerning compensated absences, bonuses, and retirement benefits, although they are supposed to be recognized when services are rendered. Furthermore, underreporting occurs, when liabilities, e.g., debt obligations, are eligible to be measured and carried at fair value, mainly through the manipulation of valuation techniques, parameters, and inputs. (Joseph T. Wells (2017) and Gerard M. Zack (2012))

Finally, as already mentioned in the asset-based schemes, the improper capitalization of revenue-related expenses is another method of financial statement fraud. Sometimes though, the opposite happens. Some companies proceed to expensing costs that should be capitalized, primarily for tax purposes or in order to increase future periods’ profits, when current earnings are satisfying. To achieve the latter goal, businesses may also keep inappropriately liabilities accounts open, so that they have the opportunity to utilize their reduction in subsequent accounting periods. (Joseph T. Wells (2017) and Gerard M. Zack (2012))

2.4.4. Other Financial Statement Fraud Schemes

In addition to the financial reporting fraud techniques that directly impact specific financial accounts and items, there are several other schemes, which are utilized by companies for financial statements’ fabrication purposes. These actions are mainly

related to a) Consolidated Financial Statements and Business Combinations, and b) Financial Statements' Disclosures.

Consolidated Financial Statements, which include the presentation of the financial amounts of a parent entity and its subsidiaries as a single entity, are required under specific circumstances, as defined by the U.S. GAAP and IFRS accounting standards. According to ASC 810 (US GAAP), consolidation is applied when the reporting – parent entity has a controlling financial interest in another entity under the “variable interest entity” or the “voting interest” model. Under IFRS 10, the basis of consolidation is the existence of “control” of an entity by the parent company. (RSM, 2020) In many cases, fraud occurs, when a company takes advantage of the “grey areas” of concepts such as “controlling financial interest” or “control” and deliberately fails to comply with the accounting standards. As a result, the reporting entity either includes companies, typically highly profitable, that are not supposed to be consolidated or excludes entities, usually reporting losses, that are required to be consolidated. Other common schemes incorporate the manipulation of intercompany activities, the recording of unjustified consolidation entries, and even the creation of fictitious transactions, in order to overstate or understate balances. As the fraud case of Enron shows, a company may create and transact with specific “off-balance sheet” or special purpose vehicles, so as to conceal debt or other liabilities and improve its financial position. Moreover, the merger or the acquisition and the first-time consolidation of a business pose a number of accounting treatment challenges, because of the significant estimations and judgments required for the valuation of the acquired companies. Timing schemes are also utilized in terms of prematurely consolidating a company before significant controlling interest or control is gained. Last but not least, failing to extensively disclose all of the related parties, their nature, and activities, and the corresponding transactions with the reporting entity in the financial statements is an additional form of financial misrepresentation. (Deloitte (2009), Gerard M. Zack (2012))

Financial Statements are required to be accompanied by the necessary disclosures and notes, which, as stated in IAS 1, contain information in addition to the four primary financial statements, including “narrative descriptions, disaggregations of the items presented in those statements and information about items that do not qualify for recognition in those statements”. Notes are of primary importance to the

users of the financial statements, as they provide a fuller, more detailed, and more effective view of an entity's financial health. Non-compliance with the disclosure requirements defined by the accounting principles, is a type of financial statements manipulation, with the aim to beautify the appearance of a company's finances and mislead the investors. Disclosure fraud can be perpetrated by omitting required financial data, providing inadequate or no details about specific items or events, misrepresenting information presented in notes as well as reporting confusing descriptions and explanations. (Gerard M. Zack (2012)) According to Deloitte (2009), Joseph T. Wells (2017), and Gerard M. Zack (2012), the most susceptible areas to fabrications are the disclosures of:

- i. Related parties' transactions,
- ii. Financial arrangements, commitments, and contingent, liabilities, which can impact materially the financial figures of a business once activated due to a future event,
- iii. Subsequent events, occurring or known after the reporting date, especially when they are related to negative events such as court judgments or litigation outcomes,
- iv. Changes in accounting estimates, judgments, and policies, typically concerning the useful life and residual value of assets, warranty obligations, inventory obsolescence, and receivables impairment,
- v. Changes in accounting principles and failure to retrospectively restate all affected financial statement items, as determined by the accounting standards.

2.5. Financial Statement Fraud Detection

2.5.1. Financial Statement Fraud Detection Models

In the last few decades, there has been an increasing interest in developing models and empirical methodologies for the detection of fraudulent financial statements, because it is an ever-challenging issue, often difficult to handle with ordinary audit processes, as Porter and Cameron (1987) suggest. From a clearly accounting-oriented perspective to more statistical and ratio-based analysis and subsequently, to the use of more sophisticated systems and machine learning algorithms, the research community has applied a variety of methods toward financial statement fraud detection.

2.5.1.1. Benford's Law

Benford's law, often referred to as "Newcomb-Benford law", "the law of anomalous numbers" or the "first-digit law" is a statistical approach, which was initially introduced by the American astronomer Simon Newcomb in 1881, and holistically developed and explained by the physicist Frank Benford in 1938. Benford's law is an observation that the occurrence of the first significant digits in large sets of data conforms to a decreasing logarithmic distribution, instead of being uniform. As Nigrini points out (2012), there is a larger bias toward small digits, i.e. the lower numbers (such as 1,2, and 3) tend to appear more frequently as the leading digits in datasets as compared to the higher ones (such as 7,8, and 9). Benford supported the phenomenon by analyzing more than 20.000 records of different types of datasets, related to population sizes, chemical compounds, baseball statistics, etc, and concluded to the following expected first, second, third, and fourth-digit frequencies as shown below:

Digit	Position in Number			
	1st	2nd	3rd	4th
0		.11968	.10178	.10018
1	.30103	.11389	.10138	.10014
2	.17609	.10882	.10097	.10010
3	.12494	.10433	.10057	.10006
4	.09691	.10031	.10018	.10002
5	.07918	.09668	.09979	.09998
6	.06695	.09337	.09940	.09994
7	.05799	.09035	.09902	.09990
8	.05115	.08757	.09864	.09986
9	.04576	.08500	.09827	.09982

Figure 2: First, Second, Third, and Fourth Digit Proportions of Benford's Law (Nigrini, 1996)

The explanation of Benford's law was a challenge for many mathematicians and statisticians for about 90 years (Durtschi et al., 2004). Hill's research has been highlighted by many academics; in his 1995 paper, the mathematician provided evidence that datasets following Benford's law are derived from a combination of distributions, i.e., "second generation" distributions.

Similar to other statistical models, Benford's law can be used as a tool for the exploration of unusual patterns and abnormalities in data. The researcher Mark J. Nigrini, through intensive research on Benford's Law since 1996, provided guidance on how the auditing community could apply the law to spot potential accounting and financial reporting fraud indicators. In his book "Benford's law", Nigrini (2012) examined the conformity of multiple real-world examples, including travel and expense claims, insurance refund amounts, accounts payable numbers, invoice and reported turnover amounts, vendor transactions, taxable income data, accounting entries, ledger balances and reported figures of published financial statements.

In his article in "The Journal of Accountancy," Collins (2017) also recommends the application of Benford's Law on general ledger data with the use of Microsoft Excel in combination with auditing processes. The CPA suggests that larger datasets display better conformity with Benford's Law, an opinion which is also mentioned by Durtschi et. al (2004) in their study. The latter further consider that, instead of sampling the data, the entire ledger accounts should be examined for more reliable results. Nigrini (2012) has also noted specific conformity requirements, that is

that the data should closely approximate a geometric sequence, the records should not represent identification numbers or labels and the dataset should not contain predefined minimum or maximum values, etc. In any case, it should be noted that “the law of anomalous numbers” could never be an explicit sign of proof of certain events, i.e., manipulation, and therefore, it could not determine the audit process or be used as a standalone method of fraud detection (Collins (2017), Shi et. al (2017))

2.5.1.2. Beneish M-Score

The M-Score is a statistical model, introduced by Messod Beneish in 1999, which is based on the analysis of the financial ratios of a company. Specifically, the formula consists of 8 financial indexes, which are calculated with data derived from the published financial statements and compare the performance of the business between two consecutive periods. These ratios, weighted by coefficients, are designed to capture either the financial reporting distortion as a result of earnings manipulation or to discover the tendency of an entity to engage in such practices (Dikmen and Küçükkocaoğlu (2010), Kamal et al. (2016)). Therefore, the Beneish Model can be utilized as an accounting forensic tool to detect potential earnings fabrication in the financial statements.

The calculation of the variables of the Beneish M-Score can be described as follows:

- Days Sales in Receivables Index (DSRI): It compares the ratio of the receivables to the days’ sales between the current and the previous year. A large increase of this index indicates a higher probability of earnings manipulation.
- Gross Margin Index (GMI): It compares the gross margin of the previous year to the one of the current year. Higher values of this variable suggest a higher likelihood of income fabrication.
- Asset Quality Index (AQI): It is the ratio of the reporting year’s non-current assets to total assets, excluding Property, plant & equipment, compared to that of the prior year. The index is positively correlated to potentially fraudulent activity.

- Sales Growth Index (SGI): It compares the current year's sales to that of the previous year. Higher values of this ratio increase the possibility of earning management, reflecting higher growth expectations.
- Depreciation Index (DEPI): It is the ratio of the rate of depreciation expense for the previous year to that of the current year. The correlation between this variable and the likelihood of manipulation is deemed to be positive.
- Sales General and Administrative Expenses Index (SGAI): This variable compares the SG&A Expenses between the reporting and the comparative period. Higher values of this ratio suggest a higher probability of earnings manipulation.
- Leverage Index (LVGI): It is calculated by dividing the total debt by the total assets of the current year by that of the previous year.
- Total Accruals to Total Assets (TATA): It is the ratio of total accruals, calculated as the change in working capital accounts other than cash, less depreciation, to the total assets of the current year. A large increase of this index might suggest a higher manipulation probability.

The combination of the eight aforementioned variables is applied with the following formula:

$$\text{M-Score} = -4.84 + 0.920 \cdot \text{DSRI} + 0.528 \cdot \text{GMI} + 0.404 \cdot \text{AQI} + 0.892 \cdot \text{SGI} + 0.115 \cdot \text{DEPI} - 0.172 \cdot \text{SGAI} + 4.679 \cdot \text{TATA} - 0.327 \cdot \text{LVGI}$$

According to the Beneish Model, companies with an M-Score of greater than -2.22 are likely to have engaged in manipulation activities. It should be noted that a five-factor version of the M-Score (with a benchmark of -2.76) may also be used, excluding SGAI, DEPI, and LVGI as the least significant variables. The formula of the 5-ratios Beneish model is modified as follows:

$$\text{M-Score} = -6.065 + 0.823 \cdot \text{DSRI} + 0.906 \cdot \text{GMI} + 0.593 \cdot \text{AQI} + 0.717 \cdot \text{SGI} + 0.107 \cdot \text{DEP}$$

2.5.2. The role of Data Science in Financial Statement Fraud Detection

Financial fraud poses a growing challenge for companies, which is becoming more and more complex, costly, and concerning as fraudsters are evolving their strategies and creating ever more innovative tactics. Despite the well-defined auditing standards, the stricter enforcement actions against fraudsters, and the broader awareness among businesses, perpetrators manage to be fast, adaptive, and inventive in terms of their devious behavior over time. Therefore, it is imperative that financial statement fraud detection and prevention advance in a more intelligent manner, so that the disastrous consequences of fraudulent reports are minimized. Towards this direction, the research community has developed data-driven fraud detection systems and models, which solely rely on data-mining techniques and machine learning algorithms, capable of classifying a company as fraudulent or non-fraudulent, providing early warning signs (red flags) of fraud, discovering the hidden patterns of fraudulent behavior or even predicting whether an enterprise is likely to commit fraud.

2.5.2.1. Related Research Approaches

Ravisankar et al. (2010) compared six machine learning classification algorithms in terms of their performance in predicting financial statement fraud, with and without feature selection, using a dataset of 35 financial values and ratios of 101 fraudulent and 101 non-fraudulent companies listed in Chinese stock exchange markets. More specifically, after pre-processing the data, using normalization and dimension reduction techniques, they initially fed the six classifiers, namely MLFF, SVM, GP, GMDH, LR, and PNN, with the whole set of variables and then repeated the same procedure with the 18 and 10 most important features, as extracted by the feature selection process, which was based on the t-statistic values. Relying on the AUC metric, the researchers observed that PNN (35/10 features) and GP (18 features) generally outperformed all other classifiers, with AUC being over 90% in all three cases. With regards to the most significant financial items in indicating the presence of financial statement fraud, 80% of the selected features are affiliated with the profitability of the enterprise (e.g. Gross profit/Total assets ratio) and 40% with the primary business income (e.g. Primary business income/Fixed assets ratio), revealing that companies tend to manipulate the profit or income values when committing financial fraud.

Kirkos et al. (2007) make also use of financial ratios in their study, applying three Data Mining techniques in a matched dataset of 76 Greek manufacturing companies in total, half of which were probably involved in financial reporting fraud. In particular, the researchers identified initially 27 ratios deriving from the financial statements line items, corresponding to eight wider “red-flag” indicator categories, including financial distress, debt, accounts receivables, inventory levels, etc. Using ANOVA, they reduced these variables to the 10 most important ones, which formed the input of the classification methods, namely Decision Tree applied with ID3 splitting algorithm and tree pruning, Backpropagation Neural Network, and Bayesian Belief Network. Although presenting the lowest general performance in the training set, the BBN outperformed the other methods in the stratified 10-fold cross-validated sample (90,3%), followed by NN (80%) and DT (73,6%). The Bayesian Belief Network associated fraudulent financial statements with the ratios Debt to Equity, Net Profit to Total Assets, Sales to Total Assets, Working Capital to Total Assets, and Altman’s Z-Score, whereas Decision Tree, marked mainly the latter mentioned as the most important variable.

Similarly, Wyrobek (2020), using a total of 298 normalized financial variables, as extracted from a dataset of 1317 financial statements and ratios of 54 fraudulent and 58 matched non-fraudulent companies, aimed to spot patterns that demonstrate various financial fraud types and unethical corporate culture in general, not exclusively in terms of financial reporting. The researchers applied a variety of machine learning techniques, including upsampling for balancing the model sample, 10-k fold validation as the data dividing process, and feature selection with the use of a genetic algorithm, in order to identify the 15 best performing variables. The dataset was fed in different algorithms, including Logistic Regression, Gradient Boosted Decision Trees, Random Forests, Deep Neural Network, Naïve Bayes Model, Linear Discriminant Analysis, and SVM. The best combination of metrics, namely accuracy, precision, recall, and type I and II errors, was performed by the XGB approach, followed by RF and Linear Discriminant Approach. As shown in the results, fraudulent companies reported higher values of cash flows from financing activities, discontinued operations, other equity, cash flow extraordinary items, and D/E ratio and lower values of interest and tax payments, restricted cash, total receivables, assets, and current liabilities, as well as net profits and SG&A expenses in comparison

with honest firms. They concluded that fraudulent businesses are generally related to significant financial and discontinued operations, lower liquidity and tax payments, and higher gross but lower net profits.

From a more statistical point of view, Spathis (2002) focused his research on 76 Greek manufacturing companies listed in ATHEX, aiming to assist auditors in detecting falsified financial statements, especially since the significant increase of Athens Exchange Stock listings and the tax reduction efforts of many businesses in the early 2000s in Greece. Using publicly available data and releases, the researcher first selected the 10 most effective out of 17 financial ratios, applying both correlation analysis and statistical significance tests. The chosen variables were a) Debt to Equity, b) Sales to Total Assets, c) Net Profit to Sales, d) Accounts Receivable to Sales, e) Net Profit to Total Assets, f) Working Capital to Total Assets, g) Gross profit to Sales, h) Inventory to Sales, i) Total Debt to Total Assets and j) Altman's Z-Score, covering all aspects of business performance, namely profitability, leverage, solvency and managerial performance and the method used was logistic regression. The univariate and multivariate tests highlighted NP/TA, WC/TA, INV/SAL, TD/TA, and Z-Score as the best predictive features, achieving an accuracy of over 84% and indicating that firms with lower profitability, liquidity and solvency, and higher debt and inventory levels are more likely to manipulate financial statements.

Lin et al. (2003) used a dataset of 160 non-fraudulent and 40 listed firms accused of financial statement fraud by SEC in 1980-1995, matched at a size and industry level, to evaluate the predictive performance of hybrid intelligent systems in comparison with the traditional logit approach. In particular, the researchers combined fuzzy logic with neural networks in an FNN model, considering its effectiveness in handling simultaneously complex decision rules from a human-perceived view. The variables fed into the algorithms were mainly associated with revenue recognition and accounting estimates, as these values tend to complicate audits. Although both FNN and logit models achieved a high accuracy rate of non-fraudulent cases, FNN outperformed with regards to the prediction of fraudulent firms, and had a lower overall error rate and lower costs related to type I and II errors, confirming the superiority of hybrid expert systems.

In their research, Glancy and Yadav (2010), based on the Interpersonal Deception and Media Richness Theories, proposed a computational fraud detection model (CFDM), which leveraged the textual information of the 10-K SEC MD&A Section filings for the detection of financial reporting fraud at the senior-management level. As they accurately pointed out in their paper, textual data can provide signs of deception, as long as the writer is aware of intentionally committing fraud. To this end, they used balanced datasets of the MD&A filings of fraudulent and non-fraudulent companies of the same size and sector, to test whether their text-mining-based model had the ability to cluster correctly the two cases. A significant process of their model was the data preparation and transformation, including text extraction techniques such as Stemming and Part of Speech (PoS), and subsequently the creation of Singular Value Decomposition Vectors (SVDs), which consisted of the frequency and term weights of the textual data. These vectors were analyzed with the use of hierarchical and expectation maximization clustering algorithms and the process was repeated for validation purposes. Altogether, their model managed to identify fraudulent firms with a confidence level of about 90%, contributing significantly to the financial fraud detection literature.

Similarly, Humpherys et al. (2010) using Natural Language Processing (NLP) techniques on a balanced MD&A filings dataset of 202 companies, extracted 24 linguistic variables, associated with terms such as affect, complexity, diversity, expressivity, no immediacy, quantity, specificity, and uncertainty, which are expected to be indicative of the discrimination of fraudulent and non-fraudulent reports. PCA was subsequently implemented to identify a reduced 10-variable model. These two sets of linguistic features were used as the decision variables of six machine learning algorithms, namely Logistic Regression, C4.5, LWL, Naïve Bayes, SVM, and JRip for the classification of the sample enterprises as fraudulent or truthful. The highest accuracy of 67% was achieved by the Naïve Bayes and C4.5 classifiers of the 10-feature model. The researchers confirmed their hypothesis that fraudulent MD&As contain significantly more active and complex, yet less lexical and content diverse language than non-fraudulent reports, indicating that managers tend to magnify their written statements and use irrelevant or confusing terms in order to deceive the readers of their financial statements.

Cecchini et al. (2010) created an accounting-based ontology using the MD&As of 122 instances to find specific terms that distinguish fraudulent from non-fraudulent firms. To test the effectiveness of their dictionary, they performed SVM to the labeled dataset of the token count vectors of each company and achieved 75% classification accuracy, using 200 and 500 concepts. Of these terms, the most negative, thus, the most representative of financial statement fraud, was found to be the phrase “Additional cost”. In a following comparison of the above text-based methodology with the quantitative Beneish’s fraud discrimination ratio approach, researchers concluded that, although individually the textual data were more powerful in classification than the financial ratios, their combination resulted in the best score in detecting financial fraud.

In the same year, Goel et al. (2010) examined the impact of not only the verbal content but also the presentation style of the textual information of the 10-Ks in fraud detection and detection of different stages of fraud. Using sets of pre-fraud, adv-fraud, and post-fraud data of 126 fraudulent and 622 non-fraudulent companies, they initially implemented mere NLP techniques such as Bag-of-Words, Pruning, and Stop words list in combination with machine learning algorithms, namely NB and SVM, achieving the highest accuracy of 71,67%. In addition, they extracted a complete list of linguistic variables, including length-based features, linguistic style markers and associated with the voice, tone, uncertainty, readability, content, and relevance of the documents. Interestingly, they concluded that fraudulent reports systematically differentiate from the non-fraudulent ones in terms of their textual content and writing style, consisting of more passive-voice sentences, more ambiguous words and phrases, a broader lexical variety, and a more complex substance in general. In their subsequent study (2016), the researchers performed sentiment analysis of the MD&A section of a matched sample of 360 companies, by employing the SVM classifier with lexicon-based attributes as well as more advanced linguistic features, as derived from the Part-of-Speech Tagging process. The textual sentiment was measured from the perspectives of polarity, subjectivity, and intensity, and the classification model with the highest accuracy was aroused from the combination of the aforementioned variable sets. The main findings of this study suggest that fraudulent MD&A sections, as compared to the truthful ones, are characterized by a greater positive and negative

sentiment magnitude and involve the significant and frequent use of adverbs and adjectives, indicating subjectivity and generally words giving emphasis.

In their study, Hajek and Henriques (2017) combined both financial and linguistic variables into an intelligent financial fraud system, aiming not only to identify the most important features in terms of financial statements fraud detection, but also to provide an accurate, yet interpretable model. In particular, using a matched dataset of 622 NYSE or Nasdaq listed US firms and obtaining data from both the 10-K annual reports and publicly available information from Reuters and Value line database, they identified 32 financial ratios and 8 linguistic variables. The financial ratios involve a variety of categories: firm size, corporate reputation, profitability, activity, liquidity, leverage and market value ratios, asset structure, and business status. The linguistic features mainly consist of the normalized frequency rates of “positive”, “negative”, and “uncertain” words and some finance-specific word classes, e.g., “litigious” or “constraining” phrases. After creating random stratified samples of data, the researchers proceeded to feature selection, in order to recognize the most significant set of variables, and applied 14 Machine Learning algorithms, including Logistic Regression, Bayesian classifiers, SVMs, Decision Trees, Neural Networks, and Ensemble classifiers. Their findings show that BBN and DTNB outperformed the other techniques, with or without feature selection, although the latter significantly improved the performance of most classification methods. Among the most important predictors of financial statement fraud were the expected revenue and the EPS growth, the insider holdings, and the stock price to earnings to EPS growth. The proportion of negative sentiment was the only significant of the linguistic variables, which were generally found to perform poorly without the contribution of financial features.

Maka et al. (2020), used a dataset of 3.500 Indian firms’ financial statements, covering a period of 5 years, which were labeled as “fraud” or “genuine” based on the auditor’s comments in the reports. Due to the fact that the fraud instances were significantly less than the non-fraud ones, the researchers trained the machine learning models with undersampled and oversampled subsets of the initial dataset. In particular, they ran 38 algorithms, including Random Forests, Neural Networks, Logistic Regression, Ensemble classifiers, etc, using 16 variables, e.g., the type of company, the total value of assets, accounts receivables, liabilities, the debt-to-equity

ratio, the ratio of investment to sales, etc, and applied feature selection to identify the most significant ones. Parallel random forest and Stochastic gradient boosting were found to be the best performing algorithms, yielding the highest values of accuracy, sensitivity, and negative precision. The variables with the greatest predictive power in terms of financial reporting fraud detection were the interest earned, the Altman Z-Score, and the debt-to-equity Ratio.

Craja et al. (2020) tested the effectiveness of both traditional classification Machine Learning and more advanced Deep Learning models in detecting financial statement fraud, combining financial, linguistic, and text variables, as extracted from a dataset of 201 fraudulent and 962 non-fraudulent companies' annual reports during 1993-2019. Specifically, the researchers evaluated the standalone and the combined predictive performance of 47 financial ratios, and 9 linguistic features, including the average negative and positive sentiment, the general tone, and the average sentence length as well as the raw textual data of the MD&A section of the financial report. The inputs were fed to Logistic Regression, SVM, Random Forests, XGB, ANN, HAN, and GPT-2 algorithms, with the last two utilizing "word2vec" pre-trained embeddings of text data, instead of traditional TF-IDF word transformations. Their main difference is that the first text preprocessing method incorporates grammar, words and sentence importance, and semantic and context features of text, in comparison with the unsophisticated Bag-of-Words approaches. This particular advantage enabled the researchers to identify fraudulent reporting "red flags" both at a word and a sentence level. All in all, RF outperformed the other algorithms with regard to financial, linguistic, and simplified text sources, followed by DL techniques, which demonstrated great predictive power from the combination of financial and advanced textual data. The results showed that the MD&A raw text was superior to specific linguistic metrics, whereas both of these types of features enhanced significantly the model performance, suggesting that the mixture of quantitative and qualitative data may be highly valuable to financial fraud detection models.

Contrary to the aforementioned supervised learning techniques, Deng and Mei (2009) designed an unsupervised clustering model for the detection of fraudulent financial statements, which included a stochastic self-organizing map (SOM) system, K-means Clustering algorithm, and the cluster validity metric "Silhouette Index". The researchers used a dataset of 100 matched Chinese listed companies' financial

statements of 1999-2006 and identified 47 financial ratios in terms of business solvency and management effectiveness, profitability, and liquidity, expense rationality, business growth, and development. These features were fed as an input to the SOM, and after the learning process, the node vector produced was clustered through the k-means algorithm. The clustering results were evaluated using the Silhouette measure. In total, 8 experiments were carried out with the last being the best performing one, achieving an average accuracy rate of 89%.

From a quite different perspective as compared to the aforementioned researchers, Throckmorton et al. (2015) mainly analyzed audio data in addition to some basic financial information for financial fraud detection. More specifically and focusing on revenue misrepresentation, they collected a dataset of 1572 public conference audio samples, which contained discussions on quarterly financial results between CEOs and financial analysts. They matched 41 out of these cases with corresponding irregular financial reports and thus labeled them as fraudulent, while the remaining were marked as honest. They identified 4 types of variables, namely 4 financial ratios, e.g., the book-to-market equity ratio or stock return volatility, 7 acoustic features, related to the pitch, loudness distress, and the hoarseness of the speaker, and 7 linguistic variables, including the proportion of singular and plural pronouns, the number of words with positive and negative sentiment, etc. and 3 baseline metrics, namely AR, F-score, and CORDIS. By applying the training data to the Bayesian-based generalized likelihood classifier GLRT, they evaluated the standalone and combined performance of the features and consequently tested the data using Logistic Regression, Naïve Bayes Classifier, KNN, and GLRT. While the linguistic variables had a very poor performance, the combination of accounting and acoustic features outperformed the baseline metrics, and the mixture of the last three offered the highest predictive power. Out of the classifiers, GLRT was the best performing, but all of them achieved improved AUC scores when the linguistic features were discarded and the informative ones were combined across categories, which highlights the importance of proper feature selection.

2.5.2.2. Machine Learning Algorithms and Data Mining Techniques

Machine learning algorithms and advanced data science methods have been extensively applied by the academic community to the detection of fraudulent financial statements, the exploration of underlying patterns, and the determination of significant features associated with financial fraud. Due to their high computational power, their continuous advancement, and their ability to handle multi-dimensional and multi-variety data, ML techniques tend to achieve higher performance as compared to traditional approaches. There are two approaches with regard to Machine Learning: a) Supervised ML, which uses labeled datasets for the classification or the prediction of a target variable, and b) Unsupervised ML, which analyzes unlabeled data, in order to discover hidden relationships between the observations. Most researchers consider financial statement fraud detection as a supervised classification problem, as it mainly involves the categorization of a company's financial statements into two classes; "fraudulent" or "non-fraudulent". In this context, the most popular algorithms can be grouped into 8 categories and are discussed in more detail in the following paragraphs:

i. Logistic Regression

Logistic Regression is a type of generalized linear model, which is applied as a binomial regression method when the target variable is binary. It is principally used to understand the relationship between one dependent categorical and one or more predictor variables – either numerical or categorical, by estimating the probability of occurrence of a specific categorical class.

ii. Bayesian Classifiers

The Bayesian network is a probabilistic graphical model that uses a directed acyclic graph (DAG) in the form of nodes and edges, to represent the conditional dependencies and independent relationships between a set of different variables. One of the simplest forms of this algorithm is the Naïve Bayes Classifier, which, assuming the independence of the input attributes, classifies a new observation according to the maximization of the calculated probability.

iii. Support Vector Machines

SVM is a supervised ML algorithm that was introduced by Vapnik in 1992 and can be applied for both regression and classification problems. When used for classification purposes, SVM finds the optimal hyperplane that best segregates the two classes of the output variable by mapping the input vectors nonlinearly in a high dimensional space. The training observations that are closest to the maximum margin hyperplane are called support vectors.

iv. Decision Trees

Decision trees distribute the input vectors into a flowchart-like tree structure, where each internal node represents a prediction attribute and each branch denotes conjunctions of features, resulting in the terminal leaves, which hold a class label of the dependent variable. The data sample is divided into subgroups until all the training data are ultimately classified. A decision tree model can be developed with the use of various algorithms, including ID3, C4.5, CART, etc.

v. Neural Networks

Neural Networks (NN) or Artificial Neural Networks (ANN) is a deep learning process that aims to identify the underlying relationships in a set of data, by imitating the behavior of the human brain. ANNs are typically composed of a collection of computing units, called artificial neurons, which are structured into layers, namely an input layer, one or more hidden layers, and an output layer. The interconnected nodes develop an input-output relationship, which is computed by some non-linear function and is dependent on the weight and the threshold associated with each neuron. The most common type of NN, used in financial statement fraud detection literature, is the Feedforward neural network, in which signals flow and are processed only in one direction. However, many academics have also utilized a Multi-layer feedforward neural network, which is bi-directional, allowing weights to be adjusted and updated through a learning process using the backpropagation algorithm.

vi. K-Nearest Neighbors

K-nearest neighbors (KNN) is a non-parametric classifier, which endeavors to predict the correct class of the input data, relying on the proximity of the data points. In other words, KNN classifies an observation into the group, which consists of the k data points with the nearest distance from the new sample point. It's worth noting that the KNN is a "lazy learner" algorithm, as it does not include a training phase.

vii. Ensemble Classifiers

Ensemble methods are techniques that apply multiple classifiers and develop the proper combinations of them, to achieve higher predictive performance. For example, Random Forests mix multiple decision trees, fed with diverse sub-samples of the initial dataset, and use averaging to boost the model's accuracy and control over-fitting. Another tree-based algorithm is Gradient Boosted Decision Trees, which converts many weak-performance trees into strong predictors, by optimizing an arbitrary differentiable loss function through successive steps in the learning process. Ensemble methods are usually more accurate in comparison with a single model; however, they are also more computationally expensive.

2.6. Financial Fraud Cases and Scandals

Over the last two decades, several financial statement fraud scandals have come to light both in the global and the Greek economic environment, which have made investors more cautious and have led regulatory authorities to stricter principles and standards regarding financial reporting. According to the ACFE, financial statement fraud costs organizations a median loss of \$954.00. It is, however, the least common (10%) of occupational fraud cases. (ACFE Report to the Nations, 2020). Some of the most serious scandals worldwide involve Enron and Parmalat, whereas the Bank of Crete and Folli Follie cases are in the spotlight of fraud incidents in Greece.

2.6.1. Enron Corporation

The Enron scandal is undoubtedly one of the most notorious and significant financial fraud cases that shook the business world and one of the main reasons for the establishment of the Sarbanes-Oxley Act, an accounting law aiming to boost the transparency and reliability of financial statements and the accounting-auditing practices in general. Enron was founded in 1985 and became one of the world's leading and most innovative energy, commodities, and services organizations, reporting \$100 billion in 2000. However, a series of scandalous events and disclosures regarding Enron turned its surprising rise into a tragic fall and subsequently to its bankruptcy on December 2, 2001 (Li, 2010). The abrupt resignation of the CEO of Enron in August 2001 shook stakeholders' confidence and made analysts suspicious, who began to investigate thoroughly Enron's financial statements and applied accounting practices. It was discovered that Enron transacted with "special purposes entities" (SPEs), which were not included in Enron's financial statements, in order to inflate earnings and hide debts and losses. Following these discoveries, Enron restated its financial statements for the previous five years, announcing losses of about \$586 million. The reputational damage among investors was huge and, in the end, the company filed for Chapter 11 bankruptcy protection. A key player in this fraud was the auditing firm of Enron, Arthur Andersen, which was accused of applying inadequate auditing duties as well as contributing to fraud through consulting services. Furthermore, the accounting company was convicted of obstructing justice, due to the destruction of all incriminating documents concerning Enron's fraud. According to Benston & Hartgraves (2002), the "creative" accounting practices used in this scandal can be summarized as follows:

- a) The accounting of not consolidating SPEs, based on the GAAP principle of 3% third-party ownership
- b) The handling of the transactions with SPEs, which served the manipulation of financial reporting figures
- c) Fair-Value Accounting for the revaluation of investments in unreliable numbers
- d) Recognition of future-period revenue as current
- e) Insufficient disclosure of information regarding related-party transactions, conflicts of interest, and costs to stakeholders

The aforementioned actions in combination with the systematic positive assurance and approval by the auditing firm set the basis of a financial fraud scandal, which eventually destroyed both Enron and Arthur Andersen.

2.6.2. Parmalat

Being described by SEC as “one of the largest and most brazen corporate financial frauds in history”, Parmalat’s case has definitely marked financial statement falsification in the European Union. The Italian dairy and food corporation, founded in 1961, managed to become a powerful player in the Italian food sector until 2003 when the company was declared bankrupt after the disclosure of financial statements manipulation (Melis, 2004). Aiming to increase its market share, Parmalat followed aggressive acquisition strategies, financed by bank loans and bonds. The formation of such a large group of companies made not only the management but also the estimation of real revenues and audit procedures too difficult. Meanwhile, Parmalat started applying unethical “creative accounting” techniques, considering that some of its subsidiaries reported losses. Probably the most distinctive feature that raised suspicions regarding financial statements fraud was the concurrent high levels of cash and debt. The big fall began when Parmalat failed to liquidate an investment of about €500 million in its subsidiary, Epicurum. On the same day, the company defaulted on a €150 million bond. Following these events, Bank of America declared that the document confirming the €3.95 billion deposit account was forged, leading Parmalat’s stock price to near zero. Consequently, the entity was officially declared insolvent, and CEO Calisto Tanzi was convicted and arrested for fraud. Parmalat utilized the following devious accounting techniques:

- a) The artificial increase of revenues and assets through a double-billing scheme
- b) The use of fake receivables as collateral for loans
- c) Recognition of fictitious assets for the overstatement of financial position
- d) The concealment of legitimate debt from investors
- e) Financial engineering techniques, in cooperation with investment bankers, for debt write-off or debt recording as equity for reporting reasons

2.6.3. Bank of Crete

According to Negakis and Tachynakis (2013), the Bank of Crete scandal is synonymous with the Koskotas scandal, because Koskotas, in cooperation with accounting professionals and government officials, was in the spotlight of this fraud. As the director of the internal control department of the bank, he stole and deposited in personal accounts funds from clients' cheques, amounting to \$1.155.000 in July 1980 and \$1.507.515 in August 1980, without recording the relevant accounting entries. Being afterward promoted as the deputy-head of the accounting department in 1981, Koskotas, along with other executive duties, was in charge of operating payment transfers both in Greece and overseas, as the authorized legal representative (Jones, 2011). This position of authority in conjunction with his arrogant personality provoked him further into misappropriating large amounts of bank reserves for his own benefit. Following these devious actions, he managed to gain complete control of the organization, became the CEO and the President of the Board of Directors of the bank, and was appointed to administrative positions individuals of his trust, in order to continue his fraudulent activity. Things began to get tough for Koskotas when in 1987 The Bank of Greece decided to carry out various branch audits. In order to cover up his crimes, he forged a number of documents with the assistance of his accountants. Additionally, in 1988 under the instructions of Koskotas, two copies of bank letters were forged, so that auditors were misled and convinced regarding the bank's millions of dollars credit balances. The systematic embezzlement gave Koskotas, among others, various business opportunities in the mass media sector, including the establishment or the acquisition of two daily newspapers, five magazines, and a radio station (Wikipedia). His wealthy, yet criminal, lifestyle came to an end when he was sentenced to 25 years in prison for a variety of felonies, including embezzlement, forgery, and obstruction of justice. Koskotas embezzled \$30.718.190 in total (Dermitzakis, 1999), taking advantage of the following circumstances:

- a) Insufficient internal mechanisms, reconciliation difficulties between bank branches due to the problematic architecture of the accounting information systems, and lack of group governance

- b) His position of authority, which enabled him to intervene in business practices, e.g., by employing people willing to cover his actions or by bribing government officials
- c) The fact that at the time there were no financial reporting requirements according to a widely approved accounting framework, like IFRS or GAAP
- d) Regulations that protected the confidentiality of bank accounts, which delayed the disclosure of his fraudulent behavior

2.6.4. Folli Follie

The Greek-based jewelry company Folli-Follie seemed to be a successful and profitable business story, constantly expanding its branch network and commercial activities and increasing its market share not only in Greece but also worldwide. However, in 2018 the report of the American hedge fund firm “Quintessential Capital Management” struck Folli Follie’s “perfect” image, as it raised significant doubts regarding the reliability of the company’s reported financial figures, the real number of its branches as well as the competence of its auditors (Kourtali,2018). QCM pointed out that FF Group actually struggled in terms of revenue, cash, profitability, and network, and that its Chinese subsidiaries were probably at the center of these issues. Following these allegations, Folli Follie announced the appointment of the auditing firms E&Y and Alvarez & Marsal for the audit of the consolidated financial statements of the FF Group and the financial statements of the Asian subsidiaries respectively. The findings of the Alvarez & Marsal report were shocking, highlighting

the chaotic deviations between reported and actual financial figures of FF Group Sourcing, as shown below:

Main Account	FS 2017	Alvarez & Marsal
Inventories	581,681,095	33,873,632
Trade Receivables	718,957,460	99,125,013
Other Receivable, Deposit & prepayment	310,742,476	7,568,415
Bank & cash balances	296,771,278	6,400,473
Trade & other payables	144,561,043	260,932,940
Revenue	1,112,348,021	116,847,155
Cost of sales	614,207,787	33,234,017
Profits	316,444,076	(44,702,304)
Retained earnings and other reserves	1,831,930,169	(-180,638,116)

Figure 3: FFG Sourcing FS 2017 Differences (FF Group, 2018)

The above picture shows clearly that the Asian subsidiary FF Group Sourcing has been manipulating its financial statements, reporting 10 times greater revenue than the actual one (\$ 1,1 billion against \$ 117 million). Furthermore, it was revealed that the founders of Folli Follie have repeatedly forged bank documents, in an attempt to create artificial receipts originating from artificial trade receivables and sales. Apart from the loss of investors' trust and the swan dive of FF Group's share price, the disclosure of fraud resulted in the conviction of 13 individuals involved in the crime, among which were the founder of Folli Follie and his son, who were held in custody. In short, FF Group's financial fraud was initially based on indirect fictitious transactions during 2007-2015, which were afterward executed directly, through a scheme named "merry-go-round", so that the revenues were repeatedly transferred between subsidiaries and were eventually reported at much higher levels in the consolidated financial statements.

2.7. Financial Statement Fraud and COVID 19

The coronavirus 2019 (COVID-19) pandemic is not only a global public health and social crisis but has also evolved into a global economic recession, affecting significantly global and regional economies, industry sectors, and businesses, as well as individuals. With most societies entering lockdown for a long period of time, corporates suddenly found themselves struggling to operate normally, generate cash, manage their expenses, and keep businesses afloat. While the economic downturn is a matter of fact for almost every corporation and industry, the pressure of meeting financial targets, and maintaining high business performance and efficiency expectations remains a priority among CFOs and managers. This may increase their tendency to proceed to unethical corporate behavior, accounts and books manipulation as well as financial statements distortion practices, considering that these cases usually arise during an economic downturn (Deloitte, 2020). According to the International Federation of Accountants (2020), internal financial reporting controls are also at greater risk and may be more easily overridden by fraudsters, as a result of the new remote working reality that COVID-19 has caused to the corporate environment.

Considering that the sharp decrease in corporate income and earnings is probably the most significant impact of COVID-19 on businesses, companies are likely to overstate their revenue, either by fabricating income accounts or by recording premature revenue, so as to boost their performance and report profits. Despite the drop in revenues due to the health crisis, organizations are still burdened with various general operating expenses. Consequently, companies might be tempted to capitalize or defer expenses, aiming to amortize them over several accounting periods (Deloitte, 2020). Another common financial statement fraud scheme in the context of the pandemic is the deliberate application of certain assumptions, provisions, valuation, and impairment methods in elements such as goodwill, financial instruments, trade receivables, and especially inventory. More specifically, as the net realizable value of inventories is possibly reduced due to the declining demand and sales (KPMG, 2020), managers may avoid valuing inventories reliably, resulting in inflated gross profit margins (Deloitte,2020). Last but not least, the disclosures in financial statements may be insufficient, as companies will be

motivated to conceal the complete and actual consequences of COVID-19 on their overall financial results (Deloitte,2020).

As the motives to commit financial statement fraud are growing during the pandemic, the regulators and the Securities and Exchange Commission (SEC) agencies respond with stricter measures, financial compliance standards, and monitoring processes. For instance, the US SEC and the US Department of Justice have proclaimed that they will emphasize their efforts in investigating, detecting, and prosecuting fraud cases associated with the coronavirus pandemic. Moreover, the Australian SEC has issued COVID-19-related financial reporting and auditing requirements and facts, so that a specific framework regarding the preparation and filing of financial statements during the pandemic is defined. As mentioned in the FAQs of ASIC's official website, the main focus regarding financial reporting will be placed on three key areas: a) Recognition and measurement of assets/liabilities b) The disclosures made in the financial statements, especially regarding the basis of assumptions, provisions as well as the financial results and business decisions analysis c) Other significant points, including the going concern evaluation and the events subsequent to the reporting period. It is obvious that regulatory authorities are proactive and highly aware of the possible fields of financial statement fraud in the wake of COVID-19.

CHAPTER 3: RESEARCH METHODOLOGY

3.1. Identifying the problem and setting the target

The first and foremost step of the research process is the definition and identification of the problem as well as the determination of the target that will be examined. The selected topic of this study is Financial Statement Fraud Detection, which can be expressed as a classification problem, i.e., a predictive modelling problem where a class label is predicted for a given example of input data. In Financial Statement Fraud Detection, each observation is to be classified in one of the following two categories: a) Fraudulent and b) non-Fraudulent. As discussed above, there is a variety of approaches that can be applied to this particular problem, including the use of numerical data, financial ratios, linguistic data and textual information.

In this research, the primary goal is to explore the patterns and insights that can be extracted from the MD&A section of the corporate financial reports and develop a classification model that can be used to determine whether the Financial Statements of a company are fraudulent or not. Text analysis in financial literature has been receiving increasing interest over the last fifteen years (Glancy et al. (2010), Humphreys et al. (2010), Cecchini et al. (2010), Goel et al. (2010), Craja et al. (2020)) and its utilization is expected to evolve as more intelligent and advanced software, tools and algorithms are being developed. In addition to the research community, text analytics are also currently applied in real-world business problems by accounting firms such as Deloitte, EY, KPMG etc.

In this study, the choice of analyzing unstructured text information, instead of structured numerical data relies on the importance of the meanings that can be derived from text messages that reflect the human thought, language use and semantics. Furthermore, given that the majority of business information mostly includes semi-structured or unstructured data, this form of data analysis has become crucial for research and modeling purposes. In financial reporting, the Management Discussion & Analysis is an unaudited section, within which management provides an overview of the company's current performance and condition, compliance with laws and standards, as well as the primary risks and challenges it is facing and its plan and actions towards them. It also includes the views of executives regarding future goals,

prospects, and projections of the reporting entity. It is thus a subjective source of information that is highly prone to manipulation considering that it includes decisions, predictions, and comments that the management could purposely present in a more favorable light with the use of specific linguistic tricks.

3.2. Specifying research questions and hypotheses

The main question that this study is attempting to answer is whether and to what extent fraud can be detected from the text of the MD&A section of the corporate Financial Statements. So, the first research question of this study is expressed as follows:

- Research Question 1: Is the analysis of the textual data reported in the MD&A section of the Financial Statements effective at detecting Fraudulent Financial Statements?

Textual analysis might also include the extraction of informative features related to the context, the structure, the use of language, the sentiment, and the general tone within a text. This process is known as ‘Feature Engineering’ and is an important step in every machine learning project, that can actually improve the performance of the developed model. So, the question arising is whether linguistic variables such as polarity, readability, sentence length, and the proportion of specific word categories (positive, negative, litigious, uncertainty) are useful in predicting fraudulent financial statements:

- Research Question 2: Do the linguistic features extracted from the MD&A text present significant predictive power in Financial Statement Fraud Detection?

Generally, it can be argued that there is a natural tradeoff between model accuracy and model interpretability in the context of Machine Learning. In most cases, more complex ‘black box’ models, such as neural networks, tend to be more accurate and less interpretable whereas, less flexible models usually achieve higher interpretability but fall short in accuracy. Depending on the problem and the available data, the best practice is to find the right balance between accuracy and interpretability. In this study, this balance is accomplished with the selection of machine learning algorithms that incorporate the ‘feature importance’ element, i.e., the score that each variable contributes to the model prediction. In text analytics problems, this component is

illustrated at a word or phrase level, and thus, in the particular research can demonstrate specific “red flags” that may signal financial statement fraud. This leads to the following research question:

- Research Questions 3: Can the proposed models provide “red flag” indicators on word and phrase level to assist with the decision-making related to the Financial Statement Fraud Detection?

3.3. Choosing the Research Design

The research methodology followed in this study is a Data Analysis approach that includes all the basic phases of a Data Science Problem:

- i) **Data Collection:** Data Collection is the process of gathering the requisite data to derive insights and turn the business problem into a probable solution. It typically involves the determination of the necessary data for the research and the review of the available data sources to collect them. Data sources can vary from surveys and questionnaires to internal databases or purchased databases from external sources. Obtained data should be carefully reviewed for their quality, meaning that they should be relevant and validated.
- ii) **Data Preparation:** Data Preparation equals to the cleaning and processing of the data, so that they have the proper format, structure, consistency and content to be appropriate for analysis in the modelling phase. It is estimated that this step usually takes around 80% of the total time in a Data Science problem. It may involve a great variety of tasks, e.g., removal of duplicate values or irrelevant observations, change of data types, handling of missing values etc. Especially in the case of text data, multiple stages of preprocessing should be applied, since the unstructured nature of words and sentences increases the complexity, in comparison with numerical data. To this end, Natural Language Processing (NLP), a field of Artificial Intelligence that makes human language intelligible to machines, is broadly used. Some of the most common NLP functions include the replacement of special characters and the conversion of specific parts to readable texts, the removal of whitespaces, punctuation,

and stopwords, Tokenization, Lemmatizing, Count-Vectorization, etc. (Figure 4). The preprocessing of data is a crucial process that improves the quality of information and results in more conclusive and accurate findings.

- iii) **Feature Engineering:** Feature Engineering is defined as the process of transforming raw data into features that better represent the underlying problem and enhance the predictive power of the model. Feature engineering comprises feature extraction, which combines variables into features for dimensionality reduction, feature construction, which manually creates features from raw data, and feature selection, which selects the variables that contribute most to the model. In text analytics, NLP integrates a number of processes for the creation of features from raw text, which better reflect the context, the meaning, the sentiment, the part of speech, and the linguistic expressions of a document. These features involve the sentence count, the fog index, the polarity and subjectivity within the text, the PoS Tagging, the TF-IDF score, n-grams extraction, etc. (Figure 4)
- iv) **Data Exploration (Exploratory Data Analysis-EDA):** Data Exploration is the approach of analyzing and investigating datasets and summarizing their main characteristics, often employing data visualization techniques. EDA enables the discovery of patterns and outliers of data, the testing of hypotheses, or the check of assumptions, and contributes to a better understanding of the variables and the relationships between them.
- v) **Modelling:** The final phase of the research design is the development of the machine learning model, which is expected to make predictions or discover patterns, depending on the underlying data science problem. Correspondingly, a supervised learning problem uses labeled datasets, which train or “supervise” algorithms into classifying data or predicting outcomes, whereas, an unsupervised learning problem is where a model looks for hidden patterns in unlabeled datasets. As discussed at the beginning of this chapter, Financial Statement Fraud Detection is a supervised classification problem, where the output variable contains two

classes, i.e., ‘Fraudulent’ and ‘non-Fraudulent’. The classification model is built using training data and is subsequently tested on unseen data. Once implemented, the machine learning model is evaluated using the appropriate metric(s). In classification problems, the common metrics used include Accuracy, Precision, Recall and F1-score.

The following figure illustrates the detailed process of the research, as it has been designed with regards to the problem of Financial Statement Fraud Detection:

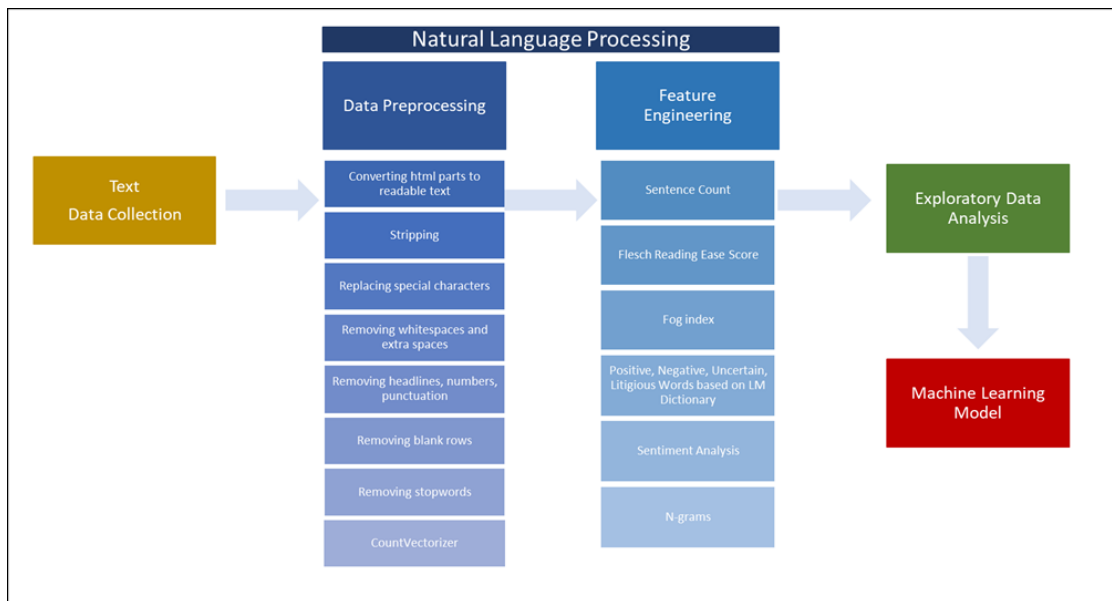


Figure 4: Flow diagram of proposed research design

CHAPTER 4: DATA ANALYSIS

Financial Statement Fraud Detection is a topic that has gained major interest from the research community. The academic framework of this problem varies from more traditional statistical models to more state-of-the-art, data-driven techniques, as discussed in the previous chapters. This particular study approaches the issue of financial reporting fraud through the application of Data Science methods on the Management's Discussion and Analysis (MD&A) section of the annual corporate financial statements (Form 10-K), which is a required public company filing with the Securities and Exchange Commission (SEC) that "provides a comprehensive overview of the company's business and financial condition and includes audited financial statements".

The MD&A section, as already highlighted, is completed in order to give investors a management's view regarding the company's current financial condition, health and results of operations, forward looking performance, goals and plans, the main risks and opportunities. Although unaudited and quite uncontrolled in terms of its content, MD&A is one of the most read sections of the financial statements. The research community has broadly used the Interpersonal Deception Theory (IDT), proposed by Burgoon and Buller (1994), as a theoretical basis for the detection of deceptive behavior in communication means, including MD&A. Although IDT mainly addresses deception in interactive and dialogic conversations between senders and receivers, it also provides a solid theoretical framework for non-interactive, asynchronous media of communication, such as the MD&A. Specifically, on the basis of IDT, as deception is defined as goal-oriented, the deceiver, having a knowledge of fraud, will have the incentive to manipulate the message's completeness, trustworthiness, sentiment and relevance, in order to deceive the reader, and thus, the MD&A should provide linguistic indicators of deceptive behavior of senior management being aware of financial statement fraud. In this direction, the research study is organized following the aforementioned phases of a Data Science problem, which are discussed in detail below.

4.1. Data Collection

The dataset used in this research study consists of US companies' annual financial statements (10-K filings) that are publicly available on the Securities and Exchange Commission's (SEC) website. The annual filing of the 10-Ks is mandatory for all public companies, in order to keep investors aware of the company's financial condition and performance during the reporting year. The research dataset requires samples from both fraudulent and non-fraudulent annual 10-K filings. Since it is quite complex to define Financial Statement Fraud, following former research papers, the issuance of an Accounting and Auditing Enforcement Release (AAER) is used as the criterion for marking a company's financial statements as 'fraudulent'. An AAER is an administrative proceeding or litigations release imposed by SEC, that contains enforcement actions against companies that have violated the accounting or auditing regulations and standards. By labelling the annual filings based on the AAER, a possible bias from subjective categorization is avoided, since AAERs are official statements published in case of financial reporting violations.

First, 716 AAER cases, which were announced over the period between May 17th, 1982 and December 31, 2018, are collected. The dataset is downloaded from the following repository <https://github.com/JarFraud/FraudDetection> and it contains observations for the period 1982-2016 that were compiled by the University of California-Berkeley's Center for Financial Reporting and Management (CFRM) and some additional cases for the period up to December 2018, that were manually added from the SEC website. This dataset will be the starting point for selecting the sample of 'Fraudulent' companies. The dataset (Figure 5) includes four columns:

- AAER No
- Company CIK, which refers to a unique number assigned to each entity that submits its filings to SEC
- Fraud year and
- Whether the AAER relates to an understatement case

It contains 1.746 observations, which means that each AAER may refer to multiple financial reporting years. Consequently, for each company the latest

reporting year is chosen and years earlier than 2001 are excluded, resulting in an initial dataset of 177 observations.

P_AAER	CIK	YEARA	UNDERSTATEMENT
1555	1089567	2001	0
1631	1379895	2001	0
1662	1060239	2001	0
1773	1123337	2001	0
1780	1126294	2001	0
1824	894501	2001	0
1839	1608109	2002	0
1848	1093285	2001	0
1923	72020	2001	0
1938	872867	2001	0
1940	718909	2001	0
1977	723527	2001	0
1979	1064539	2002	0
1984	1048237	2002	0
1999	1015610	2001	0
2033	1009304	2001	0
2075	14272	2001	0
2082	785161	2002	0

Figure 5: AAER Dataset

Subsequently, the database including the links to the files of the SEC financial reporting filings is formed, by using the library ‘python-edgar’ within Spyder IDE. This library produces tsv (tab-separated) master data files per quarter of all SEC filings since 1993, containing the following information:

- Company name (eg. TWITTER, INC)
- Company CIK (eg. 0001418091)
- Filing date (eg. 2013-10-03)
- Filing type (eg. S1)
- Filing URL for txt file (edgar/data/1418091/0001193125-13-390321.txt)
- Filing URL for html file (edgar/data/1418091/0001193125-13-390321.html)

The URL columns are populated with the phrase “<https://www.sec.gov/Archives/>”, in order to match the full existing hyperlinks. The process is run for the year 2001 and onwards, resulting in 87 index files (Figure 6). The files are then concatenated in a single csv database with the use of Alteryx Designer. (Figure 7)

```

1000032|BINCH JAMES G|4|2012-02-02|edgar/data/1000032/0001181431-12-005367.txt|edgar/data/1000032/0001181431-12-005367.txt
1000032|BINCH JAMES G|4|2012-03-02|edgar/data/1000032/0001181431-12-013915.txt|edgar/data/1000032/0001181431-12-013915.txt
1000045|NICHOLAS FINANCIAL INC|10-Q|2012-02-09|edgar/data/1000045/0001193125-12-047920.txt|edgar/data/1000045/0001193125-12-047920.txt
1000045|NICHOLAS FINANCIAL INC|4|2012-02-23|edgar/data/1000045/0001000045-12-000001.txt|edgar/data/1000045/0001000045-12-000001.txt
1000045|NICHOLAS FINANCIAL INC|8-K|2012-02-02|edgar/data/1000045/0001193125-12-035754.txt|edgar/data/1000045/0001193125-12-035754.txt
1000045|NICHOLAS FINANCIAL INC|13G/A|2012-02-14|edgar/data/1000045/0001193125-12-061110.txt|edgar/data/1000045/0001193125-12-061110.txt
1000069|EMPIRIC FUNDS, INC|4858POS|2012-01-27|edgar/data/1000069/0000894189-12-000411.txt|edgar/data/1000069/0000894189-12-000411.txt
1000069|EMPIRIC FUNDS, INC|4858POS|2012-02-16|edgar/data/1000069/0000894189-12-000758.txt|edgar/data/1000069/0000894189-12-000758.txt
1000069|EMPIRIC FUNDS, INC|497K|2012-02-03|edgar/data/1000069/0000894189-12-000549.txt|edgar/data/1000069/0000894189-12-000549.txt
1000069|EMPIRIC FUNDS, INC|497K|2012-03-28|edgar/data/1000069/0000894189-12-001698.txt|edgar/data/1000069/0000894189-12-001698.txt
1000069|EMPIRIC FUNDS, INC|497|2012-01-27|edgar/data/1000069/0000894189-12-000355.txt|edgar/data/1000069/0000894189-12-000355.txt
1000069|EMPIRIC FUNDS, INC|497|2012-02-01|edgar/data/1000069/0000894189-12-000501.txt|edgar/data/1000069/0000894189-12-000501.txt
1000069|EMPIRIC FUNDS, INC|497|2012-02-02|edgar/data/1000069/0000894189-12-000532.txt|edgar/data/1000069/0000894189-12-000532.txt
1000069|EMPIRIC FUNDS, INC|497|2012-02-10|edgar/data/1000069/0000894189-12-000665.txt|edgar/data/1000069/0000894189-12-000665.txt
1000069|EMPIRIC FUNDS, INC|497|2012-03-28|edgar/data/1000069/0000894189-12-001697.txt|edgar/data/1000069/0000894189-12-001697.txt
1000069|EMPIRIC FUNDS, INC|N-Q|2012-02-17|edgar/data/1000069/0000894189-12-000779.txt|edgar/data/1000069/0000894189-12-000779.txt
1000097|KINGDOM CAPITAL MANAGEMENT LLC|13F-HR|2012-02-14|edgar/data/1000097/0000919574-12-001491.txt|edgar/data/1000097/0000919574-12-001491.txt
1000097|KINGDOM CAPITAL MANAGEMENT LLC|13G/A|2012-02-14|edgar/data/1000097/0000919574-12-001375.txt|edgar/data/1000097/0000919574-12-001375.txt
1000097|KINGDOM CAPITAL MANAGEMENT LLC|13G/A|2012-02-14|edgar/data/1000097/0000919574-12-001381.txt|edgar/data/1000097/0000919574-12-001381.txt
1000097|KINGDOM CAPITAL MANAGEMENT LLC|13G/A|2012-02-14|edgar/data/1000097/0000919574-12-001385.txt|edgar/data/1000097/0000919574-12-001385.txt
1000097|KINGDOM CAPITAL MANAGEMENT LLC|13G|2012-01-26|edgar/data/1000097/0000919574-12-000293.txt|edgar/data/1000097/0000919574-12-000293.txt
1000151|GREEHWICH FINANCIAL SERVICES, L.L.C.|FOCUSN|2012-02-29|edgar/data/1000151/9999999997-12-006751.txt|edgar/data/1000151/9999999997-12-006751.txt
1000151|GREEHWICH FINANCIAL SERVICES, L.L.C.|X-17A-5|2012-02-29|edgar/data/1000151/9999999997-12-004570.txt|edgar/data/1000151/9999999997-12-004570.txt
1000152|WESTERN INTERNATIONAL SECURITIES, INC.|FOCUSN|2012-02-27|edgar/data/1000152/9999999997-12-003600.txt|edgar/data/1000152/9999999997-12-003600.txt
1000152|WESTERN INTERNATIONAL SECURITIES, INC.|X-17A-5|2012-02-27|edgar/data/1000152/9999999997-12-002626.txt|edgar/data/1000152/9999999997-12-002626.txt
1000177|NORDIC AMERICAN TANKERS Ltd|424B2|2012-01-18|edgar/data/1000177/0001193125-12-015665.txt|edgar/data/1000177/0001193125-12-015665.txt
1000177|NORDIC AMERICAN TANKERS Ltd|424B2|2012-01-20|edgar/data/1000177/0001193125-12-018848.txt|edgar/data/1000177/0001193125-12-018848.txt
1000177|NORDIC AMERICAN TANKERS Ltd|6-K/A|2012-01-18|edgar/data/1000177/0001193125-12-015554.txt|edgar/data/1000177/0001193125-12-015554.txt
1000177|NORDIC AMERICAN TANKERS Ltd|6-K|2012-01-18|edgar/data/1000177/0001193125-12-015564.txt|edgar/data/1000177/0001193125-12-015564.txt
1000177|NORDIC AMERICAN TANKERS Ltd|6-K|2012-01-18|edgar/data/1000177/0001193125-12-015622.txt|edgar/data/1000177/0001193125-12-015622.txt
1000177|NORDIC AMERICAN TANKERS Ltd|6-K|2012-01-18|edgar/data/1000177/0001193125-12-015725.txt|edgar/data/1000177/0001193125-12-015725.txt
1000177|NORDIC AMERICAN TANKERS Ltd|6-K|2012-01-19|edgar/data/1000177/0001193125-12-016451.txt|edgar/data/1000177/0001193125-12-016451.txt
1000177|NORDIC AMERICAN TANKERS Ltd|6-K|2012-01-23|edgar/data/1000177/0001193125-12-019085.txt|edgar/data/1000177/0001193125-12-019085.txt
1000177|NORDIC AMERICAN TANKERS Ltd|6-K|2012-02-14|edgar/data/1000177/0001000177-12-000003.txt|edgar/data/1000177/0001000177-12-000003.txt
1000177|NORDIC AMERICAN TANKERS Ltd|FWP|2012-01-18|edgar/data/1000177/0001193125-12-015706.txt|edgar/data/1000177/0001193125-12-015706.txt
1000177|NORDIC AMERICAN TANKERS Ltd|SC|13G|2012-02-14|edgar/data/1000177/0000902219-12-000313.txt|edgar/data/1000177/0000902219-12-000313.txt
1000180|SANDISK CORP|10-K|2012-02-23|edgar/data/1000180/0001000180-12-000012.txt|edgar/data/1000180/0001000180-12-000012.txt
1000180|SANDISK CORP|4|2012-01-03|edgar/data/1000180/0001242648-12-000001.txt|edgar/data/1000180/0001242648-12-000001.txt
1000180|SANDISK CORP|4|2012-01-05|edgar/data/1000180/0001242648-12-000002.txt|edgar/data/1000180/0001242648-12-000002.txt
1000180|SANDISK CORP|4|2012-01-12|edgar/data/1000180/0001242648-12-000003.txt|edgar/data/1000180/0001242648-12-000003.txt
1000180|SANDISK CORP|4|2012-02-02|edgar/data/1000180/0001242648-12-000004.txt|edgar/data/1000180/0001242648-12-000004.txt
1000180|SANDISK CORP|4|2012-02-10|edgar/data/1000180/0001242648-12-000005.txt|edgar/data/1000180/0001242648-12-000005.txt
1000180|SANDISK CORP|4|2012-02-10|edgar/data/1000180/0001242648-12-000006.txt|edgar/data/1000180/0001242648-12-000006.txt
1000180|SANDISK CORP|4|2012-02-22|edgar/data/1000180/0001242648-12-000007.txt|edgar/data/1000180/0001242648-12-000007.txt
1000180|SANDISK CORP|4|2012-02-22|edgar/data/1000180/0001242648-12-000008.txt|edgar/data/1000180/0001242648-12-000008.txt

```

Figure 6:Master Data TSV File for SEC Filings of Q1 2012

```

Index,CIK,Company Name,Cat,Fill Date,Text File,HTML File
42,1000112,CC MASTER CREDIT CARD TRUST II,10-K,2001-03-23,https://www.sec.gov/Archives/edgar/data/1000112/0000930661-01-000661.txt,https://www.sec.gov/Archives/edgar/data/1000112/0000930661-01-000661-index.html
81,1000227,TB WOODS CORP,10-K,2001-03-12,https://www.sec.gov/Archives/edgar/data/1000227/0000950116-01-000403.txt,https://www.sec.gov/Archives/edgar/data/1000227/0000950116-01-000403-index.html
90,1000228,SCHEN HENRY INC,10-K,2001-03-28,https://www.sec.gov/Archives/edgar/data/1000228/0001125282-01-001159.txt,https://www.sec.gov/Archives/edgar/data/1000228/0001125282-01-001159-index.html
96,1000229,CORE LABORATORIES N V,10-K,2001-03-15,https://www.sec.gov/Archives/edgar/data/1000229/0001000229-01-000004.txt,https://www.sec.gov/Archives/edgar/data/1000229/0001000229-01-000004-index.html
106,1000232,BOURBON BANCSHARES INC /KV/,10-K,2001-03-30,https://www.sec.gov/Archives/edgar/data/1000232/0001000232-01-000001.txt,https://www.sec.gov/Archives/edgar/data/1000232/0001000232-01-000001-index.html
114,1000235,PATRIOT BANK CORP,10-K,2001-03-29,https://www.sec.gov/Archives/edgar/data/1000235/0000893220-01-000359.txt,https://www.sec.gov/Archives/edgar/data/1000235/0000893220-01-000359-index.html
136,1000297,TRAFFIX INC,10-K,2001-03-05,https://www.sec.gov/Archives/edgar/data/1000297/0000950123-01-002014.txt,https://www.sec.gov/Archives/edgar/data/1000297/0000950123-01-002014-index.html
139,1000298,IMPAC MORTGAGE HOLDINGS INC,10-K,2001-03-30,https://www.sec.gov/Archives/edgar/data/1000298/0001017062-01-000697.txt,https://www.sec.gov/Archives/edgar/data/1000298/0001017062-01-000697-index.html
143,1000301,AMBANC HOLDING CO INC,10-K,2001-03-27,https://www.sec.gov/Archives/edgar/data/1000301/0001000301-01-000001.txt,https://www.sec.gov/Archives/edgar/data/1000301/0001000301-01-000001-index.html
190,1000377,SIMON TRANSPORTATION SERVICES INC,10-K,2001-01-12,https://www.sec.gov/Archives/edgar/data/1000377/0001000577-01-000001.txt,https://www.sec.gov/Archives/edgar/data/1000377/0001000577-01-000001-index.html
232,1000623,SCHWEITZER MAINDUIT INTERNATIONAL INC,10-K,2001-03-02,https://www.sec.gov/Archives/edgar/data/1000623/0000950144-01-003277.txt,https://www.sec.gov/Archives/edgar/data/1000623/0000950144-01-003277-index.html
241,1000683,BLONDER TONGUE LABORATORIES INC,10-K,2001-03-29,https://www.sec.gov/Archives/edgar/data/1000683/0000950116-01-000523.txt,https://www.sec.gov/Archives/edgar/data/1000683/0000950116-01-000523-index.html
249,1000694,NOVAVAX INC,10-K,2001-03-29,https://www.sec.gov/Archives/edgar/data/1000694/0000950133-01-001077.txt,https://www.sec.gov/Archives/edgar/data/1000694/0000950133-01-001077-index.html
262,1000697,WATERS CORP /DE/,10-K,2001-03-27,https://www.sec.gov/Archives/edgar/data/1000697/0000927016-01-001505.txt,https://www.sec.gov/Archives/edgar/data/1000697/0000927016-01-001505-index.html
274,1000736,CAREMARK RX INC,10-K,2001-03-21,https://www.sec.gov/Archives/edgar/data/1000736/0000950144-01-003724.txt,https://www.sec.gov/Archives/edgar/data/1000736/0000950144-01-003724-index.html
285,1000753,ADMINSTAFF INC /DE/,10-K,2001-03-16,https://www.sec.gov/Archives/edgar/data/1000753/0000950129-01-001482.txt,https://www.sec.gov/Archives/edgar/data/1000753/0000950129-01-001482-index.html
303,1000823,FIRST INDUSTRIAL SECURITIES L P,10-K,2001-03-09,https://www.sec.gov/Archives/edgar/data/1000823/0000950137-01-000257.txt,https://www.sec.gov/Archives/edgar/data/1000823/0000950137-01-000257-index.html
328,1001082,ECHOSTAR COMMUNICATIONS CORP,10-K,2001-03-13,https://www.sec.gov/Archives/edgar/data/1001082/0000950134-01-002076.txt,https://www.sec.gov/Archives/edgar/data/1001082/0000950134-01-002076-index.html
392,1001193,TRANSMETA CORP,10-K,2001-03-07,https://www.sec.gov/Archives/edgar/data/1001193/0001095811-01-000319.txt,https://www.sec.gov/Archives/edgar/data/1001193/0001095811-01-000319-index.html

```

Figure 7:Full Master Dataset of SEC Filings

Each of the 177 observations of the AAER Dataset is then mapped to the appropriate filing file link of the Full Master Dataset, by joining the two databases on the CIK and Year, with the use of Excel VLOOKUP. The choice of the year depends on the fraud year of the AAER Dataset and the reporting year of the Full Master Dataset, which may or may not be equal to the year of filing (e.g. a financial report with a filing date 16/2/2001 covers a reporting period of 01/12/1999-30/11/2000). Accordingly, a sample of 175 filing file links of non fraudulent companies for the corresponding reporting periods is collected. The combined dataset is comprised of 177 fraudulent and 175 non-fraudulent companies' data and a label column is created to indicate Fraud, where '0' corresponds to 'Non-fraudulent' and '1' to 'Fraudulent'. This dataset (Table 1) contains the following data:

- Index
- Company CIK
- Company name
- Filing type, which is “10-K” for all observations
- Filing date
- Txt File URL
- HTML File URL
- Fraudulent Label

Index	CIK	Company Name	Cat	Fill Date	Txt File	Html File	Fraudulent
6594	1013243	AAIPHAR MA INC	10-K	15/6/2004	https://www.sec.gov/Archives/edgar/data/1013243/0000950144-04-006287-144-04-006287.txt	https://www.sec.gov/Archives/edgar/data/1013243/0000950144-04-006287-index.html	1
138819	2491	ALLIANCE GAMING CORP	10-K	13/9/2004	https://www.sec.gov/Archives/edgar/data/2491/0001104659-04-027415-04-027415.txt	https://www.sec.gov/Archives/edgar/data/2491/0001104659-04-027415-index.html	1
76692	846538	ALLOU HEALTH & BEAUTY CARE INC	10-K	15/7/2002	https://www.sec.gov/Archives/edgar/data/846538/0000910680-02-000626-80-02-000626.txt	https://www.sec.gov/Archives/edgar/data/846538/0000910680-02-000626-index.html	1
33875	1092492	AXESSTEL INC	10-K	28/2/2013	https://www.sec.gov/Archives/edgar/data/1092492/0001193125-13-084624-125-13-084624.txt	https://www.sec.gov/Archives/edgar/data/1092492/0001193125-13-084624-index.html	1
41560	14272	BRISTOL MYERS SQUIBB CO	10-K	2/4/2001	https://www.sec.gov/Archives/edgar/data/14272/0000014272-01-500006-2-01-500006.txt	https://www.sec.gov/Archives/edgar/data/14272/0000014272-01-500006-index.html	1
41306	1082084	BROOKE CAPITAL CORP	10-K	14/3/2008	https://www.sec.gov/Archives/edgar/data/1082084/0000950137-08-003662-137-08-003662.txt	https://www.sec.gov/Archives/edgar/data/1082084/0000950137-08-003662-index.html	1

Table 1: Full Labelled Master Dataset of SEC Filings

The Txt File URLs are used for the collection of the MD&A section of the annual filings. Specifically, the corresponding txt documents are accessed through the SEC EDGAR Filings Extractor API ([SEC EDGAR Filings API \(sec-api.io\)](https://sec-api.io)) within Spyder IDE, are collected and saved in an Excel Worksheet. A monthly subscription has enabled the use of unlimited requests to the API and has provided the authentication key, through which the MD&A text is returned. The item parameter of the request is set to “7”, corresponding to the MD&A section.

The fetched data are stored in column ‘Data’ in a raw, text form and contain special characters, missing values, duplicates etc. (Table 2). Therefore, it is of great importance to clean and preprocess the data, as discussed in the next Subchapter (4.2.).

Index	CIK	Company Name	Cat	Fill Date	Txt File	Html File	Fraudulent	Data
6594	1013243	AAIPHAR MA INC	10- K	15/6/2004	https://www.sec.gov/Archives/edgar/data/1013243/000950144-04-006287.txt	https://www.sec.gov/Archives/edgar/data/1013243/0000950144-04-006287-index.html	1	Item 7. Management’s Discussion and Analysis of Financial Condition and Results of Operations The Company is restating its financial statements for the first, second and third quarters of both 2002 and 2003 and for the year 2002 (the “Restatement”).
138819	2491	ALLIANCE GAMING CORP	10- K	13/9/2004	https://www.sec.gov/Archives/edgar/data/2491/0001104659-04-027415.txt	https://www.sec.gov/Archives/edgar/data/2491/0001104659-04-027415-index.html	1	ITEM 7. MANAGEMENT’S DISCUSSION AND ANALYSIS OF FINANCIAL CONDITION AND RESULTS OF OPERATIONS Forward-Looking Statements
76692	846538	ALLOU HEALTH & BEAUTY CARE INC	10- K	15/7/2002	https://www.sec.gov/Archives/edgar/data/846538/000910680-02-000626.txt	https://www.sec.gov/Archives/edgar/data/846538/0000910680-02-000626-index.html	1	ITEM 7. MANAGEMENT'S DISCUSSION AND ANALYSIS OF FINANCIAL CONDITION AND RESULTS OF OPERATIONS. RESULTS OF OPERATIONS We distribute consumer personal care products and prescription pharmaceuticals on a national basis. We also manufacture upscale hair and skin care products for sale under private labels.
33875	1092492	AXESSTEL INC	10- K	28/2/2013	https://www.sec.gov/Archives/edgar/data/1092492/0001193125-13-084624.txt	https://www.sec.gov/Archives/edgar/data/1092492/0001193125-13-084624-index.html	1	ITEM 7. MANAGEMENT’S DISCUSSION AND ANALYSIS OF FINANCIAL CONDITION AND RESULTS OF OPERATIONS Forward-Looking Statements Statements in the following discussion and throughout this report that are not historical in nature are “forward-looking statements.”
41560	14272	BRISTOL MYERS SQUIBB CO	10- K	2/4/2001	https://www.sec.gov/Archives/edgar/data/14272/000014272-01-500006.txt	https://www.sec.gov/Archives/edgar/data/14272/0000014272-01-500006-index.html	1	Item 7. MANAGEMENT'S DISCUSSION AND ANALYSIS OF FINANCIAL CONDITION AND RESULTS OF OPERATIONS. Summary In September 2000, the Company announced the planned divestitures of the Zimmer and Clairol businesses. Accordingly, their results have been reported as discontinued operations
41306	1082084	BROOKE CAPITAL CORP	10- K	14/3/2008	https://www.sec.gov/Archives/edgar/data/1082084/000950137-08-003662.txt	https://www.sec.gov/Archives/edgar/data/1082084/0000950137-08-003662-index.html	1	ITEM 7. MANAGEMENT’S DISCUSSION AND ANALYSIS OF FINANCIAL CONDITION AND RESULTS OF OPERATIONS. MANAGEMENT’S DISCUSSION AND ANALYSIS OF FINANCIAL CONDITION AND RESULTS OF OPERATIONS.

Table 2: Dataset with raw text data from API

4.2. Data Preparation

The data extracted from the API require extensive preparation, in order to have the appropriate format and content for modelling purposes. The preparation of data is employed in Spyder IDE with the use of Python.

A sample of the text data before any processing is shown in the image below:

The image shows a snippet of raw text data with numerous HTML entities. The text is a financial report section titled 'ITEM 7. MANAGEMENT'S DISCUSSION AND ANALYSIS OF FINANCIAL CONDITION AND RESULTS OF OPERATIONS'. The entities include non-breaking spaces () and spaces (). The text describes the company's financial condition, operations, and results for 2012 compared to 2011. Key figures mentioned include revenue of \$2,780.5 million in 2012 versus \$2,507.3 million in 2011, and earnings per diluted share of \$2.00 in 2012 versus \$1.42 in 2011. The text also mentions the company's retail and direct segments, store locations, and financial services.

Figure 8:Raw text data sample before preprocessing

First, any rows or columns containing NA values are dropped, using Panda's `dropna()` method. As it can be seen in Figure 8, there are many HTML special characters, e.g. “ ”, which are removed using the Python library “BeautifulSoup”.

After keeping in a Dataframe, i.e. a data structure that organizes data into a 2-dimensional table of rows and columns, only the ‘Data’ and ‘Fraudulent’ columns and converting the data type of ‘Data’ from ‘object’ to ‘string’, further data cleaning processes are applied:

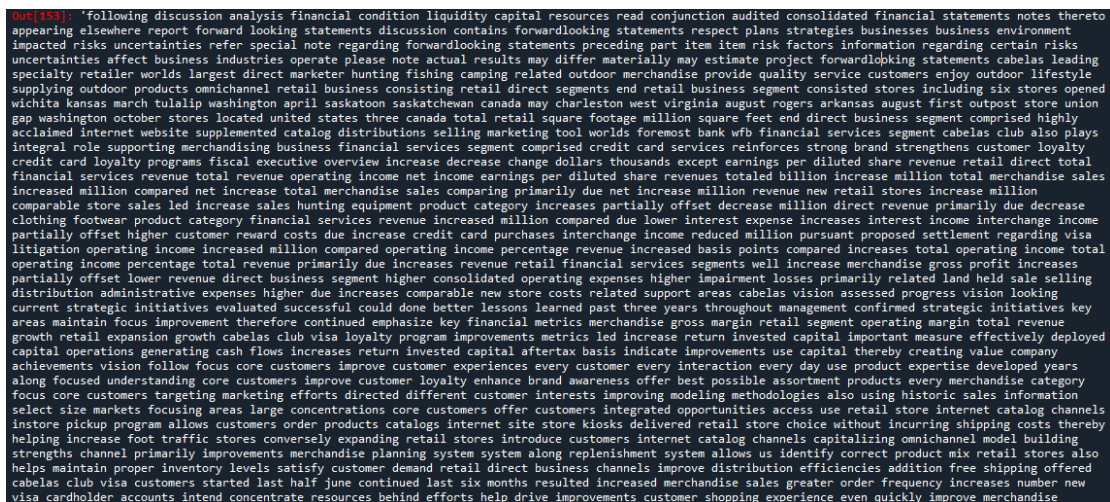
- a. Text Stripping using the `strip()` method: The text is ‘stripped’ with the removal of both the leading and the trailing characters.
- b. Lowercase conversion using the `lower()` method: All the text should be in same case (Upper or lower) for data analysis purposes. Therefore, all words are converted to lower case.
- c. Removal of specific special characters including ‘\n’, ‘&’, ‘\xa0’, using `replace()` method.
- d. Removal of white and extra spaces, applying specific functions.
- e. Removal of numbers, punctuation and blank rows, using `replace()` method in combination with regular expressions.
- f. Removal of the title ‘Item 7. management’s discussion and analysis of financial condition and results of operations’ with the `re.sub()` function.

Since the title is common for every observation, it does not add significance to the content.

- g.** Removal of the phrases ‘the year ended (month)’, which refers to the reporting period and is frequently repeated in annual filings, without adding significant information.
- h.** Removal of stop words, utilizing the stop words list of NLTK corpus. Stop words are commonly used words such as ‘the’, ‘is’, ‘a’, ‘in’, which are considered useless and are typically ignored by search engines and typical tokenizers. These words do not add much meaning to a sentence and they should therefore be removed, so that no valuable space or processing time is consumed.
- i.** Removal of duplicate values using drop_duplicates() method.
- j.** Lemmatization using WordNet’s function WordNetLemmatizer.
Lemmatization is the process of reducing a word to its Lemma.

The “cleaned” data have been stored in the “Data Edited” column of the dataset.

A sample of the text data after applying the aforementioned preprocessing techniques is illustrated in the following image:



following discussion analysis financial condition liquidity capital resources read conjunction audited consolidated financial statements notes thereto appearing elsewhere report forward looking statements discussion contains forwardlooking statements respect plans strategies businesses business environment impacted risks uncertainties refer special note regarding forwardlooking statements preceding part item item risk factors information regarding certain risks uncertainties affect business industries operate please note actual results may differ materially may estimate project forwardlooking statements cabelas leading specialty retailer worlds largest direct marketer hunting fishing camping related outdoor merchandise provide quality service customers enjoy outdoor lifestyle supplying outdoor products omnichannel retail business consisting retail direct segments end retail business segment consisted stores including six stores opened wichita kansas march tulalip washington april saskatoon saskatchewan canada may charleston west virginia august rogers arkansas august first outpost store union gap washington october stores located united states three canada total retail square footage million square feet end direct business segment comprised highly acclaimed internet website supplemented catalog distributions selling marketing tool worlds foremost bank wfb financial services segment cabelas club also plays integral role supporting merchandising business financial services segment comprised credit card services reinforces strong brand strengthens customer loyalty credit card loyalty programs fiscal executive overview increase decrease change dollars thousands except earnings per diluted share revenue retail direct total financial services revenue total revenue operating income net income earnings per diluted share revenues totaled billion increase million total merchandise sales increased million compared net increase total merchandise sales comparing primarily due net increase million revenue new retail stores increase million comparable store sales led increase sales hunting equipment product category increases partially offset decrease million direct revenue primarily due decrease clothing footwear product category financial services revenue increased million compared due lower interest expense increases interest income interchange income partially offset higher customer reward costs due increase credit card purchases interchange income reduced million pursuant proposed settlement regarding visa litigation operating income increased million compared operating income percentage revenue increased basis points compared increases total operating income total operating income percentage total revenue primarily due increases revenue retail financial services segments well increase merchandise gross profit increases partially offset lower revenue direct business segment higher consolidated operating expenses higher impairment losses primarily related land held sale selling distribution administrative expenses higher due increases comparable new store costs related support areas cabelas vision assessed progress vision looking current strategic initiatives evaluated successful could done better lessons learned past three years throughout management confirmed strategic initiatives key areas maintain focus improvement therefore continued emphasize key financial metrics merchandise gross margin retail segment operating margin total revenue growth retail expansion growth cabelas club visa loyalty program improvements metrics led increase return invested capital important measure effectively deployed capital operations generating cash flows increases return invested capital aftertax basis indicate improvements use capital thereby creating value company achievements vision follow focus core customers improve customer experiences every customer every interaction every day use product expertise developed years along focused understanding core customers improve customer loyalty enhance brand awareness offer best possible assortment products every merchandise category focus core customers targeting marketing efforts directed different customer interests improving modeling methodologies also using historic sales information select size markets focusing areas large concentrations core customers offer customers integrated opportunities access use retail store internet catalog channels instore pickup program allows customers order products catalogs internet site store kiosks delivered retail store choice without incurring shipping costs thereby helping increase foot traffic stores conversely expanding retail stores introduce customers internet catalog channels capitalizing omnichannel model building strengths channel primarily improvements merchandise planning system system along replenishment system allows us identify correct product mix retail stores also helps maintain proper inventory levels satisfy customer demand retail direct business channels improve distribution efficiencies addition free shipping offered cabelas club visa customers started last half june continued last six months resulted increased merchandise sales greater order frequency increases number new visa cardholder accounts intend concentrate resources behind efforts help drive improvements customer shopping experience even quickly improve merchandise

Figure 9: Text data sample after preprocessing

A further significant NLP technique that has been applied in the text data is Tokenization using CountVectorizer() method, which converts a collection of text documents to a sparse matrix of token counts, so that the data are understandable by the machine learning algorithms. This method enables the creation of ngram ranges, a

feature which is broadly used for the creation of different classification models in this study, and will be analyzed in detail in the modelling phase of Subchapter 4.6.

4.3. Feature engineering

Feature engineering is defined as the process of transforming raw data into variables, so that they can be used as features for modelling purposes. Model features are the inputs that machine learning models use during training and prediction steps. The performance of the model relies on a precise set of features, and thus, feature engineering is a vital step in every data science project.

In text analytics, feature engineering typically includes the extraction of features that represent the content, context, sentiment, semantics, tone and part of speech of the document. Humpherys et al. (2010) have found that the MD&A section of fraudulent Financial Statements contains more “activation” language, imagery and pleasantness cues, group references and less lexical diversity than non-fraudulent reports. According to Glancy and Yadav (2010), the stress of the preparer falsifying financial reports affects the writing and provides specific linguistic cues in a low presence, high rehearsability and low synchronicity media. Goel and Uzuner (2016) suggest that fraudulent MD&A sections contain a higher degree of sentiment, indicated by a more pronounced use of positive and negative sentiment and more subjectivity and intensity clues in the form of adjectives and adverbs. Accordingly, the significance of negative sentiment in corporate annual reports is also highlighted in the study of Hajek and Henriques (2017). Following the guidelines of the existing studies, specific linguistic features have been extracted.

Using Python Textstat library, which helps determine readability, complexity and grade level of a text document, the variables below are calculated:

1. Sentence Count (variable name: `sentence_count`), which returns the number of sentences in the given text.
2. Flesch Reading Ease Score (variable name: `flesch_ease`), which returns a value depending on the readability in a document.

The following table presents an example of the levels of reading difficulty for different ranges of Flesch Reading Ease Score:

Score	Difficulty
90-100	Very Easy
80-89	Easy
70-79	Fairly Easy
60-69	Standard
50-59	Fairly Difficult
30-49	Difficult
0-29	Very Confusing

Table 3: Flesch Reading Ease Score Levels (Wikipedia)

The formula for the Flesch Reading Ease Score is as follows:

$$206.835 - 1.015 \left(\frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left(\frac{\text{total syllables}}{\text{total words}} \right)$$

Figure 10: Flesch Reading Ease Score Formula (www.readable.com)

3. The Fog Scale (Gunning Fog Formula) (variable name: fog_index), which is a formula generating a grade level between 0 and 20, estimating the education level required to understand a particular text. For example, a Gunning Fog Score of 6 is easily readable for sixth-graders.

The formula for the Gunning Fog Score is as follows:

$$0.4 \left[\left(\frac{\text{total words}}{\text{total sentences}} \right) + 100 \left(\frac{\text{complex words}}{\text{total words}} \right) \right]$$

Figure 11: Gunning Fog Score Formula (www.readable.com)

It should be noted that the above three variables were created in the preparation phase, before removing the title, the numbers, the punctuation and the stop words. This choice was made on the basis that these readability scores should require the full content of the sentences in order to be more accurately calculated. It would be therefore false to base these features on incomplete sentences.

4. Lexicon-Based Sentiment Analysis: A lexicon-based (or dictionary-based) sentiment analysis uses pre-determined lists of words that are labelled in association with a specific sentiment, i.e., positivity, negativity, ambiguity, in order to gain insights into a document's content. Loughran and McDonald (2011) concluded that applying a general sentiment dictionary to sentiment analysis of Financial Statements is not as informative as using a domain-

specific lexicon. Consequently, they formed a dictionary adapted to the financial industry, which contains seven main sentiment categories, including:

- a. Positive, which is comprised of words with good connotations
- b. Negative, which is comprised of words with bad connotations
- c. Litigious, which is comprised of litigation related words
- d. Uncertainty, which is comprised of words indicating imprecision
- e. Strong modal, which is comprised of words expressing certainty of an action
- f. Weak modal, which is comprised of words expressing uncertainty of an action
- g. Constraining, which is comprised of words related to constraints

In this study, the updated version of the LM Dictionary (covering 1993-2021) has been employed in Spyder IDE to extract for each MD&A observation the positive, negative, uncertainty and litigious word counts. Accordingly, the variables 'pos_count', 'neg_count', 'unc_count', 'lit_count' are created.

5. Sentiment Analysis using the nltk.sentiment package: NLTK Library has a built-in, pretrained sentiment analyzer, called the Vader Sentiment Intensity Analyzer, which can automatically calculate the polarity scores of text data in terms of valence, i.e. positive or negative, and intensity, i.e. how strong the sentiment is. Hence, the following features are created:
 - a. 'pos': positive score component
 - b. 'neg': negative score component
 - c. 'neu': neutral score component
 - d. 'compound': the sum of the three score components
6. Sentiment Analysis using TextBlob library: TextBlob returns the polarity and the subjectivity of a sentence. Polarity score lies between [-1,1], where -1 defines the most negative sentiment, whereas 1 identifies the most positive sentiment. Subjectivity lies between [0,1] and measures the amount of

personal opinion and factual information contained in a text. A higher subjectivity score indicates a more subjective, personal content rather than objective, factual information. Two additional variables named ‘polarity’ and ‘subj’ are then added.

The following table summarizes the total features extracted in this phase:

Variable	Description
sentence_count	The number of sentences within a given text
flesch_ease	The Flesch Reading Ease Score within a given text
fog_index	The Gunning Fog Score within a given text
pos_count	Frequency count of occurrence of words on the positive wordlist of LM Dictionary within a given text
neg_count	Frequency count of occurrence of words on the negative wordlist of LM Dictionary within a given text
unc_count	Frequency count of occurrence of words on the uncertainty wordlist of LM Dictionary within a given text
lit_count	Frequency count of occurrence of words on the litigious wordlist of LM Dictionary within a given text
pos	The positive sentiment score within a given text according to the Vader Sentiment Intensity Analyzer
neg	The negative sentiment score within a given text according to the Vader Sentiment Intensity Analyzer
neu	The neutral sentiment score within a given text according to the Vader Sentiment Intensity Analyzer
compound	The compound sentiment score within a given text according to the Vader Sentiment Intensity Analyzer
polarity	The polarity score within a given text according to TextBlob
subj	The subjectivity score within a given text according to TextBlob

Table 4: Summary of the features created in the Feature Engineering step

The above features are carefully selected after the creation and visualization of the Pearson correlation matrix, as illustrated in the next Subchapter (4.4).

4.4. Data Exploration

Exploratory Data Analysis (EDA) is an important stage in a Data Science project, during which visual representations and summary statistics are applied on data, so as to observe their main characteristics, discover underlying patterns, derive useful insights, spot anomalies and check assumptions. The most common tools for EDA in text classification problems include the creation of word clouds, bar charts, histograms and correlation matrices.

4.4.1. Word Clouds and Bar Charts

A word cloud is a graphical representation of the frequency of the terms present in a text body. It is a collection of words depicted in different sizes, letters and colors. The bigger and bolder the word, the more often it appears within the source text.

Using the WordCloud library in Spyder IDE, a word cloud with the 100 most frequent three or four-word phrases in the MD&A section of the annual reports labelled as **Fraudulent** is created and presented below:



Figure 12: Word Cloud of 100 most common phrases of Fraudulent MD&A

For a more comprehensive review of the insights, bar plots are created using the Plotly Python Graphing Library. The following bar chart presents the frequency of 20 most common three or four-word phrases of the MD&A section of the annual reports labelled as **Fraudulent**:

Frequency of 20 Most Common Phrases | Class : Fraudulent

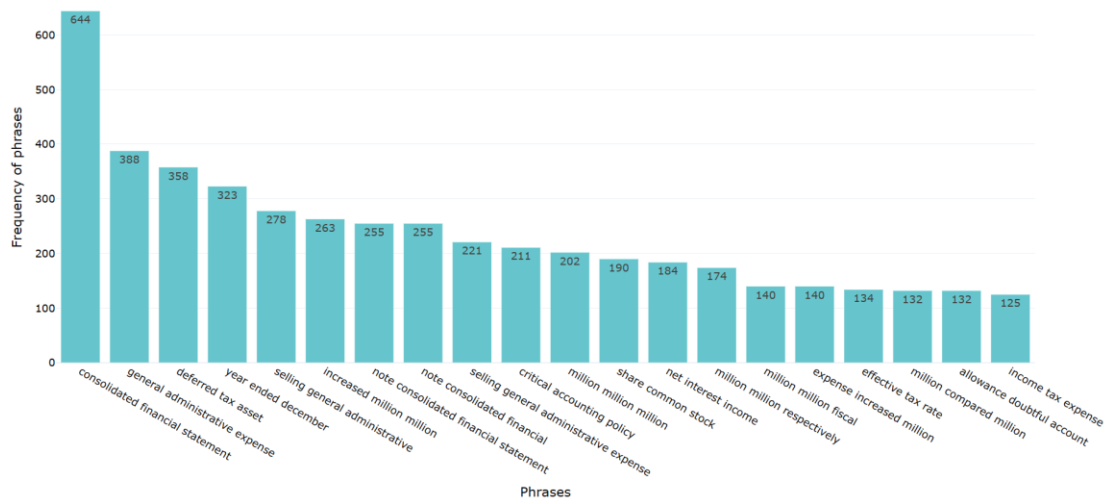


Figure 13: Bar chart of 20 most common phrases of Fraudulent MD&A

From the above plots it is obvious that the most common term of the ‘Fraudulent’ MD&A sections is the phrase ‘consolidated financial statements’. However, this item does not add much information to the content of the annual reports. On the contrary, the phrases ‘general administrative expense’, ‘deferred tax asset’, ‘critical accounting policy’, ‘share common stock’, ‘net interest income’, ‘effective tax rate’ and ‘allowance doubtful account’ are of particular interest, as they are more industry - specific and appear quite frequently in the MD&A sections.

Accordingly, the word cloud with the 100 most frequent three or four-word phrases in the MD&A section of the annual reports labelled as **non-Fraudulent** is illustrated in the following figure:



Figure 14: Word Cloud of 100 most common phrases of non-Fraudulent MD&A

The following bar chart depicts the frequency of 20 most common three or four-word phrases of the MD&A section of the annual reports labelled as **non-Fraudulent**:

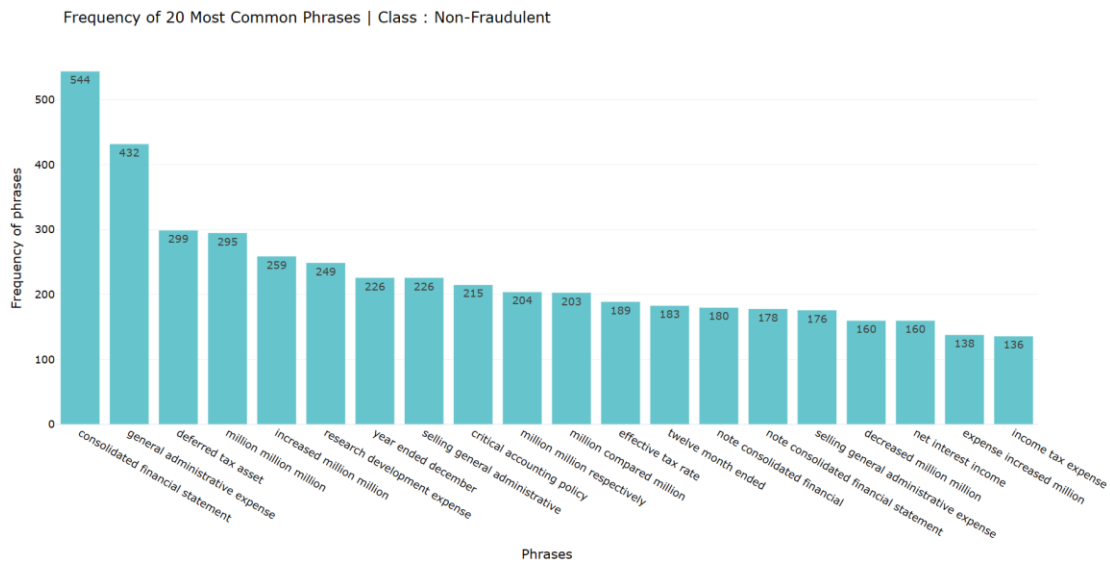


Figure 15: Bar chart of 20 most common phrases of non-Fraudulent MD&A

In the case of the ‘non-Fraudulent’ MD&A sections, the most frequently appeared phrase is also the element ‘consolidated financial statement’. In general, the most common terms are quite similar with the ones of the ‘Fraudulent’ MD&A sections, which can be attributed to the fact that there are specific accounting areas (e.g., general administrative expenses or deferred taxation) which are always widely discussed in the Financial Statements.

The terms that are repeatedly mentioned in the ‘Fraudulent’ but not in the ‘non-Fraudulent’ MD&A sections include ‘share common stock’ and ‘allowance doubtful account’. The latter phrase is a quite interesting finding, as it relates to the provision of expected credit losses as a result of the doubtful debts. In times of financial distress and continuous losses, managers may be tempted to avoid recording the appropriate provision for uncollectible accounts, which would result in an additional expense amount and thus reduced profits. This action could probably be considered as fraudulent, since it violates the financial reporting standards.

The direct comparison of the two classes (Fraudulent and Non-Fraudulent) is illustrated in the following graph:

Phrases - Comparison: Fraudulent | Non-Fraudulent

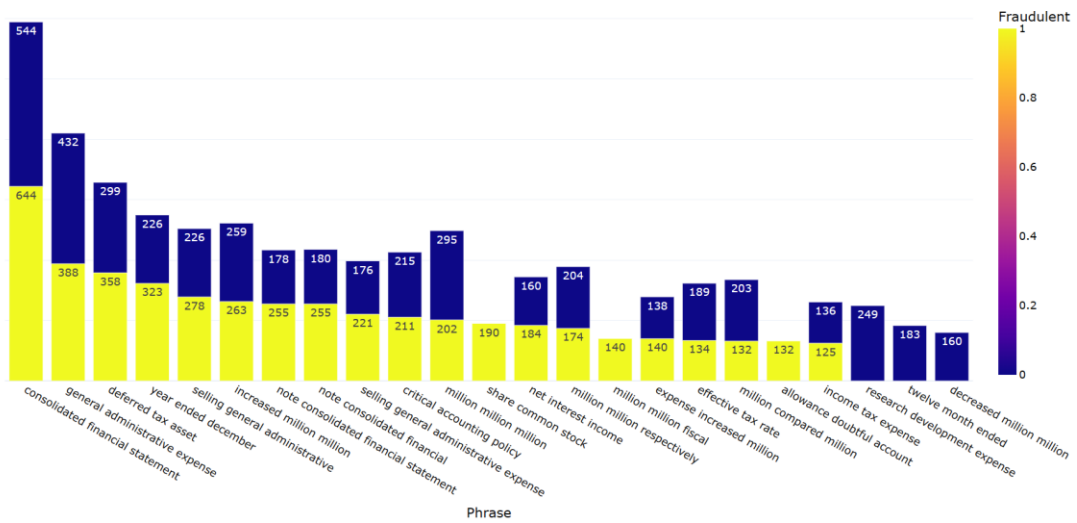


Figure 16: Bar chart of 20 most common phrases: Comparison of the two classes

4.4.2. Polarity Histogram

In statistics, a histogram is representation of the distribution of numerical data, where the data are binned and the count for each bin is represented. Using the Plotly Python Graphing Library, the following histogram is created:

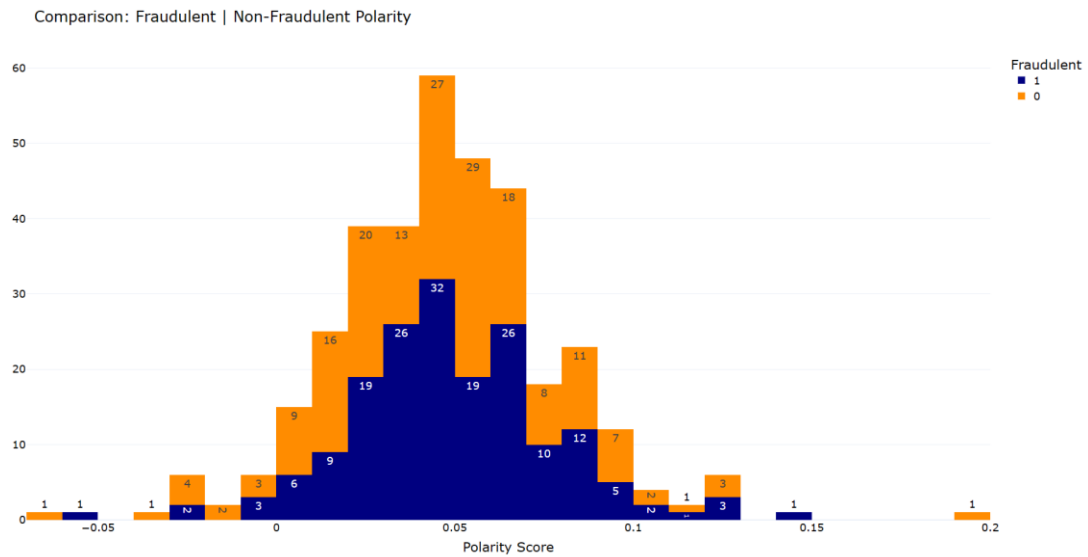


Figure 17: Histogram of Polarity scores: Comparison of the two classes

The histogram above shows the distribution of the polarity scores for the Fraudulent (in blue) and non-Fraudulent (in orange) MD&A sections of the sample. For both classes, it is pretty obvious that the overall sentiment is mainly positive, since the majority of the observations present a polarity score higher than zero. The two distributions are very similar and approach a normal distribution.

The corresponding graph for the distribution of the subjectivity scores is shown below:

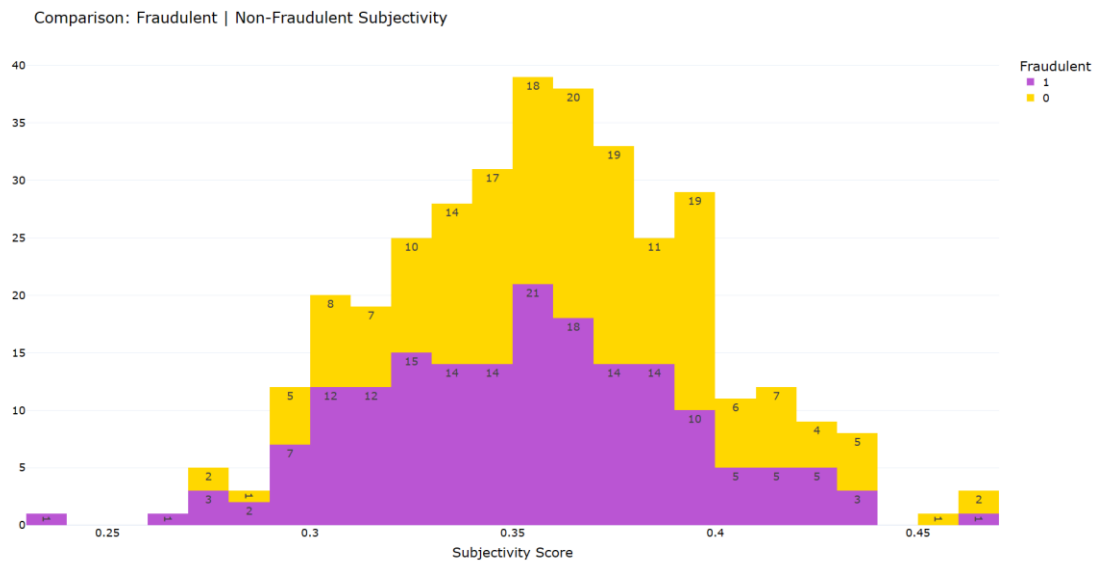


Figure 18: Histogram of Subjectivity scores: Comparison of the two classes

In both cases, the average subjectivity scores are close to 0.35, which is a relatively low value, indicating a higher degree of factual information within the MD&A section of the annual filings.

The similarities of the above distributions for both variables highlight that in this study, polarity and subjectivity ratios are probable not indicative factors in predicting fraudulent financial statements.

4.4.3. Correlation Matrix

The subchapter 4.3. discusses the process of feature engineering, which resulted in the creation of 13 new features in the dataset. Before moving on to the modelling phase of the research, it is essential to make the right selection of the variables to be used as the model inputs. In many cases, not all features are useful in building a robust machine learning model, so feature selection plays a vital role in every data science project.

Feature selection can be applied by calculating and analyzing the pairwise correlation of the dataset features. In statistics, correlation describes the extend to which two variables are linearly related, i.e., move in coordination with one another. In data science, features with high correlation typically have the same effect on the dependent variable, which makes it redundant to include all of them in the dataset. The use of only one of the highly correlated features is sufficient, and thus, the others

can be eliminated, since they are not expected to add much new information to the analysis.

The pairwise variable correlation is measured using the Pearson correlation coefficient, which returns a value between -1 and 1, where -1 indicates a total negative correlation, 0 means no correlation and 1 is total positive correlation. An absolute correlation coefficient of >0.7 signals high correlation. The correlation matrix of the 13 linguistic features is visualized with the Plotly Python Graphing Library, as shown in the figure below:

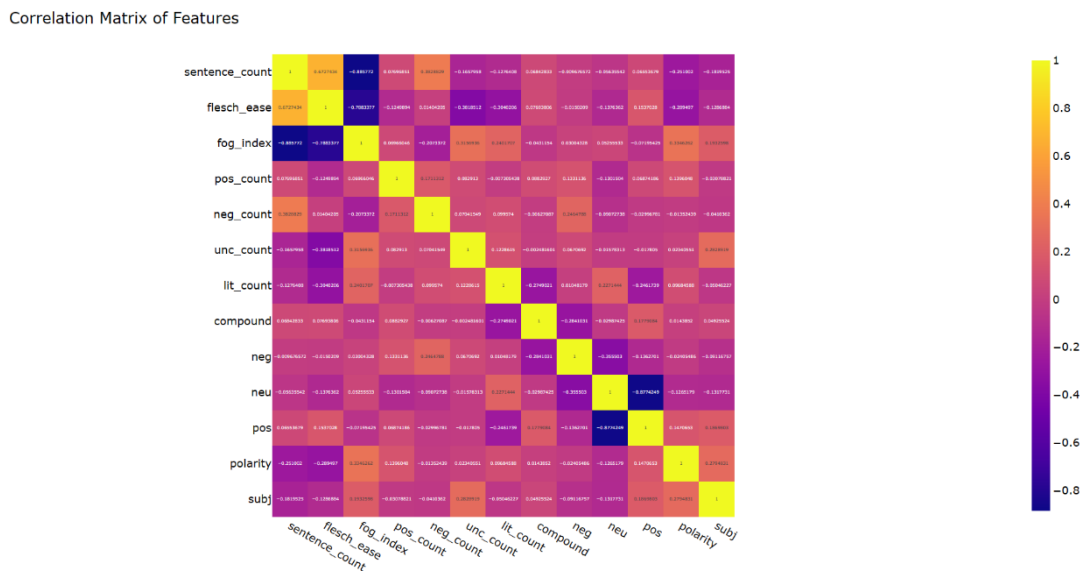


Figure 19: Correlation Matrix of the Linguistic Features

The above correlation matrix demonstrates a strong negative correlation between “fog_index” and “sentence_count” ($r=-0.886$), between “fog_index” and “flesh_ease” ($r=-0.788$) and between “neu” and “pos” ($r=-0.877$). Consequently, the variables “fog_index” and “neu” are excluded from the dataset.

4.5. Modelling

The final stage of a data science project is the process of training, developing and testing a machine learning model, which aims to provide meaningful findings and answers to the research hypotheses of the problem. It typically involves the partition of the input data to a training and test dataset, the creation of model(s) with the application of machine learning algorithms on the training dataset, the evaluation of the model(s) using the test dataset and the determination of the “best” solution by comparing the metrics between alternative methods.

In this study, two kinds of machine learning models have been developed. The first one uses as input data the linguistic features created in the “Feature engineering” step and selected after the Exploratory Data Analysis, with the aim to provide insights on the first and the second research questions. The second model is trained on the full textual MD&A data, which have been processed in the “Data Preparation” phase and are stored in the “Data_Edited” variable. The modelling on the complete MD&A document is related to the assessment of the first and the third research questions. Both techniques represent supervised classification models, which assign a class label: “Fraudulent” or “non-Fraudulent” (output variable) to every new observation, based on its set of features (input variables).

In the literature, the most commonly used classification algorithms involve Logistic Regression, Support Vector Machine, Decision Trees, Bayesian Classifiers, Ensemble Methods and Neural Networks. In this study the Random Forest algorithm has been utilized in both machine learning models developed. Random Forests, based on the concept of ensemble learning, combine the output of multiple decision trees, built on different random data samples with replacement, to reach a final result, that better solves the problem and improves the performance of the model. In addition to the built-in bagging method that they apply, Random Forests select the best split feature among a random subset of features, instead of considering every single variable. These two characteristics lead to a lower correlation, bias and variance across decision trees, resulting in more accurate predictions. They can be used for both Classification and Regression problems. Despite the fact that they are generally more complex and require more resources, computational power, and training time in comparison with other techniques, they are quite popular due to their performance and overall benefits. In particular, their reduced risk of overfitting, their flexibility to well with both categorical and continuous data without the requirement of normalization, their ability to handle missing values and calculate efficiently feature importances contribute to high-accuracy robust models. The `RandomForestClassifier()` provided by the Scikit-learn library is used in both models.

The efficiency and strength of a model should not be based exclusively on the data utilized for its development but instead should be evaluated on new, unseen data. For that reason, the initial dataset is split into a training and a test set using Scikit-learn `train_test_split()` method. The proportion of the partition is determined by

setting the parameters 'test_size' or 'train_size'. Furthermore, the random forest algorithm uses by default the bagging ensemble method, which applies bootstrapping on the training dataset, by selecting approximately the 2/3 of the observations, and setting aside the 1/3 of them (out-of-bag sample) for cross validation purposes. This element is controlled by the parameter 'bootstrap' of the Scikit-learn RandomForestClassifier(), whose default value 'True' is selected in both models.

As illustrated above, the selection of the hyperparameters is of utmost importance in the modelling stage. Indeed, models with different hyperparameters can produce completely different results on the same input dataset. The process of choosing the optimal combination of hyperparameters that maximizes the performance of a model is called 'Hyperparameter tuning'. It is a crucial step in every machine learning project, and thus it is also applied in this research for both models. The hyperparameters of the RandomForestClassifier() algorithm that are tested for their optimal values to each model are described as follows:

- **n_estimators**: The number of trees in the forest. Model time complexity increases as the value of this parameter grows. Its default value is 100.
- **max_features**: The number of features to consider when looking for the best split. Its default value is 'auto' and it can generally have four values: 'auto', 'sqrt', 'log2' and 'None':
 - In case of auto: considers $\text{max_features} = \sqrt{\text{n_features}}$
 - In case of sqrt: considers $\text{max_features} = \sqrt{\text{n_features}}$, (same as auto)
 - In case of log2: considers $\text{max_features} = \log_2(\text{n_features})$
 - In case of None: considers $\text{max_features} = \text{n_features}$
- **max_depth**: The maximum depth of the tree. It is one of the most important hyperparameters with regards to the accuracy of the model; as the depth of the tree increases, the model accuracy increases up to a certain limit, over which it starts to decrease gradually due to the overfitting of the model. Its default value is 'None', which specifies

that the nodes are expanded until all leaves are pure or until all leaves contain less than `min_samples_split` samples.

- `min_samples_split`: The minimum number of samples required to split an internal node. By increasing its value, the number of splits in the decision tree reduces, and therefore overfitting is prevented. However, the value should not be extremely large, causing the model to underfit. Generally, the value of this parameter should be set between 2 and 6. Its default value is 2.
- `min_samples_leaf`: The minimum number of samples required to be at a leaf node. It helps to reduce overfitting when there is an ample number of parameters. Its default value is 1.

In this study, hyperparameter tuning is operated using Scikit Learn `RandomizedSearchCV()`, which checks randomly-selected combinations of hyperparameter values and identifies the best-performing one for the particular model. The values tested per each hyperparameter are shown in the following table:

Hyperparameter	Value
n_estimators	200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000
max_features	'auto', 'sqrt'
max_depth	10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None
min_samples_split	2, 4, 5
min_samples_leaf	1, 2, 4

Table 5: Hyperparameter tuning values

Subsequently, the model is fitted on the training dataset using the optimal hyperparameters determined in the Hyperparameter tuning procedure. The efficiency of the model is then evaluated using the test dataset. There are multiple metrics which can be used to measure the predictive performance of a model, including accuracy, sensitivity, specificity, F1-score, AUC score etc. The selection of the most appropriate metric depends on the nature of the business problem. In Financial Statement Fraud Detection, according to Hajek and Henriques (2017) the cost of failing to classify

correctly fraudulent financial statements (FN rate) is higher than the cost of predicting non-fraudulent fillings as fraudulent (FP rate). Therefore, a higher sensitivity is preferred over a higher specificity or a higher precision, because an increase of the False Negatives (fraudulent companies classified as non-fraudulent) is more costly than an increase of the False Positives (non-fraudulent companies classified as fraudulent). The metrics used in the two models of this study include accuracy, sensitivity, specificity, precision and F1-score, the definitions of which are described as follows:

Metric	Formula	Description
Accuracy	$(TP+TN)/(P+N)$	The percentage of correctly classified samples.
Sensitivity	$TP/(TP+FN)$	The number of companies correctly classified as fraudulent as a percentage of all fraudulent companies.
Specificity	$TN/(TN+FP)$	the number of companies correctly classified as non-fraudulent as a percentage of all non-fraudulent companies.
Precision	$TP/(TP+FP)$	The number of companies correctly classified as fraudulent as a percentage of all companies classified as fraudulent.
F1-score	$2*(precision*recall)/(precision +recall)$	The harmonic mean of the precision and recall.

Table 6: Classification metrics

Where:

- TP: The number of fraudulent companies classified as fraudulent
- TN: The number of non-fraudulent companies classified as non-fraudulent
- FP: The number of non-fraudulent companies classified as fraudulent

- FN: The number of fraudulent companies classified as non-fraudulent
- P: The number of companies identified as fraudulent
- N: The number of companies identified as non-fraudulent

Accuracy and F1-score are calculated using Scikit-Learn methods `accuracy_score` and `f1_score` respectively, whereas Sensitivity, Specificity and Precision are computed based on the Confusion Matrix, which is generated with Scikit-Learn method `confusion_matrix`.

The final step of the modelling process, after the performance assessment, is to summarize the results, interpret the findings and draw meaningful conclusions with respect to the questions of the research. To this end, the technique of ‘Feature Importance’ can be utilized, so as to determine and understand which variables are critical in predicting the output variable. The `RandomForestClassifier()` algorithm has an integrated attribute, called “`feature_importances_`”, which relies its estimations on “Gini Importance” or “Mean Decrease in Impurity (MDI)”. Specifically, it computes how much each feature contributes to decreasing the impurity measure, when this feature is used in a split. Therefore, a feature with higher importance means that it is more likely to be used in a split and is more informative than the other features. This approach is also used in this study; in the first model the most relevant linguistic features are discovered, whereas in the second model, the most significant MD&A phrases are highlighted.

4.6. Results

4.6.1. Classification Model: Linguistic Features

In the first modelling method, the aim is to build a model to detect fraudulent financial statements, using the linguistic features of the MD&A section, as calculated in the Feature Engineering phase. In particular, the input variables include: sentence count, Flesch Reading Ease Score, frequency count of occurrence of negative, positive, uncertainty, and litigious words, negative, positive, and compound sentiment scores, and polarity and subjectivity. The output variable is a binary indicator of whether the financial statements are fraudulent ($y=1$) or non-fraudulent ($y=0$). As already discussed, the `RandomForestClassifier()` is used, and a train-test split method is employed, using 25% of the data for testing. After performing the hyperparameter

tuning process, using the RandomizedSearchCV to optimize the performance of the model, the following ‘optimal’ hyperparameters are obtained:

Hyperparameter	Value
n_estimators	600
max_features	'sqrt'
max_depth	None
min_samples_split	4
min_samples_leaf	4

Table 7: Optimal hyperparameters of the first model

The prediction metrics generated by the developed classification model are shown in the table below:

Metric	Value
Accuracy	62.92%
Sensitivity	61.36%
Specificity	64.44%
Precision	62.79%
F1-score	63.74%

Table 8: Prediction metrics of the first model

The Random Forest model has achieved an accuracy of 62.92%. Sensitivity is 61.36% which means that the model correctly identifies 61.36% of the actual fraudulent financial statements as fraudulent. Specificity is 64.44% which means that the model correctly identifies 64.44% of the actual non-fraudulent financial statements as non-fraudulent. Precision is 62.79% which means that out of all the cases that the model predicted as fraudulent, 62.79% of them are actually fraudulent. F1-score is 63.74% which is a measure of a model's accuracy that considers both precision and recall.

The feature importances derived from the Random Forest Model, which indicate which input variables have the greatest impact on the model predictions, are illustrated in the following bar chart:

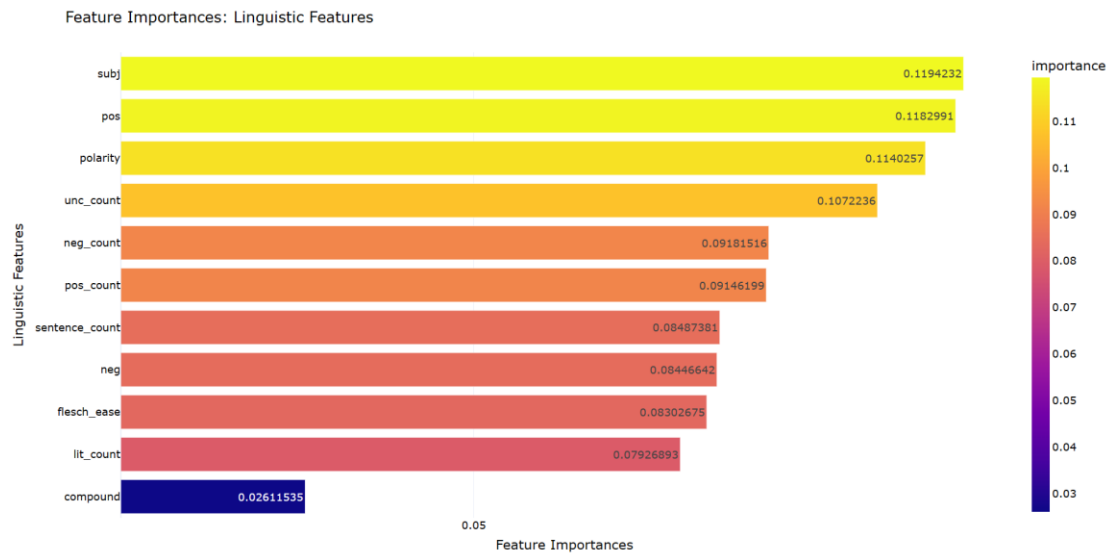


Figure 20: Feature importances of the first model

As it can be viewed in the above diagram, the highest importances relate to the subjectivity (0.1194) and positive sentiment score (0.1183). This suggests that these two variables have the strongest relationship with the target variable (fraudulent/non-fraudulent financial statements) in the dataset used. It could be explained that subjectivity variable, which measures the level of subjectivity in the text, and positive sentiment score, which measures the level of positive sentiment in the text, could be useful indicators of fraudulent financial statements. For example, a higher level of subjectivity in financial statement text might indicate that the statement is more likely to contain misleading or biased information, while a higher level of positive sentiment might indicate that the statement is overly optimistic and might not accurately reflect the financial performance of the company.

Overall, the implemented model provides insightful findings to the first and second research questions. Concretely, it appears that the linguistic features extracted from the MD&A text present some level of predictive power in Financial Statement Fraud Detection. The Random Forest model achieved an accuracy of 62.92% which is considered a decent performance, but not optimal. Furthermore, sensitivity is 61.36% which could be improved. The feature importances indicate that the subjectivity variable and positive sentiment score have the greatest impact on the model's predictions and thus are considered to be the most important features in the dataset used. This suggests that these linguistic features are informative and have a relationship with the target variable (fraudulent/non-fraudulent financial statements).

4.6.2. Classification Model: MD&A Text

In the second modelling phase, the focus of the classification model is the full MD&A section of financial statements, with the use of different n-gram ranges to extract significant insights from the text. N-grams are a sequence of N words that occur together in a document. They are useful in text mining and natural language processing tasks, because they are able to capture the context of words in a sentence, and thus provide highly informative features. They are capable of understanding the meaning of a word within the context of the words that come before and after it, which can provide more information than a single word alone.

In this study, three different versions of a Random Forest model have been developed, each trained on a different set of n-grams. The n-grams have been produced using the CountVectorizer, which is a common tool for feature extraction in text analysis. This technique uses tokenization, which splits the input text into words or subwords (n-grams) and then converts the tokens into a numerical representation, known as a bag-of-words representation, which can be used as the input to the machine learning model. In particular, the first model uses unigrams (n-grams with range (1,1)), the second model uses unigrams, bigrams, and trigrams (n-grams with range (1,3)), and the third model uses bigrams, trigrams and 4-grams (ngrams with range (2,4)). Again, the target variable is a binary indicator representing whether the financial statements are fraudulent ($y=1$) or non-fraudulent ($y=0$). The algorithm applied is the RandomForestClassifier() and the dataset is split, using 80% of the data for training, and 20% as a test subset.

4.6.2.1. Classification Model: Unigrams

The first model is trained on the single words extracted from the MD&A section of the financial statements. The hyperparameter tuning technique identifies the following hyperparameters as the best performing:

Hyperparameter	Value
n_estimators	800
max_features	'sqrt'
max_depth	100
min_samples_split	5
min_samples_leaf	1

Table 9: Optimal hyperparameters of the Unigrams model

The performance of this model is evaluated on the basis of the following metrics:

Metric	Value
Accuracy	73.24%
Sensitivity	72.73%
Specificity	73.68%
Precision	70.59%
F1-score	74.67%

Table 10: Prediction metrics of the Unigrams model

The Random Forest Classification model built on the unigrams is performing relatively well, with an overall accuracy of 73.24%. Sensitivity implies that 72.73% of the actual fraudulent financial statements were correctly classified as fraudulent. As with the first model, specificity is higher, which means that the model correctly classifies 73.68% of the actual non-fraudulent financial statements. Precision has the lowest value in comparison with the other metrics, and it indicates that 70.59% of the instances classified as fraudulent are actually fraudulent. Finally, F1-score is 74.67% which is the harmonic mean of precision and recall.

The unigrams that are considered to be the most informative to the detection of fraudulent financial statements are highlighted in the following bar plot, which presents the importances of the 20 most significant features:

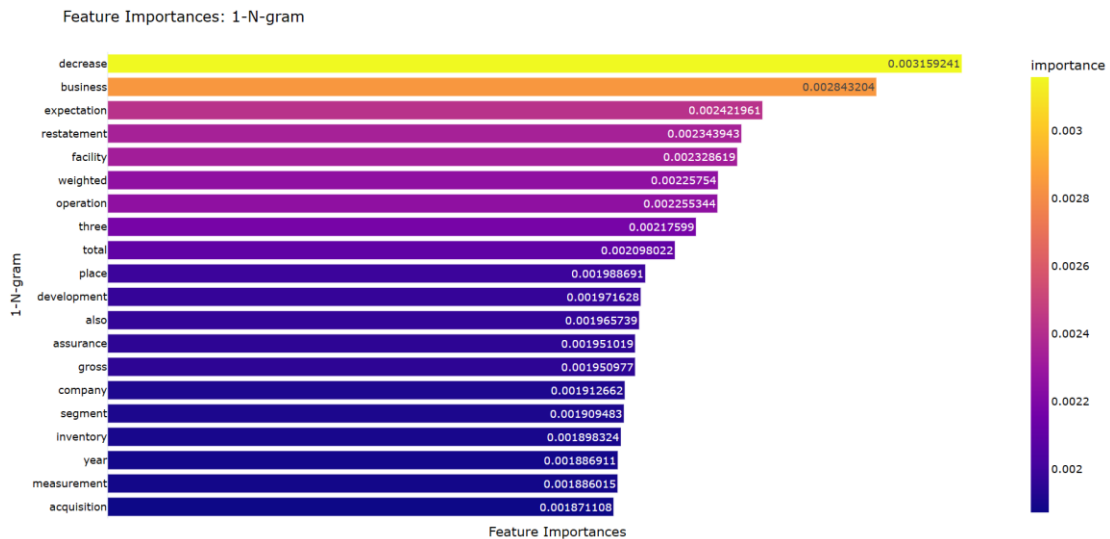


Figure 21: Feature importances of the Unigrams model

The above figure demonstrates that the words ‘decrease’ (0.03), ‘business’ (0.028), ‘expectation’ (0.0024) and ‘restatement’ (0.023) have the largest effect on the model, given the particular dataset used. This could suggest that the use of these words could indicate the presence of fraud. More specifically, the word ‘decrease’, having a negative meaning, might signify that the reporting company is experiencing losses and financial difficulties, which could be a red flag for falsified reporting. A possible explanation for the word ‘expectation’ is that the management might overstate the expected performance and future prospects of the business, in order to influence investors' opinions and make the entity appear more favorable. Finally, the word “restatement” could denote that the company had to revise its previously issued financial statements, possibly so as to correct prior period errors, omissions, or misstatements.

4.6.2.2. Classification Model: Unigrams, Bigrams and Trigrams

In the second model, one, two or three-word phrases have been used as input variables to the Random Forest Classifier. The most optimal hyperparameters for this model are shown in the following table:

Hyperparameter	Value
n_estimators	1.200
max_features	'sqrt'
max_depth	30
min_samples_split	5
min_samples_leaf	2

Table 11: Optimal hyperparameters of the Unigrams/Bigrams/Trigrams model

The accuracy, sensitivity, specificity, precision and F1 scores are described below:

Metric	Value
Accuracy	74.65%
Sensitivity	71.05%
Specificity	78.79%
Precision	79.41%
F1-score	74.29%

Table 12: Prediction metrics of the Unigrams/Bigrams/Trigrams model

The unigrams/bigrams/trigrams model outperforms the unigrams model in terms of accuracy, specificity and precision, however, presents lower sensitivity and F1-score, which are considered more important in the problem of Financial Statement Fraud Detection. Particularly, the model has correctly classified 74.65% of the observations in both classes (fraudulent and non-fraudulent), while it has recognized as fraudulent the 71.05% of the actual fraudulent financial statements. The specificity of 78.79% implies a greater capability of correctly identifying the actual non-fraudulent fillings, whereas the precision means that 79.41% of the financial reports labelled as 'fraudulent' are actually fraudulent. The F1-score of 74.29% is evidence of relatively low false positives and false negatives.

The n-grams associated with the greatest feature importances are revealed in the graph below:

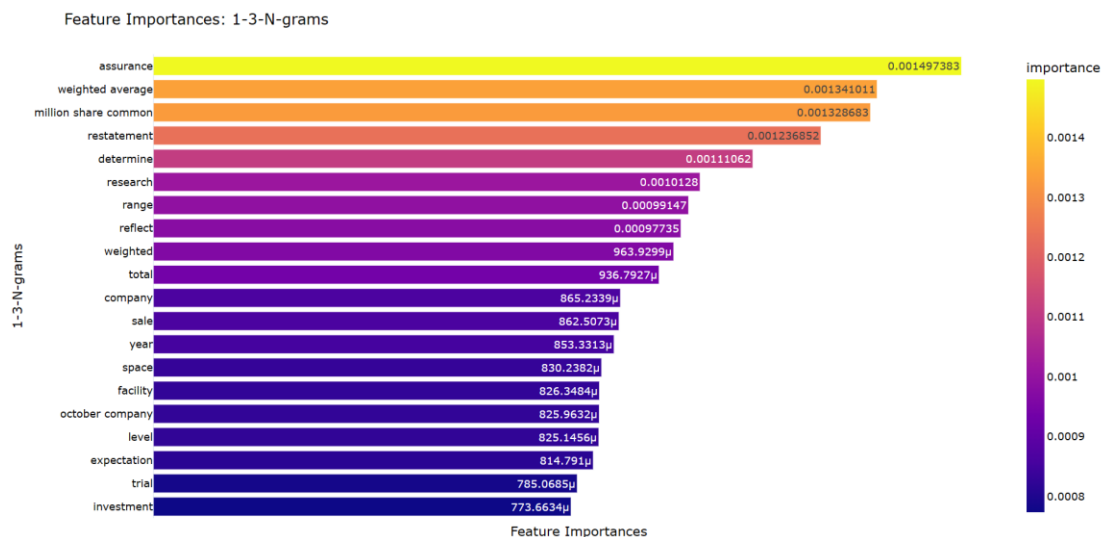


Figure 22: Feature importances of the Unigrams/Bigrams/Trigrams model

Out of the 20 variables presented in the above bar chart, the n-grams ‘assurance’ (0.0015), ‘weighted average’ (0.00134), ‘million share common’ (0.00133) and ‘restatement’ (0.00124) contribute the most in predicting fraudulent financial statements. Although this model also considers two or three-word phrases, it is prominent that the most important features are single words. With regards to the interpretation of the importances, the word ‘assurance’ could indicate an extravagant effort of the company to provide a sense of security or confidence in their financial statements, perhaps to cover reporting inaccuracies or omissions. The phrase ‘weighted average’ could suggest that financial calculations, which are complex or difficult to understand, have been used in the measurement of specific accounts. Finally, the phrase "million share common" probably relates to the reference of the market value of the company’s shares, which might be included in the MD&A section so as to compensate or justify the profitability and the position of the reporting entity.

4.6.2.3. Classification Model: Bigrams, Trigrams and 4-grams

The last model utilizes two-, three- and four-word phrases for the classification of the fraudulent and non-fraudulent annual fillings. The RandomizedSearchCV features the following values for the hyperparameters:

Hyperparameter	Value
n_estimators	800
max_features	'sqrt'
max_depth	100
min_samples_split	5
min_samples_leaf	1

Table 13: Optimal hyperparameters of the Bigrams/Trigrams/Fourgrams model

This model achieves the highest prediction metrics of interest (Accuracy, Sensitivity and F1-Score) in comparison with the former approaches:

Metric	Value
Accuracy	76.06%
Sensitivity	74.29%
Specificity	77.78%
Precision	76.47%
F1-score	76.71%

Table 14: Prediction metrics of the Bigrams/Trigrams/Fourgrams model

Based on the above metrics, the Bigrams/Trigrams/Fourgrams classification model has a quite competent performance, reaching an overall accuracy of 76.06%, which means that 76.06% of the time the model correctly predicts whether a financial statement is fraudulent or not. Sensitivity, also known as recall, denotes that out of all the fraudulent reports, the model correctly identifies 74.29% of them. The value of specificity suggests that the model classifies accurately the 77.87% of the total non-fraudulent instances. The precision metric of 76.47% is equivalent to the proportion of the actual fraudulent reports out of all observations classified as fraudulent by the model. Finally, the F1-score, which takes into account both precision and recall, is 76.71%.

The input features that stand out as the most useful in predicting the existence of financial statement fraud are as follows:

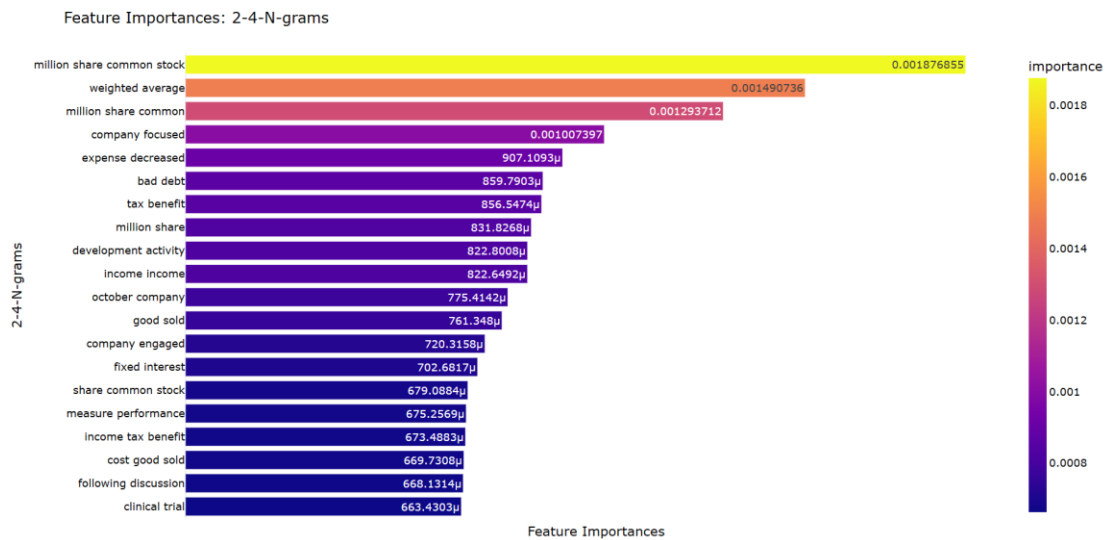


Figure 23: Feature importances of the Bigrams/Trigrams/Fourgrams model

The depicted n-grams with the highest scores, include phrases that have been previously identified in the former models presented, e.g., ‘million share common stock’, ‘weighted average’ and ‘million share common’. Of particular interest are the bigrams ‘bad debt’ and ‘tax benefit’, because they both relate to accounting practices that can be easily manipulated and are often associated with financial fraud. In detail, ‘bad debts’ refer to accounts receivable that are likely to be uncollectible. Generally, management might intentionally misclassify doubtful receivables to performing accounts, in order to avoid recording an expense allowance of bad debts, which would decrease the company’s assets and profits of the period. ‘Tax benefit’ corresponds to a reduction in the amount of taxes that a business is obliged to pay, due to certain deductions or credits. It may be linked to the deliberate increase of specific expenditures, aiming to reduce the taxable income. In short, both terms are probably fair indicators of fraudulent financial statements, as they could lead to reporting falsifications due to their subjective nature.

In conclusion, the three developed classification models achieved an accuracy of over 73%, with the third model being the best performing at 76.06%. Given that the cost of failing to classify correctly fraudulent financial statements (FN rate) is generally considered to be higher than the cost of predicting non-fraudulent fillings as fraudulent (FP rate), the sensitivity metric is deemed to be more important than the specificity, and thus a higher value of sensitivity is preferred. All models resulted in a sensitivity of over 71%, however, in all cases specificity was found to have a greater value, thus a better capability of detecting non-fraudulent financial statements.

Overall, all models seem to perform quite decently, which confirms the first research question, i.e., the textual data reported in the MD&A section of the Financial Statements are effective at the Financial Statement Fraud Detection. Additionally, all three models have highlighted key phrases, such as 'decrease', 'restatement', 'expectation', 'assurance', 'million share common', 'bad debt' and 'tax benefit', which could probably indicate the existence of fraud within the annual filings, and may thus be considered as 'red flags' for further investigation. This answers the third research question, i.e., the proposed models provide "red flag" indicators on word and phrase level to assist with the decision-making related to the Financial Statement Fraud Detection.

CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS

5.1. Discussion

Financial statement fraud detection is a crucial aspect of ensuring the integrity and stability of financial markets. With the advancement of technology and the increasing amount of data available, data science has emerged as a powerful tool for detecting and preventing financial fraud. An interesting field of data science is text analysis, which can be highly effective for detecting financial statement fraud, particularly when applied to the Management Discussion and Analysis (MD&A) section of financial statements. The MD&A section of financial statements contains valuable information about a company's performance and financial condition, including information about its financial results, risks and uncertainties, and future prospects. Text analysis can be used to extract information from the MD&A section of financial statements and identify patterns and anomalies that may indicate fraud. For example, natural language processing (NLP) techniques can be used to identify specific words or phrases that are commonly associated with fraudulent activity. Additionally, sentiment analysis can be used to determine the overall tone of the MD&A section, which can provide insight into the company's management's level of honesty and transparency. Overall, text analysis can be a valuable technique for detecting financial statement fraud by providing a more in-depth understanding of the information contained in the MD&A section.

To this end, this research project aimed to utilize intelligent NLP and Machine Learning Algorithms to discover the content of the MD&A section of the annual SEC filings and assess its efficiency in predicting fraudulent financial statements. There were two types of models developed; the first one used linguistic variables related to the context, sentiment, and overall tone of the MD&A section, while the second one was trained to the actual words and phrases within the document. Both models were developed on a Random Forest Classifier, which is known for its ability to handle a large number of features and high dimensional data, its robustness to overfitting, and its ability to give feature importance. The feature importance function of Random Forest enabled the identification of specific 'red-flag' indicators at a word and phrase

level, which could be very helpful in the early discovery of irregularities and misstatements within the financial statements.

The first model, having as inputs the linguistic variables, achieved a mediocre accuracy of 62.92% and a sensitivity of 61.36%. While these metrics may not be as high as desired, it is important to consider the complexity of the written linguistics, especially in formal documents and forms like annual filings. Perhaps the use of another data sample, the selection of a different combination of features, the alternative tuning of the hyperparameters, or the use of another machine-learning algorithm might have led to greater performance. In broad terms, the linguistic variables derived from the MD&A text, including the sentiment, the use of positive, negative, uncertainty, litigious words, subjectivity, and polarity, seem to have some level of predictive abilities in detecting financial statement fraud.

In the second model, three versions with three different ranges of n-grams (Unigrams/ Unigrams,Bigrams,Trigrams/ Bigrams,Trigrams,Fourgrams) as input variables were developed. All three models were found to have an impressive predictive performance, achieving over 71% in terms of both accuracy and sensitivity. The third model version, using two-, three- or four-word phrases of the MD&A section outperformed the others, predicting 76.06% of the total observations correctly. This could indicate that the use of multiple- instead of single-word phrases might be more efficient at capturing the contextual information of a document, understanding the actual meaning behind the linguistics, detecting text patterns, and handling language variations and misspellings. Furthermore, the models were able to reveal specific phrases as “red-flag” indicators, which could have a significant impact on the predictions and could assist auditors or other authorities in the process of determining whether an annual financial report requires further investigation. To conclude, the analysis of the text data included in the MD&A section of the financial statements, with the use of advanced Machine Learning models and Natural Language Processing (NLP) techniques, appears to be highly successful at detecting financial statement fraud.

5.2. Limitations of the study

As with all research studies, this one also has its limitations and weaknesses. First, the dataset used in this study was relatively small and consisted of a balanced sample of fraudulent and non-fraudulent firms. Thus, the diversity and representativeness of the population, the capability of the model to detect more complex patterns and generalize well to new data is questionable. In addition, this study has focused on companies registered with the U.S. Securities and Exchange Commission (SEC), and their annual filings, which are publicly available via APIs. This limits the applicability of the research findings to private entities, or entities that are not publicly traded.

Another weakness of this project is the fact that the sensitivity of all models, which is considered the most important metric in financial statement fraud detection, was lower, compared to specificity or precision. As with the majority of previous studies (Craja et. al (2020)), this one has also exhibited a greater predictive power in detecting non-fraudulent filings, instead of fraudulent ones. In other words, the developed models have resulted in a higher rate of false negatives, meaning that actual fraudulent cases are more likely to be misclassified. Given that the cost of failing to classify correctly fraudulent financial statements (FN rate) exceeds the cost of predicting non-fraudulent filings as fraudulent (FP rate), it would be beneficial to keep the sensitivity at a higher rate. In fact, the cost of a false negative can be extremely high, as it can lead to financial losses for investors and other stakeholders, reputational damage for the company, and legal and regulatory penalties. Notwithstanding, it should be noted that the sensitivity of over 71% of the three n-gram models is still considered acceptable in the field.

Finally, it is important to highlight that the specified words or phrases, that were identified by the models as the most significant features in detecting fraudulent financial statements, alone can't confirm fraud or be considered as a signal of fraud. Instead, they should be handled as a red flag for a deeper examination of the corresponding financial reports, taking into account additional factors as well, e.g., the context in which these terms are used. This indicates that reaching conclusions by interpreting these features in isolation could be misleading.

5.3. Recommendations for future research

Future research in the field of financial statement fraud detection using text data from the MD&A section could focus on several key areas in order to improve the performance of the models and overcome the limitations of the current study, i.e., the issue of lower sensitivity. Some potential future directions include:

1. Incorporating more diverse datasets: Expanding the dataset to include a greater diversity of companies, both in terms of size, industry, and location, and different reporting period filings, e.g., quarterly or semiannual reports, could help to improve the predictive power of the models and the generalizability of the results;
2. Combining text data with other data sources: Incorporating other types of data, such as financial data and ratios, and using a multivariate approach to detecting fraud, could contribute to the performance of the models;
3. Using more complex techniques: Using more sophisticated methods, such as neural networks, ensemble methods or deep learning models, could enhance their ability to capture the underlying patterns and relationships in the data and provide more accurate results;
4. Incorporating domain knowledge: Incorporating domain knowledge into the model, such as the specific accounting principles, taxonomies and industry practices, could be useful in increasing the interpretability of the findings and arriving at meaningful solutions;
5. Utilizing alternative word extraction and text analytics techniques: Instead of the n-grams approach followed in the current study, using word embeddings, TF-IDF, or transformer models to extract features from the text data, could result in more robust and accurate models.

References

Articles

Al-Hashedi, K.G., & Magalingam, P. (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Comput. Sci. Rev.*, *40*, 100402.

Alm, J., & Torgler, B. (2011). Do Ethics Matter? Tax Compliance and Morality. *Journal of Business Ethics*, *101*, 635-651.

Anti-Fraud Collaboration (2021). Mitigating the Risk of Common Fraud Schemes: Insights from SEC Enforcement Actions.

<https://antifraudcollaboration.org/mitigating-the-risk-of-common-fraud-schemes-insights-from-sec-enforcement-actions/>

Ashok, K.G. (2015). Creative accounting. *Indian Journal of Management Science (IJMS)*, *5(2)*, 59-64.

Association of Certified Fraud Examiners (2020). Report to the Nations 2020 Global Study on Occupational Fraud and Abuse. Corporate Finance Institute.

<https://acfepublic.s3-us-west-2.amazonaws.com/2020-Report-to-the-Nations.pdf>

Baralexis, S. (2004). Creative accounting in small advancing countries: The Greek case. *Managerial Auditing Journal*, *19*, 440-461.

Beneish, M.D. (1999). The Detection of Earnings Manipulation, *Financial Analysts Journal*, *55 (5)*, 24-36.

Benkraiem, R., Uyar, A., Kılıç, M., & Schneider, F. (2021). Ethical behavior, auditing strength, and tax evasion: A worldwide perspective. *Journal of International Accounting, Auditing and Taxation*, *43*, 100380.

Benston, G.J., & Hartgraves, A.L. (2002). Enron: what happened and what we can learn from it. *Journal of Accounting and Public Policy*, *21*, 105-127.

Burgoon, J.K., & Buller, D.B. (1994). Interpersonal deception: III. Effects of deceit on perceived communication and nonverbal behavior dynamics. *Journal of Nonverbal Behavior*, *18*, 155-184.

- Cecchini, M., Aytug, H., Koehler, G.J., & Pathak, P. (2010). Making words work: Using financial text as a predictor of financial events. *Decis. Support Syst.*, 50, 164-175.
- Cerullo, M.J., & Cerullo, V. (1999). Using neural networks to predict financial reporting fraud: Part 1. *Computer Fraud & Security*, 1999, 14-17.
- Ciocan, C.C. (2018). Creative Accounting And Fraud: A Comparative Approach. *Management Strategies Journal, Constantin Brancoveanu University*, 42(4), 157-163.
- Craja, P., Kim, A., & Lessmann, S. (2020). Deep learning for detecting financial statement fraud. *Decis. Support Syst.*, 139, 113421.
- Cressey, D.R. (1950). The Criminal Violation of Financial Trust. *American Sociological Review*, 15, 738.
- DE JESUS, T.A., Pinheiro, P., Kaizeler, C., & Sarmiento, M. (2020). Creative Accounting or Fraud? Ethical Perceptions Among Accountants. *International Review of Management and Business Research*, 9, 58-78.
- Deloitte (2009). Sample listing of fraud schemes.
<https://www2.deloitte.com/content/dam/Deloitte/in/Documents/risk/Corporate%20Governance/Audit%20Committee/in-gc-fraud-schemes-questions-to-consider-noexp.pdf>
- Deloitte (2020). Forensic Focus on COVID-19: Financial statement fraud. Available:
<https://www2.deloitte.com/content/dam/Deloitte/us/Documents/finance/us-forensic-focus-on-covid-19.pdf>
- Deloitte (2020). The risk of financial statement fraud in the wake of COVID-19. Available: https://www2.deloitte.com/content/dam/Deloitte/in/Documents/finance/in-fa-FinancialStatement_06.05.20_v2.pdf
- Deng, Q., & Mei, G. (2009). Combining self-organizing map and K-means clustering for detecting fraudulent financial statements. *2009 IEEE International Conference on Granular Computing*, 126-131.
- Dermitzakis, S. (1999). Il Penal Code, Special Section (Embezzlement) AP. 899/1999 Vol. E, 1220, Pinika, Vol. 12.

- Di Lullo, C. (2006). The Sanctity of Numbers, *Journal of Financial Service Professionals*, 60(3), 8-11.
- Dikmen, B., & Küçükkocaoğlu, G. (2010). The Detection of Earnings Manipulation: The Three Phase Cutting Plane Algorithm using Mathematical Programming. *Journal of Forecasting*, 15(4), 357–380.
- Du Toit, E. (2008). Characteristics of Companies with a Higher Risk of Financial Statement Fraud: A Survey of the Literature. *South African Journal of Accounting Research*, 22, 19 - 44.
- Durtschi, C., Hillison, W., & Pacini, C. (2004). The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data. *Journal of Forensic Account*, 5, 17-34.
- EY (2020). Preventing and Detecting Fraud. https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/topics/assurance/assurance-pdfs/ey-preventing-and-detecting-fraud.pdf
- Fan, C.S., Lin, C., & Treisman, D. (2010). Embezzlement Versus Bribery. *Microeconomics: Welfare Economics & Collective Decision-Making eJournal*.
- Glancy, F.H., & Yadav, S.B. (2011). A computational model for financial reporting fraud detection. *Decis. Support Syst.*, 50, 595-601.
- Goel, S., Gangolly, J., Faerman, S.R., & Uzuner, Ö. (2010). Can Linguistic Predictors Detect Fraudulent Financial Filings. *Journal of Emerging Technologies in Accounting*, 7, 25-46.
- 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.
- Hájek, P., & Henriques, R. (2017). Mining corporate annual reports for intelligent detection of financial statement fraud - A comparative study of machine learning methods. *Knowl. Based Syst.*, 128, 139-152.
- Hill, T.P. (1995). A Statistical Derivation of the Significant-Digit Law. *Statistical Science*, 10, 354-363.

- 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. *Decis. Support Syst.*, *50*, 585-594.
- International Federation of Accountants (2010). IAASB Handbook ISA 200. https://www.ifac.org/system/files/publications/files/ISA-200-Revised_2016.pdf
- International Federation of Accountants (2010). IAASB Handbook ISA 210. https://www.ifac.org/system/files/publications/files/ISA-210-Revised_2016.pdf
- International Federation of Accountants (2010). IAASB Handbook ISA 240. https://www.ifac.org/system/files/publications/files/ISA-240-Revised_2016.pdf
- International Federation of Accountants (2010). IAASB Handbook ISA 300. https://www.ifac.org/system/files/publications/files/ISA-300-Revised_2016.pdf
- International Federation of Accountants (n.d.). ISA 705: Modifications to the Opinion in the Independent Auditors' Report. https://www.ifac.org/system/files/publications/files/ISA-705-Revised_0.pdf
- Kamal, M.E., Salleh, M.F., & Ahmad, A. (2016). Detecting Financial Statement Fraud by Malaysian Public Listed Companies: The Reliability of the Beneish M-Score Model. *Jurnal Pengurusan UKM Journal of Management*, *46*, 23-32.
- Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data Mining techniques for the detection of fraudulent financial statements. *Expert Syst. Appl.*, *32*, 995-1003.
- KPMG (2020). COVID-19: Potential impact on financial reporting.
- Lin, J.W., Hwang, M.I., & Becker, J.D. (2003). A fuzzy neural network for assessing the risk of fraudulent financial reporting. *Managerial Auditing Journal*, *18*, 657-665.
- Li, Y. (2010). The Case Analysis of the Scandal of Enron. *International Journal of Biometrics*, *5*, 37.
- 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.
- Maka, K., Pazhanirajan, S., & Mallapur, S. (2020). Selection of most significant variables to detect fraud in financial statements. *Materials Today: Proceedings*.

- Marilena, Z., & Corina, I. (2012). Embellishment of Financial Statements Through Creative Accounting Policies and Options. *Procedia - Social and Behavioral Sciences*, 62, 347-351.
- McVay, S. (2006). Earnings Management Using Classification Shifting: An Examination of Core Earnings and Special Items. *The Accounting Review*, 81(3), 501-531.
- Melis, G., & Melis, A. (2004). Financial Reporting, Corporate Governance and Parmalat. Was it a Financial Reporting Failure?
- Montgomery, D.D., Beasley, M.S., Menelaides, S.L., & Palmrose, Z. (2002). Auditors' New Procedures for Detecting Fraud; ED's Proposed Changes Address Fraudulent Financial Statements. *Journal of accountancy*, 193, 63.
- Mulford, C.W., & Comiskey, E.E. (2002). The Financial Numbers Game: Detecting Creative Accounting Practices.
- Ngai, E.W., Hu, Y., Wong, Y.H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decis. Support Syst.*, 50, 559-569.
- Nieschwietz, R.J., Schultz Jr., J.J., Zimbelman, M.F. (2000). Empirical research on external auditors' detection of financial statement fraud. *Journal of Accounting Literature*, 19, 190-246
- Nigrini, M. (1996). A taxpayer compliance application of Benford's Law. *Journal of the American Taxation Association*, 18(1), 72-91.
- Porter, B., & Cameron, A. (1987). Company fraud – what price the auditor?. *Accountant's Journal*, 12, 44-47.
- PwC (2013). Understanding a Financial Statement Audit.
<https://www.pwc.com/im/en/services/Assurance/pwc-understanding-financial-statement-audit.pdf>
- Ravisankar, P., Ravi, V., Raghava Rao, G., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decis. Support Syst.* 50, 491-500.

- Rezaee, Z. (2005). Causes, Consequences, and Deterrence of Financial Statement Fraud. *Critical Perspectives on Accounting*, 16(3), 277-298.
- Rubasundram, G.A. (2015). Perceived “Tone From the Top” During A Fraud Risk Assessment. *Procedia. Economics and finance*, 28, 102-106.
- Sabau, A., Safta, L., Miron, G. and Monica Violeta, A. (2020). Manipulation of Financial Information through Creative Accounting: Case Study at Companies listed on the Romanian Stock Exchange. *Rsep Conferences* 18, 64-80.
- Sadgali, I., Sael, N., & Benabbou, F. (2019). Performance of machine learning techniques in the detection of financial frauds. *Procedia Computer Science*, 148, 45-54.
- Shi, J., Ausloos, M., & Zhu, T. (2017). Benford’s law first significant digit and distribution distances for testing the reliability of financial reports in developing countries. *Physica A*, 492, 878–888.
- Spathis, C. (2002). Detecting false financial statements using published data: some evidence from Greece. *Managerial Auditing Journal*, 17, 179-191.
- Stuart, R. (2020). Consolidations at a Glance. *RSM US LLP*.
- Stolowy, H., & Breton, G. (2004). Accounts Manipulation: A Literature Review and Proposed Conceptual Framework. *Review of Accounting and Finance*, 3, 5-92.
- Throckmorton, C.S., Mayew, W.J., Venkatachalam, M., & Collins, L.M. (2015). Financial fraud detection using vocal, linguistic and financial cues. *Decis. Support Syst.*, 74, 78-87.
- Vasek, M., & Moore, T.W. (2015). There’s No Free Lunch, Even Using Bitcoin: Tracking the Popularity and Profits of Virtual Currency Scams. *Financial Cryptography and Data Security*.
- Wang, J., Liao, Y., Tsai, T., & Hung, G. (2006). Technology-based Financial Frauds in Taiwan: Issues and Approaches. *2006 IEEE International Conference on Systems, Man and Cybernetics*, 2, 1120-1124.

Wei, W., Li, J., Cao, L., Ou, Y., & Chen, J. (2013). Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web*, 16, 449-475.

Wyrobek, J. (2020). Application of machine learning models and artificial intelligence to analyze annual financial statements to identify companies with unfair corporate culture. *International Conference on Knowledge-Based Intelligent Information & Engineering Systems*.

Zdanowicz, J.S. (2009). Trade-Based Money Laundering and Terrorist Financing. *Review of Law & Economics*, 5, 855 - 878.

Zhou, W., & Kapoor, G. (2011). Detecting evolutionary financial statement fraud. *Decis. Support Syst.*, 50, 570-575.

Books

Νεγκάκης, Χ., Ταχυνάκης, Π. (2013). Σύγχρονα Θέματα Ελεγκτικής και Εσωτερικού Ελέγχου. Εκδόσεις Διπλογραφία.

Jones, M. (2011). *Creative Accounting, Fraud and International Scandals*. John Wiley Sons Ltd.

Kevin, A. (2003). *Uncovering Creative Accounting*. Pearson Education Limited.

Naser, K. H. (1993). *Creative Financial Accounting: Its Nature and Use*. Prentice Hall.

Nguyen, K. (2010). *Financial statement fraud: motives, methods, cases and detection*. Dissertation.com.

Nigrini, M. (2012). *Benford's Law*. John Wiley Sons Ltd.

Wells, J. (2017). *Corporate Fraud Handbook*. John Wiley Sons Ltd.

Zack, G. (2012). *Financial Statement Fraud*. John Wiley Sons Ltd.

Websites

American Institute of Certified Public Accountants (n.d.). Auditing. <https://aaahq.org/>

ASIC (2020). COVID-19 implications for financial reporting and audit - frequently asked questions. <https://asic.gov.au/regulatory-resources/financial-reporting-and-audit/covid-19-implications-for-financial-reporting-and-audit-frequently-asked-questions-faqs/>

Association of Certified Fraud Examiners. (n.d.). Fraud 101: What is Fraud? <https://www.acfe.com/fraud-resources/fraud-101-what-is-fraud>

Association of Chartered Accountants (n.d.). <https://www.accaglobal.com/>

Collins, J.C. (2017). Excel and Benford's Law to detect fraud. *Journal of Accountancy*. <https://www.journalofaccountancy.com/issues/2017/apr/excel-and-benford-s-law-to-detect-fraud.html>

Corporate Finance Institute. (n.d.). Corporate Fraud. <https://corporatefinanceinstitute.com/resources/esg/corporate-fraud/>

Criminal Lawyer Group. (n.d.). Securities and Commodities Fraud. <https://www.criminallawyergroup.com/practice-areas/securities-and-commodities-fraud/>

IFAC, (n.d.). Reporting and Fraud Risk Arising from COVID-19 Pose Significant Challenges for Professional Accountants. <https://www.ifac.org/knowledge-gateway/contributing-global-economy/discussion/reporting-and-fraud-risk-arising-covid-19-pose-significant-challenges-professional-accountants>

JarFraud (2022) Fraud Detection. <https://github.com/JarFraud/FraudDetection>

SEC EDGAR Filings API <https://sec-api.io>

The Law Dictionary. (n.d.). Bribery. <https://thelawdictionary.org/bribery/>

US Department of Justice. (n.d.). Criminal Resource Manual 826: US Code, Title 18, Crimes and Criminal Procedure, 18 U.S.C. § 1344. <https://www.justice.gov/archives/jm/criminal-resource-manual-826-applicability-18-usc-1344>

Wikipedia. (n.d.). Corruption. <https://en.wikipedia.org/wiki/Corruption>

Wikipedia. (n.d.). Embezzlement. <https://en.wikipedia.org/wiki/Embezzlement>

Wikipedia. (n.d.). Flesch–Kincaid Readability Tests. https://en.wikipedia.org/wiki/Flesch%E2%80%93Kincaid_readability_tests

Wikipedia. (n.d.). Insurance Fraud. https://en.wikipedia.org/wiki/Insurance_fraud

Wikipedia. (n.d.). Σκάνδαλο Κοσκωτά.

https://el.wikipedia.org/wiki/%CE%A3%CE%BA%CE%AC%CE%BD%CE%B4%CE%B1%CE%BB%CE%BF_%CE%9A%CE%BF%CF%83%CE%BA%CF%89%CF%84%CE%AC

Appendix

The Python code developed in this study can be found in the following Github repository:

<https://github.com/morfoulaisd/FinancialStatementFraud>