fraudsters and limit the damage caused by financial fraud, innovative fraud detection methods are required that are capable of identifying indications of the continuously fraudulent activities. The research described in this thesis examined the contribution of text analysis to detecting indications of fraud in the annual reports of companies de. In an annual report, a company presents its financial results and activities from the previous year. In addition, the annual report contains textual explanations. In case of e annual report does not provide a fair view of the financial position of the company. For the research described in this thesis, a total of 1,727 annual reports have been collected, 402 are of the years and companies in which fraudulent activities took place, and which have an impact on the information disclosed in the annual report. Since it is assumed companies do not engage in such fraudulent activities, the majority, 77%, of the annual reports in the data set are not fraudulent. Furthermore, the composition of this data set o account the possibility that the year the annual report is concerned with, the sector in which the company operates and the size of the company may affect the information in the annual report. Therefore, for each fraud annual report, non fraudulent annual reports from the same year, of a company in the same sector, and of a comparable size are added to the data set. A method for the automatic extraction of information from annual reports has been the data set. A method for the automatic extraction of information from annual reports has been proposed to obtain the data needed to compile the data set. By applying this method, the year, sector and size of the company can be determined for a large number of annual reports. The approach has also been used to extract the Management Discussion & (MD&A) section, which is the part of the annual report the research in this thesis focuses on. In the MD&A section, the company provides information concerning the its nce and the activities of the preceding year and the expectations for the following year. The first models developed for the research described in this thesis, analyze the texts ng the words (unigrams) in the MD&A section. These word counts are normalized by the term frequency-inverse document frequency (TF-IDF) method that takes into account h of the text and the frequency with which a word occurs in the data set. The most informative words are determined using the chi-square feature selection model. This ation of the text is the input of the machine learning algorithms Naive Bayes (NB) and Support Vector Machine (SVM). These algorithms learn patterns to classify the annual s `fraud' or `no fraud'. The NB model classifies the texts based on probability calculations. The SVM model uses unigrams to create a vector space. Subsequently, the SVM n determines the most optimal hyperplane in the space that separates the 'fraud' and 'no fraud' texts. The NB model shows the best performance. The percentage of correctly d annual reports is 89%. Subsequently, the NB and SVM models based on unigrams are expanded with the linguistic features of the text found to be informative in the previous concerning the detection of fraud or deception in text. We subdivided the linguistic features into six categories. The first category comprises groups of two consecutive words). The second category consists of features that describe the general properties of the text, such as the total numbers in the text. The third category describes the complexity of using measures for the complexity of the words and sentences. The fourth category focuses on the grammatical aspects of the text. The fifth category consists of measures that e the readability of a text, expressed as the number of years of education needed to understand the text. The final category concerns, the Linguistic Inquiry and Word Count tool that is used to extract psychological features, such as emotions, from the text. Furthermore, we developed a new type of feature that reflects the grammatical relations words. The classification results show that only the addition of bigrams to the SVM model improves the result slightly. The other categories of linguistic features do not improve . The latest development in machine learning is the use of deep learning. By using networks consisting of several layers complex patterns can be found. A Convolutional Neural (CNN) model has been found successful in research that classifies texts in domains other than fraud. With this model, each word in the text is represented by a vector (word ng). Word embeddings aim to include the semantic relationships between words, in addition to the representation of the individual words. Due to the limited computer capacity during the research described in this thesis, experimentation with the CNN model was performed using 40% of the original data set. In order to compare the results with the models, an NB model was also developed on this smaller data set. The results are significantly lower, but also show that the CNN model achieves slightly better results than odel. The results show that text analysis can contribute to the detection of indications of fraud. However, it is desirable to further enhance the performance of the models. This chieved either by improving their performance or by adding additional sources of information. Future research may experiment with machine learning algorithms, such as the a combination of various algorithms. The model for text can be expanded with models that use the company's financial information. More textual information may be added to l from the annual report itself or from other documents of the company. To catch fraudsters and limit the damage caused by financial fraud, innovative fraud detection methods red that are capable of identifying indications of the continuously changing fraudulent activities. The research described in this thesis examined the contribution of text analysis ing indications of fraud in the annual reports of companies worldwide. In an annual report, a company presents its financial results and activities from the previous year. In the annual report contains textual explanations. In case of fraud, the annual report does not provide a fair view of the financial position of the company. For the research d in this thesis, a total of 1,727 annual reports have been collected, of which 402 are of the years and companies in which fraudulent activities took place, and which have an n the information disclosed in the annual report. Since it is assumed that most companies do not engage in such fraudulent activities, the majority, 77%, of the annual reports ta set are not fraudulent. Furthermore, the composition of this data set takes into account the possibility that the year the annual report is concerned with, the sector in which any operates and the size of the company may affect the information disclosed in the annual report. Therefore, for each fraud annual report, non fraudulent annual reports from year, of a company in the same sector, and of a comparable size are added to the data set. A method for the automatic extraction of information from annual reports has been to obtain the data needed to compile the data set. By applying this method, the year, sector and size of the company can be determined for a large number of annual reports. oach has also been used to extract the Management Discussion & Analysis (MD&A) section, which is the part of the annual report the research in this thesis focuses on. In the ection, the company provides information concerning the its performance and the activities of the preceding year and the expectations for the following year. The first models d for the research described in this thesis, analyze the texts by counting the words (unigrams) in the MD&A section. These word counts are normalized by the term y-inverse document frequency (TF-IDF) method that takes into account the length of the text and the frequency with which a word occurs in the data set. The most informative e determined using the chi-square feature selection model. This representation of the text is the input of the machine learning algorithms Naive Bayes (NB) and Support Vector (SVM). These algorithms learn patterns to classify the annual reports as `fraud' or `no fraud'. The NB model classifies the texts based on probability calculations. The SVM model es unigrams to create a vector space. Subsequently, the SVM algorithm determines the most optimal hyperplane in the space that separates the 'fraud' and 'no fraud' texts. The el shows the best performance. The percentage of correctly classified annual reports is 89%. Subsequently, the NB and SVM models based on unigrams are expanded with the e features of the text found to be informative in the previous research concerning the detection of fraud or deception in text. We subdivided the linguistic features into six s. The first category comprises groups of two consecutive words (bigrams). The second category consists of features that describe the general properties of the text, such as the bers in the text. The third category describes the complexity of the text using measures for the complexity of the words and sentences. The fourth category focuses on the ical aspects of the text. The fifth category consists of measures that determine the readability of a text, expressed as the number of years of education needed to understand the final category concerns, the Linguistic Inquiry and Word Count (LIWC) tool that is used to extract psychological features, such as emotions, from the text. Furthermore, we d a new type of feature that reflects the grammatical relations between words. The classification results show that only the addition of bigrams to the SVM model improves the ghtly. The other categories of linguistic features do not improve the result. The latest development in machine learning is the use of deep learning. By using networks consisting l layers complex patterns can be found. A Convolutional Neural Network (CNN) model has been found successful in research that classifies texts in domains other than fraud. model, each word in the text is represented by a vector (word embedding). Word embeddings aim to include the semantic relationships between words, in addition to the ation of the individual words. Due to the limited computer capacity available during the research described in this thesis, experimentation with the CNN model was performed % of the original data set. In order to compare the results with the previous models, an NB model was also developed on this smaller data set. The results are significantly lower, show that the CNN model achieves slightly better results than the NB model. The results show that text analysis can contribute to the detection of indications of fraud. However, ble to further enhance the performance of the models. This may be achieved either by improving their performance or by adding additional sources of information. Future may experiment with machine learning algorithms, such as the CNN or a combination of various algorithms. The model for text can be expanded with models that use the company's financial information. More textual information may be added to the model from the annual report itself or from other documents of the company. To catch fraudsters and limit the damage caused by financial fraud, innovative fraud detection methods are required that are capable of identifying indications of the continuously changing fraudulent activities. The research described in this thesis examined the contribution of text analysis to detecting indications of arch described in this thesis examined the contribution of text analysis to detecting indications of fraud in the annual reports of companies worldwide. In an annual report, a presents its financial results and activities from the previous year. In addition, the annual report contains textual explanations. In case of fraud, the annual report does not fair view of the financial position of the company. For the research described in this thesis, a total of 1,727 annual reports have been collected, of which 402 are of the years panies in which fraudulent activities took place, and which have an impact on the information disclosed in the annual report. Since it is assumed that most companies do not such fraudulent activities, the majority, 77%, of the annual reports in the data set are not fraudulent. Furthermore, the composition of this data set takes into account the y that the year the annual report is concerned with, the sector in which the company operates and the size of the company may affect the information disclosed in the annual herefore, for each fraud annual report, non fraudulent annual reports from the same year, of a company in the same sector, and of a comparable size are added to the data set. A or the automatic extraction of information from annual reports has been proposed to obtain the data needed to compile the data set. By applying this method, the year, sector of the company can be determined for a large number of annual reports. The approach has also been used to extract the Management Discussion & Analysis (MD&A) section, the part of the annual report the research in this thesis focuses on. In the MD&A section, the company provides information concerning the its performance and the activities ceding year and the expectations for the following year. The first models developed for the research described in this thesis, analyze the texts by counting the words (unigrams) D&A section. These word counts are normalized by the term frequency-inverse document frequency (TF-IDF) method that takes into account the length of the text and the y with which a word occurs in the data set. The most informative words are determined using the chi-square feature selection model. This representation of the text is the input achine learning algorithms Naive Bayes (NB) and Support Vector Machine (SVM). These algorithms learn patterns to classify the annual reports as `fraud' or `no fraud'. The el classifies the texts based on probability calculations. The SVM model uses unigrams to create a vector space. Subsequently, the SVM algorithm determines the most optimal ne in the space that separates the 'fraud' and 'no fraud' texts. The NB model shows the best performance. The percentage of correctly classified annual reports is 89%. ently, the NB and SVM models based on unigrams are expanded with the linguistic features of the text found to be informative in the previous research concerning the detection or deception in text. We subdivided the linguistic features into six categories. The first category comprises groups of two consecutive words (bigrams). The second category of features that describe the general properties of the text, such as the total numbers in the text. The third category describes the complexity of the text using measures for the ty of the words and sentences. The fourth category focuses on the grammatical aspects of the text. The fifth category consists of measures that determine the readability of a ressed as the number of years of education needed to understand the text. The final category concerns, the Linguistic Inquiry and Word Count (LIWC) tool that is used to extract gical features, such as emotions, from the text. Furthermore, we developed a new type of feature that reflects the grammatical relations between words. The classification results t only the addition of bigrams to the SVM model improves the result slightly. The other categories of linguistic features do not improve the result. The latest development in learning is the use of deep learning. By using networks consisting of several layers complex patterns can be found. A Convolutional Neural Network (CNN) model has been ccessful in research that classifies texts in domains other than fraud. With this model, each word in the text is represented by a vector (word embedding). Word embeddings aim e the semantic relationships between words, in addition to the representation of the individual words. Due to the limited computer capacity available during the research d in this thesis, experimentation with the CNN model was performed using 40% of the original data set. In order to compare the results with the previous models, an NB model developed on this smaller data set. The results are significantly lower, but also show that the CNN model achieves slightly better results than the NB model. The results show analysis can contribute to the detection of indications of fraud. However, it is desirable to further enhance the performance of the models. This may be achieved either by g their performance or by adding additional sources of information. Future research may experiment with machine learning algorithms, such as the CNN or a combination of lgorithms. The model for text can be expanded with models that use the company's financial information. More textual information may be added to the model from the annual elf or from other documents of the company. To catch fraudsters and limit the damage caused by financial fraud, innovative fraud detection methods are required that are capable fying indications of the continuously changing fraudulent activities. The research described in this thesis examined the contribution of text analysis to detecting indications of the annual reports of companies worldwide. In an annual report, a company presents its financial results and activities from the previous year. In addition, the annual report textual explanations. In case of fraud, the annual report does not provide a fair view of the financial position of the company. For the research described in this thesis, a total annual reports have been collected, of which 402 are of the years and companies in which fraudulent activities took place, and which have an impact on the information d in the annual report. Since it is assumed that most companies do not engage in such fraudulent activities, the majority, 77%, of the annual reports in the data set are not nt. Furthermore, the composition of this data set takes into account the possibility that the year the annual report is concerned with, the sector in which the company operates ize of the company may affect the information disclosed in the annual report. Therefore, for each fraud annual report, non fraudulent annual reports from the same year, of a in the same sector, and of a comparable size are added to the data set. A method for the automatic extraction of information from annual reports has been proposed to obtain needed to compile the data set. By applying this method, the year, sector and size of the company can be determined for a large number of annual reports. The approach has also d to extract the Management Discussion & Analysis (MD&A) section, which is the part of the annual report the research in this thesis focuses on. In the MD&A section, the provides information concerning the its performance and the activities of the preceding year and the expectations for the following year. The first models developed for the described in this thesis, analyze the texts by counting the words (unigrams) in the MD&A section. These word counts are normalized by the term frequency-inverse document y (TF-IDF) method that takes into account the length of the text and the frequency with which a word occurs in the data set. The most informative words are determined using quare feature selection model. This representation of the text is the input of the machine learning algorithms Naive Bayes (NB) and Support Vector Machine (SVM). These ns learn patterns to classify the annual reports as `fraud' or `no fraud'. The NB model classifies the texts based on probability calculations. The SVM model uses unigrams to

TEXT MINING TO DETECT INDICATIONS OF FRAUD IN ANNUAL REPORTS WORLDWIDE

Marcia Fissette

Promotie Commissie

Voorzitter/secretaris	Prof. Dr. T.A.J. Toonen
Promotor	Prof. Dr. Ir. B.P. Veldkamp
	Prof. Dr. Ir. T. de Vries
Leden	Prof. Dr. C.A.W. Glas
	Prof. Dr. M. Junger
	Prof. Dr. A. Shahim
	Prof. Dr. M. Pheijffer
	Prof. Dr. R.G.A. Fijneman
	Dr. R. Matthijsse

Marcia Fissette
Text mining to detect indications of fraud in annual reports worldwide
PhD thesis, University of Twente, Enschede, The Netherlands
ISBN: 978-90-365-4420-7
DOI: 10.3990/1.9789036544207
Printing: Ridderprint BV | www.ridderprint.nl

TEXT MINING TO DETECT INDICATIONS OF FRAUD IN ANNUAL REPORTS WORLDWIDE

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. T.T.M. Palstra,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
donderdag 21 december 2017 om 16.45 uur

door

Marcia Valentine Maria Fissette
geboren op 31 mei 1986
te Maastricht, Nederland

Dit proefschrift is goedgekeurd door promotoren:
Prof. Dr. Ir. B.P. Veldkamp
Prof. Dr. Ir. T. de Vries

Contents

1 Introduction

The world is repeatedly faced with major financial fraud cases. The fraud schemes and fraudulent activities used vary from case to case and develop over time. Therefore, fraud detection methods should be capable of detecting indications of fraud schemes that have not been found before. The still increasing amount of data available, the improved computer capacity, and the continued development of new and innovative techniques provide opportunities for the development of such advanced fraud detection methods. Text mining may be one of the new methods that aid in the detection of indications of financial fraud.

1.1 Financial fraud

Financial fraud is the deliberate act of deceiving others for unlawful financial gain. The financial damage caused by such frauds is enormous. Fraud affects various parties including investors, creditors and employees. In addition to the financial consequences, the trust of these parties in the company, or companies in general, is damaged. Eventually, the fraudulent activities may also lead to bankruptcy.

The world has seen various major financial fraud cases in the past decades. Many of them concern the US companies, including the infamous Enron and WorldCom cases. However, as other major fraud cases show, financial fraud is a worldwide phenomenon. Few examples of these cases are the Italian Parmalat, Dutch Ahold, Indian Satyam, and the Japanese Olympus. In the subsequent paragraphs, we briefly describe this selection of major fraud cases and their consequences.

The Enron case is often mentioned when describing financial fraud. This case is one of the largest fraud schemes and led to several consequences. Not only Enron filed for bankruptcy, but also the auditor (then one of the big–5 accountant firms) was dissolved. The fraud resulted in one of the largest class action settlement of 7.2 billion dollars, which is approximately 10% of the 78 billion dollars that vanished from the stock market. The former president was sentenced to 24 years of imprisonment. Until the fraud was detected in 2001,

the energy company Enron was considered one of the most innovative and profitable companies. However, Enron applied a range of complex 'creative' accounting activities that inflated the profits. Subsidiaries were established to manage risks and hide debts.

In 2002, the WorldCom scandal followed Enron. This fraud also led to bankruptcy, a 6.1 billion dollar class action settlement and the conviction of the management. The CEO was sentenced to 25 years in prison. The telecom company WorldCom improperly recorded profits. As opposed to Enron, these improper accounting activities were basic in nature. After the fraudulent bookings were revealed, the company needed to restate 3.85 billion dollars. The stock price dropped to 10 cents.

In the same period a fraud case was reported in Europe. In the Italian dairy and food company Parmalat, more than 14 billion dollars vanished. The company hid debts and losses by using a range of fraudulent accounting activities. Fake transactions inflated the revenues, the receivables of the fake sales were used to borrow money from banks, fake assets were reported and debts were moved off the balance sheet or were reported as equity. The investors lost their money and the company went bankrupt. The CEO was convicted of fraud and money laundering and was sentenced to 10 years in prison.

The Ahold case also took place during the same period, and was regularly compared to the cases of Enron, WorldCom and Parmalat. Although, contrary to the latter cases, there was no personal gain in Ahold, fraudulent activities took place. The supervision on the subsidiary US Foodservice was deemed inadequate. Ahold reported higher net sales primarily due to inflated sales at US Foodservice. Furthermore, the company improperly consolidated joint ventures and made other accounting errors, resulting in a total overstatement of 30 billion dollars in the net sales during the fiscal years 2000 to 2002. Due to this, imprisonment and penalties were imposed on the management.

In India, a fraud case, called the Enron of India, was reported at the information technology company Satyam in the period from 2002 to 2008, following which the founder and chairman of the company resigned, confessing to the annual statement fraud. He admitted to manipulating the bookings for the total amount of 1.47 billion dollars by reporting non-existing assets in the financial statements and omitting debts. Fake invoices and tax returns inflated the revenue and the profit. When the fraud was discovered, the shares dropped. Investors lost 2.2 billion dollars. The founder was fined and sentenced to seven years of imprisonment. Nine others were found guilty, including two partners of the auditing firm. The auditing firm was also fined 6 million dollars.

Over the entire period of the aforementioned cases, fraud occurred at Olym-

pus in Japan, lasting until 2010. The company covered up losses on investments since 1991. Estimates indicate that a total of 4.9 billion dollars were not accounted for. In 2011, the newly appointed CEO revealed the improper accounting practices. The stock market value dropped by 75 to 80%. Olympus reduced its work force and production plants. The chairman, the former executive vice–president and the auditing officer were sentenced to 2.5–3 years of imprisonment. Olympus was fined approximately 7 million dollars.

These fraud cases show that financial reporting fraud creates a false impression of a company's financial strength. The management of the companies directly or indirectly benefits from such fraud. The management's bonus can be directly linked to the business results. Good company performance mitigates the management's risk of getting laid off. In addition, for companies listed on a stock exchange, good results may increase the stock price, indirectly affecting the management's reward. Financial reporting fraud is also considered management fraud since the prime responsibility for the information in financial reports lies with management.

Furthermore, the briefly described fraud cases illustrate that companies engage in various fraudulent schemes and the activities vary. Fraudulent activities can be classified into three categories: (1) companies engaged in the falsification of the underlying documents of the annual report; (2) made false estimations and judgments; (3) omitted significant information. These activities may affect the company's performance disclosed in the annual reports. The question arises whether this effect is detectable in the text of the annual reports.

1.2 Text mining

Text mining is the task of identifying patterns in textual data. Text mining can be interpreted as a sub task of data mining, which is the extraction of information from data. In general, this data consists of structured tables containing the information. In text mining the data is comprised of texts. In text mining tasks the textual input is first transformed to a structured representation. Subsequently, information or patterns are extracted from this representation.

Text classification is a text mining task that assigns a textual object to one of two or more classes (Manning and Schütze, 1999). The textual object may be a word. In that case, an example of a text classification task is assigning the word to its part–of–speech, for example the word 'bicycle' should be assigned to the class 'noun'. The textual object may also be an entire text.

Identifying the author of a text from a limited set of authors is an example of text classification.

Text classification models consist of a data set of textual objects, a procedure for transforming the data into a structured representation and a training procedure that learns the patterns from the representations. First, each object in the data set is labeled with one class. The data set should be split into two sets: a training set and a test set. The test set is used to determine the performance of the model developed using the training set. Therefore, the test set should consist of textual objects that are not included in the training set (Manning and Schütze, 1999). Subsequently, each text object is transformed to a structured representation. This process is usually referred to as feature extraction. Typically, the textual input is represented by a vector of weighted word counts (Manning and Schütze, 1999). In that case, each distinct word in the data set is a feature. The training procedure of the model is defined by a machine learning algorithm. The trained model is used to classify unseen objects into one of the classes. Various machine learning algorithms exist that differ in how the characteristic patterns are learned from the features. The Naive Bayes (NB) algorithm constructs a model based on probabilities. The support vector machine (SVM) places each object in a space that is defined by the features. Neural networks (NN) construct a network in which the input nodes are defined by the features and the output nodes are the class labels. The NB and SVM algorithms are used in Chapters 4 and 5 of this thesis. Chapter 6 exploits an advanced NN.

Text mining brings a number of advantages. It allows a number of documents to be analyzed in a short period of time. In addition, the text mining models can detect patterns that are not noticeable to people. A machine can also concentrate on the aspects of texts that are difficult for people to focus on. For example, it is difficult for people to ignore the meaning and focus on the linguistic information concerning how something is said (Pennebaker et al., 2003). These advantages make text mining an interesting subject of research when textual information is available.

1.3 Research question

In financial frauds, the numbers in financial overviews are falsified. Consequently, methods for detecting indications of fraud focus on these numbers. However, a large part of a company's information is textual in nature. In fact, when a company communicates financial information, the financial numbers are accompanied by texts. The textual information is included in various

documents, such as business plans, codes of conduct, legal agreements, procedures and memos. In addition, people communicate via e–mails or chat. Furthermore, companies communicate to the outside world and its stakeholders through the company's website, press releases and financial reports.

Annual reports are among the most important types of financial reports. Annual reports contain information concerning the company's performance and activities in the preceding year. This information is provided in the financial statements that present the financial performance using numerical information. These statements are accompanied by texts explaining the numbers in the financial statements, the company's activities and the expectations for the future. Companies worldwide, especially those listed on a stock exchange, publish annual reports in English that are easily accessible via the internet. On this level of disclosure, we define fraud, following the definition of the Committee of Sponsoring Organizations of the Treadway Commission (COSO), as the 'intentional material misstatement of financial statements or financial disclosures or the perpetration of an illegal act that has a material direct effect on the financial statements or financial disclosures' (Beasley et al., 2010).

Major frauds that affect the numbers in the financial overviews of the company may affect the textual information as well. The textual information that explains the results and the activities with which those results were achieved, such as the Management Discussion and Analysis (MD&A) section of the annual report, has to be in accordance with the falsified numbers. As more people are able to understand the textual information as opposed to the financial statements, texts in the annual reports may have a greater reach and impact on the stakeholders. This is the reason companies may use texts intentionally to deceive the stakeholders. In these ways, fraud directly influences texts in the annual report. The influence of fraud on texts may also be unintentional. In large fraud schemes, management knew or should have known that improper accounting activities have taken place. This was the case for the examples described in Section 1.1. The management writes, or is at least responsible for, the texts in the annual report. Authors of texts may unconsciously leak information into the texts when lying (Wang and Wang, 2012). The word usages in false stories may differ from truthful ones (Newman et al., 2003). Therefore, indications of fraudulent activities may unintentionally end up in the texts of annual reports.

Compared to the analysis of the numerical information, texts offer several possible advantages. The textual information may provide more information than the financial numbers as it includes explanations of the numbers and expectations for the future. Another advantage of texts over numbers is that

texts are less subject to the financial rules and regulations. As a result, there is more freedom in the presentation of textual information. The choices made regarding the disclosure, formulation or omission of the textual information may be influenced by the presence of fraud in the company to a larger extent than the numerical information.

The availability and reach of annual reports, the potential effects of fraud on the texts of annual reports and the advantages of text mining techniques provide an opportunity to research the possibility of text mining techniques as an advanced fraud detection method. These considerations, therefore, result in the following main research question:

Can text mining techniques contribute to the detection of indications of fraud in annual reports worldwide?

1.4 Structure of the thesis

Chapter 2 of this thesis provides a brief theoretical background about the origin, purpose and use of annual reports worldwide. Furthermore, the data collection steps performed to gather the annual reports that were used to answer the research question are explained. Chapters 3-6 each describe a text mining technique that may contribute to the detection of fraud. Each chapter is readable on its own. Therefore, some overlap in the information provided in the chapters may exist.

Chapter 3 suggests a method for the automatic extraction of textual information from annual reports. The approach is applied to extract various types of information from the annual reports to demonstrate the possibilities and limitations. This text analysis method may contribute to the detection of fraud in annual reports in two ways. First, textual information from a large number of annual reports can be retrieved and structured for analysis, including fraud detection. Second, the extracted textual information can be used as a data preparation step for text mining. The latter is the application of choice in this thesis. By applying the automatic extraction approach, the data could be collected, as explained in Chapter 2, which is necessary for the research described in Chapters 4-6.

In Chapter 4, a baseline model is developed to assess whether simple textual features in a text mining model can detect indications of fraud in the MD&A section of annual reports of companies worldwide. Two established machine learning methods, Naive Bayes and Support Vector Machines, are used in the development of the baseline model.

Chapter 5 explores a wide range of linguistic features to extend the baseline model of Chapter 4. Various categories of linguistic features of texts are described and explained in the context of fraud and lie detection. For each category, the added value for the baseline model is determined. This approach reveals which types of textual features contribute in a text mining model for the detection of indications of fraud in annual reports.

In Chapter 6 a Deep Learning approach is explored. This state-of-the art machine learning model differs from the text mining models in the previous chapters. Instead of the simple textual features and the sets of linguistic features, the texts are represented by word embeddings that capture relations between words. The performance of the deep learning Convolutional Neural Network, is compared to a text mining model that uses simple textual features and the Naive Bayes machine learning method. In this comparison we assess the suitability of the method for the detection of fraud in the MD&A sections of annual reports of companies worldwide.

The thesis concludes with an answer to the main research question, a discussion of these results and suggestions for further research.

2 Annual reports of companies worldwide

The research described in the thesis uses annual reports as the data source. In this chapter we give further information on the data and the data collection process. Section 2.1 provides a background on annual reports, the rules and regulations and the changes in annual reports in the past decades. The recent debates concerning financial reporting are discussed. Finally, more background information on the Management Discussion and Analysis (MD&A), which is the part of the annual report that this research focuses on, is provided. Section 2.2 explains the data collection process that resulted in the data set used in the research described in the remaining Chapters of the thesis. Section 2.3 refers to the location where the data set is archived.

2.1 Annual reports

Financial reports are concerned with the disclosure of the financial position and achievements of a company. In financial reports, companies need to account for their results and the pursued policies and procedures. This may include information concerning the business activities, the organizational structure, the mission statement, the most important suppliers and customers, financial analysis, explanations for important transactions, such as takeovers and investments, personnel and remuneration policy and expectations for the future, including expected developments (Klaassen et al., 2008). Annual reports provide information on the activities of the companies in the preceding year. The annual report is considered an important source of information as it contains information from the past, and may also contain the management's expectations for the future.

In general, annual reports contain financial statements accompanied by explanatory textual information. Financial statements typically include four parts. The first part is the balance sheet that reflects the financial position of the company in terms of its assets, liabilities and equity. The second part is the income statement that shows the income, expenses and profits of the preceding year. The third part lists the changes in equity. The final part, which is the

cash flow statement, reports the cash and cash equivalents of the company. The textual information includes the notes to the financial statements that explain specific items of the financial statement. In addition, an annual report usually starts with a letter to the shareholders in which the board provides a short overview of the company's operations and financial results. On top of that, annual reports contain an MD&A or operating and financial review explained in Section 2.1.4 in more detail. Furthermore, the auditor's report as briefly described in Section 2.1.2 is included in annual reports.

The type and amount of information disclosed may depend on the size of the company and the sector the company operates in. The annual reports of large companies can be very extensive. Typically, reports consist of 40-60 pages. However, only a limited part of the content is concerned with mandatory information. Some companies provide a lot of information voluntarily to aid clear communication. The increased interest in the stock exchanges has led to an increased requirements for information, that companies try to meet (Klaassen et al., 2008). Another development that affects the size and contents of annual reports is integrated reporting, in which companies report on their activities concerning corporate social responsibility. The Global Reporting Initiative (GRI) developed general standards that include environmental, social and economic aspects. The environmental aspects are about the impact on the environment and planet. The social aspects are concerned with people, such as employment and social security, and working conditions. Economical aspects define the financial contribution to society, such as knowledge gained from the research and education of people (Klaassen et al., 2008). The use of integrated reporting is under development, as described in Section 2.1.3.

Companies have several stakeholders who are interested in their annual reports for various purposes. Investors consider selling or buying shares. Banks decide on the provision and withdrawal of credit. Suppliers are interested in the solvency of the company. Furthermore, job applicants may gather information to decide when to accept or refuse a job offer and employees may consider resignation. Regarding the provision of information opposing interest exists between the organization and the stakeholders. Usually, companies are reluctant to provide information that can be useful for their competitors. As a consequence, plans and expectations for the future are unlikely to be disclosed in detail (Klaassen et al., 2008). The disclosure of financial information is subject to rules and regulations, providing a minimum level of information that companies need to disclose.

2.1.1 Rules and regulations

The current form of annual reports and the corresponding rules, regulations and supervisory bodies are fairly new. The first form of regulation that existed was self-regulation. After the Second World War, the Western countries established standard setters to set rules for financial reporting. However, these rules were only recommendations, as opposed to the enforceable rules that exist today. Since 1970, as a result of the increasing globalization, the standard setters have been working on the harmonization of the international rules (Hoogendoorn et al., 2004; Klaassen et al., 2008). At the European level, in 2001, this resulted in the foundation of the International Accounting Standards Board (IASB) that develops the IFRS. Initially, the IASB was responsible for drafting general rules. Over the years, the general rules have been transformed into extensive and complex rules (Hoogendoorn et al., 2004). At the same time as the start of harmonization of the rules in Europe, the Financial Accounting Standards Board (FASB) was founded. The FASB is responsible for the publication of the United States Generally Accepted Accounting Principles (US GAAP) (Hoogendoorn et al., 2004).

Currently, based on the existing rules and regulations, the world can be roughly divided into two parts. The IFRS and the US GAAP are the two major sets of accounting rules. The IFRS have been obligatory for listed companies in the EU since 2005 and are now applied worldwide in several countries, including Russia, Australia, New Zealand, Canada and several Asian and South American countries (IFRS, 2017). In China the Chinese Accounting Standards, which are largely converged with IFRS, are applied. The Unites States applies US GAAP. The major difference between IFRS and US GAAP is that IRFS is principle-based, while US GAAP is rule based. From 2007, non-US companies listed on the US stock exchange no longer have to reconcile their financial reports with US GAAP if they applied IFRS.

Another milestone in the history of the financial reporting practice was the acceptance of the Sarbanes–Oxley Act (SOX). SOX became effective in 2002 after several fraud cases were reported, including WorldCom and Enron. This legislation aims at enforcing reliable entrepreneurship. The management of the company must take individual responsibility for the company's financial reports. Since the independence of the accountant is important for the reliability of the auditor's opinion, SOX includes standards that define the auditor's independence. The act includes a rule that specifies non-control activities that are prohibited for companies listed on a stock exchange (Hoogendoorn et al., 2004). Furthermore, the law prescribed additional reporting requirements. Lee et al. (2014) found that the implementation of SOX resulted in an

increase in the length of the MD&A section of 10-K annual reports.

Within the rules and regulations, companies have the possibility of choosing between alternatives. In some cases, choices are necessary. For example, within IFRS, estimates of provisions, which is a liability for which the timing or the amount is uncertain, need to be made. This freedom of choice allows for creative accounting. Even though choices are within the limits of the rules and regulations, as a result of various accounting scandals, the term 'creative accounting' has a negative connotation. A specific example of the distinction between legitimate and illegitimate practices is 'earnings management', which is legitimate and 'earnings manipulation', in which a company does not comply with the rules and regulations. The choices that are made may influence the decisions made by the stakeholders, which may have economic consequences, such as the stock price and provision of credit (Hoogendoorn et al., 2004).

2.1.2 Financial supervision

The rules and regulations provide the opportunity to control the financial disclosures according to these rules. This financial supervision aids the compliance with the rules and regulations. One such control is the auditor's report, which expresses an opinion on the validity and reliability of a company's financial statements. The auditor's conduct a risk inventory to conclude whether the financial statements provide a true and fair view and are in accordance with the rules and regulations, such as IFRS and US GAAP (Hoogendoorn et al., 2004; Klaassen et al., 2008).

For companies listed on a stock exchange, financial supervisory bodies also monitor their activities, including the financial reports. The largest financial supervisory body is the Securities and Exchange Commission (SEC) in the United States. Listed companies are required to provide information using standard forms (Hoogendoorn et al., 2004). There are also specific forms for the annual report: form 10-K for the US companies and 20-F for foreign companies listed on a stock exchange in the US. Until 2002, companies that did not disclose their financial trading activities in time filed on form 10-K405. There is no difference between forms 10-K and 10-K405. Therefore, when we refer to form 10-K, this includes form 10-K405. Until 2009, an abbreviated 10-K form, 10-KSB, existed for the small US companies. All forms are publicly available through the EDGAR database on the SEC's website. In addition, the SEC publishes statements concerning the acceptability of reporting practices as defined by the Securities Exchange Act of 1934, which includes US GAAP, on its website. The acceptable practices as well as the convictions leading to subsequent steps are published. The latter statements are published as the

'Accounting and Auditing Enforcement Releases' (AAERs). These releases describe the actions taken against the company, the auditor of the company or the company's officers that violated the reporting rules. No other financial supervisory body has such extensive publicly available documentation.

2.1.3 Recent debates

The types of information that should be included in an annual report and what constitutes a good annual report is a subject of ongoing debate. A regularly recurring topic is integrated reporting. In addition, the media reports mentioned the process of convergence of IFRS and US GAAP. Finally, digital developments, such as XBRL and the inclusion of information in the annual report concerning the digital safety of the company, are discussed. We conclude this section with a comment on the impact of these developments on annual reports and the work of the accountant.

Integrated reporting

In the past decade efforts were made to include information concerning value creation in corporate reports. This includes the strategy of a company, the expected risks and the companies risk management. Sustainability should be a component of the strategy. The inclusion of this type of information in annual reports was deemed insufficient in 2012 (FD, 2012a,c). However, in early 2013, integrated reporting was expected to be the standard for listed companies over the next ten years (AN, 2013). At that point in time, companies in Europe and the US presented some elements of integrated reporting in their annual reports, but this information lacked unity (FD, 2013b). At the end of 2013, the disclosure of risk information showed improvement, but did not meet the expectations (FD, 2013g). Furthermore, an increase in the application of integrated reporting was observed (FD, 2013e). The research of annual reports of 4.100 companies worldwide revealed that half of the companies paid attention to sustainability. However, only a limited number of companies reported on the influence of the scarcity of raw material and climate change on the financial results (FD, 2013c). Although 1 in 10 companies believed that they applied integrated reporting, only few subjects of integrated reporting were considered.

The international framework for integrated reporting developed by the International Integrated Reporting Council (IIRC) should aid companies in applying integrated reporting. A draft version of the framework was released in April 2013. The first version of the framework was released at the end of

2013. Annual reports were tested against the draft framework for integrated reporting. The best scores were achieved for the disclosure of risks and opportunities. The explanation of strategical decisions improved. On the contrary, the explanation of value creation scored the lowest. The annual reports contained very few measurable performance indicators, a limited information on governance and the amount of future oriented information was limited (AC, 2013). The framework received some criticism. The framework lacked measurable indicators (FD, 2013b). Furthermore, the application of the framework is voluntary. The discretion to disclose information lies solely with companies. However, due to the specific request by means of the framework, not disclosing certain information is a risk (FD, 2014d).

Besides the development of this framework, the European Union proposed, in the spring of 2013, a directive concerning the disclosure of non-financial information, including sustainability, diversity, human rights and anti–corruption, for companies that have more than 500 employees (FD, 2013d; VK, 2013). Companies that omit information must explain why they do not publish the information. In April 2014, the EU directive was adopted by the European Parliament (AN, 2014a). In October 2014, the Council of the European Union also adopted the directive. The member states had two years to translate the directive into national legislation (AN, 2014b).

At the end of 2015 and 2016, integrated reporting was applied more regularly. Nowadays, Dutch listed companies pay more attention to non financial information. Corporate reports include more information concerning sustainability (FD, 2015). The interest in integrated reporting increased (AM, 2016). At the end of 2016, companies were better at applying integrated reporting. Although some countries have national rules, integrated reporting is still self-regulated (FD, 2016b,a). It is posed that standard setters and regulators should draw up directives for the disclosure of non financial information similar to directives that exist for the disclosure of financial information (FD, 2016a).

Convergence of IFRS and US GAAP

Annual reports are not only affected by the addition of new frameworks, but the existing rules and regulations are also subject to change. The major topic in this area in the past few years has been a change in IFRS 9, which dictates that banks have to take losses on credits sooner (FD, 2013a, 2014a; FT, 2014; FD, 2014b; AC, 2014b). Furthermore, the definitions provided by the International Accounting Standards Board (IASB) are a subject for discussion. Unambiguous definitions are required (AC, 2014a; NU, 2016). In 2002

the IASB and the Financial Accounting Standards Board (FASB) began the process of convergence of the IFRS rules defined by the IASB and the United States Generally Accepted Accounting Principles (US GAAP) published by the FASB. The IASB and FASB worked on several convergence projects (AT, 2013). In 2008 the SEC proposed the adoption of the IFRS to replace US GAAP. However, full convergence has not been reached (AN, 2014c).

Digital developments

Digital developments affect the financial reports in several ways. First, the development of eXtensible Business Reporting Language (XBRL) affects the way in which financial information is communicated. XBRL can be used to exchange the information in a structured manner, enabling stakeholders to analyze the information more efficiently Hoogendoorn et al. (2004). XBRL tags are increasingly included in financial reporting and the US GAAP taxonomy includes the use of XBRL (FT, 2013; AC, 2015). The SEC exploits the XBRL tags to develop a tool that automatically triggers alerts (FT, 2013). Secondly, the digital developments affect companies. The theft of digital information and cyber–attacks are major risks for companies. It is argued that companies should include information about their digital safety in their annual reports (FD, 2014c).

Changes in annual reports and accounting

The previously discussed changes in corporate reporting generally require the disclosure of additional information. In 2013, a survey of the IASB in Africa, Asia, Europe and North America showed that the preparers of the annual reports indicated that additional requirements lead to disclosure overload (AA, 2013). Standard setters need to take this consequence into account to prevent the disclosure of too much information, which remains a challenge. In 2016, the report of the Institute of Chartered Accountants in England and Wales (ICAEW) showed that corporate reports were too large and complicated (AW, 2016).

The increased importance of the disclosure of non-financial information also affects the work of the accountant. Providing assurance in the areas of corporate governance, risk management and corporate social responsibility may be an expected audit service (FD, 2013f).

The recent changes discussed in this section have been obtained from media reports from the past few years, since the beginning of the research for this thesis. The annual reports collected for this research belong to the years prior

to the beginning of the research. Therefore, these developments may not yet apply to the annual reports in the data set used in this research.

2.1.4 The Management Discussion and Analysis

The MD&A is the part of the annual report in which the management provides a written overview of the company's performance and activities of the preceding year as well as a discussion of the expectations and goals of the following year. This information is included in all annual reports, but the depth of the information provided differs per company and annual report (Brown and Tucker, 2011). The SEC requests the following information in forms 10-K and 20-F, respectively:

Discuss registrant's financial condition, changes in financial condition and results of operations. The discussion shall provide [..] such other information that the registrant believes to be necessary to an understanding of its financial condition, changes in financial condition and results of operations (Reporting, 1934).

The purpose of this standard is to provide management's explanation of factors that have affected the company's financial condition and results of operations for the historical periods covered by the financial statements, and management's assessment of factors and trends which are anticipated to have a material effect on the company's financial condition and results of operations in future periods (Securities and Exchange Commission, 2014d).

In annual reports on form 10-K, the MD&A section is usually item 7 and termed 'Management's Discussion and Analysis of Financial Condition and Results of Operations' (Securities and Exchange Commission, 2014c). In form 20-F, the MD&A section is usually included in item 5, termed 'Operating and Financial Review and Prospects' (Securities and Exchange Commission, 2014d). In annual reports in other formats than the SEC forms the information is disclosed in sections termed 'Financial Review', 'Operational and Financial Review' or 'Review of Operations'. The information may also be included in the free format message or report from the board or the management of the company to the shareholders, customers and/or employees.

The MD&A section is arguably the part of the annual reports that is the most read and cited by analysts (Balakrishnan et al., 2010; Li, 2010; Rogers and Grant, 1997). Rogers and Grant (1997) found that the texts in annual reports provide almost twice the information that is provided by the financial

statements (Balakrishnan et al., 2010). A survey of the Association for Investment Management and Research (AIMR) among financial analysts revealed that 86% of the respondents found the management's discussion an important factor to assess the value of the company (Balakrishnan et al., 2010).

Although the MD&A section is an unaudited part of the annual report, it may be affected by changes in rules and regulations and the demands of the stakeholders. The length of MD&A sections in 10-K annual reports increased after SOX became effective . However, no differences in language use between annual reports before and after SOX were found (Lee et al., 2014). Li (2010) concluded that there was no change in the information content of the forward-looking statements made in the MD&A sections over time.

2.2 Data collection and preparation

Section 2.2.1 describes the steps taken in the collection process to obtain the data set for the research described in this thesis. The limitations of the steps are also mentioned. A short summary of the resulting data set is provided in Section 2.2.2. To prepare the collected annual reports for the text analysis, several data preparation steps were taken. Section 2.2.3 summarizes these steps.

2.2.1 The collection process

Several steps were taken to obtain the data set consisting of annual reports from companies and years when frauds were reported and from companies not involved with fraudulent activities. Information from several sources was gathered to collect annual reports and information concerning the fraudulent activities. Subsequently, additional information was extracted from the annual reports themselves. This information is needed to create a well balanced data set. Figure 2.1 summarizes the data collection process. In this section, we explain each of the steps in more detail.

We only know that a company was involved in fraudulent activities when it has been convicted for fraud and the conviction has been made public. The first step of the data collection process, therefore, is gathering information about fraud convictions. We searched for fraud cases up to 15 years prior to the beginning of the data collection process (from 1999 to 2013). One of the sources for this type of information are media reports. A news article about a fraud is questionable as a source for determining the existence of fraud in a company. The research, therefore, only includes cases that were discussed by several newspapers over a period of at least several days. With

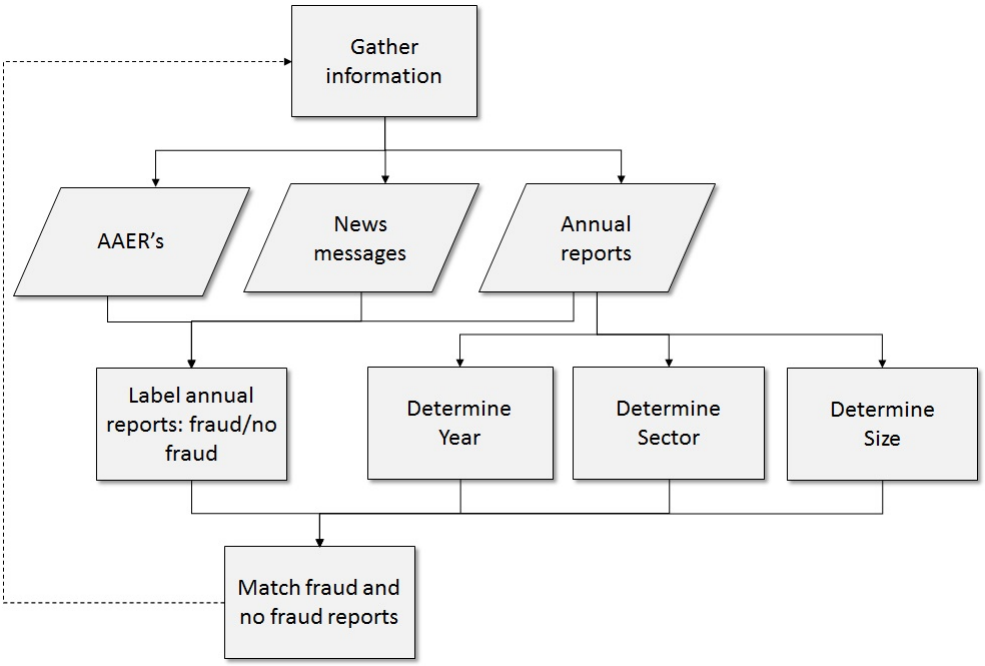


Figure 2.1: A schematic overview of the annual report collection process.

this approach, errors may be corrected after a couple of days or by the same or other news sources. The news messages were gathered with search queries in LexisNexis, which has access to various news sources. This resulted in a collection of 624 news messages. Another source of information on fraud were the AAERs published by the SEC. Similar to news messages, for one fraud case multiple AAERs may be published. We selected the AAERs concerning litigation in the period from 1999 to 2013. In this period, 2.232 litigation AAERs were published. Including the complements to the AAERs the total number of documents published amounts to 2.555. From this total set we selected the AAERs that explicitly contain the word 'fraud', which resulted in 1.354 AAERs. A total of 519 of AAERs are likely to discuss annual reports since these releases contain the word '10-K', '20-F' or 'annual report'. Figure 2.2 summarizes the total number of litigation AAERs, the AAERs concerning fraud and the AAERs concerning fraud and annual reports per year.

The second step of the information gathering process is the collection of annual reports themselves. This step is combined with the labeling of annual reports in the categories 'fraud' and 'no fraud'. Only the cases for which persons were convicted or companies fined are included in the data set as

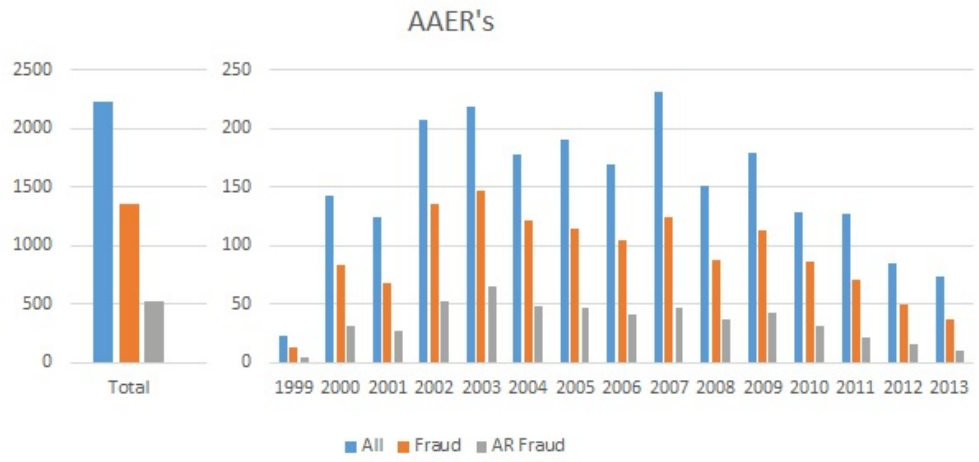


Figure 2.2: The number of AAERs.

fraud cases. Cases discussed in newspapers for which fraud was not proven or that were still under investigation are not included in the data set. The fraud cases are selected by reading the news messages and AAERs. By reading the cases in detail, the periods during which the frauds occurred and may have affected the annual reports were noted. Subsequently, the annual reports of the companies and the information about the affected years were gathered. Annual reports are publicly available via the internet. These are downloaded from company websites, general websites such as 'www.annualreports.com' and the EDGAR database of the SEC. Not all annual reports of the years affected by fraud are available. For 427 of the fraudulent annual reports found in the news messages and AAERs, 403 annual reports could be retrieved. The annual reports for the 'no fraud' category are collected from the same sources as the fraudulent reports. A total of 5.259 files was downloaded from the website 'www.jaarverslag.com' that contains annual reports from various companies. The SEC website has master index files that contain an overview of the filings available. These master index files are grouped in a SAS database (WRD.US, 2014). We converted the SAS database, containing the filings in the period from 1993 to 2012, to a SQL database. From this overview we randomly selected 10% of the annual report filings as a potential source of the no-fraud annual reports. This totals to 12.000 annual reports. Only annual reports of companies not mentioned in the news or AAERs are labeled as no-fraud. Note that the absence of news messages and AAERs about fraud does not prove the absence of fraud in the company. We label annual reports as no-

fraud following the presumption of innocence that one is considered innocent unless proven guilty.

Three types of information are extracted from the collected annual reports. First, the fiscal year for which the financial performance is disclosed in the annual report is selected. The fiscal year is not necessarily the same as the calendar year. The fiscal year is always included in the annual report, either in the title of the report or at the beginning of the report. Chapter 3 describes the method used to automatically extract the fiscal year from annual reports on forms 10-K, 10-KSB and 20-F. The second type of information extracted for each annual report is the sector the company operates in. The sectors are defined as the 14 divisions of corporation (AD offices) used by the SEC at the time of data collection in 2013 (Securities and Exchange Commission, 2014a). Table 2.1 lists the AD offices. Each company that files with the SEC has a Standard Industrial Classification (SIC) code that indicates the company's type of business. The SIC code defines the AD office to which the annual report is assigned for review. Chapter 3 describes the automated method applied to extract the SIC code from annual reports on forms 10-K, 10-KSB and 20-F. The annual reports not filed with the SEC were assigned manually to one of the divisions, based on the prime business of the company, which can be found online or in the annual report itself. The number of employees is reported in annual reports. Chapter 3 also describes the method for automatically extracting this type of information from the annual report. The information is retrieved manually for the annual reports for which automatic extraction was not possible.

The year, sector and number of employees are used to match the fraud reports with the no-fraud reports to form a balanced data set. First, matching on the year of the annual report is important due to the changes in rules and regulations that may affect the contents of the annual reports. Furthermore, it is conceivable that other events occur that are specific to a period of time and that influence the information disclosed in annual reports in general, such as the 2008 financial crisis. Secondly, matching on sector is important to account for industry specific subjects disclosed in the annual report. These subjects may include sector specific topics or events that occurred in the sector in the fiscal year. The sector matching is specifically important for text mining research since the word usage in financial documents may vary across sectors. In this research, we want to determine whether differences between fraud and no-fraud annual reports exist, and do not aim at detecting the differences that exist between sectors. Finally, the fraud and no-fraud reports are matched based on the company size. The size of the company may affect the level of

AD office	Description
1	Healthcare and Insurance
2	Consumer Products
3	Information Technologies and Services
4	Natural Resources
5	Transportation and Leisure
6	Manufacturing and Construction
7	Financial Services I
8	Real Estate and Commodities
9	Beverages, Apparel and Mining
10	Electronics and Machinery
11	Telecommunications
12	Financial Services II
2&3	Combination of 2 and 3
All	Non-operating

Table 2.1: The AD offices (Securities and Exchange Commission, 2014a).

detail of the information disclosed in the annual report (Aerts, 2001). The number of employees is an indication of the company size. Three categories of company sizes are distinguished (European Commission, 2014). Optionally, a fourth category can be distinguished, which is defined as micro companies that employ 1 to 9 people. Small companies employ less than 50 people. Medium-sized companies employ 50-250 people. Large companies employ more than 250 people. A large variety exists within the category of large companies. Large companies may also employ over 100.000 people. It is conceivable that differences exist in the disclosure of financial information by companies employing 500 people and companies employing 50.000 people. Therefore, we made smaller categories for matching the companies of similar sizes based on the number of employees. Table 2.2 lists the sizes of the smaller subcategories.

The matching process reflects the assumption that most companies are not involved in fraudulent activities. For each annual report in the 'fraud' category, several annual reports of the 'no fraud' category were collected for the data set. Furthermore, the data collection reflects that in the world, more number of small companies exist (Bureau, 2014). For each annual report of a small company in the 'fraud' category, five annual reports of the 'no fraud' category are included in the data set. For each annual report of a medium-sized company that was affected by fraud, four no-fraud annual reports are

included. For each annual report of a large company affected by fraud, three no-fraud annual reports are collected. If the initial selection of annual reports does not contain sufficient annual reports that meet the matching criteria, additional annual reports are obtained through company websites or the EDGAR database. The information gathering and extraction steps are repeated until for all fraud annual reports a sufficient number of no-fraud annual reports are collected.

Category	**Subcategory size**
Small	1-9 10-49
Medium	50-249
Large	250-999 1.000-9.999 10.000-99.999 >100.000

Table 2.2: Company size on the basis of the total number of employees.

2.2.2 The data set

The resulting data set consists of 403 fraudulent and 1.325 non fraudulent annual reports. Table 2.3 summarizes the number of fraudulent and non fraudulent reports per category of the company size. One of the fraudulent reports of a large company was later removed due to the extremely low digital quality of the report.

	Fraud	No fraud
Small	36	180
Medium	44	176
Large	323	969
Total	403	1.325

Table 2.3: The total number of fraudulent and non fraudulent annual reports collected per category of the company size.

The collected annual reports concern the period from 1999 to 2011. Detection and investigation of possibly fraudulent activities for 2012 and 2013

were not completed at the moment of data collection. Purda and Skillicorn (2010) determined that the average period between the end of the fraud and the publication of an AAER is three years. This also explains why our fraud data set contains less reports in the more recent years. Figure 2.3 shows the distribution of the fraudulent annual reports over the years.

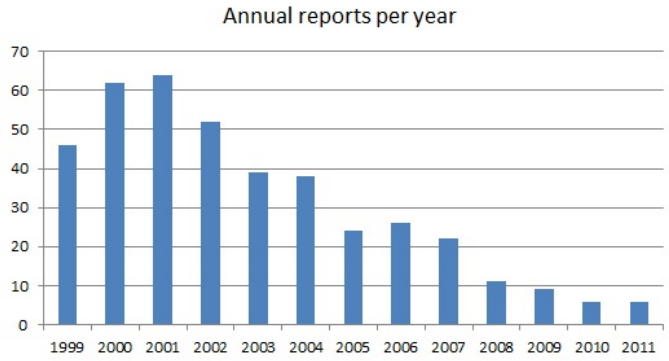


Figure 2.3: Number of annual reports in the fraud data set per year.

While there are more number of small- than large-sized organizations in the world, the number of annual reports selected is much higher for large companies. Out of the selected annual reports, 80% belong to large companies. This observation, does not imply that large organizations commit more fraud. A possible alternative explanation is that large companies are more often subject to review.

2.2.3 Data preparation

The annual reports in the data set need pre-processing to make them suitable for text analysis. The preparation steps taken depend on the format of the annual report. The annual reports on forms 10-K, 10-KSB and 20-F are in html format. These texts are machine-readable. Other annual reports are available in pdf format and may not be machine-readable. Optical character recognition (OCR) converges these documents to machine-readable texts. The conversion may not be 100% accurate, resulting in some typographical errors. ABBYY FineReader, arguably the most accurate commercial OCR tool, was used to perform the conversion (Boschetti et al., 2009; Kumar et al., 2013).

In the research discussed in this thesis, the focus is on the MD&A section of the annual report as described in Section 2.1.4. Chapter 3 describes an approach for automatically extracting the MD&A sections from annual reports

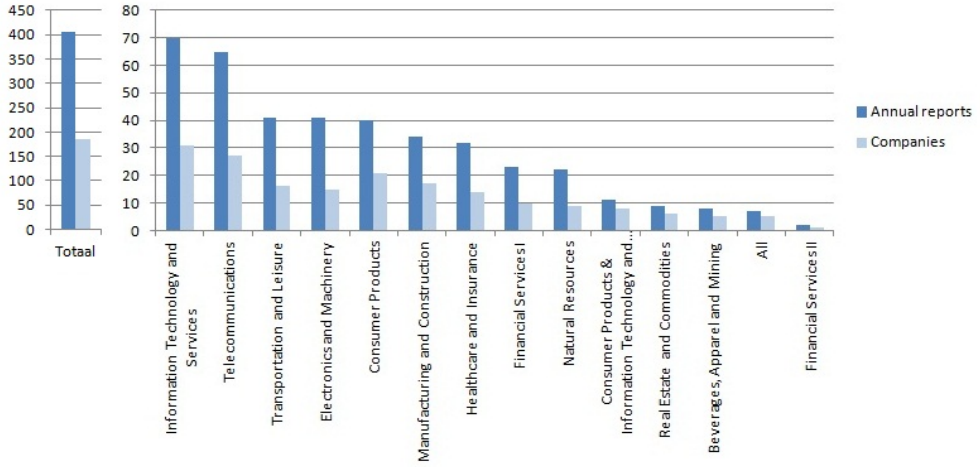


Figure 2.4: Number of companies and annual reports in the fraud data set per sector.

on forms 10-K, 10-KSB and 20-F. The MD&A sections of the annual reports for which the automated extraction method was not accurate and the annual reports in other formats were extracted manually.

The html documents contain html tags that a computer processes as all the other words in the texts. These html tags are for structuring purposes and are not part of the content of the text. Therefore, we need to exclude the html tags prior to the texts analysis. We applied the Python package 'BeautifulSoup' to remove the tags from the texts. Unfortunately, methods that automatically remove html tags are not flawless. As a result, some of the MD&A section from html annual reports may still contain html tags. In addition to html tags, the html documents contain tables. Tables should be excluded from the documents because tables are used to communicate numerical information, while we focus on the textual information. Chapter 3 describes the method for automatically detecting the tables so that they can be removed. In manual extraction, tables are already excluded.

Figure 2.5 summarizes the data preparation steps.

2.3 Data archive

The data were collected and stored in the cloud. The full data set can be downloaded from https://surfdrive.surf.nl/files/index.php/s/m34LCElefSj6M8y. The annual reports are stored in two zip files. One file contains the annual reports

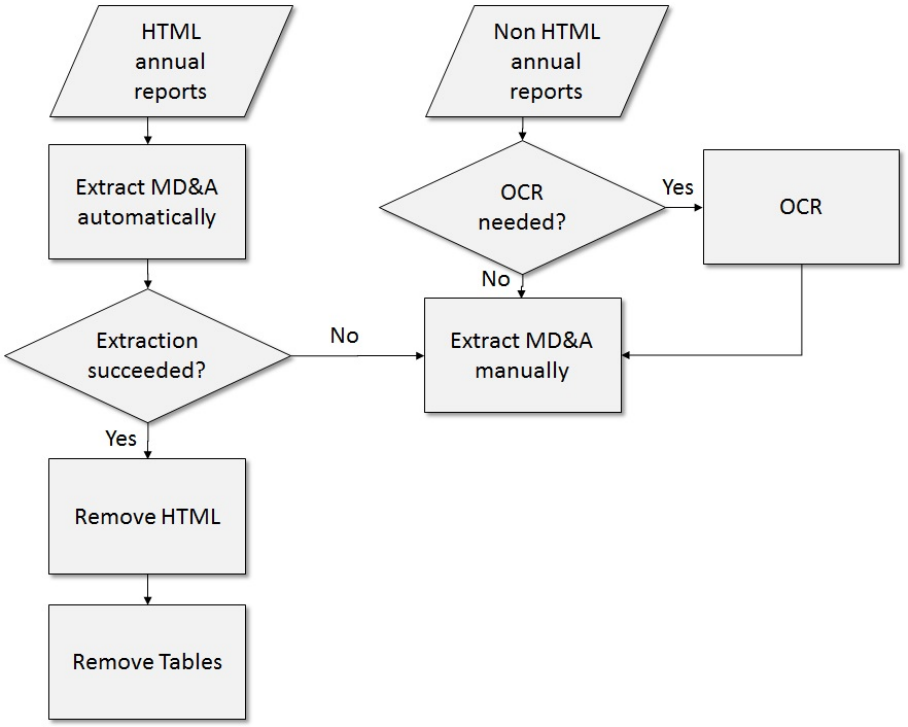


Figure 2.5: A schematic overview of the data preparation steps.

in the 'fraud' category and the other file in the 'no fraud' category. The included Excel file, 'Data_overview', lists the file names of the annual reports with the corresponding years and company names. The company names of the non fraudulent annual reports were extracted with the automatic method proposed in Chapter 3 of this thesis.

The published collection of fraud and no-fraud annual reports can directly be used to develop innovative methods that may detect indications of fraud in annual reports. The data collection may be extended with fraud cases that were revealed after 2013 using the procedure described in Section 2.2.

3 Automatic extraction of textual information from annual reports

Abstract

In this paper, we demonstrate a fairly straightforward and practical approach for automatically locating information in a large number of annual reports. The approach is based on an intuitive principle: To locate information in a document we need to model where the relevant piece of information starts and ends. The approach is used to extract several types of information from annual reports. These examples show that the modeling of the starts and ends depends on the complexity of the information to be extracted. However, for all types of information it holds that understanding the layout of the document is important for the automatic extraction of information from that document.

3.1 Introduction

With the ever increasing amount of textual information present within companies and online, people are searching for ways to efficiently process these huge amounts of text. In recent years, researchers and companies have taken steps to standardize financial reporting, including the development of the International Financial Reporting Standards (IFRS) and eXtensible Business Reporting Language (XBRL) (Markelevich et al., 2015). Such standardization may facilitate the exchange of information and allow the automated analysis of the financial information. This is required since reading all information available in search for specific information becomes intractable soon. In this paper, we demonstrate an approach for automatically locating information in annual reports to overcome the tractability problem when analyzing the textual information in large numbers of annual reports.

The annual reports contain various types of information of interest to various stakeholders. This makes automating the information extraction approach challenging. By focusing such an approach on the specific type of information that someone searches, may reduce the complexity. This requires knowing the goal of the information search. For a technical approach this means knowing

the type or format of the information that someone is looking for. The type of information to be located can be very specific, such as the year of the annual report, or more broadly, such as a section of the annual report. Note that we focus on locating information that a person is consciously looking for, opposed to finding hidden insights or finding information that people did not know they were looking for.

An automated approach to extract information allows repetition of the information search task for multiple annual reports. Such an automated approach may serve several goals, of which many are similar to the motivation for having a standardized format, such as XBRL. The approach allows the structuring of unstructured information for the construction of a database to perform traditional data analysis or data mining. Sections extracted from documents may be used in further text analysis research. Including only specific sections in the research data provides focus to the research and may improve the research results. This way, prior knowledge about which sections are relevant for the task at hand can be incorporated in the research. Furthermore, being able to identify specific sections provides additional research opportunities. For example, text analysis research may include the examination of the amount of information that each section contributed to the result. Sections that do not contribute to a better result may be omitted from the research model, which may result in a less complex or faster model. Additionally, the automatic selection allows the inclusion of many more documents in a much shorter period of time, so that a larger data set can be used.

The approach that we demonstrate in this paper is fairly straightforward and practical to use. The complexity of the model depends on the type of information that needs to be localized. In this paper, we search for several types of information in annual reports. This includes very specific information, such as the year of an annual report, specific sections of the report, and the identification of tables. These examples show for which types of information the approach is successful and what the limitations are.

This paper is organized as follows. Section 3.2 outlines the literature related to automatic extraction of information from texts. Section 3.3 describes the annual report data used for developing the methods in this research. Section 3.4 presents the methods and results for several types of information that may be extracted from annual reports. Section 3.5 discusses the pros and cons of the method.

3.2 Literature

The research on textual information extraction from annual reports is limited. However, the extraction of specific information from documents is the subject of research in a wide range of other domains. In this section, we first describe the research regarding information extraction in annual reports. Secondly, we give an overview of the textual information tasks explored in previous research in the various other domains.

Leinemann et al. (2001) developed an approach to automatically extract financial information from annual reports on form 10-K because investors interested in information from the balance sheet prefer immediate online access over having to read the entire report. They extracted the sections that contain financial information from annual reports filed by US companies with the Securities and Exchange Commission (SEC) on form 10-K. The semi structure of the 10-K reports allows for the identification of the starts and ends of these financial sections. Subsequently, Leinemann et al. (2001) used keyword identification to locate the financial items to be extracted. The detected financial data is then transformed into XML format to make it machine understandable. Heidari and Felden (2016) provided structure to the footnotes of the financial statements by developing a method that automatically assigns the sentences in the footnotes to pre-defined topic categories to support the analysis of texts in annual reports. Although we only found these two research papers, the automatic extraction of information from annual reports might be applied more often but not be described in the scientific literature. We therefore discuss the literature of automatic extraction of information in general.

In other domains researchers developed methods to automatically extract information from various types of textual sources. We distinguish three types of textual information extraction tasks based on the complexity of the source and the information searched for that can be found in the literature. The first type comprises of information extraction tasks for which all words of a very short text are assigned to an element. For example, identifying the elements of an address, such as street names. Secondly, we distinguish the tasks of finding predefined elements in short texts, such as extracting the name of a company for which someone worked from a curriculum vitae. Third are the tasks that locate information in larger documents with varying content, such as websites.

The first type of textual information extraction task we discuss, is the identification of all elements in a very short text. One of the domains for this category is the identification of address elements, such as 'street' and 'city', in address text fields. The goal of the address identification is de-duplication of stored addresses and collection of addresses in data warehouses for subse-

quent analysis. For address text fields, it is roughly known which elements it likely contains. The complexity of the identification of address elements is that the order of the elements vary and that not all possible address elements will occur in all addresses. However, Agichtein and Ganti (2004) found that within batches of addresses from the same source, the order of elements is the same for each address. This information is useful to fasten the segmentation process. Another challenging aspect for the identification of the elements of an address is that the number of words an element contains varies. For example, a street name can be one word or consist of multiple words.

Borkar et al. (2001) used a supervised Hidden Markov Model (HMM) to segment addresses stored in large corporate databases as a single text field. Borkar et al. (2001) used a labeled set of addresses, in which each address element is tagged with the element name, to learn the parameters of the HMM. Agichtein and Ganti (2004) used an unsupervised method to build an HMM to segment addresses. Instead of labeling addresses they used reference tables of a data warehouse, in which each column corresponds to an address element, to train the HMM.

Borkar et al. (2001) and Agichtein and Ganti (2004) applied their methods for segmenting addresses to bibliographical references. This type of references includes elements, such as author names, title of the paper, the year of publication and title, volume and number of the journal. Some references may include a web address or postal address. Several similarities between addresses and bibliographical references exist. Both have a limited number of words and each word needs to be assigned to an element. Similar to addresses the order of the elements in a bibliographical reference may vary and not all elements need to be present in all references. Also, references from the same source usually have the same ordering of the elements. Anzaroot and McCallum (2013) and Kluegl et al. (2012) used conditional random fields (CRF's) instead of HMM's because they offer more flexibility in the design of the input features. Anzaroot and McCallum (2013) applied their model to 1.800 citations from PDF research papers on physics, mathematics, computer science and quantitative biology. By automatically segmenting bibliographical references the research literature can be organized more easily to provide insight into the landscape of science and specific research areas. Kluegl et al. (2012) make use of context specific information, which is information that documents have in common as a result of the process in which the document is created, such as filling in templates. Adding this type of information improves the results of segmenting bibliographical references.

The second type of textual information extraction task is concerned with

finding specific, predefined elements in short texts. Similar to addresses and bibliographical references, curriculum vitae (CV's) and job postings contain predefined elements. However, the texts of CV's and job postings are a little more extensive. Kluegl et al. (2012) were able the apply their CRF model to identify the time span and company for which the author of the CV worked. Mooney (1999) developed a model for finding information in newsgroup messages or web pages. The model is created by taking pairs of documents and corresponding filled templates to induce pattern-match rules. These rules are subsequently used to fill the slots of templates for unseen documents. Each rule consists of three patterns: a pattern that must match the text immediately preceding the slot, the pattern that matches the actual slot and a pattern that must match the text immediately following the slot. This approach was tested on job postings to create a database of available jobs.

The third type of textual information extraction task we distinguish in this paper focuses on locating information in larger documents, such as websites. The model of Mooney (1999) is not specifically designed for finding information on websites. Other researchers, such as Hahn et al. (2010) and Yang et al. (2009), focused on information extraction from websites. Websites contain substantially more text with a larger variety of elements and format compared to addresses and CV's. However, websites contain meta data in the form of html tags to aid automatic information extraction. The goal of Hahn et al. (2010) is to allow users to query Wikipedia like a structured database. They made use of the fact that Wikipedia articles do not only consist of free text, but also contain structured information in the form of templates. Unfortunately, multiple templates are used for the same types of information. Therefore, using the structure of the templates for locating information is not straightforward. Hahn et al. (2010) manually build an ontology from the 550 most commonly used info box templates of the English Wikipedia. Yang et al. (2009) developed a Markov Logic Network (MLN) to extract information from any web forum, such as post title, post author, post time and post content. MLN's model data by combining rules in first order logic formula's with probabilities. Each formula has a weight which indicates the strength of the constraint. The formulas in the model of Yang et al. (2009) represent relations between html elements, in which information from an individual page and the entire website are incorporated. Yang et al. (2009) labeled 20 websites of various categories as input for learning the model.

One of the difficulties of an automatic approach for structuring and automatically extracting information from texts is the dependence of the domain and type of document. Each approach needs to be tailored to the specific

information request. However, as the previously described research shows, similarities in the approach for each domain exists. The researchers make use of the known structure of the textual information searched for and the structure of the source searched in. The previously mentioned types of textual information sources all have some form of structure. The researchers all search for elements of which they know to be present in the text. Furthermore, in each problem setting the models search for the boundaries between the elements or between an element and the non-relevant text. Therefore, automatic extraction of textual information can be achieved by modeling the boundaries. In this paper, we present a straightforward and practical approach for locating information in annual reports based on modeling the boundaries between the information searched for and the remaining text. Similar to websites, annual reports contain a larger amount of text than addresses or CV's. As opposed to websites, annual reports do not have the high level of meta data as websites have. Nevertheless annual reports do roughly contain the same types of information across companies. This holds especially for annual reports filed on forms with the SEC, such as form 10-K. As a result of rules and regulations it is known which types of information are or should be present in the text. This ranges from the presence of specific information concerning the content to the inclusion of specific sections.

3.3 Data and sample selection

This section describes annual reports as a data source. First, the sample of annual reports and the source of this sample is described. Secondly, the structure of the selected annual reports is described in more detail. Finally we describe the limited pre-processing steps required before applying the methods discussed in Section 3.4.

3.3.1 The sample

In this research, we use annual reports filed with the Securities and Exchange Commission (SEC) in the United States, to illustrate the possibilities and challenges of specific textual information extraction. The SEC provides a clear structure for annual reports in forms 10-K for U.S. based companies and 20-F for foreign companies (Securities and Exchange Commission, 2014c,d). However, deviations from this structure are possible. Some sections can be omitted when specific criteria are met. The forms 10-K and 20-F are intended as a guide to meet the requirements of the SEC, not as a form to be filled in. The annual reports can differ between companies and between years. The

automatic extraction of specific information in annual reports is therefore not straightforward.

The annual reports are retrieved from the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system, which stores annual reports of organizations that file their reports with the SEC. These reports are freely available via the EDGAR website (Securities and Exchange Commission, 2014b). In the EDGAR website one can search for specific files or organizations using the company name, Central Index Key (CIK) or file number. For a more general query can be searched for state, country or Standard Industrial Classification (SIC).

The data set collected consists of 1.609 annual reports on forms 10-K, 10-KSB and 20-F in the period from 1999 to 2011. Research on annual reports mainly focuses on reports on form 10-K only, which are filed by U.S. based companies. The form 10-KSB is an abbreviated form 10-K that was filed by small companies until 2009. The forms 20-F are filed by companies based in other countries than the U.S. To be able to extract information about more organizations the selection of annual reports for this research includes annual reports on forms 10-KSB and 20-F as well as 10-K. Table 3.1 summarizes the number of annual reports per form type selected for this research. The next section, section 3.3.2, describes the structure of the forms 10-K, 10-KSB and 20-F in more detail.

Form type	**Nr of reports**	**Percentage of reports**
10-K	1.334	82,9 %
10-KSB	194	12,1 %
20-F	81	5,0 %
Total	1.609	100,0 %

Table 3.1: The number of annual reports per form type selected for this research

3.3.2 The structure of annual reports

This section shortly describes the structure of annual reports on forms 10-K, 10-KSB and 20-F required in the subsequent sections. These subsequent sections explain the automatic extraction of specific textual information.

The digital forms downloaded from the EDGAR website contain some general information before the actual start of the form. This information consists of information about the form and the company, including the form type, the

filing date, the conformed period of the report, the company name, Central Index Key (CIK code), Standard Industrial Classification (SIC code), business address and if applicable former company names. Figure 3.1 shows the beginning of the submission file of Google to the SEC in 2015.

```
<SEC-DOCUMENT>0001288776-15-000008.txt : 20150209
<SEC-HEADER>0001288776-15-000008.hdr.sgml : 20150209
<ACCEPTANCE-DATETIME>20150206194031
ACCESSION NUMBER:               0001288776-15-000008
CONFORMED SUBMISSION TYPE:      10-K
PUBLIC DOCUMENT COUNT:          14
CONFORMED PERIOD OF REPORT:     20141231
FILED AS OF DATE:               20150209
DATE AS OF CHANGE:              20150206

FILER:

        COMPANY DATA:
                COMPANY CONFORMED NAME:                 Google Inc.
                CENTRAL INDEX KEY:                      0001288776
                STANDARD INDUSTRIAL CLASSIFICATION:     SERVICES-COMPUTER PROGRAMMING, DATA PROCESSING, ETC. [7370]
                IRS NUMBER:                             770493581
                STATE OF INCORPORATION:                 DE
                FISCAL YEAR END:                        1231

        FILING VALUES:
                FORM TYPE:              10-K
                SEC ACT:                1934 Act
                SEC FILE NUMBER:        001-36380
                FILM NUMBER:            15586408
```

Figure 3.1: The beginning of the submission text file of Google to the SEC in 2015.

The form 10-K is divided in four parts consisting of a total of 15 items. The four parts divide the report in the subjects 'Company overview and description', 'Business Discussion and Financial Data', 'Insiders and Related Parties' and 'Full financial Statements and Footnotes'. Until March 2009 smaller companies could file their reports on form 10-KSB (10-K for Small Businesses). The form 10-KSB is similar to form 10-K but contains less detailed information. The form 10-K can be downloaded from the SEC the website, https://www.sec.gov/about/forms/form10-k.pdf (Securities and Exchange Commission, 2014c).

The form 20-F is divided in three parts consisting of a total of 19 items. The third part contains the financial statements. There is no exact reconciliation between the items of form 10-K and form 20-F. Only some of the sections on form 20-F directly correspond to sections on the form 10-K. Items 17 and 18 'Financial Statements and Supplementary Data' in form 20-F correspond in form 10-K to item 8 'Financial Statements and Supplementary Data'. Item 5 'Operating and Financial Review and Prospects' of form 20-F can be found in form 10-K as item 7 'Managements Discussion and Analysis of Financial Condition and Results of Operations'. The form 20-K can be downloaded from the SEC website, https://www.sec.gov/about/forms/form20-f.pdf (Securities and Exchange Commission, 2014d).

3.3.3 Data pre-processing

For documents to be processed they must be machine readable. The annual reports in forms 10-K and 20-F downloaded from the Edgar-website are in html format, which is a format that can be processed by a computer. Therefore, the pre-processing steps required for the methods developed for this research are limited.

The pre-processing steps required concern the presence of html tags. Since annual reports on forms 10-K and 20-F downloaded from the Edgar-website are in html format, they contain html tags that are not part of the content of annual reports. These html tags are omitted in this research. We use the Python package BeautifulSoup to perform the recognition and removal of the html tags in the text. However, some of the html tags may provide useful indications for finding specific parts of information in the document. For example, the html tags that indicate the start and end of tables. Section 3.4.4 discusses this in more detail.

3.4 Methods and results

The annual report contains a lot of textual information. In this research, we identified four categories of information to be located. The first category is specific information, which are short and clear pieces of information, such as a year or company name. The second category consists of specific sections. The specific section used for this research is the section 'Managements Discussion and Analysis' section. The third category discussed are referenced sections. Larger documents, such as annual reports, contain references to other parts of the document, in the form of 'See Section X on pages Y-Z for subject S'. In this example 'subject S' is what we termed a referenced section. The last category are tables. We do not consider numerical tables to be textual information. However, in html documents tables may be used as a markup tool instead of presenting numbers. Therefore, the task of locating tables does not only consists of identifying the table but also determining what type of information the table contains.

To locate information in a document a basic intuitive principle applies. We need to know where the relevant piece of information starts and ends. This means that we model the boundaries between the information searched for and the remaining text. Depending on the type of information that needs to be identified the start and end markers vary. These start and end markers are the characteristic keywords that the methods needs to locate the requested information. These markers are the technical substitute of the knowledge that

humans use to locate information. In the remainder of this research paper we refer to this intuitive principle with the term 'information location principle'.

We apply the information location principle to the four categories of information to be located in annual reports. Each of the next sections elaborates on one of the categories. The extraction of specific information using the information location principle is described in section 3.4.1. Section 3.4.2 explains how the information location principle can be applied to locate a specific sections in a document. Section 3.4.3 elaborates on the requirements for locating the referenced sections. Section 3.4.4 describes locating the text tables in annual reports.

3.4.1 Extracting specific information

Specific information is the answer to a straightforward question and consists of one to several words. For example, the question 'what year is this annual report about?' is straightforward and the answer consists of a single word. An automated approach for locating this type of information allows providing the answers to straightforward questions without having to search the document manually or even opening the document. In this way, information can be extracted from a large number of documents. For example with the purpose of sorting documents or selecting documents with specific characteristics, such as selecting documents from a specific year.

In this section, we discuss the information location principle for automatically finding specific information. The start and end locations of the specific information in the documents are modeled. The way in which these starts and end locations can be modeled depends on how the specific information is incorporated in the document searched in. We describe four examples of specific information that exists in annual reports, namely the year of the annual report, the sector in which the company operates, the company name and the number of employees that work within the company. These examples show how the way in which information is incorporated in documents influences the approach and possibility for locating specific information.

Automatically extracting the year that the annual report is about is straightforward because the year information is included in the general information before the actual start of the forms 10-K and 20-F, as shown in Figure 3.1. The structure of the beginning of the document is similar for all reports. This structure contains multiple dates, including the date of the period of which an annual report is about and the date on which the report was filed. The start of this information can easily be identified by the phrases 'Conformed period of report' and 'Filed as of date'. The end of the information is straightforward

because the length of the date information is known. The dates consist of eight characters: four for the year, two for the month and two for the day. Therefore, the first four digits after the phrase 'Conformed period of report:' indicate the year the annual report is about. The result of the automatic retrieval of the report year is 100%.

Similar to the year of an annual report, identifying the sector is straightforward because the information is included in the structured information at the beginning of the document. The start of the sector information can be identified with the phrase 'Industrial classification'. Since we know that the information is on top of the forms 10-K and 20-F, as seen in Figure 3.1, the first instance of 'Industrial classification' that occurs in annual reports is the begin marker of the information to be extracted. Identifying the end point of this information is slightly more challenging compared to the year of an annual report because the length of the sector information varies. The sector information does not only contain the SIC code which is a 4-digit code to classify industries, but also a short description of the industry. However, this part of an annual report is still very structured, so the information following the sector information should be the same for each report. Therefore, the end marker for the sector information is the next information in an annual report, which should be the 'IRS number, Internal Revenue Service (IRS)'. This is true for 98,4 % of the annual reports. Table 3.2 shows the possible end markers for the sector information in annual reports.

End marker	**Nr of reports**	**Percentage of reports**
IRS number	1.583	98,4 %
Filing values	8	0,5 %
Fiscal year	10	0,6 %
State of	8	0,5 %
Total	1.609	100,0 %

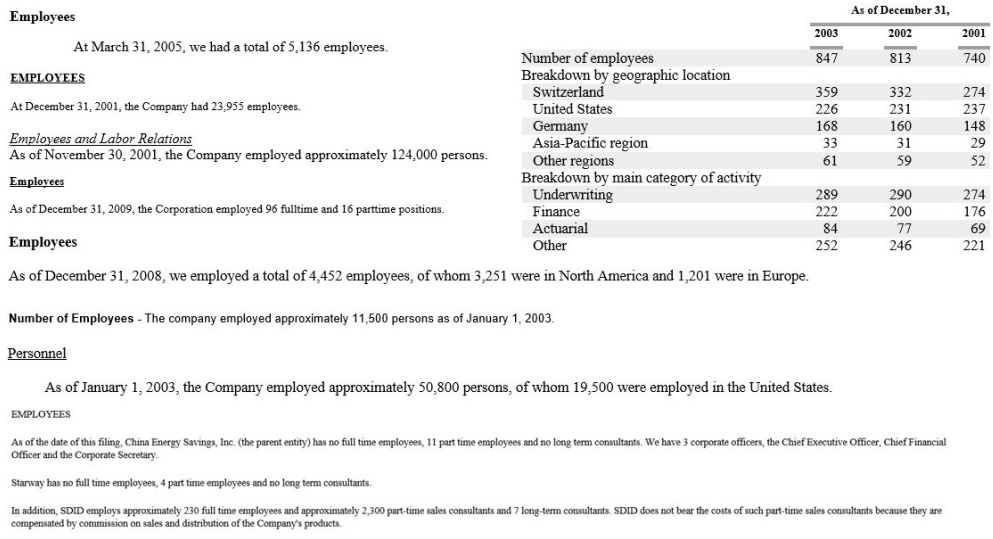
Table 3.2: The end markers per annual report for the sector information.

The approach for extracting the company name is similar to the extraction of the year because the name is also part of the general information at the top of the form. The start identifier for the company name is the phrase 'Company Conformed Name'. Of course the length of company names varies. The end of the company name can be recognized in an annual report because the company name is followed by the phrase 'Central Index Key'. However, locating each instance of a company name in the report is more difficult because abbreviated

forms of the company name may be used throughout the text. The conformed company name is the full name of the company. In the example of Google the conformed company name is 'Google N.V.', while in the title the name 'Google Inc.' is used and throughout the report mostly 'Google'. A further complicating factor for recognizing the company name is that the term Google is not only used to refer to the company but is also part of product names, such as 'Google+' or as part of the website link. Recognizing a company name is within the field of Named Entity Extraction which is beyond the scope of this research.

The examples so far show how to extract information that for each annual report is located in the same position, either between two identifiable phrases or after an identifiable phrase and a fixed number of characters. An other example of specific information is the number of employees that works within the company. This information is included in the text of an annual report, but the format differs per annual report. Figure 3.2 gives an overview of some of the ways in which the number of employees is reported in different annual reports. The number of employees is either reported in a table or within the text. Different headers are used for the paragraph containing the number of employees, such as 'Employees', 'Number of employees' and 'Personnel'. Within the text also different terms are used to refer to the employees. Examples of these terms are 'employees', 'persons' and 'parttime positions'. In some annual reports the number of employees is reported in several numbers. This occurs when the number of full-time and part-time employees are reported separately or the numbers are reported per business division. Because of all these variations, modeling the boundaries between the number of employees and remaining text can not be achieved by one start and end maker for all annual reports. Only two main characteristics can be defined for all annual reports. First, the information searched for is a number. Secondly, the number is within the proximity of one of the following words: 'employee', 'employees' and 'personnel'. The varieties between annual reports could all be modeled and the model than needs to examine several possibilities. So, to be able to locate the number of employees more techniques than only the information location principle are required.

Overall, to automatically locate specific information in annual reports we make use of the structure of annual reports. This structure aids the modeling of the start and end location of the specific information. So, it can be concluded that, in the absence of a clear structure additional information is required to locate the specific information.



Employees

At March 31, 2005, we had a total of 5,136 employees.

EMPLOYEES

At December 31, 2001, the Company had 23,955 employees.

Employees and Labor Relations
As of November 30, 2001, the Company employed approximately 124,000 persons.

Employees

As of December 31, 2009, the Corporation employed 96 fulltime and 16 parttime positions.

Employees

	As of December 31,		
	2003	2002	2001
Number of employees	847	813	740
Breakdown by geographic location			
Switzerland	359	332	274
United States	226	231	237
Germany	168	160	148
Asia-Pacific region	33	31	29
Other regions	61	59	52
Breakdown by main category of activity			
Underwriting	289	290	274
Finance	222	200	176
Actuarial	84	77	69
Other	252	246	221

As of December 31, 2008, we employed a total of 4,452 employees, of whom 3,251 were in North America and 1,201 were in Europe.

Number of Employees - The company employed approximately 11,500 persons as of January 1, 2003.

Personnel

As of January 1, 2003, the Company employed approximately 50,800 persons, of whom 19,500 were employed in the United States.

EMPLOYEES

As of the date of this filing, China Energy Savings, Inc. (the parent entity) has no full time employees, 11 part time employees and no long term consultants. We have 3 corporate officers, the Chief Executive Officer, Chief Financial Officer and the Corporate Secretary.

Starway has no full time employees, 4 part time employees and no long term consultants.

In addition, SDID employs approximately 230 full time employees and approximately 2,300 part-time sales consultants and 7 long-term consultants. SDID does not bear the costs of such part-time sales consultants because they are compensated by commission on sales and distribution of the Company's products.

Figure 3.2: An overview of various ways in which the number of employees is reported in annual reports.

3.4.2 Extracting specific sections

Instead of the answer to a specific question, a whole section of an annual report may be of interest. Contrary to specific information, the length of a section of a document varies between a few and a few hundred words. The information location principle can also be applied to find a specific section. In this paper, we applied the principle to locate the 'Management's Discussion and Analysis' (MD&A) section because this section is claimed to be the most read part of an annual report (Li, 2010).

A section in annual reports starts with a section header. The end of a section can be recognized by the section header of the next section in the report. The section headers differ between the forms 10-K and 20-F. From the standard forms of the SEC we know that in forms 10-K the header should be 'item 7 Managements Discussion and Analysis of Financial Condition and Results of Operations' and in forms 20-F 'Item 5 Operating and Financial Review and Prospects' (Securities and Exchange Commission, 2014c,d). The sections following these are 'Item 7A. Quantitative and Qualitative Disclosures About Market Risk' and 'Item 5. Operating and Financial Review and Prospects' for form 10-K and form 20-F respectively. The examination of the table of contents for a small subset of annual reports shows a few deviations. Item 7 may be followed by Item 8 instead of Item 7A. The title of Item 8 in these forms

10-K can be 'Item 8 Consolidated Financial Statements and Supplementary Data' or 'Item 8 Financial Statements and Supplementary Data'. Another deviation is the numbering of the sections, the MD&A section is for some annual reports numbered as Item 6. For some annual reports the section header is shortened to only the two words 'Financial Statements'. Other deviations are probably the result of technical deficiencies, such as missing white spaces and apostrophes in 'management's'. The model of the start and end markers is summarized in two tables. Table 3.3 summarizes the recognition of the start of the sections, while Table 3.4 summarizes the end location. Four words are specific enough to recognize the section headers and to capture all variations in the section headers.

Word nr.	**Possible words** (separated by ;)
1	Item; item7
2	5; 6;, 7; 7management's; 7managements; management's; managements
3	Discussion; management's, managements, operating
4	Discussion; and

Table 3.3: Section starts.

Word nr.	**Possible words** (separated by ;)
1	Item; item7a
2	6; 7; 8; 7a; 7aquantitative
3	And; financial; consolidated; quantitative; directors
4	And; financial; statements; statement; quantitative; senior

Table 3.4: Section ends.

The begin and end markers not only respond to the section headers but also the table of contents or references to the sections. The table of contents is avoided by starting the search after the first 450 words in annual reports. By making use of the format of annual reports, we determined empirically that the MD&A section starts well after 450 words, while the table of contents will fall within the first 450 words. A reference to the section can be recognized because the references to sections are preceded by one of the words 'see', 'in', 'also' or 'and', while the actual section headers are not preceded by one of these words. Additionally, the section header within a reference may be

placed between quotation marks. As a result, the word of the section header starts with a quotation mark. The start and end markers do not contain these quotation marks so words starting with a quotation mark do not match with any of the markers. For example, the approach differentiates between the words *'Item* and *Item*. Therefore, these references are automatically ignored in the search for the MD&A section.

Table 3.5 shows the results of the extraction of the MD&A section from the annual reports. For 96% of the annual reports the MD&A section was extracted using the information location principle. In summary, this result is achieved by making use of the extracted knowledge about the structure of annual reports augmented by general rules about the recognition of references to extract a specific section from annual reports.

Result	**Nr**	**Percentage**
Extracted	1.544	96,0 %
Not extracted	65	4,0 %
Total	1.609	100,0 %

Table 3.5: Section results.

3.4.3 Extracting referenced sections

Large documents, such as annual reports, contain references to other parts of the document, in general formulated as 'See Section X on pages Y through Z for subject S'. As humans we are able to locate 'subject S' using the presented section and page numbers. In this section the information location principle is applied to automate this search task. The page numbers in the reference could be used to automatically locate the referenced sections in annual reports. We selected 80 references from the MD&A section that contain page numbers and refer to information relevant for this section. Figure 3.3 shows four examples of such references. For example, annual reports on form 10-K may include a report to the shareholders which contains information required by form 10-K.

The page numbers within the references are a clear measurable component not subject to interpretation. Such components can be identified automatically by applying the approach for locating specific information as discussed in Section 3.4.1. The page numbers are extracted by identifying the phrase of the form 'pages Y through Z' first. Within the identified phrase we apply the location information principle to extract the page numbers. The start page

The information required by Item 7 is incorporated by reference from pages 1 through 32 of the 2002 Annual Report to Shareholders.

The information set forth under the caption "Results of Operations and Financial Condition" on pages 28 through 41 of the Company's 1999 Annual Report to Shareholders is hereby incorporated by reference in this document in answer to this Item.

The information called for in this Item 7 is incorporated herein by reference from the section of Park's 2007 Annual Report captioned "FINANCIAL REVIEW," on pages 21 through 35.

Incorporated herein by reference to "Management's Discussion & Analysis" on pages 32 through 39 of the Company's 2000 Annual Report to Stockholders.

Figure 3.3: Four example references to management's discussion and analysis information.

number can easily be identified because the number follows the word 'pages'. Similarly, the end page number follows the word 'through'.

The next step is to locate the pages referred to in the annual report file. Locating pages using the page numbers requires the files to have page numbering. We make use of the html format to identify the page numbers. The html tag <PAGE> indicates the start of a new page. The page number is located before this tag. To be able to find the page numbers after removing the html tags, a page start is marked with the identifier 'MARKPAGESTART' before removing the html tags.

Some of the annual report files consist of multiple html documents that each have a new page numbering. As a result, the page numbers referred to may occur multiple times within the annual report file. Therefore, we need to determine which of the documents in the file contains the referenced section. The html format is also useful to identify the various documents in the file because the start of a new document is marked by the html tag <DOCUMENT>. This tag is replaced with the identifier 'MARKDOCUMENTSTART' before removing the html tags. Some of the documents may be excluded from the search for the referenced section because its page numbers do not fall within the range of page numbers in the reference. Only documents for which the last page number is larger than or equal to the end page number of the reference may contain the referenced section, otherwise the document is too short to contain the referenced section. Likewise, the first page number of the document must be smaller or equal to the begin page number of the reference. If after this exclusion of documents still multiple documents remain that do match with the range of page numbers in the reference, two other intuitive principles are applied. First, the referenced section will be located after the reference because it is located in an appendix to the form 10-K from which

the reference is obtained. Secondly, the begin page of the referenced section is likely to contain the phrase 'management's discussion' or 'results of operations'. Therefore, the referenced section is extracted from the document that contains one of these phrases on the first page of the referenced pages.

Table 3.6 shows the results of the extraction of the 80 referenced sections. The used approach finds the correct section for 37,5% of the references. Unfortunately, many of the annual report files do not include page numbering. As a result the referenced sections could not be extracted. The extraction of a referenced section using the suggested approach requires the presence and localization of two types of information. First, the information location principle utilizes the structure of the reference to extract the page numbers referred to. Secondly, by making use of the html information the page numbers of the annual report may be located.

Result	**Nr**	**Percentage**
Extracted	30	37,5 %
Not extracted	50	62,5 %
Total	80	100,0 %

Table 3.6: Reference extraction results.

3.4.4 Extracting tables

Besides textual information, annual reports contain tables. These tables give an overview or summary of numbers that mainly present financial information. Automatically locating the tables in text documents allows the extraction of the information in these tables. This may be useful for finding information of which we know that it can be found in a table, such as the financial information. On the contrary, in text mining solutions tables containing only numerical information will most likely be excluded because the focus is on the textual information. In this case, being able to locate tables is useful for removing them from the texts.

In this section we discuss the information location principle for automatically finding tables in the MD&A section of annual reports. This means that we need to know where a table starts and ends. Defining these starts and ends is straightforward, since annual reports are in html format and therefore contain html tags indicating the start and end of tables, <TABLE> and </TABLE> respectively. Before removing the html tags in the pre-processing

phase we placed identifiable markers at the place of these html tags to preserve the table start and end information. The start of the table is marked with the identifier 'MARKTABLESTART' and the end of the table with 'MARKTABLEEND'.

Although the html format has the big advantage that html tables can be located using the html tags, the format presents a challenge. Html tables are also used for document mark-up purposes. A table for mark-up purposes contains text instead of numerical overviews. As a consequence, when the information location approach locates a table it still needs to decide whether the table found is an actual table or is a table for mark-up purposes. To perform this task automatically we define rules to distinguish texts from numerical overviews. The main difference between texts and numerical information is the percentage of numbers present in that piece of information. Figure 3.4 shows the distribution of the percentage of numbers present in texts outside html tables in annual reports. For most annual reports, less than 10 % of the words in texts outside html tables are numbers. This results in the first rule: if less than 10 % of the words in a html table are numbers then the html table most likely contains text and is used for mark-up purposes. Figure 3.4 furthermore shows that for texts outside html tables the percentage of numbers is never more than 35 %. Therefore, the second rule is: if more than 35 % of the words in an html table are numbers then the html table most likely is an actual table.

Additional information is required to make an accurate decision for html tables for which the percentage of numbers lies between 10 % and 35 %. We add two rules for this category of html tables. The first rule is based on common sense and knowledge about the size of tables in annual reports. On average a page in an annual report contains approximately 800 words. A numerical table does not consist of more than two pages. Therefore, html tables containing more than 1.600 words are most likely texts instead of actual tables. The second rule makes use of common sense and knowledge about the consequences of removing the html tags in tables. In html tables the cells of a table are defined with the html tag <TD>. After removing these html tags the numbers in a table are only separated with a white space. In texts however, numbers are usually separated by punctuation marks, such as comma's and semicolons. Therefore, when the html table contains consecutive numbers not separated by a punctuation mark the table most likely is an actual table. An alternative approach for identifying consecutive cells containing numbers creates identifiers for the tag <TD> before removing the html and search for the existence of these markers between numbers. Which approach to use

depends on the goal of the table extraction task. In case the actual tables are required, the identifiers may be useful to locate information within that table. However, if the mark-up tables need to be used for its text, the presence of the identifiers is undesirable.

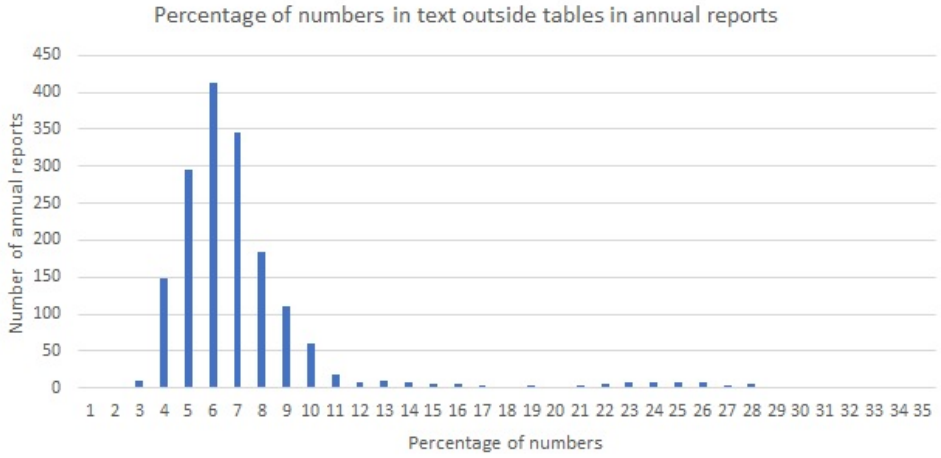


Figure 3.4: The percentage of numbers in the texts outside the html tables in annual reports.

The result of the decision whether the table is an actual table or used for mark-up purposes is examined manually for 339 tables extracted from the MD&A sections of 11 annual reports. For 99,1 % of the tables the approach makes the correct decision. For all tables that the approach identifies as being an actual table the approach is correct. For 0,9 % of the tables the approach incorrectly decided that the table is used for mark-up purposes. An examination of the incorrect decision shows that the mistakes are made for small tables for which the descriptions in the left column consist of a relatively large number of words, and therefore have a low percentage of numbers. Table 3.7 summarizes the results.

In summary, the majority of the tables are identified by using the html information supplemented with knowledge about the distribution of numbers in the text and the general structure of tables.

	Actual table		Table for mark-up		All tables	
Result	**Nr**	**Percentage**	**Nr**	**Percentage**	**Nr**	**Percentage**
Correct	98	100,0 %	238	98,8 %	336	99,1 %
Incorrect	0	0,0 %	3	1,2 %	3	0,9 %
Total	98	100,0 %	241	100,0 %	339	100,0 %

Table 3.7: Table decision results.

3.5 Discussion and conclusion

In this paper, we demonstrated a fairly straightforward and practical approach for automatically locating information in annual reports. The approach provides the possibility to efficiently extract information from a large number of annual reports. The implementation of the approach depends on the type of information that needs to be located. The approach was tailored to the goal of the information search.

We searched for several types of information in annual reports. The approach successfully located the majority of these types of information. Depending on the type and complexity of the information that needs to be identified the modeling of the start and end markers varies. The types of information we extracted in this paper show that even with the presence of structure in the document the automatic extraction of textual information may be challenging. However, for all types holds that understanding the layout of a document is important for the automatic extraction of information from that document. The layout provides clues that aid the process of locating the information searched for.

The approach is most successful when the start and end locations of the information searched for are clear. The presence of structure in documents facilitates this clarity. In annual reports this is especially true for the specific information at the beginning of the annual report forms. In the absence of a clear structure the approach needs additional information. This information may vary from clearly defined, such as a list of words that precede section headers in a reference to distinguish the start of a section from a reference, to more complex, such as the percentage of numbers in texts.

The less structured an annual report is and the more complex the necessary additional knowledge and rules are, the more challenging the automatic extraction of textual information using the information location principle is. A lack of clear structure may be compensated with additional knowledge and

rules. The examples in this paper show that the complexity of the additional knowledge and rules increases when the type of information to be extracted is more complex in terms of variety in length and position in an annual report. Specific information located at the top of each annual report is easier to extract than specific information that can be positioned anywhere in the document in various formats, such as the number of employees. Similarly, modeling the starts and ends of sections is more challenging than the starts and ends of the specific information. The position and length varies more for sections than for specific information that consists of one to several words. Therefore, the extraction of sections requires more information about the possible start and end markers than is necessary for the extraction of specific information.

The complexity of the information location approach further increases with the complexity of the information request. The request for specific information is clearly defined as a straightforward question. The request for extracting a specific section defines the section to extract. The information request itself may be less clearly defined, as is demonstrated by the example of the extraction of referenced sections. In this example, the information in the reference defines the information request. As a consequence, the information that may be used in the model to locate the referenced section is limited to the information present in the reference.

A lack of information required in the extraction process limits the possibilities of the information location approach. In that case, other possibilities need to be examined. In other cases, the required information is not clearly present or absent but is unforeseen or not evident, such as the additional information required to distinguish actual, numerical, tables from tables used for mark-up purposes.

The information location principle is practical to use and applicable to several types of information in annual reports. However, tailoring the approach to a new information search goal requires knowledge about the document and costs time. A machine learning approach may be able to automatically learn the starts and ends markers of the information searched for. A drawback of most machine learning techniques is that in order to learn, they need a set of documents for which the information to be extracted is known. For example, to let a machine automatically learn a model for extracting a specific section from an annual report it will need as input a set of documents with the resulting specific section itself or the start and end boundaries of the section. Preparing this input, which means annotating the documents for all information that may need to be extracted from the documents, may be very time consuming. Furthermore, it is argued that the advantage of a machine learn-

ing approach is that the method can be applied to new data sets. However, for each new information search goal the machine learning process, that includes data pre-processing, learning and testing, needs to be repeated. Models developed to locate the year of an annual report are not able to locate the number of employees. Future research may implement a machine learning approach and compare the results and time required with the approach demonstrated in this paper to determine which approach is more efficient and practical to use.

The information location approach for extracting the number of employees and locating referenced sections may be improved. For these types of information requests the knowledge about the structure of an annual report documents is insufficient to locate the information. More information is required to model where the relevant piece of information starts and ends. Adding knowledge and rules that reflect the variety of formats in which the number of employees is reported in an annual report may allow the automatic extraction of this information. The result of the extraction of referenced sections based on page numbers may be improved by adding other indications, such as the title of the document that includes the section searched for.

The findings of this paper emphasize the usefulness of a standardized reporting format, such as XBRL. The suggested approach provides an alternative method for the automatic extraction of information from annual reports that lack such standard tags and to extract information that may not be captured by tags of a standard format.

4 Text mining to detect indications of fraud in annual reports worldwide

Abstract

Fraud affects the financial results presented in the annual reports of companies worldwide. Analysis performed on annual reports focuses on the quantitative data in these reports. However, the amount of textual information in annual reports increased in the past decade with companies using the reports to project themselves. The texts provide information that is complementary to the financial results. Therefore, the analysis of the textual information in annual reports may provide indications of the presence of fraud within a company. This piece of research examines the possibility of using text mining techniques to detect indications of fraud in annual reports worldwide. The results of the simple baseline model that only looks at the individual words are very promising.

4.1 Introduction

Fraud is a worldwide phenomenon affecting all types of companies, from energy company Enron in the US and dairy and food corporation Parmalat in Italy to the optics and reprography manufacturer Olympus in Japan. The variety of companies engaging in fraudulent activities developed a wide range of fraud schemes. Enron hid debt and overstated profits by using loopholes in the accounting rules and improper financial reporting. The shares dropped from $90.56 per share in August 2000 to penny-stock at the end of 2001 when the fraud was revealed (Craig and Amernic, 2004). The management of Parmalat provided false accounting information by altering documents, falsifying administrative and distorting communication and data processes (Di Castri and Benedetto, 2005; Okike, 2011). Recalculation of accounting information of the first nine months of 2003 revealed that the debt was 14 billion and 300 million euros, which is eight times the reported debt. The true income turned out to be 1.300 million euros less than the reported income (Di Castri and Benedetto, 2005). Olympus covered up the losses and overstated the value of the businesses it acquired (Morgan and Burnside, 2014). The scheme resulted

in an estimated total of 4.9 billion dollars that were not accounted for in the financial statements (Tabuchi, 2011).

As the real life examples demonstrate, a wide range of fraudulent activities can affect a company's financial report. The opportunity to engage in such activities without the presence of fraud being evident from the financial reports is a result of the nature of the preparation of the financial reports. The preparation process requires making estimations, judgments and choices. Choices need to be made regarding the accounting methods to be used, when to recognize economic events and how to determine the amount of income reported (Revsine, 1991, 2002). Companies may decide to take advantage of this freedom of choice. The accounting practice liberty could be misused by fraudsters to deliberately make false estimations and judgments. Documents and other data that form the basis for the preparation of the financial reports could be falsified which has the consequence that the financial reports present an untrue and unfair view of the company's financial condition. Instead of misrepresenting information, a company can also intentionally omit significant information, which may have the same effect as providing an untrue view of its financial condition. In this paper, we make no distinction between different types of fraudulent activities.

Besides the wide range of fraudulent activities that affect the financial report, other types of fraud may occur within a company that do not have an effect on the financial statements. For example, an employee may steal money from the company by claiming travel expenses that were not actually made. This type of fraud has financial consequences for the company, but the financial effect may not be large enough to alter the financial statements. These types of fraud are not included in the definition of fraud used in this research. This research follows the definition of fraud by the Committee of Sponsoring Organizations of the Treadway Commission (COSO), which defines fraud as 'intentional material misstatement of financial statements or financial disclosures or the perpetration of an illegal act that has a material direct effect on the financial statements or financial disclosures' (Beasley et al., 2010). The latter part of the definition does not define who is perpetrating the illegal act. The perpetration may not be done by the company's management that is responsible for the financial statements and disclosures. However, due to the material effect of the perpetration, the management of the company is expected to be reasonably aware of some suspicious activities within their company. The definition is in line with the UK Companies Act 2006, which holds directors responsible for the financial statements by stating: the directors of a company must not approve accounts unless they are satisfied that

the accounts present a true and fair view of the assets, liabilities, financial position and profit or loss of the company.

With the disclosure of the financial statements in the annual report, the quantitative financial information is accompanied by textual information. In contrast to the financial information providing some details about the previous year, some of the text is about the company's expectations for the future. This especially applies to the Management Discussion and Analysis (MD&A) section, which may contain forward-looking information. In this regard, the textual information provides information that is complementary to the financial information. Therefore, examining the text incorporated in annual reports for detecting indications of fraud in them can enhance the methods used for fraud detection. Another advantage that makes the text in annual reports suitable for fraud detection is that the texts are not subjected to as many rules as the financial information, giving the company's management more freedom in their textual disclosures. Furthermore, the reach of textual information is greater than that of the financial information. More people understand the textual information in the annual reports than the quantitative information. Companies and regulators realized this in the past decade, leading to an increase in the amount and change of usage of the texts. Therefore, the text may be considered an important source of information in fraud detection procedures. The textual information ultimately has the power to influence and, in case of fraud, deceive a greater number of people.

The textual disclosures in annual reports can be analyzed by humans reading the text. However, computers are able to process a higher number of reports in a shorter amount of time. Although, unlike humans, computers may not able to understand the content of the text, they are better equipped to extract the more abstract linguistic information from it. For humans, it is challenging to ignore the message of the text and focus only on the linguistic information that provides clues on how something is said (Pennebaker et al., 2003). This type of information may contain indications of fraud that would be missed by humans. The field of text mining utilizes the computer's ability to derive characteristics from a large number of textual documents. Through machine learning models the computer is able to find patterns based on these characteristics that are specific to predefined categories, such as the category of 'fraudulent reports'.

As a result of the increased awareness of the importance of textual information in annual reports and the expansion of computer capabilities, the texts in financial reports draw the attention of researchers. The research related to the texts focuses on the annual reports filed by the US companies with the Securities and Exchange Commission (SEC), which are referred to as annual

reports on form 10-K (10-K). However, fraud is not limited to the US companies but occurs worldwide, affecting annual reports worldwide. Indications of fraud may appear in the textual information, regardless of the format of the annual report that include those texts. Therefore, this research examines the possibilities of using textual information to detect fraud in annual reports worldwide. The research includes annual reports in all types of formats and sizes filed in various countries.

A previous research conducted analyzing the text in the MD&A section of annual reports showed promising results. Therefore, this research focuses on the MD&A section to determine whether it can also be successfully used to detect the indications of fraud in annual reports worldwide. Despite the similarity, the data set used in this research differs from the previous studies on several points. Firstly, this research is not limited to the annual reports on form 10-K but includes annual reports from several countries in various formats. Secondly, previous research used a limited number of 10-K reports, possibly due to the time consuming manual extraction of subsections such as the MD&A. However, this research uses an approach for automatic extraction of the MD&A, allowing more annual reports to be processed. Thirdly, it uses a data set that has a more realistic distribution of annual reports affected by fraud and those that are not. The data set reflects that the majority of the annual reports are not affected by fraudulent activities. For a more accurate depiction of reality, the data set reflects that more small than large companies exist. Finally, since this is a more recent study, the fraud cases and other annual reports included in the data set are more recent as well.

The current research uses the extensive and reality approaching data set containing annual reports of companies worldwide to develop a baseline model comprising simple textual features. The baseline model is used to provide an answer to the research question:

Can a text mining model be developed that can detect indications of fraud in the management discussion and analysis section of annual reports of companies worldwide?

The remainder of this article is organized as follows: Section 4.2 provides an overview of the research conducted on annual reports with special attention to risks and the detection of indications of fraud. In Section 4.3 the data collection and processing steps are described. Section 4.4 briefly describes the machine learning methods that are used to develop the baseline text mining model. Section 4.5 provides an overview of the performance of the baseline model. Lastly, Section 4.6 discusses the results and limitations of the baseline

model and presents suggestions for further research.

4.2 Previous research on the detection of fraud in annual reports

The need for annual financial reporting arose from the increasing complexity of business arrangements. The goal of the financial report is to provide information for further analysis (Revsine, 1991). Revsine (1991, 2002) explains that management learned that the way people perceive the information in the financial report can be controlled as a result of flexible reporting rules. Even with financial reporting rules management gets the opportunity to present a financial report that lacks transparency. Ndofor et al. (2015) show that this lack in transparency may be enough for management to commit fraud. However, other incentives reinforce feeling the urge to commit fraud.

Researches found several possible incentives for committing financial fraud. Ndofor et al. (2015) found that management having stock options enhances the likelihood of financial fraud. The research of Dechow et al. (1996) showed that the main motive for manipulating earnings is to attract external financing at low costs (Dechow et al., 1996). Beneish (1999), however, could not replicate the results of Dechow et al. (1996), but found that managers are more likely to sell their company before the public discovery of the earnings manipulation. Risks of fraud are greater when a company has a board dominated by a few individuals and management has override controls (Wang et al., 2011). Agostini and Favero (2013) propose that financial statement fraud is caused by an excess of power in the hands of the managers or the pressure on performance exerted by investors and analysts. Beasley et al. (2010) summarize the motivations for fraud cited by the SEC which include meeting earnings expectations, concealing a deteriorating financial condition, increasing the stock price, bolstering financial performance of equity or debt financing and increasing the management compensation. Rezaee (2005) analyzed well known fraud cases including Enron and Worldcom and found that participation, encouragement, approval and knowledge of top management exists in the majority of the fraud cases. Economic incentives are the main motivation to commit financial statement fraud. The probability of fraud is the highest when the companies financial results are deteriorating which is in line with the results of (Beasley et al., 2010).

4.2.1 Traditional detection of fraud

The incentives for committing financial fraud may lead to fraudulent activities. These activities can be summarized in fraud schemes. Most of the fraud schemes affecting financial reports include overstating revenues and assets and misstating expenses (Dechow et al., 2011; Spathis et al., 2002). The characteristics of regularly encountered and well known fraud schemes are used to develop procedures and methods to detect indications of fraud. The Sarbanes-Oxley Act of 2002 contains provisions regarding the quality, integrity and reliability of financial reports that make top management accountable for their companies disclosures. These provisions include the need for the top management to certify that the disclosures are accurate and complete. A provision regarding the textual disclosures in the annual report dictates that the MD&A section should discuss and fully disclose critical accounting estimates and policies (Rezaee, 2005). The ASB SAS standard requires auditors to assess the risk of fraud in each audit. The three main categories of 'red flags' are: management's characteristics and the attitude of the management toward the internal control system; industry conditions; operating characteristics and financial stability (Spathis et al., 2002). Detecting irregularities that have a material effect on the companies financial statements is considered a responsibility of the auditor (Persons, 1995). Investors deem the auditors and the regulators responsible for fraud detection. Investors who actively incorporate the risk of fraud in their investment decisions use the red flags to assess the risk of fraud themselves (Brazel et al., 2015). Besides the detective function, auditor involvement has a preventive function. The research of Ndofor et al. (2015) reveals that extensive monitoring by the audit committee reduces the likelihood of fraud.

Despite the guidelines, detecting indications of fraud during an audit is challenging owing to the lack of adequate knowledge about the characteristics of fraud, a lack of experienced auditors due to the infrequency of occurrence of fraud and the efforts of a company to hide the fraud (Spathis et al., 2002). The KPMG fraud survey of 2008 (Forensic, 2008) shows that none of the large frauds was detected by the external audit and only 4% by the internal audit. The report to the nations on fraud and abuse of the ACFE in 2016 showed larger but still limited fraud detection frequencies by internal and external audits, 16,5% and 3,8% respectively (ACFE, 2016b). As a result of the challenges prevailing in the detection of fraud indications, a range of models is being developed that can aid the auditors in detecting fraud.

4.2.2 Management and financial information to automatically detect fraud

Several researchers have developed automated models that try to detect indications of fraud. The majority of these models focus on management characteristics and financial ratios. Persons (1995) used 10 financial statement variables that are regularly used to assess the companies financial condition. With a linear regression model, Persons (1995) showed that a subset of these variables is suited to detect indications of fraud in the financial statement. Beneish (1997) used measures for incentives, for example, the percentage of shares held by management and financial statement measures to detect earnings management for companies violating the Generally Accepted Accounting Principles (GAAP). In a later research Beneish et al. (1999) used only financial variables to detect such GAAP violations. Note that financial statements can also be misleading when they are in accordance with GAAP. Razali and Arshad (2014) combine the model of Beneish et al. (1999) with Altman's Z-score (Altman et al., 2000), indicating the bankruptcy risk based on financial measures to determine the likelihood of fraud in the annual reports of Malaysian companies. They found that the likelihood of fraud is lower for companies with board effectiveness, which is measured by comparing the contents of the annual reports with the Malaysian code of corporate governance. However, they found no significant relation between the likelihood of fraud and the board size, contrary to the results of Wang et al. (2011). Kaminski et al. (2004) extracted 21 financial ratio's from annual financial statements of seven years from 79 fraudulent companies and 79 matched non fraudulent companies. The results of their discriminant analysis show that the financial ratio's only have a limited ability to detect fraud. Hoogs et al. (2007) use financial ratio's and financial metrics as input for their genetic algorithm that detects fraud in the financial statements of companies reporting to the SEC. With this approach, they achieve results 4% above the chance level. Ravisankar et al. (2011) were able to detect fraud in the financial statements of Chinese companies by applying several machine learning techniques with financial ratio's as input. Huang et al. (2012) developed a neural network approach using financial statement variables to aid credit providers in assessing the reliability of Taiwanese financial statements. Their model provides the risk indicators that need further investigation. Perols (2011) tried several machine learning algorithms with 42 financial variables and ratio's. Only six of these 42 variables are found to be informative for fraud detection in all algorithms. Grove and Basilisco (2008) demonstrate that financial information alone is not sufficient to detect fraud. They therefore combine financial ratio's with corporate governance informa-

tion. The limited detection capabilities of financial ratio's alone could also be improved by incorporating textual information.

4.2.3 Textual information to automatically detect fraud

The amount of textual information in the annual report increased in the past half-century. For example, in the UK the amount of text increased by 375% between 1965 and 2004 (Beattie et al., 2008). Goel and Gangolly (2012) reported an increase in the median numbers of 15.991 in 1994 to the median of 55.000 words in 2007 for 10-K annual reports. The increase includes regulatory as well as voluntary information, with the largest part of the increase attributable to the latter. The usage of the annual report changed over time. From mostly communicating accounting information, companies began using it to express their corporate identity (Beattie et al., 2008; Lee, 1994). At the same time, the computer capabilities advanced, allowing for more efficient exploration of textual data (Smith and Taffler, 1999). With the increase in the amount and usage of text in annual reports and enhanced computer power, the interest in the text and possibilities for analysis increased.

Impression management

Li (2006) counted the frequency of words related to risk and uncertainty in 10-K reports to predict lower future earnings. He also found that the tone of the forward looking statements in 10-K and 10-Q reports is positively associated with future earnings when the tone is extracted using the Naïve Bayes Machine Learning approach (Li, 2010). However, the measures for tone based on the three commonly used dictionaries can not be used to predict future performance. Li et al. (2011) considered management discussions of 10-K annual reports for words used to refer to the companies competition to construct, a measure of the competition, which in earlier research was based on financial ratios. Li et al. (2011) found that managers strategically distort competitive information in their publications. Smith and Taffler (1999) introduced the term 'accounting bias' to explain the phenomenon in which negative results are explained in technical accounting terms and positive results in terms of cause and effect. This bias may be the underlying mechanism of impression management. Impression management hypothesizes that textual information is used by the management to conceal bad results or emphasize good performance (Merkl-Davies and Brennan, 2007). Clatworthy and Jones (2003) found this bias in 100 UK listed companies by analyzing the Chairman's statement in the annual report. Management has the tendency to blame the external

environment while taking credit for good performance. This tendency holds for companies with deteriorating performance as well as the ones with good results. In their follow up research Clatworthy and Jones (2006) found that the unprofitable companies use fewer personal references, more passive sentences, and focus more on the future in the chairman's statement compared to that by the profitable ones. Aerts (2005) further explores the internal cognitive and external processes that lead to impression management.

Financial distress

Several researchers examined the possibility of using text to predict bankruptcy. Cecchini et al. (2010) created dictionaries of keywords based on concepts of WordNet, a lexical database of English in which words are organized into sets of synonyms, from the MD&A section of 10-K reports. The created dictionary is extended with two word and three word phrases. These dictionaries are used as input for a machine learning algorithm to discriminate between 78 bankrupt and 78 non bankrupt companies, which succeeded for 80% of the 156 companies. Smith and Taffler (1999) demonstrated that the chairman's statement in the annual report contains information about the future of the company, which allows a discrimination between bankrupt and the financially healthy companies, based on words. Balakrishnan et al. (2010) used words to predict market performance. The results demonstrate that the text contains information regarding the under-performance and out-performance of companies. Hájek and Olej (2013) combine financial ratio's with word categories that indicate the sentiment in the annual reports of US companies allowing for detection of financial distress. Their results show that adding sentiment information leads to a more accurate prediction of financial distress than when the prediction is based on financial information alone. In a subsequent research with more sentiment word categories Hajek et al. (2014) showed that a change in the development of a company influences the textual sentiment information in the annual report.

Detection of lies

The impression management and financial distress studies confirm that the financial performance of a company influences the textual information. However, impression management studies show that the way in which bad performance is disclosed may be completely truthful. It is conceivable that the texts in case of fraud are deceptive. In deception detection research deception is defined as a deliberate attempt to mislead others (DePaulo et al., 2003). Al-

though the definition of fraud shows overlap with the definition of deception, the focus of deception detection research is on facial expressions and emotions of people telling lies. A large discrepancy exists between detecting spoken lies and written misstatements. The major differences are the lack of non verbal cues and the time available to create a credible story. The story in well-written documents such as the annual report is usually thought through. Ultimately, the story is not created to cover up the misstatements in the annual report but is concealed in all processes throughout the business. The annual report is the final step in the business processes to be affected by the misstatements. However, deception detection research does include cues that may also affect written lies. DePaulo et al. (2003) found that liars provide less details than truth tellers and to a lesser extent, liars include fewer ordinary imperfections and unusual contents in their stories than truth tellers. Burgoon et al. (2003) performed deception detection research on chat texts and concluded that the language use by liars differs from that of the truth tellers. Newman et al. (2003) showed that non content words are a suitable means to detect liars in several contexts. Their automated approach based on word counts outperformed humans. Zhou et al. (2004a) showed that lie detection based on textual cues is feasible. However, the relevant cues that enable distinguishing truthful texts from the lies depends on the context.

Detection of fraud in textual information

The results from deception detection research can be applied to fraud research to determine whether these are suitable to detect fraud in textual information. A difficulty with applying deception detection theory to annual reports is that annual reports are likely to be written by other persons than only those who are directly involved in the fraud scheme. The fraudulent activities can still influence the text in the annual report since the results and expectations likely result from the fraud scheme. However, the writer of the reports may not be the one who is lying consciously. Wang and Wang (2012) argue that liars leak information in the MD&A section of 10-K annual reports. They counted words known from deception detection theory of Larcker and Zakolyukina (2012) to detect fraud in a limited data set consisting of five companies of which three are fraud cases, the well known Enron, Worldcom and Xerox. However, Skillicorn and Purda (2012) show that fixed word lists are only weakly predictive or not predictive of fraud. Humpherys et al. (2011) implemented the deception cues of Zhou et al. (2004a) as input for several machine learning techniques to detect fraud in 10-K annual reports. They demonstrated that a subset of the cues with the machine learning methods 'Decision Tree' and 'Naïve Bayes'

contribute to the detection of fraud. Goel and Gangolly (2012) systematically tested linguistic cues of Humpherys et al. (2011) and additional linguistic cues on 10-K annual reports and concluded that significant differences exist between fraudulent and non fraudulent reports in terms of sentence complexity, reading difficulty and the use of negative words, passive voice, uncertainty markers and adverbs.

Several researchers have gone a step further in detecting fraud in textual information. Instead of using predefined lists of cues and words these studies use machine learning to extract the characteristics that are most informative for distinguishing between fraudulent texts and non fraudulent texts automatically. Glancy and Yadav (2011) counted the words of the MD&A section of 69 fraudulent and 69 non fraudulent reports on form 10-K and the shorter form 10-KSB. The most informative words are input for their word clustering model that automatically creates two clusters, one with fraudulent reports and the other with the non fraudulent reports. The model was tested on 11 fraudulent and 20 non fraudulent reports. Even for this limited data set they were able to assign the majority of the reports to the correct cluster with an accuracy that is well above chance level. Goel et al. (2010) exploited the text of the entire 10-K reports for the data set consisting of 405 fraud reports of 126 companies and 622 non fraudulent reports of 622 companies. They extract word counts and several linguistic features such as the average word length, average sentence length and the frequency of proper nouns. The performance of their machine learning models increased after incorporating linguistic features compared to only using the word counts. Purda and Skillicorn (2010) looked at the MD&A section of 4.895 annual and quarterly reports, in forms 10-K or 10-KSB and 10-Q or 10-QSB respectively, of which 1.038 of the reports were affected by fraud. The reports were of 189 companies which all were, at some point in time, affected by fraud. The non fraudulent reports are the reports of the same 189 companies as the fraudulent reports, but outside the fraud period. Their results show that by using word counts of automatically generated lists of informative words more reports are classified correctly as being fraudulent or non fraudulent than when only quantitative variables are used. In following this method, classifying the non fraudulent reports immediately prior to and after the fraudulent period is the most challenging part. In a follow-up research, (Purda and Skillicorn, 2012, 2015) demonstrate that the classification results are better for automatically generated lists of informative words than for pre-defined lists of words that are said to be indicative of deception, negativity, uncertainty and litigious activity. Cecchini et al. (2010) applied a method that creates dictionaries of word concepts from WordNet,

two-word phrases and three-word phrases from the MD&A section to distinguish between reports of fraudulent and non fraudulent companies. With this method they were able to discriminate between bankrupt and non bankrupt companies with an accuracy of 80%. The same approach is able to discriminate between the 10-K reports of 61 companies that committed fraud and 61 who did not commit fraud with an accuracy of 75% for the 122 10-K reports.

The research that uses automatically extracted lists of informative words for distinguishing between fraudulent and non fraudulent reports achieve promising results. These good results are obtained for the text of the entire report and for the MD&A section only. However, the research only uses 10-K annual reports, which means that the results only hold for the US companies that file with the Securities and Exchange Commission. The question remains whether the text in annual reports can be used to detect fraud worldwide. Furthermore, in most of the previous studies, a limited number of annual reports is included in the data set. This research applies a method for automatic extraction of the MD&A section, allowing more annual reports to be processed and included in the data set. Another notable characteristic of many of the data sets used in studies that classify annual reports as being fraudulent or not is the distribution of fraudulent and non fraudulent reports. This distribution is often 50% fraudulent and 50% non fraudulent. However, in reality fraud is not committed by half of the existing companies. The data set for this research reflects that the majority of the companies is not affected by fraud. Additionally, the composition of the data set reflects that in the real world, more small than large companies exist. The next section, Section 4.3, describes the data collection process.

4.3 Data selection

To test whether the automatic analysis of the textual information in the management discussion and analysis section of annual reports provides indications of fraud, we collect annual reports which are known to be affected by fraud and those which are not. The annual reports affected by fraud form the 'fraud set', while the other reports are the 'no fraud set'. Not knowing whether fraud took place does not necessarily mean that the annual report is not affected by fraud at all. Fraud might have taken place, but may not have yet been detected. However, following the presumption of innocence, we assume that these annual reports are not affected by fraud.

The next section describes the data collection process and the resulting set of annual reports. We subsequently extract the management discussion and

analysis section from these reports. Section 4.3.2 describes the procedure to automatically retrieve the management discussion and analysis sections.

4.3.1 Annual reports

In general, annual reports are prepared to provide shareholders with information about the companies activities and financial performance in the preceding year. However, the exact basis for preparation of annual reports differs for organizations worldwide because the requirements vary per country. The largest difference exists between the annual reports of US companies that file with the SEC and of those that are not US companies or do not file with the SEC.

Different sets of rules apply when preparing the financial statements and the annual report depending on the stock exchange to which the organization is listed or the country of origin. As of 2005, companies listed to a European stock market need to comply with the International Financial Reporting Standards (IFRS). Companies in the United States must comply with the US Generally Accepted Accounting Principles (US GAAP). The IFRS differs from the US GAAP. The IFRS is considered principles based, while US GAAP is more rule based.

Besides the different sets of rules that are used for annual reports across the world, the formats differ as well. Organizations listed to a US stock exchange must fill in forms that need to be filed with the SEC. The US companies file on form 10-K (Securities and Exchange Commission, 2014c), while non-US companies listed to a US stock exchange file on form 20-F (Securities and Exchange Commission, 2014d). The disclosure on form 20-F is less extensive than that in form 10-K due to exemptions allowed for foreign companies (Higgins, 2003). Until 2009, small US businesses filed on an abbreviated 10-K form, referred to as 10-KSB. After 2009, small business also filed on form 10-K. European annual reports are known to have a freer format. The European Union has the Accounting Standards Board (ASB) that drafts the IRFS. Each country subsequently interprets and implements these laws and regulations that defines the frequency and structure of the disclosures. The applicable regulations for a company further depend on the industry of the company and whether the company is listed to a stock exchange.

Scientific research that analyzes annual reports generally focuses on annual reports on form 10-K filed with the SEC by US companies (Humpherys et al., 2011; Purda and Skillicorn, 2010). To test whether it is possible to detect indications of fraud using text analysis for organizations worldwide, regardless of the annual report format or the set of accounting rules used, annual reports in formats other than 10-K will be selected. As fraud occurs in various

industries for companies of all sizes, no selection based on industry or size is made (Beasley et al., 2010). Although the companies selected for this research are located in English as well as non-English-speaking countries, all selected annual reports are written in English. As a result of the internationalization process an increasing number of companies publishes their annual reports in English. In 2003 50% of the companies from non-English speaking companies published their reports in English (Jeanjean et al., 2010).

We selected the fraud cases in the period from 1999 to 2013 from news messages and the Accounting and Auditing Enforcement Releases (AAER's) published by the SEC. Fraud cases are only selected from news articles when the case is described by multiple articles in various newspapers and the investigation has proved the presence of fraud. Cases which are still being investigated or for which fraud could not be concluded from the investigation are not selected.

The SEC takes enforcement actions against firms that it identifies as having violated the financial reporting requirements of the Securities Exchange Act of 1934, which includes US GAAP. The details of the actions are published in AAER's and are publicly accessible on the SEC website. The SEC publishes an AAER when enforcement actions are taken against a company, a company's auditor or officers for having violated the financial reporting rules of the Securities Exchange Act of 1934, which includes US GAAP. For one fraud case multiple AAER's can be published. From the approximately 2.500 AAER's in the period from 1999 to 2013, we selected the AAER's that contain the words 'fraud' and a term indicating that an annual report is affected. These terms are '10-K', '20-F' or 'annual report'. In the rest of the world there is no financial supervisory board with such extensive public documentation of the annual reports and enforcement actions. The European Union has the ASB that drafts laws and regulations, but no umbrella committee exists that requires standard forms for filings, takes enforcement actions and publishes the reporting violations and annual reports. Because the SEC filings and enforcements are very well documented, especially in comparison with the rest of the world, most of the selected annual reports are in 10-K or 20-F format.

For each annual report in the fraud set we collect annual reports of companies similar to the companies in the fraud set, but for which no fraudulent activities are known. The latter category is referred to as the no-fraud reports. We match the fraud and no-fraud reports on year, sector and number of employees. The year is defined as the fiscal year for which the annual report describes the organizations activities and financial performance. The fiscal year of an organization is not necessarily the same as the calendar year.

We defined the sectors as the divisions of corporation finance (A/D offices) to which the annual reports are assigned for review by the SEC. To which office an organizations annual report is assigned depends on the Standard Industrial Classification (SIC) codes that indicate the company's type of business. Approximately 445 SIC codes are assigned to 14 offices[1]. The non-SEC annual reports are assigned to a division, based on the companies main business.

The number of employees, obtained from the annual reports themselves, is used as a rough indication of the size of the organization. Small companies employ up to 50 people, medium-sized companies up to 250 and large companies employ more than 250 people (European Commission, 2014). The company size has an influence on the level of details contained in the annual report (Aerts, 2001). Therefore, the annual reports of fraudulent companies are matched to companies of a similar size. To reflect that more smaller than larger companies exist in the world we match the annual reports from smaller companies to more no-fraud annual reports not affected by fraud than annual reports from larger companies (Bureau, 2014). For each annual report of a small company affected by fraud, we put five no-fraud annual reports in the data set. For each medium-sized company affected by fraud, we collect four no-fraud annual reports. For the fraudulent annual reports of large companies, we collect three no-fraud annual reports.

The selection process results in 402 annual reports in the fraud set. The matching process results in 1.325 annual reports which do not contain known fraud. The resulting data set contains annual reports in the period from 1999 to 2011. We refer to Chapter 2 for detailed overviews of the number of annual reports per year, sector and company size.

4.3.2 Selecting the 'Management Discussion and Analysis' section

It is argued that the MD&A section is the most read part of the annual report (Li, 2010). Previous research on 10-K reports showed promising results. Therefore, the MD&A section is a good starting point to determine whether text mining is a suitable means for detecting indications of fraud in annual reports worldwide.

The MD&A section is an unaudited part of the financial statements. Inclusion of this section is mandatory in the US and recommended in the UK

[1]Healthcare and Insurance, Consumer Products, Information Technologies and Services, Natural Resources, Transportation and Leisure, Manufacturing and Construction, Financial Services I, Real Estate and Commodities, Beverages, Apparel and Mining, Electronics and Machinery, Telecommunications, Financial Services II, Office of Structured Finance, Office of Global Security Risk

(Clatworthy and Jones, 2003). The MD&A must cover certain topics but the depth at which these topics are discussed is upon a company's discretion (Brown and Tucker, 2011). In forms 10-K, the MD&A section is included in item 7. Wholly-owned subsidiaries may omit item 7 when meeting certain criteria (Securities and Exchange Commission, 2014c). In forms 20-F, the management discussion is included in item 5 (Securities and Exchange Commission, 2014d). A description of the information that should be included in the management discussions of form 10-K and 20-F are respectively:

Discuss registrant's financial condition, changes in financial condition and results of operations. The discussion shall provide [..] such other information that the registrant believes to be necessary to an understanding of its financial condition, changes in financial condition and results of operations (Reporting, 1934).

The purpose of this standard is to provide management's explanation of factors that have affected the company's financial condition and results of operations for the historical periods covered by the financial statements, and management's assessment of factors and trends which are anticipated to have a material effect on the company's financial condition and results of operations in future periods (Securities and Exchange Commission, 2014d).

The MD&A section is not always explicitly present in the free format annual reports. Some of these types of reports also discuss the financial condition and results of the past year in the letter to the shareholders and do not have a separate management discussion section. Many of the free format reports do include review sections which describe the financial results of the past year. These sections most closely correspond to the MD&A section in the form 10-K.

Hundreds of annual report are subject to analysis for the research described in this paper. The manual extraction of the MD&A sections is a labor-intensive task. Automating this task saves time and allows for future annual reports to be included quickly. The structure of the annual reports on forms 10-K and 20-F provides the opportunity to develop an automated approach. We developed an algorithm that is able to detect the start of the MD&A section based on the first four words of the section header. The algorithm recognizes the end of the MD&A section on the basis of the first four words of the section header that follows the MD&A section. The algorithm takes into account the slight variations that may exist in the section headers and the fact that forms 10-K and 20-F have different headers. Using this approach, we were able to extract

the MD&A section automatically for 96% of the annual reports on forms 10-K and 20-F. We extracted the MD&A sections manually for the annual reports for which automatic extraction was not possible. The MD&A sections form the input for the development of the text mining model described in this paper.

The cosine similarity is one of the most common measures to define document similarity (Manning and Schütze, 1999; Tan et al., 2005). Cosine similarity is a simple measure that does not take word order into account. The cosine similarity score lies between 0 and 1, where a score of 1 means that the compared texts contain the exact same words. The cosine similarity scores obtained by comparing MD&A sections of reports on form 10-K and reports in other formats show that the MD&A sections are similar regardless of the format of the report. Figure 4.1 presents the cosine similarities for the comparison between reports on form 10-K with the reports in other formats. Secondly, Figure 4.1 shows the cosine similarities between MD&A sections that are both from reports on form 10-K and the similarities between MD&A sections that are both from reports other than form 10-K. The distribution of similarity scores is similar for all types of comparisons. The 10-K reports are only marginally similar to other 10-K reports than to reports in other formats.

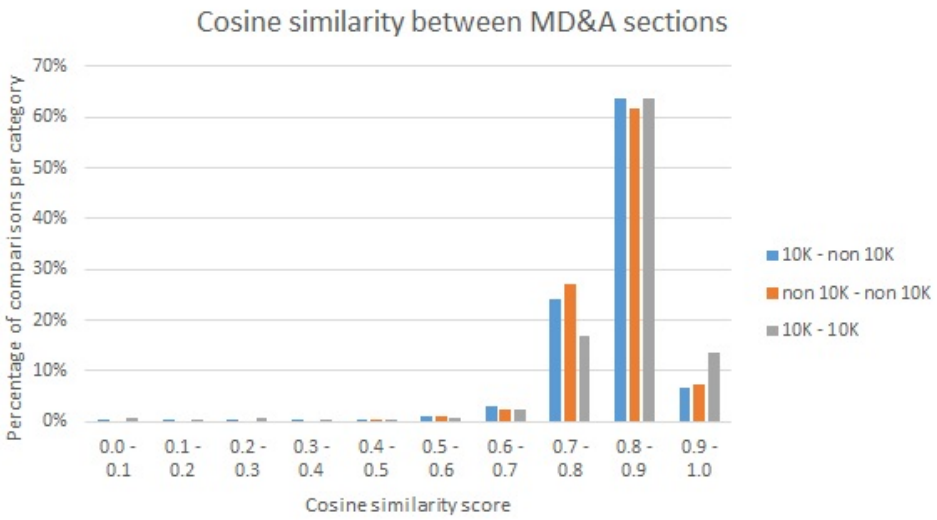


Figure 4.1: Cosine similarity scores indicate the similarity between two MD&A sections from reports on form 10-K and reports in other formats.

4.4 The text mining model

The ability to detect indications of fraud in the MD&A section of annual reports implies that differences between the texts of fraudulent and non fraudulent MD&A sections exist. These differences can be obtained by comparing the texts. However, a simple comparison is therefore not enough as the differences between the MD&A sections of companies are expected to be subtle. With the increase in reporting norms companies adopted a more similar reporting practice to meet the standards (Beattie et al., 2008). Furthermore, companies reuse parts of the MD&A section of previous years in their latest annual report. Forces such as habits, traditions and formal procedures result in financial report texts to be similar year after year, with only minor changes (Aerts, 2001). To determine whether this applies to the data used in this research, the cosine similarity scores for pairs of MD&A sections are calculated for several categories. Figure 4.2 gives an overview of the distribution of the similarity scores per category. The results of the first category, comparing the MD&A sections of reports affected by fraud with the MD&A sections of reports not affected by fraud, show that the differences between the fraudulent and non fraudulent MD&A sections are indeed subtle. The second category compares the MD&A sections of fraudulent reports with the MD&A sections of other fraudulent reports. The third category performs a comparison between MD&A sections of non fraudulent reports. The results of the comparisons of the second and third category are similar to the first, which indicates that the subtle differences found can not all be attributed to the presence or absence of fraudulent activities. The similarity scores show that the differences between fraudulent reports and the differences between non fraudulent reports are comparable to the differences between fraudulent and non fraudulent reports.

Based on the cosine similarity scores, we can not conclude that indications of fraud can be detected in the text of the MD&A section of annual reports. Text mining techniques execute a more complex analysis to extract the characteristics of texts. We therefore developed a text mining model to determine whether a text mining model is able to detect indications of fraud in the MD&A section. The creation of a text mining model involves several steps. The pre-processing step prepares the text for feature extraction and selection. The selected features are the input for the machine learning technique. The selected features and the machine learning technique together form the text mining model. The following three sections describe each of the steps. Section 4.4.1 provides an overview of the data pre-processing steps. This is followed by a description of the feature extraction and selection steps in Section 4.4.2. Sec-

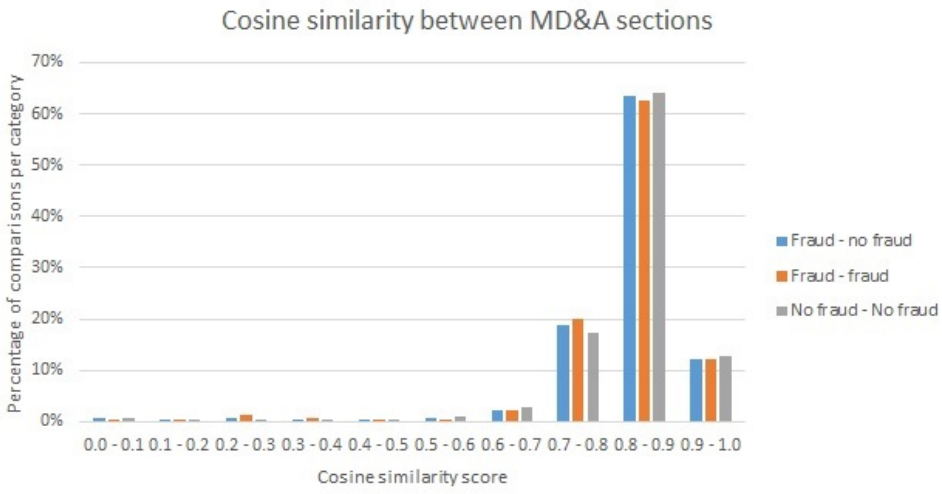


Figure 4.2: Cosine similarity scores indicate the similarity between two MD&A sections. The graph shows per category the percentage of similarity scores for 10 intervals of the similarity scores possible.

tion 4.4.3 explains the machine learning techniques applied in the text mining model.

4.4.1 Data pre-processing

The text mining model must be able to detect whether the annual report is affected by fraudulent activities based on the text. Following the principle 'garbage in, garbage out', the quality of the textual features that are the input for the model is important. Prior to the feature extraction process several steps are performed that further a correct extraction of textual features. These steps take into account the elements that make up an annual report and the elements of texts in general.

Companies use different elements to communicate information in their annual reports, including graphs, figures and tables. Therefore, before performing text analysis we need to define which parts of the MD&A section we consider to be text. Graphs and figures are excluded because their primary way to convey information is not based on text. The main message of information communicated through tables in annual reports is numerical. Tables are therefore also excluded for text analysis. However, numbers are also used within sentences. In this case we will treat the numbers as words.

The features that are the input of text mining models are based on words and sentences. The definition of a sentence and a word seems straightforward for humans. From a computer perspective the recognition of sentences and words is more challenging because punctuation marks are ambiguous. For example, a period does not necessarily indicate the end of a sentence, it may denote an abbreviation or be part of an e-mail address. The processes of identifying sentences and words are referred to as sentence tokenization and word tokenization, respectively (Jurafsky and Martin, 2000; Manning and Schütze, 1999). In this research we used the tokenizers from the Natural Language Toolkit (NLTK) for Python (Bird and Klein, 2009). The tokenizer is able to recognize the punctuation marks that are part of a word such as the periods in abbreviations and the apostrophe in words such as 'couldn't', which is split in 'could' and 'n't'. The remaining punctuation marks that do indicate pauses in sentences or the end of a sentence are excluded because these are not part of the words. The period indicating the end of a sentence is not part of the last word in that sentence.

From the tokenization process, the computer knows how to split the characters of a text so that only separate words remain. However, some of these words need adjustments or be excluded from the feature extraction and selection step. First, a computer defines two words as being different when one of the words starts with a capitalized character, such as most first words in English sentences, and the other word is written only in lower case characters. For example, the words 'Two' and 'two' are different to a computer. To circumvent this all characters are transformed to lowercase characters. Secondly, many of the annual reports in format 10-K or 20-F are saved as html files. These reports therefore contain html tags that are not part of the information in the annual report but which the computer sees as all the other words. The Python package 'BeautifulSoup' is able to recognize and remove these tags from the text. Finally, a specific group of words in all annual reports that a computer can not identify as such is the company name. The occurrence of a company name is specific to the annual report of a company and will not be an indication of fraudulent activities. It is by definition an irrelevant feature for recognizing fraud and will therefore be omitted from the input of the text mining model. Recognizing company names within a text is a specific instance of text mining, referred to as Named Entity Recognition (Jurafsky and Martin, 2000). Named Entity Recognition for company names is beyond the scope of this research. We therefore chose a pragmatic approach. We identified for each annual report the full company name and if appropriate the short name regularly used to refer to the company. For example, 'DT Industries' goes by

'DTI' as well. The company names are replaced by an identifiable placeholder to keep the information that the company name was mentioned.

4.4.2 Feature extraction and selection

Essentially, text is a chain of words. Therefore, it is not surprising that the most regularly used features in text mining are based on the individual words, in this context often referred to as unigrams (Jurafsky and Martin, 2000). Word counts are the most popular type of feature in text mining. The word counts are often normalized for the length of the text they occur in because the word counts of longer texts are higher by definition. In this research 'term frequency-inverse document frequency' (TF-IDF) is applied as a normalization step of the word counts. This normalization step takes into account the length of the text and the commonality of the word in the entire data set (Manning and Schütze, 1999). The word counts are normalized for the length of the MD&A and for the fact that some words occur more frequently in all annual reports. Equations 4.1 to 4.3 provide the mathematical description of the term frequency-inverse document frequency. Note that by using "TF+" the words that have a zero score for IDF, which are the words that occur in all MD&A's, are not ignored.

$$\text{TF-IDF}(word) = \text{TF} + \text{TF} * \text{IDF} \tag{4.1}$$

$$\text{TF}(word) = \frac{\text{Number of times } word \text{ appears in a document}}{\text{Total number of words in the document}} \tag{4.2}$$

$$\text{IDF}(word) = log_e \frac{\text{Total nr of documents}}{\text{Number of documents that contain } word} \tag{4.3}$$

After the data pre-processing steps described in the previous section the computer is able to identify the single words in the texts. As is typical for a set of documents, all annual reports together contain thousands of different words. Having such a high number of features has several drawbacks for the machine learning process. Learning a model from a large matrix is difficult for the machine learning algorithms - a phenomenon that is known as the curse of dimensionality (Tan et al., 2005). In theory, a model is able to classify all

instances correctly if the model has seen examples of all possible combinations of features. In proportion to the number of features, only a limited number of examples to learn from are available. Furthermore, when the dimensionality of the feature matrix increases, the feature matrix becomes sparse. In a sparse matrix many of the feature values are zero. The feature matrix based on words is sparse because every MD&A section does not contain all the words of all MD&A sections in the data set. Many of the words occur infrequently which makes them a less reliable source of information for machine learning models that use the frequency of words to make estimates and construct a model (Manning and Schütze, 1999). Finally, with an increase in features the amount of time needed for the machine learning algorithm to learn the features and construct the model increases.

Feature selection reduces the drawbacks of large feature sets by selecting only a subset of the extracted features. Feature selection in text classification is possible without loss of information because the text is likely to contain words that are redundant or even irrelevant for distinguishing between fraudulent and non fraudulent reports. For example, words may occur at the same frequency in fraudulent reports as in non fraudulent reports. Stop words, the most common words of a language, are likely to occur on a regular basis in all of the MD&A sections regardless of the presence of fraud. Examples of stop words are 'the', 'it' and 'by'. The stop words are excluded using the stop word list of NLTK (Bird and Klein, 2009). Secondly, the number of distinct features is further reduced by stemming the words. Stemming is the removal of the inflectional endings of words, resulting in the stem of the word (Manning and Schütze, 1999; Jurafsky and Martin, 2000). For example the words 'talked', 'talker' and 'talking' all have the stem 'talk'. Stemming reduces the number of distinct words and increases the frequencies while maintaining the meaning of the words. The Porter stemmer is a well known stemmer empirically shown to be effective and is therefore worthwhile to apply in a baseline model (Cecchini et al., 2010; Glancy and Yadav, 2011; Porter, 1980). The implementation of the Porter stemmer in NLTK is applied to obtain only the stems of the words and reduce the number of features (Bird and Klein, 2009). Some of the remaining words can be removed by using common sense. Words that appear only in one MD&A section in the entire data set are not informative. Such a word is not characteristic for either the fraudulent or non fraudulent annual reports. These words will therefore not be used as features. The dimensionality of the feature matrix can further be reduced using the chi squared method. This mathematical approach tests the dependence between two events. For the feature selection process the chi squared formula calculates the dependence

between each feature and the two classes 'fraud' and 'no fraud'. The features with a high dependence with one of the classes are selected. Initially, the top 1.000 most informative features are selected as input for the machine learning algorithm. This top 1.000 is increased with the next most informative features in steps of 1.000 until 24.000 to find the optimal number of features, which is the lowest number of features for which the best result is achieved (Manning and Schütze, 1999).

4.4.3 Machine learning

Machine learning algorithms give a computer the ability to build a model by learning characteristics from data and subsequently using the learned model to make predictions of new unseen data samples. For the construction of a machine learning model, the data is split into a training and test set. The machine learning algorithm uses the training set to learn the characteristics and build the model. The test set is used to assess the performance of the model. Several mathematical implementations exist that differ in how they learn the characteristics and construct the model. The Naïve Bayes classifier (NB) and Support Vector Machine (SVM) have been proven successful in text classification tasks in several domains (Cecchini et al., 2010; Conway et al., 2009; Glancy and Yadav, 2011; Goel et al., 2010; He and Veldkamp, 2012; Joachims, 1998; Manning and Schütze, 1999; Metsis et al., 2006; Purda and Skillicorn, 2015). Therefore, this research uses these two types of machine learning approaches to develop text mining models that can detect indications of fraud in the management discussion and analysis section of annual reports. The results of the models will be compared to determine which of the two approaches is the best baseline model for further development. The remainder of this section explains how the data is split into the training and test set and subsequently describes the rationale of the NB classifier and Support Vector Machine.

The data is randomly split into the development set and validation set. The development data set consists of 70% of the data, the other 30% is saved as a validation set to evaluate the performance of the final model. The goal of the current research is to develop a baseline model that can be extended with additional features. Therefore, the 30% validation set will be saved for testing the extended model. The research described in the current paper uses only the selected 70% development set. Stratification is performed so that both sets contain the same distribution of fraudulent and non fraudulent annual reports as the original data set. In the original data set, 23% of the reports are fraudulent. By performing stratified sampling the percentage of

fraudulent reports in the training and sets is also 23%. For the development of the baseline model we need to split the 70% data set into a training set and test set. Using stratified 10-fold cross-validation, the training set is randomly split in 10 partitions called folds. Each fold in turn is used as a testing set while the remaining 9 folds are used as the training set (Russell and Norvig, 2003). Using this method each annual report is used for testing exactly once. Stratified cross-validation ensures that each fold has the same distribution of fraudulent reports as the original data, i.e., 23%.

The NB classifier applies Bayes' theorem which describes the probability of an event given several conditions related to that event (Manning and Schütze, 1999; Russell and Norvig, 2003; Tan et al., 2005). Formula 4.4 shows the mathematical description of Bayes theorem, where the probability of event A given condition B is calculated. The Bayes' theorem assumes independence between the conditions. However, the method performs well even for the tasks for which the independence assumption does not hold. Applied to the task of text classification, Bayes' theorem calculates the probability that text belongs to a category given the words in the text. Formula 4.5 describes the calculation of the probability that the annual report belongs to the Fraud category given the words in the MD&A section. Formula 4.6 calculates the probability that the annual report belongs to no fraud category. The classification decision is made by comparing the probability that the annual report belongs to fraud category to the probability that the report belongs to the no fraud category. A comparison of the two formulas shows that the denominator is the same, so it is not relevant in the comparison. To make the classification decision only the numerator is relevant, as is expressed in Formula 4.7.

$$\mathrm{P(A|B)} = \frac{\mathrm{P(A)P(B|A)}}{\mathrm{P(B)}} \tag{4.4}$$

$$\mathrm{P(Fraud|words)} = \frac{\mathrm{P(Fraud)P(words\ |Fraud)}}{\mathrm{P(words)}} \tag{4.5}$$

$$\mathrm{P(No\ fraud|words)} = \frac{\mathrm{P(No\ fraud)P(words\ |No\ fraud)}}{\mathrm{P(words)}} \tag{4.6}$$

$$c = \arg\max_{y} \mathrm{P}(y) \prod_{i=1}^{n} \mathrm{P}(w_i|y), \text{ where } y = \{\text{Fraud, No Fraud}\} \tag{4.7}$$

The SVM is a machine learning method that takes a completely different approach than the NB classifier. Instead of calculating probabilities, an SVM maps each instance in the training set as a point in space and finds a hyperplane that separates the instances of the two categories. More precisely, an SVM finds the hyperplane with the maximum margin between the two categories (Joachims, 1998; Tan et al., 2005; Russell and Norvig, 2003). The features define the space. Each feature is a dimension of the space. Figure 4.3 visualizes the maximum hyperplane for a set with two categories and two features. The rationale behind the choice for the hyperplane with the maximum margin is that this hyperplane will generalize better to unseen instances. An unseen instance may be located at a point in space that is closer to the hyperplane than the previously seen instances. The chance that a new instance is located just at the incorrect side of the hyperplane, and therefore is classified incorrectly, is larger for smaller margins. For the text classification task each word is a dimension in the space. A text classification task that includes thousands of words as features results in a feature space with a very high dimension. Each MD&A is a point in that space. The SVM will find the hyperplane that separates the fraud and no fraud MD&A sections in this word space.

4.5 Results

The performance of the NB and SVM text mining models is measured using six performance measures generally used for assessing the performance of classification models: accuracy, F1 measure, sensitivity, specificity, precision and recall (Tan et al., 2005; Manning and Schütze, 1999). These measures are based on the confusion matrix that summarizes the number of instances that are classified correctly and incorrectly per category. Table 4.1 shows the confusion matrix for the classification of 'fraud' and 'no fraud' annual reports. The fraudulent reports that are assigned by the model to the 'fraud' category are the 'true frauds' (TF). The non fraudulent reports assigned to the 'no fraud' category form the 'true no frauds' (NF). The 'false frauds' (FF) are the no fraud reports that the model assigned to the 'fraud' category. The 'false no frauds' (FN) are opposite to the fraudulent reports assigned to the 'no fraud' category.

The six performance measures are calculated from the counts in the confusion matrix. The definitions of each measure are given in formulas 4.8 through

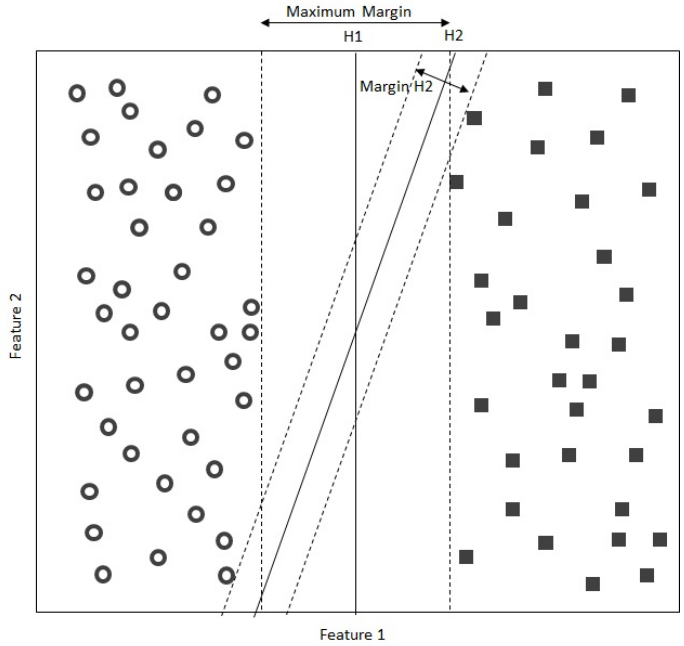


Figure 4.3: The Maxmimum Margin is the Margin for hyperplane H1. The hyperplane H2 separates the data sample but is much smaller than the margin for H1.

4.13 below. Accuracy measures how many of the instances are classified correctly. Sensitivity, also known as the true positive rate (TPR), measures the proportion fraud instances that are correctly classified as fraud. The higher the sensitivity, the better the model is at detecting the presence of fraud. Specificity, or true negative rate (TNR), measures the proportion of the no fraud instances that are correctly classified as such. A high level of specificity means that the model is good at correctly rejecting annual reports for further fraud investigations. Precision, also referred to as positive predictive value (PPV), measures the proportion of the instances assigned to the fraud category that are actually fraud instances. Recall, or negative predictive value (NPV), calculates the proportion of instances assigned to the no fraud category that truly are no fraud. The measures precision and recall provide information on the reliability of the models outcome. A high reliability gives the ability to follow up on the result with a low risk of investing too much or too less effort in further investigations. A high precision means that the model gives few false alarms, which prevents putting time and effort in further investigating

companies where no fraud is present. The higher the recall the more likely it is that no fraud cases are missed when not performing further investigations for annual reports classified as 'not fraudulent' by the model. The F1 measure combines precision and recall into a measure of accuracy. Essentially, the F1 measure is a weighted average of precision and recall.

		True Category	
		Fraud	No Fraud
Assigned Category	Fraud	TF	FF
	No fraud	FN	TN

Table 4.1: Confusion matrix

$$Accuracy = \frac{TF + TN}{TF + TN + FF + FN} \tag{4.8}$$

$$Sensitivity = \frac{TF}{TF + FN} \tag{4.9}$$

$$Specificity = \frac{TN}{TN + FF} \tag{4.10}$$

$$Precision = \frac{TF}{TF + FF} \tag{4.11}$$

$$Recall = \frac{TN}{TN + FN} \tag{4.12}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4.13}$$

4.5.1 Machine learning results

Figure 4.4 shows the results of the six performance measures for the NB and SVM models. For both types of models the optimal number of features is

around 10.000 unigrams. With an accuracy of 89% the NB model outperforms the SVM that achieves an accuracy of 85%. However, the F1 measure is similar for both models (around 86%), which means that for both models precision and recall are reasonably high. The deviation for the precision is around 10% for the 10 folds. The NB model has the highest value for recall (91%) but performs less well in terms of precision (81%). The SVM shows the opposite result, the model performs better in terms of precision (89%) while the value for recall is lower (84%). The sensitivity is substantially higher for the NB model than the SVM model at 68% and 40% respectively. The deviation in sensitivity for the NB is 5% while for the SVM it is 7% for the 10 folds. The difference in specificity between the two types of models is much smaller - 95% for the NB and 98% for the SVM.

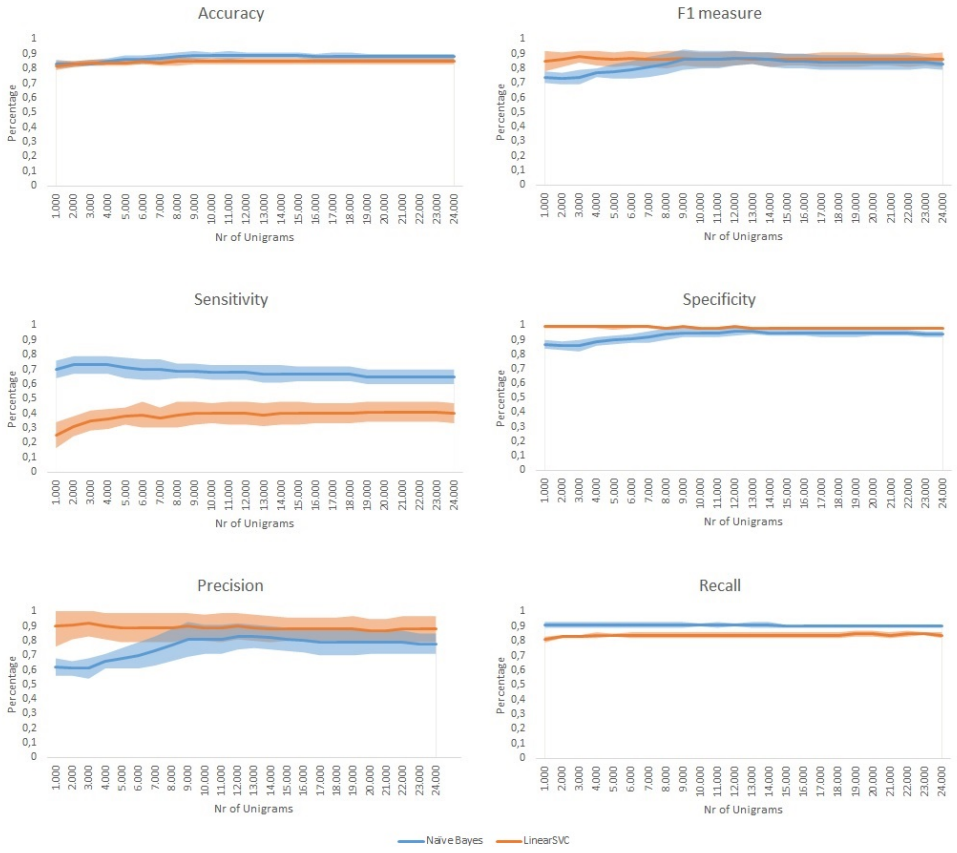


Figure 4.4: Performance of the Naïve Bayes and Support Vector Machine models.

4.6 Discussion and conclusion

The results show that it is possible to use text mining techniques to detect indications of fraud in the management discussion and analysis section of annual reports of companies worldwide. The models developed are suitable baseline models for further research on the use of text mining to detect indications of fraud in annual reports. The results are promising for the models based on both machine learning algorithms, NB and SVM. For both algorithms the optimal result is achieved with 10.000 features. Increasing the number of features does not improve the results.

The NB model is better at detecting fraudulent reports than the SVM (sensitivity is higher for NB). SVM is only slightly better in correctly rejecting non fraudulent reports for further investigations when following up on the result (specificity is higher). When following up on the result of the NB model, there is a higher chance of putting too much effort in investigating companies for fraud while no fraud is present (precision is lower than for SVM) compared to the SVM model. However, with the NB model the chance of missing fraud cases by not further investigating the reports classified as non fraudulent is lower than for the SVM model (recall is higher for NB). The F1 measure is similar for both algorithms. The preferred method depends on the desired result of a cost benefit assessment. With the SVM approach less will be invested in further investigations, however more fraud cases might be missed. The NB model misses less fraud cases, but more investigations would have to be conducted. However, which is a more cost-friendly option warrants further debate. Missing a large fraud case that would affect the annual report may prove to be more costly than investigating additional potential fraud cases that turn out to be false alarms.

The results of the baseline model described in this paper are promising. However, the research has some limitations that may affect the results. In the process of splitting the data into the 10-folds, the sector takes into account the same distribution of fraud and no-fraud cases in each fold. For the entire data set, the fraud and no-fraud cases are matched on the three variables - sector, year and company size. These values are not considered in the process of splitting into folds. As a consequence the matched fraud and no-fraud cases may not occur in the same fold. Secondly, the data set contains, for some companies, the annual reports of multiple years. As a result of the random partitioning during the 10-fold cross-validation the annual report for one company could be in the training set while the report for another year of the same company could be in the test set of the same fold. The MD&A sections of one company of several years are often similar. We therefore hypothesize that

it is easier to correctly classify an MD&A section when the MD&A section of the same company in a different year is included in the training set of the same fold. Thirdly, the automatic extraction of the MD&A section may not be perfect for a limited number of annual reports. As a result, for some of the reports more text than just the MD&A may be included while for others part of the MD&A might be omitted. Finally note that the number of non 10-K annual reports is limited due to the availability of known fraud cases and corresponding annual reports. More non 10-K annual reports are preferred to detect indications of fraud in annual reports of companies worldwide.

Additional research can assess the effects of the random 10-fold splitting process. By comparing the performance measures for the MD&A sections for which for the same company an MD&A section of another year is included in the training set with the performance measures of the MD&A sections to which this does not apply, we determine the effects of the composition of the train and test data within one fold on the overall result. For each of the three variables that were used to match the fraud and non fraudulent annual reports that are ignored by the random 10-fold splitting process we also need to assess whether and how they influence the results of the simple baseline model. Further research may also improve the performance of the baseline model that we described in this paper. The sensitivity for Naïve Bayes, in particular, shows room for improvement. The textual information of annual reports provides several opportunities to explore for expanding the baseline model to increase the performance. Firstly, the text contains more information than just single words. It consists of stylistic and grammatical information. Examples of stylistic information are counts of punctuation marks and the variety of words used in the text. Grammatical information includes part of speech tags and collocations of multiple words. By adding these types of information to the baseline model, the model may be better able to distinguish between MD&A sections from companies engaging in fraudulent activities and the non fraudulent MD&A sections. The baseline model uses the text of the MD&A sections. The annual report contains more textual information that can be included in a text mining model. The additional texts may contain additional indications of fraudulent activities.

5 Linguistic features in a text mining approach to detect indications of fraud in annual reports worldwide

Abstract

Previous research showed that the text in annual reports provides indications of the presence of fraud within companies worldwide. The text mining model developed in that piece of research used word unigrams as input features. However, this is a limited way of looking at texts. Besides the individual words, texts contain grammatical information, vary in complexity and may capture the psychological processes of the writer. The research discussed in this paper includes these types of information to determine whether they add value to the text mining model using word unigrams to detect indications of fraud in the annual reports of companies worldwide. The results show that word unigrams capture the majority of the information. The additional information provided by the linguistic features is very limited.

5.1 Introduction

Year after year, companies have been witnessing increased financial losses owing to fraud. Although the exact costs of fraud cannot be measured, estimates indicate that fraud results in a loss of 5% of expenditure (Gee et al., 2017). Overall 5% of the global Gross Domestic Product for 2016 amounts to 4.39 trillion US dollars. To combat financial fraud, researchers have been developing methods that can detect indications of fraud. The fraud detection research began with a focus on the quantitative information and predefined risk variables. The researchers use various financial ratios developed to measure a company's financial condition (Persons, 1995; Spathis et al., 2002; Kaminski et al., 2004; Kirkos et al., 2007). Examples of risk variables include the number of subsidiaries of a company, whether the company changed its CEO, and wether the company has a poor reputation (Fanning and Cogger, 1998; Bell and Carcello, 2000). In recent years, the focus of fraud detection research shifted from the quantitative information and risk variables to the use of textual information (Cecchini et al., 2010; Glancy and Yadav, 2011; Purda and

Skillicorn, 2015) Various factors contribute to making texts an interesting subject for fraud detection methods. Textual information provides information that is complementary to the quantitative information and reaches a larger audience. Furthermore, the expansion of computer capabilities enhances the possibilities for automated text analysis.

In our previous research on text mining to detect indications of fraud in annual reports worldwide, we developed a baseline model that uses word unigrams as input features (Fissette et al., 2017b). This bag-of-words approach is a commonly used and successful method to process texts. However, taking only the single words of a text into account is a limited way of looking at texts. Although, texts primarily constitute words, texts also contain other types of linguistic information. One such type of information is the grammatical rules that define the order of the words. In addition, texts vary in length and complexity. Further, the texts may represent the psychological processes of the writer. These psychological processes may be captured by the word usage of specific word groups as defined by Pennebaker et al. (2007). A bag of words approach ignores these types of information.

In text analysis research, the implicit assumption is made that the choice of words and sentence structures is an unconscious process. Pennebaker et al. (2003) explain that it is difficult for people to ignore the message of the text and concentrate on the linguistic information. The assumption is questionable for texts in annual reports since texts are written carefully, with adequate time and likely by various people. However, while lying people try to control what they say, they still leak information inadvertently (Newman et al., 2003). Furthermore, we assume that hiding fraudulent activities is such a complex thought process that can not be characterized by a limited number of linguistic features that are easy to manipulate. As it is not clear whether fraudulent and non fraudulent reports can be differentiated based on linguistic features, it is, at this point, unlikely that words and sentences are chosen consciously with the intention to hide fraud.

We expand the unigrams baseline text mining model with various types of linguistic features to determine the added value of such information for detecting fraud in annual reports. We again use the extensive and reality approaching data set that contains the annual reports in all types of formats and sizes of companies worldwide. Just like unigrams, the linguistic features are extracted from the Management Analysis and Discussion (MD&A) section of the annual reports. Using this approach, we will answer the research question:

Can linguistic features add value to a text mining model using words that can detect indications of fraud in the management discussion and analysis section of annual reports of companies worldwide?

The remainder of this research paper is organized as follows. Section 5.2 provides an overview of text analysis research in business documents and in deception theory. Section 5.3 describes a method that includes the data, the extraction of the features and the machine learning algorithms used. Section 5.4 provides an overview of the performance of the text mining model with various linguistic features. Lastly, Section 5.5 discusses the results of using linguistic features and presents suggestions for further research.

5.2 Literature

The research examines texts in financial documents and deception theory experiment with a combination of various types of linguistic features. We identified six categories of linguistic features from these research papers. The subsequent sections each describe a category of features and their role in the detection of fraud or deception. Section 5.2.1 mentions the effect of general features to describe the entire text. Section 5.2.2 describes the features that represent the complexity of the sentences and words used in the text. Section 5.2.3 captures the grammatical features. Section 5.2.4 describes several features that assess the readability of the text. The category comprising psychological process features is explained in Section 5.2.5. Section 5.2.6 explains the n-grams, such as word unigrams we used in our previous research, and how these are used in text analysis research.

5.2.1 Descriptive features

Descriptive features summarize the general properties of the entire text. We have identified three types of descriptive features. The first type is the length of the text that can be expressed by the number of sentences and the number of words. As can be expected, the correlation analysis of Moffitt and Burns (2009) showed that the number of words and sentences are highly correlated. The length of a text may be an indication of fraud or deception since creating a false story takes cognitive effort that may result in shorter texts. On the other hand, deceivers may need more words to create a plausible and convincing story. For example, Zhou et al. (2004a) and Zhou et al. (2004b) showed that deceptive e-mail messages contained a higher number of words and sentences. Similarly, Burgoon et al. (2003) demonstrated, using mock scenes, that

deceivers used longer messages when communicating about their experience. Afroz et al. (2012) used the number of words and sentences as a feature in their text mining model to detect deceptive adversarial documents. However, Tatiana Churyk et al. (2008) conclude that the MD&A sections of fraudulent 10-K reports are shorter than truthful ones. They argue that companies are less likely to elaborate when they are trying to hide fraudulent activities.

The second type of descriptive feature is the type-token ratio, also called lexical diversity. This feature captures the variety of words used in the text. Tatiana Churyk et al. (2009) and Lee et al. (2013) found that the MD&A section of fraudulent 10-K reports contains a higher number of words and fewer unique words than the MD&A sections of non fraudulent 10-K reports. Mbaziira and Jones (2016) showed that deceptive e-mails demonstrate less lexical diversity than truthful e-mails. However, ambiguity exists in the relation between the length and lexical diversity of the text and deception. Larcker and Zakolyukina (2012) determined that in conference calls, the deceptive statements are shorter and have a larger lexical diversity. This difference can be explained by the fact that the writer of an annual report has more time to create the deceptive text than a speaker on a call. Zhou et al. (2004b) found that lexical diversity is an important feature in their machine learning algorithm for detecting deception in computer-mediated communication. The deceptive messages showed less lexical diversity (Zhou et al., 2004a). In the research of Goel et al. (2010), the lexical diversity is one of the predictive features for fraud in 10-K annual reports. Fuller et al. (2011) applied the number of words and sentences and lexical diversity in their machine learning model to detect deceptive utterances in the statements of persons of interest.

The third type of descriptive features describes the number of quantitative references in the text. The research of Clatworthy and Jones (2006) showed that by analyzing the chairman's statements of the UK listed companies, profitable companies were more likely to quantify their performance in the text.

5.2.2 Complexity features

Complexity features represent the elements that constitute the text. A larger number of elements and more complex elements result in a text with higher complexity. Texts consist of sentences and sentences are made up of words. Therefore, the complexity of the texts lies in the words and sentences. For all complexity features, the assumption holds that longer words and sentences are an indication of a more complex text (Zhou et al., 2003).

The complexity features may describe the complexity of a sentence. Examples of such features are the average sentence length and the amount of

punctuation used in a sentence. In the research of Goel et al. (2010) the standard deviation of the sentence length is one of the predictive features for fraud in 10-K reports. The results of Tatiana Churyk et al. (2008) showed that the MD&A sections of fraudulent 10-K reports contain less words per sentence. Furthermore, the fraudulent MD&A sections contain fewer colons and semicolons (Tatiana Churyk et al., 2008; Lee et al., 2014; Tatiana Churyk et al., 2009; Lee et al., 2013). Similarly, in deceptive computer-mediated messages, the number of punctuations was less than that in truthful messages (Zhou et al., 2004a). However, Goel and Gangolly (2012) did not find a significant difference in formatting styles, such as the use of caps or punctuation, between fraudulent and non fraudulent annual reports.

Complexity also depends on the complexity of the words used, such as the average word length and the number of syllables per word. Words made up of three or more syllables are considered complex words. Moffitt and Burns (2009) show that, as expected, the average word length, the average number of syllables per word and the rate of six-letter words are highly correlated.

Most linguistic research that takes into account the complexity features use both sentence complexity features and word complexity features. The text mining model of Afroz et al. (2012) that detects deceptive adversarial documents includes the number of long words, the number of syllables per word, the average number of words per sentence and the number of short and long sentences. In computer-mediated communication, the deceptive messages contained less punctuation, fewer long sentences and fewer syllables per word (Zhou et al., 2003). Therefore, Zhou et al. (2004b) used these features in their machine learning model to detect such messages. The experiments with deceptive messages of Burgoon et al. (2003) showed a similar result. The messages of deceivers were less complex in terms of the number of large words, the number of syllables per word and the average sentence length. Moffitt and Burns (2009) showed with a two-tailed independent sample t-test that the MD&A sections of fraudulent 10-K reports contain more six letter words, have a larger average word length, a higher average number of syllables per word and a higher rate of three-syllable words than the MD&A sections of non fraudulent 10-K reports.

5.2.3 Grammatical features

A word unigrams machine learning model ignores the grammatical information of texts. This grammatical information can be captured by separate features. The grammatical features summarize the types of word groups in the text or the construction of sentences. The word groups, referred to as part-of-speech,

consist of words having similar grammatical elements, such as verbs and nouns.

Pronouns are the most studied word groups in the financial fraud and deception research. Clatworthy and Jones (2006) demonstrated that profitable companies use significantly more personal references in their chairman's statement than the unprofitable ones. Particularly, the occurrence of the word 'our' is higher. Fraudulent MD&A sections in 10-K reports contain fewer self-references than non fraudulent reports (Tatiana Churyk et al., 2008). Similarly, Mbaziira and Jones (2016) showed that deceptive e-mails contain less self-references in order to avoid accountability. In addition, the deceptive messages in conference calls expressed less self-reference. The deceivers used more impersonal pronouns. The use of third-person pronouns was ambiguous (Larcker and Zakolyukina, 2012; Throckmorton et al., 2015). Zhou et al. (2004b) concluded that individual references and group references are important features in machine learning models that detect deception in computer-mediated messages. Such messages contain less self-references and more group references (Zhou et al., 2004a, 2003). However, in computer-mediated messages regarding audit-related questions, the deceivers used more first person singular pronouns (Lee et al., 2009). Afroz et al. (2012) showed that the number of personal pronouns, adverbs and adjectives are some of the most important features to distinguish between truthful and deceptive adversarial statements. Fuller et al. (2011) used the counts of first person singular, first person plural and third person pronouns together with the number of verbs and modifiers in a machine learning model to detect deceptive statements of the persons of interest. Minhas and Hussain (2014) argued that the number of pronouns, adjectives and adverbs contribute to a difference between fraudulent and non fraudulent 10-K reports. The model also selected the number of articles, negations and prepositions.

As these results show, besides the pronouns, researchers take into account various word groups to distinguish between truthful and deceptive texts. Verbs, modal verbs, modifiers and function words are other important word groups in the detection of deceptive computer mediated messages (Zhou et al., 2004b). Zhou et al. (2004a) conclude that deceivers use a higher number of verbs, modal verbs, modifiers and nouns, which is consistent with the result that deceivers write longer messages. However, the deceivers in the mock scene of Burgoon et al. (2003) used less modifiers. The research concerning fraud detection in 10-K annual reports also includes several word groups. Goel and Gangolly (2012) used function words in their machine learning model to detect fraudulent 10-K reports. Furthermore, this model included adverbs as a key feature since a greater use of adverbs indicates the presence of fraud. The ma-

chine learning model of Goel and Uzuner (2016), which includes nouns, verbs, adjectives and adverbs, corroborates this result. Fraudulent 10-K reports contain more adverbs and adjectives. This is contrary to the deceptive messages in the research of Burgoon et al. (2003), which showed less adjectives and adverbs, and the fraudulent scientific reports, which contain less adjectives (Markowitz and Hancock, 2014). Other grammatical features examined in the research conducted to distinguish between fraudulent and non fraudulent 10-K reports, are modal verbs, conjunctions and to-be verbs (Goel et al., 2010). Moffitt and Burns (2009) showed with a two-tailed independent sample t-test that fraudulent 10-K reports contain more conjunctions than non fraudulent reports.

Sentences have a tense that expresses their time orientation. Research revealed that the MD&A sections of fraudulent 10-K annual reports use less present tense than those of non fraudulent reports (Lee et al., 2014; Tatiana Churyk et al., 2009; Lee et al., 2013; Moffitt and Burns, 2009). However, the result seems ambiguous since deceptive computer-mediated messages concerning audit-related questions contained more present tense than truthful messages (Lee et al., 2009). Nonetheless, the model of Minhas and Hussain (2014) selected the variables concerning the past, present and future focus to differentiate between fraudulent and non fraudulent 10-K reports. Li (2008) showed that a higher number of future tense instances in the MD&A section and notes to the financial statements relate to less persistent earnings. Likewise, Clatworthy and Jones (2006) concluded that companies that are unprofitable focus more on the future than on the past in their chairman's statement.

5.2.4 Readability scores

According to the obfuscation theory, the management of a company tries to hide bad news by making the financial reports more difficult to read (Li, 2008; Othman et al., 2012). The assumption is that longer annual reports are less readable (Li, 2008; Miller, 2010). However, text length is a very simplified way of determining the readability of a text. Various readability scores that capture more factors that may determine the reading difficulty of an English text are applied to financial documents and in deception theory.

Both the Flesch-Kincaid Grade Level (F-K), Equation 5.1, and the Flesch Reading Ease (FRE) score, Equation 5.2, take into account the number of words and the number of sentences, but have different weighting factors (Flesch, 1948; Kincaid et al., 1975). The FRE score produces a score between 0 and 100, where documents that have a score of 0 are the most difficult to read

and the documents with a score of 100 are easy to read. The F-K score calculates the number of years of education required to understand the text. A document with a low FRE score has a high Grade Level score. Markowitz and Hancock (2016) concluded, by calculating the FRE scores of scientific papers, that fraudulent scientific papers are less readable. In the mock scene experiment of Burgoon et al. (2003), the deceivers used less complex messages, measured by the F-K. This result corresponds to a result obtained with the complexity features as described in Section 5.2.2. In the pilot study of Othman et al. (2012), the FRE score showed that, on average, the chairman's statement of Malaysian companies is more complex for fraudulent companies.

$$\text{F-K} = 0.39 * \frac{\text{nr of words}}{\text{nr of sentences}} + 11.8 * \frac{\text{nr of syllables}}{\text{nr of words}} - 15.59 \tag{5.1}$$

$$\text{FRE} = 206.835 - 1.015 * \frac{\text{nr of words}}{\text{nr of sentences}} + 84.6 * \frac{\text{nr of syllables}}{\text{nr of words}} \tag{5.2}$$

Similar to the F-K and FRE scores, the Gunning Fog Index (Fog) (Equation 5.3) and SMOG Grading (Smog) (Equation 5.4) scores take into account the number of words and sentences and the number of syllables (Gunning, 1969; Mc Laughlin, 1969). The complex words in these equations refer to words having three or more syllables. These scores indicate the number of years of education required to comprehend the text. The Dale-Chall (Dale) score is a similar measure; however, instead of a syllable count they composed a word list that defines what constitutes a difficult word Dale and Chall (1948). Li (2008) calculated the Fog Index for the MD&A section and notes to the financial statements of 10-K reports. Li (2008) concluded that the reports of companies with lower earnings are more difficult to read and that the earnings in the next one to four years are more persistent for companies with easier to read reports. Miller (2010) found that less readable annual reports are associated with lower levels of trading.

$$\text{Fog} = 0.4 * \frac{\text{nr of words}}{\text{nr of sentences}} + 100 * \frac{\text{nr of complex words}}{\text{nr of words}} \tag{5.3}$$

$$\text{Smog} = 1.043 * \sqrt{30 * \frac{\text{nr of complex words}}{\text{nr of sentences}}} + 3.1291 \tag{5.4}$$

$$\text{Dale} = 0.1579 * \frac{\text{difficult words}}{\text{words}} * 100 + 0.0496 * \frac{\text{nr of words}}{\text{nr sentences}} \tag{5.5}$$

The Linsear Write Readability Formula (LWRF), just like the Fog, Smog and Dale scores, takes into account the number of words, sentences and syllables to estimate the years of education required. To calculate the LWRF score, the following two steps are performed on a sample of 100 words form the text (Shedlosky-Shoemaker et al., 2009).

1. Calculate Equation 5.6, where the easy words consist of two syllables and the complex words of three syllables.
2. If the result is larger than 20, divide by 2; otherwise subtract 2 and divide by 2

$$\frac{\text{nr of easy words} + 3 * \text{nr of complex words}}{\text{nr of sentences}} \tag{5.6}$$

The Automated Readability Index (ARI) and Coleman-Liau Index (CL) scores, Equations 5.7 and 5.8 respectively, also calculate the US grade level required to comprehend the text (Senter and Smith, 1967; Coleman and Liau, 1975). Instead of the number of syllables, these scores take into account the number of characters in a word.

$$\text{ARI} = 4.71 * \frac{\text{nr of characters}}{\text{nr of words}} + 0.5 * \frac{\text{nr of words}}{\text{nr of sentences}} - 21.43 \tag{5.7}$$

$$\begin{aligned} \text{CL} = {} & 0.0588 * \text{avg. nr of characters per 100 words} \\ & -0.296 * \text{avg. nr of sentences per 100 words} - 15.8 \end{aligned} \tag{5.8}$$

Whereas the previously described readability features are developed for the English language, the LIX (Equation 5.9) and RIX (Equation 5.10) scores are applicable to other languages as well. Long words in the LIX and RIX scores are words that consist of more than six characters (Anderson, 1983). The RIX score is a simplified version of the LIX score.

$$\text{LIX} = \frac{\text{nr of words}}{\text{nr of sentences}} + \frac{\text{nr of long words * 100}}{\text{nr of words}} \tag{5.9}$$

$$\text{RIX} = \frac{\text{nr of long words}}{\text{nr of sentences}} \tag{5.10}$$

The correlation analysis performed by Moffitt and Burns (2009) confirmed that the FRE, Fog, Smog, Dale, LWRF, ARI, LIX and RIX scores are consistent with each other. Their two-tailed independent sample t-test showed that the MD&A section of a truthful 10-K report is easier to read than the MD&A section of a fraudulent report wherein readability is measured with the FRE score. However, the other seven readability scores did not show a significant difference between fraudulent and non fraudulent reports. Guay et al. (2015) found that companies have more voluntary disclosures when filing more complex 10-K reports, measured with the Flesch Kincade, Fog, Smog, ARI LIX and RIX scores. The strength of this relation depends on the performance of the company. Butler and Kešelj (2009) concluded that the readability scores Fog, FRE, and F-K, calculated for annual reports perform well as features in a machine learning model that predicts the performance of the next year. Goel et al. (2010) used the F-K, FRE, Fog, Smog, ARI, CL and Lix scores as features in a machine learning model to detect fraud in 10-K annual reports.

5.2.5 Psychological process features

The concept that people express thoughts and intentions through language dates back to the beginning of psychological research (Tausczik and Pennebaker, 2010). Pennebaker et al. (2007) developed a program, the Linguistic Inquiry and Word Count (LIWC), that counts words in 80 psychologically relevant categories to extract the peoples' thoughts and intentions from the

written text. Many researchers applied LIWC to detect fraud or deceptive messages. In this section, we summarize the psychological constructs that are considered in the fraud and deception detection research.

One of the most prevalent thoughts expressed in texts is the sentiment people have about the subject they are writing on. Researchers posed that the emotions expressed could reveal the truthfulness of the written statements. Most research focuses on positive versus negative sentiment. Markowitz and Hancock (2016) found that fraudulent scientific papers contain less positive emotion words. The deceptive messages in conference calls contain less positive and more negative words compared to truthful messages (Larcker and Zakolyukina, 2012; Throckmorton et al., 2015). The same effects are found for 10-K reports. The hypothesis testing procedure of Goel and Gangolly (2012) showed that the likelihood of the presence of fraud increases with a greater use of negative words. Fraudulent reports contain less positive words (Tatiana Churyk et al., 2008, 2009; Lee et al., 2013, 2014). However, Goel and Uzuner (2016) showed that the fraudulent texts contain more positive and negative emotion words. Li (2010) drew a completely different conclusion. He concluded that positive and negative sentiment dictionaries, such as LIWC contains, do not work well for corporate financial statements. Instead, he suggested a machine learning model to predict the tone of sentences in forward looking statements of 10-K reports. Such an approach is beyond the scope of the research discussed in this paper.

In addition to positive and negative word lists, LIWC contains word lists that express sentiments such as anxiety, anger, sadness and affect. Goel and Uzuner (2016) took these categories into account to distinguish between fraudulent and non fraudulent MD&A sections and concluded that the categories anxiety, anger and sadness decrease the performance of their model. On the contrary, Tatiana Churyk et al. (2009) and Lee et al. (2014) found that fraudulent 10-K reports contain more anxiety words. In conference calls, the deceptive messages contain more anger, anxiety and swear words, and less words assent-related words (Larcker and Zakolyukina, 2012). Swear and assent are specific to informal language. Therefore, we do not consider these categories in the analysis of annual reports. Zhou et al. (2004a) conclude that deceptive computer mediated messages contain more affective language. The number of affect words is argued to be an important feature for detecting deception (Zhou et al., 2004b). The category of affect words is also included in the model of Afroz et al. (2012) to detect deceptive adversarial statements and in the model of Minhas and Hussain (2014) to detect fraudulent 10-K reports.

Researchers tried to differentiate between deceptive and truthful texts based

on the amount of certainty expressed in the text. In computer-mediated messages regarding audit related questions and scientific papers, the deceivers use more certainty words (Lee et al., 2009; Markowitz and Hancock, 2014). Zhou et al. (2003) conclude that deceptive computer mediated messages contain more non-immediate and uncertain language, measured by fewer self-references and more modal verbs. Zhou et al. (2004b) expand this research by adding additional features, including certainty words as a measure of non-immediacy. They argue that the results of Zhou et al. (2003) may be different in another task setting. Larcker and Zakolyukina (2012) also find contradictory results for detecting deceptive messages in conference calls. CEOs use fewer certainty words, whereas CFOs use more certainty words. Throckmorton et al. (2015) use the certainty and tentative words to capture a lack of conviction as a feature in their model to detect deception in conference calls. Afroz et al. (2012) measure the level of uncertainty with the number of tentative words and modal verbs in adversarial statements. Tatiana Churyk et al. (2008) found that the MD&A sections of fraudulent 10-K reports have a lower number of certainty words. The hypothesis testing procedure of Goel and Gangolly (2012) revealed that fraudulent 10-K reports contain a higher number of uncertainty words.

Fuller et al. (2011) included both the number of tentative and causation words in their machine learning model for detecting deceptive statements of the persons of interest. Most researchers find that deceptive and fraudulent texts contain more causation words. In deceptive computer mediated messages regarding audit related questions and scientific papers, deceivers use more causation words (Lee et al., 2009; Markowitz and Hancock, 2014). Lee et al. (2014) and Moffitt and Burns (2009) determined that this result also holds for the MD&A sections of fraudulent 10-K reports. However, Tatiana Churyk et al. (2009) contradict this result, stating that fraudulent reports contain fewer causation words (Lee et al., 2013).

Several other psychological constructs are included in the deception and fraud detection research. This includes words related to perceptual and cognitive processes to detect fraud in 10-K reports (Minhas and Hussain, 2014). The LIWC tool distinguishes several categories of cognitive processes, including discrepancies, inhibition, inclusive and exclusive words (Pennebaker et al., 2007). Li (2008) looked at the usage of inclusive and exclusive words in the MD&A section of 10-K reports and found that MD&A sections containing more exclusive words are more difficult to read. Lee et al. (2009) found that deceptive computer-mediated messages regarding audit-related questions contain more insight words. Zhou et al. (2004b) conclude that perceptional information may be a relevant cue for detecting deception. Further, spatiotemporal

information does not contribute to this task (Zhou et al., 2004a,b). Moffitt and Burns (2009), with the help of the two-tailed independent sample t-test, find that the MD&A sections of fraudulent 10-K reports have more words related to achievement.

5.2.6 Word n-grams

N-grams can be word n-grams or character n-grams where 'n' indicates the number of characters or words per feature. In a word unigrams model, each feature corresponds to one word. The occurrence of these words is counted for each document in the data set. The most used feature in text mining models is word unigrams, followed by word bigrams. Character n-grams are used less frequently. In the research of Butler and Kešelj (2009), the word n-gram model outperformed the character n-gram model in predicting the performance of a company for the next year, based on the annual report of the preceding year.

Word n-grams are regularly combined with other linguistic features. Fornaciari and Poesio (2012) combine unigrams with LIWC features in Italian court statements to distinguish between truthful and deceptive statements. Heydari et al. (2015), Karami and Zhou (2015) and Ott et al. (2011) apply a combination of unigrams and the LIWC features to detect spam in reviews. Sun et al. (2016) only use bigrams and trigrams to detect review spam. Balakrishnan et al. (2010) conclude that their word unigrams model captures more information to predict market performance than the fog-index, a readability score explained in Section 5.2.4, and the psychological process features risk sentiment and tone. Dong et al. (2016b) combine word unigrams with psychological process features, such as positive, negative, anger, anxiety and sadness words, in a collection of stock market opinions in social media messages to predict financial fraud. Dong et al. (2016a) combine unigrams with grammatical features, such as modal verbs and personal pronouns, and complexity features, such as sentence length, word length and punctuation, and a readability feature to detect fraud in the MD&A section of companies that report to the Securities and Exchange Commission (SEC) in the US on form 10-K.

Various researchers experimented with the models of word n-grams and linguistic features to detect fraud in 10-K reports. Glancy and Yadav (2011) succesfully distinguished between fraudulent and non fraudulent 10-K reports, based on word unigrams for a limited number of reports. Goel et al. (2010) combined word unigrams with complexity and grammatical features, such as word length, sentence length and proper nouns, to detect fraud in 10-K reports. Purda and Skillicorn (2010) applied a model of word unigrams to 10-K annual reports and quarterly reports on 10-Q. Cecchini et al. (2010) com-

bined word bigrams and trigrams and the dictionaries of word concepts from WordNet to distinguish between 10-K reports of fraudulent and non fraudulent companies. Minhas and Hussain (2014) combined LIWC with keyword features and concluded from their analysis of 10-K reports that the differences in keyword usage between fraudulent and non-fraudulent reports can be quite subtle.

5.3 The method

We have developed various text mining models to determine which combination of linguistic features can add value to the baseline model that uses word unigrams to detect indications of fraud in annual reports worldwide. In the literature, we have found contradictory results for the relation between linguistic features and deception or fraud. We have input each category of linguistic features separately and in combination with unigrams to the machine learning algorithms to ascertain whether the algorithms find a relation between the linguistic features and indications of fraud in MD&A sections. Section 5.3.1 describes the data collection used to develop the models. Section 5.3.2 lists the sets of linguistic features tested during this research and explains the feature extraction process. The features are the input of machine learning algorithms. In this research, we have used the same machine learning algorithms as used for the development of the word unigrams baseline model (Fissette et al., 2017b). By keeping all the factors of the baseline model constant, except for the input features, we have determined the added value of the linguistic features. Section 5.3.3 describes the machine learning algorithms and the way in which the data is used to develop the models.

5.3.1 Data selection

The data set is identical to the data set we used to develop the baseline model, using word unigrams (Fissette et al., 2017b). The data set comprises annual reports from companies worldwide. Rules and regulations for annual reports vary for different countries. As a result, the formats, accounting rules and other substantive requirements for the annual reports also vary. Nonetheless, all annual reports selected for this research are written in English. The selected annual reports present the results of the fiscal years for the period from 1999 to 2011. The total of 1.725 annual reports in the data set belong to one of two categories, 'fraud' and 'no fraud'. The number of annual reports in these categories are 402 and 1.325, respectively.

The fraud cases were selected from news messages and the Accounting and Auditing Enforcement Releases (AAER's) of the Securities and Exchange Commission (SEC) that publish fraud cases for which the evidence was sufficient to prove the presence of fraud. Furthermore, the fraud in the selected cases affected the annual report of the company.

The annual reports in the 'no fraud' category are the reports of companies for which no news message or AAER about a conviction for fraud was found. We should note that this does not exclude the possibility of the presence of fraud. For this research, we assume that since there is no conviction, no fraudulent activities that affect the annual report took place. For each annual report in the 'fraud' category, the 'no fraud' category contains three to five annual reports of companies with the same fiscal year, size and sector as the report in the 'fraud' category. The annual reports of smaller companies have more matched non fraudulent texts than those of the larger companies to reflect the real world situation that has more small companies than large ones. The number of employees, extracted from the annual reports, is used as an indication of the company size. The sectors are the divisions defined by the SEC.

From all the annual reports in the data set, we extracted the MD&A section. Our baseline model using word unigrams for annual reports worldwide developed on the MD&A section achieved good results (Fissette et al., 2017b). For the majority of the annual reports we automated the MD&A extraction task. The MD&A sections of the remaining reports were retrieved manually. For more details on the data selection and extraction of the MD&A section we refer to the Data Selection section of Fissette et al. (2017b).

5.3.2 Feature extraction

To achieve an optimal quality of linguistic features, several data-cleaning steps were performed before extracting the features from the MD&A sections. Focusing on the textual properties of the MD&A section, we excluded graphs, figures and tables from further analysis. Secondly, all words were converted to lower case characters since a computer sees two words as different words when one of the words starts with a capitalized character, such as most first words in English sentences, and the other word is written only in lower case characters. For example, the words 'Two' and 'two' are different to a computer. Thirdly, many of the annual reports in the format 10-K or 20-F are saved as html files. These reports, therefore, contain html tags that are not part of the information in the annual report but are seen in the same way by the computer as all other words. The Python package 'BeautifulSoup' is able

to recognize and remove these tags from the text.

We extracted the linguistic features that are described in Section 5.2 with two tools. The first tool we used is Python. The Natural Language Toolkit (NLTK) for Python contains several functions needed to calculate some of the descriptive, complexity, grammatical and readability features (Bird and Klein, 2009). Many of these features are based on words and sentences. NLTK contains tokenizers to split texts into sentences and sentences into words (Jurafsky and Martin, 2000; Manning and Schütze, 1999). Although the recognition of sentences and words is clear from a human perspective, for computers this task is less straightforward. Tokenizers are able to resolve the lack of clarity resulting from punctuations marks, such as distinguishing periods used in abbreviations from periods that indicate the ending of a sentence and recognizing that 'couldn't' is comprised of the two words 'could' and 'n't'. Besides tokenizers, the NLTK package contains a part-of-speech tagger that assigns a word category, such as 'verb' or 'noun', to each word. We used the Maximum Entropy tagger that is trained on the Treebank corpus for tagging the words in the MD&A sections (Santorini, 1990). The Python Textstat package calculates statistics from texts, including a count of the number of syllables and the calculation of eight readability scores. The input for Textstat is the entire text of the MD&A section. The second tool that we used to extract the features is LIWC, which is used in many pieces of the research described in Section 5.2 (Pennebaker et al., 2007; Tausczik and Pennebaker, 2010). The input of the LIWC tool is the entire MD&A section.

Section 5.2 introduced six categories of linguistic features applied in the linguistic analysis of financial text documents and in deception theory. The subsequent sections provide an overview of the linguistic features applied in this research and the extraction method for each feature. The features that are extracted in Python are listed with the equation used to calculate them. For features that are extracted using LIWC, the name of the feature in LIWC is provided. All features were normalized in Python before being input in the machine learning algorithm as explained in Section 5.3.3. In the last section, we introduce a new feature, not seen in the previous literature, that combines the properties of various linguistic features.

Descriptive features

Table 5.1 lists the six descriptive features. Three of them are about the length of the text and measure it using the number of sentences and the number of words. The fourth feature measures the lexical diversity. The final two features capture the quantitative references in the text.

Descriptive feature	Feature calculation
Number of sentences	Count of the nr of sentences.
Number of words	Count of the nr of words
Logarithm of number of words	The logarithm of the count of nr of words
Type-token ratio (Lexical diversity)	Nr of distinct words / Nr of words
Number of quantitative references	Nr of quantitative references/ Nr of words
Numbers	'number' in LIWC

Table 5.1: Overview of the descriptive features.

Complexity features

We identify eight complexity features, listed in Table 5.2. Three features cover the complexity of the sentences in the text. These are the average sentence length, the standard deviation of the sentence length and the pausality. Five of the eight features describe the complexity of the words. The features average word length, rate of six-letter words and the percentage of long words, where long words consist of more than 15 characters, are concerned with word length. The literature does not provide a definition of a long word. Following Bird and Klein (2009), we consider words consisting of more than 15 characters to be long words. The complexity of words is also expressed using the number of syllables, which is captured by the percentage of syllables and the percentage of complex words, where complex words comprise three or more syllables.

Complexity feature	Feature calculation
Average sentence length	Nr of words/ Nr of sentences
Standard deviation sentence length	Std of avg sentence length
Average word length	Nr of characters/ Nr of words
Pausality	Nr of punctuation marks/Nr of sentences
Percentage of long words	Nr of long words (>15 characters)/ Nr of words
Rate of six-letter words	Nr of words (>= 6 characters)/ Nr of words
Percentage of complex words	Nr of complex words (>= 3 syllables)/ Nr of words
Percentage of syllables	Nr of syllables/ Nr of words

Table 5.2: Overview of the complexity features.

Grammatical features

Most of the grammatical features in Table 5.3 concern the occurrence of word groups. The part-of-speech tags of the words determine to which word group does the word belong. The word groups 'to-be verbs' and the features extracted with LIWC are determined by using dictionaries. The equations for emotiveness and redundancy were taken from the research of Zhou et al. (2004b). Eight of the features focus on the use of pronouns and are all measured by LIWC. The three features regarding tense information are also extracted with LIWC.

Grammatical feature	Feature calculation
Percentage of verbs	Nr of verbs/ Nr of words
Percentage of auxiliary verbs	Nr of auxiliary verbs/ Nr of words
Percentage of to-be verbs	Nr of to-be verbs/ Nr of words
Percentage of modal verbs	Nr of modal verbs/ Nr of words
Percentage of nouns	Nr of nouns/ Nr of words
Percentage of proper nouns	Nr of proper nouns / Nr of words
Percentage of modifiers	(Nr of adjectives + Nr of adverbs)/ Nr of words
Common adverbs	'adverb' in LIWC
Emotiveness	Nr of modifiers / (Nr of nouns + Nr of words)
Percentage of wh-words	Nr of wh-word/ Nr of words
Pronouns	'pronoun' in LIWC
Personal pronouns	'ppron' in LIWC
Impersonal pronouns	'ipron' LIWC
First person singular	'i' in LIWC
First person plural	'we' in LIWC
Second person pronouns	'you' in LIWC
Third person singular	'shehe' in LIWC
Third person plural	'they' in LIWC
Frequency of function words	'funct' in LIWC
Redundancy	Nr of function words/ Nr of sentences
Articles	'article' in LIWC
Past focus	'past' in LIWC
Present focus	'present' in LIWC
Future focus	'future' in LIWC
Prepositions	'preps' in LIWC
Conjunctions	'conj' in LIWC
Negations	'negate' in LIWC
Quantifiers	'quant' in LIWC

Table 5.3: Overview of the grammatical features.

Readability scores

Table 5.4 shows the readability scores. These scores were calculated with the Python package Textstat, except for the LIX and RIX formula's, which were calculated in Python using the equations given in Table 5.4.

Readability feature	Feature calculation
Flesch-Kincaid Grade Level	'flesch_kincaid_grade' in Textstat
Flesch Reading Ease score	'flesch_reading_ease' in Textstat
Automated Readability Index	'automated_readability_index' in Textstat
Coleman-Liau Index	'coleman_liau_index' in Textstat
Gunning Fog Index	'gunning_fog' in Textstat
SMOGGrading	'smog_index' in Textstat
Linseair Write formula	'linsear_write_formula' in Textstat
Dale-Chall readability formula	'dale_chall_readability_score' in Textstat
LIX formula	(Nr of words/ Nr of sentences)/ (Nr of long words*100/ Nr of words)
RIX readability	Nr of long words / Nr of sentences

Table 5.4: Overview of the readability features.

Psychological process features

Table 5.5 lists the psychological process features that are potentially relevant for detecting fraud in annual reports worldwide. All these psychological process features were obtained with LIWC. The features discrepancies, time, money and work were not considered in the previous literature discussed in Section 5.2. However, their relation to other relevant features and the data makes them interesting to consider. The 'discrepancies' includes words such as 'could', 'would' and 'should', which are related to the tentativeness and certainty categories and express a level of conviction. The time category may add information about the use of time on top of the tense features. The money and work dictionaries may be relevant due to the subject of the annual reports.

Word n-grams

The word n-grams we used are unigrams, bigrams and a combination of both. Word unigrams are the counts of the single words, excluding stop words and words that occur only once in the total data set. The word counts were normalized using the 'term frequency-inverse document frequency' (TF-IDF) (Manning and Schütze, 1999). This normalization step takes into account the length of the MD&A and the frequency of occurrence of the word in the total

Psychological process feature	Feature calculation
Affect words	'affect' in LIWC
Positive emotion words	'posemo' LIWC
Negative emotion words	'negemo' LIWC
Anxiety	'anx' in LIWC
Anger	'anger' in LIWC
Sad	'sad' in LIWC
Cognitive processes	'cogmech' in LIWC
Insight	'insight' in LIWC
Cause	'cause' in LIWC
Discrepancies	'discrep' in LIWC
Tentativeness	'tentat' in LIWC
Certainty	'certain' in LIWC
Inhibition	'inhib' in LIWC
Inclusive	'incl' in LIWC
Exclusive	'excl' in LIWC
Perceptual processes	'percept' in LIWC
Time	'time' in LIWC
Work	'work' in LIWC
Achieve	'achieve' in LIWC
Money	'money' in LIWC

Table 5.5: Overview of psychological process features.

data set. More details on the extraction of unigrams can be found in Fissette et al. (2017b). The word bigrams are the counts of two consecutive words. The word bigrams were extracted and counted with the Python package Scikit-Learn. Bigrams that occur only once in the total data set were excluded. TF-IDF was applied as a normalization step for the word counts. The approach for extracting a combination of word unigrams and bigrams features is the same as the extraction of unigrams and bigrams separately. For each n-gram category we used the chi-squared method to select the most informative features. We repeated this procedure by increasing the selected number of features with steps of 5.000 to determine the optimal number of features for which the machine learning algorithms perform well.

Relation features

The features listed in the previous sections each represent one specific linguistic property. Word unigrams, as we explained before, is a limited way of looking at texts as it omits other types of linguistic information, such as grammar. By combining word unigrams with the grammatical features, information about

grammar, on an abstract level, is added to the text mining model. Word bigrams represent the relations between two consecutive words. Therefore, word bigrams capture some notion of grammatical information as well.

Natural language parsers divide sentences in to grammatical parts that constitue the sentence. These parts explicitly contain words and their grammatical information in relation to the sentence they occur in. The Stanford parser is a natural language parser for English. It identifies grammatical relations between all words in a sentence (De Marneffe et al., 2006). As opposed to word bigrams, the words do not have to follow each other. The parser is able to deduce the relation between words that are further apart in a sentence. For example, the sentence 'At this time, the impact of any such effect is not known or estimable.' includes the relations:

- det(time-3, this-2)
- nsubjpass(known-13, impact-6)
- nsubj(estimable-15, impact-6)
- nmod:of(impact-6, effect-10)
- conj:or(known-13, estimable-15)

The grammatical relations are an interesting type of feature as they combine the words and grammatical information at once and capture more meaningful relations between words than bigrams. These relations contain more information than the words, bigrams and grammatical information separately. We extract these relations for all sentences in the MD&A sections of all annual reports in the data set. Subsequently, we eliminate the numbers that indicate the position of the word or punctuation mark in the sentence from these relations. These numbers are specific to the sentence and not informative of the grammatical relations. We treated each relation as a word unigram, which means that we counted all the relations in each text. Similarly to word unigrams and bigrams, these relation counts were normalized using TF-IDF.

5.3.3 Machine learning

The development of a machine learning model requires a data set for building the model and a data set on which the performance of the model can be tested. Therefore, the total data set is split randomly into a development set and a validation set. The development set comprises 70% of the data, while the remaining 30% is set apart to evaluate the performance of the final model, after the development is completed. When splitting the set into the development and validation sets, stratification was performed to maintain the distribution of fraudulent and non fraudulent annual reports of the total set

in both the development and validation set. As a result, the percentage of fraudulent annual reports in all sets is 23%. In this research, we applied various combinations of linguistic features. The experimentation with these features was done on the development set. The 30% validation set was saved to test the models with the features that had the best performance during the development. Therefore, during the development, we needed to assess the performance of the models on the development set. Hence, we split the 70% development set into a training set and test set. By using stratified 10-fold cross-validation, the development set was split into a training and a test set, while maintaining the distribution of fraudulent and non fraudulent reports, 10 times so that each annual report was used for testing once (Russell and Norvig, 2003).

The two machine learning algorithms used to develop the baseline model are the Naïve Bayes classifier (NB) and Support Vector Machine (SVM) (Fissette et al., 2017b). Both the algorithms take a different approach. The NB algorithm calculates the probabilities that an annual report belongs to the two categories 'fraud' and 'no fraud' , based on the occurrence of the features in each of the categories in the training set. The annual report is subsequently assigned to the category that has the highest probability given the features. The SVM algorithm maps each annual report of the training set as a point in space, where the dimensions of the space are defined by the features. The SVM calculates the hyperplane with the maximum margin that separates the points of the two categories in the training set. The category to which the annual reports in the test set are assigned depends on the location of the annual report in the feature space. This location will be at the 'fraud' or 'no fraud' side of the hyperplane. For more details on these machine learning algorithms, we refer to Fissette et al. (2017b); Tan et al. (2005); Joachims (1998); Manning and Schütze (1999).

5.4 Results

Each category of linguistic features formed the input of the machine learning algorithms. The development set was used to experiment with the feature categories. Section 5.4.1 provides the results for the descriptive, complexity, readability and psychological features. Section 5.4.2 describes the performance of the machine learning algorithms using the word bigrams. The results of the machine learning model with the relations as input features are provided in Section 5.4.3. Finally, 30% of the total data set was set apart to test the performance of the combination of features and machine learning algorithm

that performed best on the training set, on the new unseen data. These results are presented in Section 5.4.4.

The performance of the models is measured with six performance measures. The most used measure is accuracy, which indicates the percentage of annual reports assigned to the correct category. The second and third measures are recall and precision. Recall shows the percentage of annual reports assigned to the no fraud category which are truly non fraudulent reports. Precision is the percentage of annual reports assigned to the fraud category which are truly fraudulent. Recall and precision indicate to what extent the outcome of the models can be trusted. The fourth measure, the F1 score, combines recall and precision to generate one value to represent the reliability of the model. The fifth performance measure, sensitivity, calculates the percentage of fraudulent annual reports that are correctly classified by the model as fraudulent reports. On the contrary, the sixth performance measure, specificity, measures the percentage of non fraudulent annual reports classified correctly.

5.4.1 Descriptive, complexity, grammatical and readability features

We applied both NB and SVM to test whether the addition of the descriptive, complexity, grammatical and readability features improves the results of the NB and SVM models using only unigrams as features. The SVM model with these linguistic features performed at chance level. The accuracy of the NB model with the linguistic features was even below the chance level. The best result achieved with the NB model with 10.000 word unigrams was 89%, which is in line with the results obtained in our previous research (Fissette et al., 2017b). A combination of word unigrams with the linguistic feature categories does not affect the accuracy. The accuracy results of the NB and SVM model with the various linguistic feature categories are shown in Figure 5.1. The green line represents the chance level of 77%.

The NB models with the linguistic features score low on sensitivity and precision, which means that the model has difficulties detecting the fraudulent annual reports. This result corresponds to the values for accuracy that are close to the chance level. A model achieves the chance level by assigning the majority of the annual reports to the 'no fraud' category, but makes mistakes in classifying the fraudulent reports, resulting in low values for sensitivity and precision. The addition of the linguistic features to the SVM does not improve the results obtained with 10.000 unigrams. Only two additional annual reports are classified correctly.

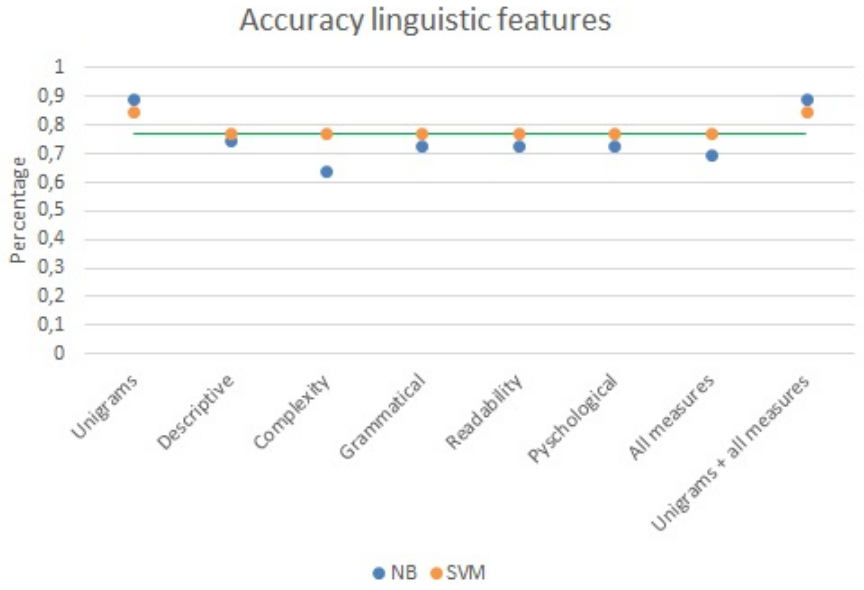


Figure 5.1: Accuracy and variance in accuracy of the Naive Bayes model with unigrams, descriptive, complexity, grammatical and readability features.

5.4.2 Word bigrams

We created three types of feature sets that contain bigrams. The first type consists of bigrams only. For the second one, we combined unigrams and bigrams and subsequently determined the most informative features from this set using the chi-squared method. For the third one, we combined the 10.000 most informative unigrams with the top 10.000 most informative bigrams separately. Each of the feature sets is used as input in both the NB and SVM algorithms. The results are summarized in Figure 5.2 and Figure 5.3. Figure 5.2 shows the results of the six performance measures for the top 5.000 to the top 30.000 most informative features of the first and second sets. Figure 5.3 shows the results of the third type of feature set. The green line in this figure indicates the 77% chance level.

The accuracy of the NB models with word bigrams only as input features is far below the chance level. To keep Figure 5.2 readable, this result is omitted from the graphs. As Figure 5.3 shows, the SVM achieved an accuracy of 83% when using 5.000 bigrams and 86% when using 30.000 bigrams. The former

is lower, while the latter is slightly higher than the accuracy of the SVM model using 10.000 unigrams, which achieved an accuracy of 85%. The values for the measures precision, specificity and the F1 score are higher than those achieved with SVM model using unigrams. However, the sensitivity and recall are slightly lower for SVM model with bigrams only.

The accuracy of the NB model with unigrams and bigrams combined is also below the chance level when the top features are selected by the chi squared method. The model achieved the chance level for the top 30.000 most informative features. This result is omitted from Figure 5.2. When using the combination of the most informative unigrams and bigrams in the SVM model, the accuracy increased to 89%, which is an increase of 4% compared to the SVM model only using the top 10.000 unigrams. Similar to bigrams only, the measures precision, specificity and the F1 score increased compared to the SVM model using unigrams only. The sensitivity and recall are also slightly higher.

As can be seen in Figure 5.3, combining the top 10.000 unigrams with the top 10.000 bigrams separately in the NB model resulted in an accuracy of 88%, which is slightly lower than the accuracy of the NB model with only the top 10.000 most informative unigrams. Figure 5.3 shows that in the SVM model, the combination of the top 10.000 unigrams and top 10.000 bigrams separately achieved an accuracy of 89%, which is equal to the result achieved with combining unigrams with bigrams and subsequently selecting the most informative features.

5.4.3 Relation features

Similar to bigrams, we created several types of feature sets that include the relation features and used these sets in both the NB and SVM algorithms. The first type of feature set consists of relations only. For the second type, we combined the relations with unigrams. We used the same approach as for bigrams to make this combination. The unigrams and relations are combined and the most informative features from this list are subsequently selected using the chi squared method. We also combined the top 10.000 most informative unigrams with the top 10.000 most informative relations, separately selected with the chi squared method. Finally, we combined unigrams, bigrams and relation features. We again make this combination using the two approaches. First, by combining all features and subsequently selecting the most informative ones. Secondly, by combining the top 10.000 most informative unigrams with the top 10.000 bigrams and top 10.000 relations separately. These results are included in Figure 5.2 and Figure 5.3.

The NB model with the relations as input achieves up to the top 28.000 most informative features, an accuracy equal to the chance level. With 29.000 features, the accuracy slightly increases to 79%. These results are omitted from Figure 5.2. The accuracy of the SVM model is stable around 81%, but is lower than the 85% accuracy achieved with unigrams. The values for sensitivity and recall are slightly lower, while specificity and precision are slightly higher. The F1 score is approximately the same. The SVM results are included in Figure 5.2.

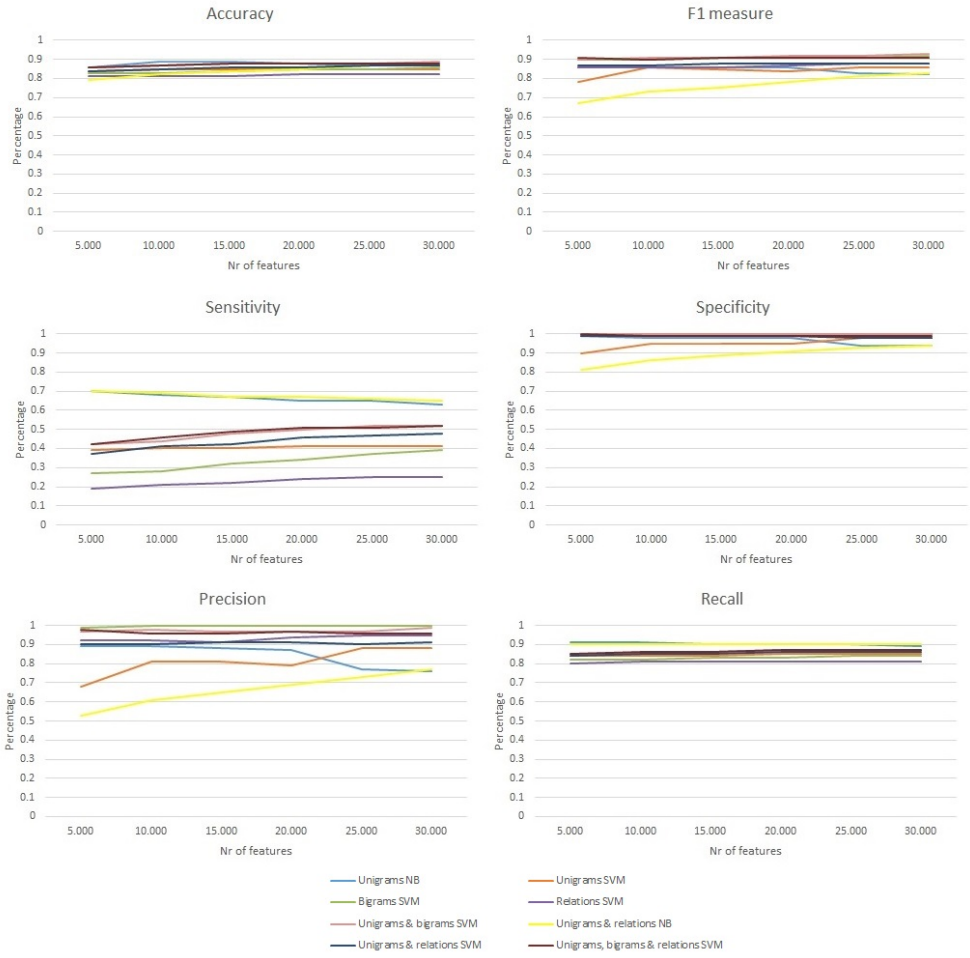


Figure 5.2: Performance of the Naive Bayes model with unigrams and Support Vector Machine models with unigrams, bigrams and relations as features.

The NB model with a combined input of unigrams and relations results in an accuracy that is slightly better than the accuracy for unigrams only when the input contains at least 25.000 features. The values for sensitivity and recall are a higher, but the specificity, precision and F1 measure are lower. The performance of the SVM that combines unigrams and relations increases for 15.000 features or more up to 87%. The sensitivity, recall and precision increase as well. Until 20.000 features, specificity is higher, but is stable for 20.000 features or more. Moreover, the F1 score is a bit higher. Figure 5.2 shows these results.

If we select the top x features as a combination of unigrams, bigrams and relations the performance of the NB is below the chance level. The accuracy of the SVM model with a combination of unigrams, bigrams and relations is higher than unigrams only, but equal to bigrams only. The same holds for the other five performance measures. The results for the SVM model with this combination are presented in Figure 5.2. When we take the top 10.000 unigrams and combine these with the top 10.000 bigrams and top 10.000 relations separately, the NB model achieves an accuracy of 88%, which is still slightly lower than the accuracy achieved with unigrams only. The SVM model achieves an accuracy of 89%. However, the SVM achieved the same performance with the combination of unigrams and bigrams. These result are summarized in Figure 5.3.

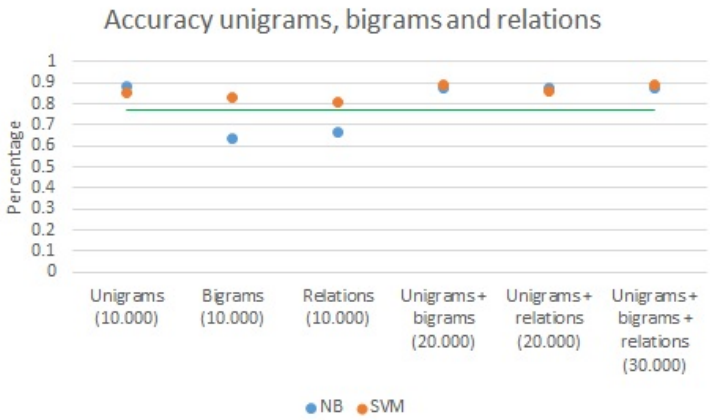


Figure 5.3: Accuracy of the Naïve Bayes and Support Vector Machine models with combinations of unigrams, bigrams and relations as features.

5.4.4 Result on test set

The models that achieved the highest level of accuracy on the development set were the NB with the top 10.000 unigrams, the SVM model with the top 10.000 unigrams combined separately with the top 10.000 bigrams and the SVM model with the top 30.000 unigrams and bigrams. These three models achieved an accuracy of 89% on the development set. Of the models, the latter model has the highest specificity, precision and F1 score, but the lowest sensitivity and recall. The NB unigrams model achieved the highest sensitivity and recall, but the lowest specificity and F1 score. From all the models, the SVM with the top 30.000 unigrams and bigrams achieved the highest F1 score.

Table 5.6 shows the performance on the test set of the three models that performed best on the development set. The NB model has the highest sensitivity and recall, but has the lowest precision and F1 score. The SVM model with the top 10.000 unigrams and top 10.000 bigrams, a total of 20.000 features, achieved the highest accuracy and F1 score. The SVM model with the top 30.000 of the combined unigrams and bigrams scored the highest on specificity and precision, but lowest on accuracy, sensitivity and recall.

Model	Accuracy	Sensitivity	Specificity	Precision	Recall	F1
NB 10.000	0.89	0.72	0.95	0.81	0.92	0.86
SVM 20.000	0.90	0.6	0.99	0.95	0.89	0.92
SVM 30.000	0.87	0.45	1.00	0.98	0.86	0.91

Table 5.6: Overview of the test results for the three best performing models for the development set.

5.5 Discussion and conclusion

The results of the test set show that, in general, the SVM model with the top 10.000 unigrams combined with the top 10.000 bigrams as input features has the best performance. The accuracy and F1 scores take into account the correctness of the model's assignment of the annual reports to the fraud and 'no fraud' categories. However, the NB model scores the best on sensitivity and recall. A high sensitivity means that the model is good at detecting fraudulent reports. A high recall indicates that, if the model assigns an annual report to the 'no fraud' category, this result is reliable. So if we choose not to further investigate the annual reports assigned to the 'no fraud' category by the NB model, we would not miss too many fraud cases. As fraud investigations are

costly, this may be cost-effective. However, the precision of the NB model is a bit lower, the model assigns to the fraud category annual reports that are not actually fraud. When the choice is made to further investigate annual reports assigned to the fraud category, non fraudulent cases are subject to further investigation as well, which incurs additional costs. This may not render the NB model as the most cost-effective one; however, it would miss the least number of fraud cases, and therefore can be considered the safest model to rely on. The SVM model with 30.000 unigrams and bigrams has the lowest sensitivity, so it detects the least number of fraud cases. However, if this model classifies an annual report as being fraudulent, the model is most likely to be accurate owing to its high level of precision.

Using only bigrams in an NB model does not perform well. As may be expected from this result, adding bigrams to the NB unigrams model does not improve the results. Bigrams are informative features in the SVM model. The addition of bigrams to the SVM unigrams model increases the results up to an accuracy that is slightly higher than the NB unigrams model. Bigrams add a limited amount of additional information to unigrams. This may be explained by the fact that unigrams are part of bigrams. The same applies to the relation features in which the words are part of the relation. As a result, an overlap between unigrams, bigrams and relations exists. In the selection of the most informative features, the words may make a bigram or relation relevant, while the additional information that bigram or relation captures may not be useful for distinguishing between fraudulent and non fraudulent annual reports, or may not add to the information provided by unigrams. The result of a model with the combination of 10.000 unigrams and bigrams is lower than the result of the unigrams model that has the same number of features. An explanation for this outcome is that, during the selection of the most informative features, bigrams are selected instead of some unigrams, and these bigrams do not provide additional information while the unigrams that are excluded do. The results of the model that combine the most informative unigrams and bigrams separately, as shown in Figure 5.3 confirm this explanation. The performance is higher when the most informative features per feature category are combined separately than when the most informative features are selected from the feature categories at once. Adding the relations to unigrams and bigrams does not improve the results.

None of the five linguistic feature categories — descriptive, complexity, grammatical, readability and psychological processes — add value to the text mining model that uses words to detect indications of fraud in the MD&A section of annual reports of companies worldwide. Several possible explanations

exist for the lack of contribution of the linguistic features to the text mining model. The research discussed in Section 5.2 states that the linguistic features are relevant for detecting fraud or deception. Many of these researchers tested the usefulness of the various features statistically. However, the features that have statistical power may not be relevant in machine learning algorithms, while variables without statistical power may be relevant (Zhou et al., 2004b). Furthermore, the previous research showed ambiguous directions for the relation between deception or fraud and the linguistic features. Researchers found that an increased use of specific features indicate fraud or deception, while other researchers concluded the opposite. This ambiguity may exist because the relation between the feature and fraud may differ per fraud case or due to an unstable relation between the linguistic features and fraud or deception. Another explanation is that, since annual reports are a formal document written by multiple authors, the documents contain a variety of linguistic styles of which none is prevalent, eliminating the presence of linguistic cues that indicate fraud. Moreover, the LIWC dictionaries developed for other domains, may not be appropriate for formal documents, such as annual reports. The number of features that a machine learning model needs to perform a subtle task of detecting indications of fraud in texts is high. The word unigrams model uses 10.000 features to achieve the most optimal result. The linguistic feature categories comprise 72 features. Compared to unigrams, this is a very limited number of features. These features may not capture sufficient information to distinguish between fraudulent and non fraudulent annual reports. The addition of the linguistic features to unigrams does not affect the result; the performance remains the same as achieved with unigrams only. This indicates that unigrams already contain a lot of information. Apparently, the count of words is more informative than the counts of specific word groups based on a dictionary of the grammatical properties of the words. Some of the words within a group may be relevant to distinguish between fraudulent and non fraudulent annual reports while other words in the same group are not. Unigrams are able to make this distinction. In essence, word unigrams are more fine-grained than the higher level linguistic features, making them more suitable to capture the subtle differences between fraudulent and non fraudulent reports.

More research is needed to substantiate the explanations of the results of this research. Firstly, the relation between the linguistic features and fraud in a machine learning approach needs validation to determine if an unambiguous relation exists. The relationship features need to be explored in more detail. This is a new type of feature that may need fine tuning or may be relevant in

another domain. Subsequent research may also examine the effects of multiple authors on the linguistic features of the text. Such research may determine whether the linguistic differences between documents decrease when they are written by multiple people. It may also be worthwhile to explore the effects of an ensemble of machine learning algorithms. The various algorithms develop different models and use the features differently. For example, bigrams are suitable in an SVM model, while not informative in the NB model. A combination of algorithms and features may reinforce each other's results. Finally, the automated data extraction process and the Python package 'BeautifulSoup' that excludes html tags used to obtain the MD&A sections from the annual reports in this research are not flawless. This may have affected the quality of the extracted linguistic features. In particular, the parsers may produce less accurate results and therefore affect the grammatical features. Research may confirm whether the extraction and data cleaning process had an influence on the linguistic feature extraction.

6 Deep learning to detect indications of fraud in the texts of annual reports worldwide

Abstract

We have applied a Naive Bayes (NB) and a Convolutional Neural Network (CNN) model to the Management Discussion and Analysis sections from the annual reports of companies worldwide to determine whether a deep learning model can detect indications of fraud in texts. The texts are represented by word embeddings obtained with a pre-trained Word2Vec. A simple CNN model with only one convolutional layer and little parameter tuning detects more fraud cases than the NB model, 23% opposed to 10%. However, the CNN model makes more mistakes in the identification of non fraudulent reports. Notwithstanding, the results of the CNN model are slightly more reliable those of the NB model.

6.1 Introduction

Fraudsters are innovative and find new ways to engage in fraudulent activities. New technologies are more often used by fraudsters to commit fraud than by fraud investigators to detect the fraudulent activities (KPMG, 2016). Companies and fraud investigators need to exploit the latest technologies to combat fraud and keep up with the innovativeness of fraudsters.

In the past decades, researchers experimented with the state-of-the-art tools to identify clever and automated methods to detect indications of fraud. The application of statistical tests reveals the financial ratios that may be indicative of the presence of financial fraud in a company (Persons, 1995; Spathis et al., 2002). In the subsequent years, the financial ratios formed the input for machine learning techniques, such as Decision Trees, Neural Networks and Bayesian Belief Networks, K-nearest neighbors and Support Vector Machines (SVM) (Kotsiantis et al., 2006; Kirkos et al., 2007). In particular, neural networks were a popular machine learning method to detect fraud based on financial information (Green and Choi, 1997; Fanning and Cogger, 1998; Bhattacharya et al., 2011; Huang et al., 2012).

Besides financial information, researchers experimented with the textual information in financial documents, mostly annual reports, to detect indications of fraud. The machine learning methods exploited to this end are SVM, NB and Clustering (Cecchini et al., 2010; Goel et al., 2010; Glancy and Yadav, 2011; Purda and Skillicorn, 2015; Fissette et al., 2017a,b). In these models, the text is represented as a bag-of-words. The frequency of occurrence of the words forms the input of the machine learning algorithms. This approach disregards the order of the words and other grammatical information. The results of the previous research show that the texts in financial documents seem to reveal some indications of fraud.

The models in the text mining research that experiment with fraud detection applied various machine learning algorithms. The newest and popular state-of-the-art machine learning technique is deep learning. Deep learning models use neural networks containing more than one hidden layer, opposed to the neural networks popular for experimentation with financial information to detect fraud (Green and Choi, 1997; Fanning and Cogger, 1998; Bhattacharya et al., 2011; Huang et al., 2012). The deep learning models are successfully used for various types of text analysis research tasks. Instead of the bag-of-words approach, the words are represented as vectors. The location of words vectors in the vector space is such that words having any semantic relation are located close to each other in the vector space.

Deep learning techniques are, to our knowledge, not yet applied to analyze financial texts. To be able to combat fraud with the latest state-of-the-art techniques, we wish to apply deep learning to detect indications of fraud in annual reports worldwide. Therefore, we formulate the research question:

Can a deep learning model be developed that can detect indications of fraud in the management discussion and analysis section of annual reports of companies worldwide?

To answer this research question, we apply a simple deep learning model for a limited number of annual reports of companies worldwide. Furthermore, we compare the results of the deep learning model with the results of a bag-of-words NB model that detects indications of fraud in the management discussion and analysis (MD&A) section to determine whether the state-of-the art deep learning model has the potential to outperform the methods that have been shown to be promising by various previous research studies.

This research paper is organized as follows. Section 6.2 provides an overview of deep learning text classification research. Section 6.3 describes a method that includes the data and the deep learning approach applied in this research.

Section 6.4 presents the results of the deep learning text mining approach and the bag-of-words NB model. Finally, Section 6.5 discusses the results of the deep learning approach for detecting indications of fraud in annual reports, compares the results to the NB model and provides suggestions for further research.

6.2 Deep learning models for text classification

In recent years, the state-of-the-art deep learning models are a popular subject of research. Many researchers applied various deep learning models to a range of text classification tasks. The deep learning models vary in the type of architecture used. Furthermore, the input of the deep learning models for text analysis varies from character level to word level embeddings. The majority of the word embeddings are learned using the Word2Vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014) model. In this section, we briefly describe the types of deep learning architectures and subsequently outline the deep learning research conducted in various text classification domains. The majority of deep learning models for text classification are tested on data sets developed for sentiment classification. Section 6.2.1 describes these models. Many of these models are also applied to classify question types and detect topics in short texts. Section 6.2.2 provides a brief overview of these results. Further, Section 6.2.3 outlines various text classification tasks performed by a limited number of researchers. Finally, Section 6.2.4 concludes with our observations in the literature about deep learning for text classification.

The two key architecture types are Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) (Yin et al., 2017). CNNs are hierarchical networks that consist of convolutional and pooling layers. A convolutional layer reduces the dimensionality of the data to aid generalization. A convolution can be interpreted as the automatic extraction of features from the input data. In a CNN for text analysis the features are extracted by applying a filter to a window of n words. The 'n' is referred to as the filter size. A convolutional layer may consist of multiple filters of various sizes. The pooling layer reduces the dimensionality by retaining the most important information. For example, a max pooling layer outputs only the largest values produced by a convolutional layer. Dropout regularization, which randomly disables hidden units in the network during training, may be applied to prevent the model to overfit the training data. RNNs model the data sequentially, by remembering information from the past, collected in a previous learning step. Two types of RNNs can be identified: long short-term memory (LSTM) and gated recurrent

unit (GRU). An LSTM remembers the past information by storing relevant information in a separate cell of the network. This memory cell has three gates that identify whether information is stored in the memory, forgotten or is outputted (Chung et al., 2014). A GRU does not have a separate memory cell; however, it does have gates that control the flow of past information.

6.2.1 Sentiment classification

Sentiment classification is the text analysis task for which several deep learning models have been developed. These models aim to detect the sentiment in short texts, varying from sentences to reviews. The models categorize the texts into two or five sentiment categories. The two categories distinguish positive and negative sentiment. The five sentiment categories are 'very positive', 'positive', 'neutral', 'negative' and 'very negative'. Related problems are classifying texts as objective or subjective and stance classification. In the latter classification task, texts are classified into three categories 'supportive', 'neutral' and 'unsupportive', or into two categories 'for' and 'against' (Chen and Ku, 2016) .

Kim (2014) developed a CNN with one convolutional layer, a max-over-time-pooling layer and a fully connected layer with a dropout of 0.5 and a softmax output to produce the classification. The input are Word2Vec word embeddings that are pre-trained on Google news. The convolutional layer has three filter sizes: 3, 4 and 5. The CNN was applied to detect the sentiment of movie reviews, product reviews and news messages and to detect the subjectivity or objectivity in sentences from movie reviews and movie plots. The size of the data sets varied from 3.775 to 11.855 sentences. With the Stanford Sentiment Treebank (SST) movie review data set, Kim (2014) achieved an accuracy of 48.0% for the five sentiment categories and an accuracy of 88.1% for the two sentiment categories. The accuracies achieved for the product review, and news messages were 85.0% and 89.5%, respectively. The accuracy obtained for the subjectivity data set was 93.4%. Kalchbrenner et al. (2014) achieved comparable results on the SST movie review data set with a CNN that has two convolutional layers, and wide filters of sizes 5, 7 and 10. The performance of their model was higher for sentiment analysis on 1.6 million twitter messages. Johnson and Zhang (2014) demonstrated with a similar CNN, but with self learned word representations, that CNNs outperform SVMs for sentiment detection in the IMDB movie reviews and Amazon product reviews. Ma et al. (2015) extended the model of Kim (2014) by including information on a word's parents, grandparents, great-grandparents and siblings from a dependency parse tree. Their results for the SST movie reviews are comparable

to those of Kim (2014) and Kalchbrenner et al. (2014). Chen et al. (2015) incorporated information about the user, the topic, the content and comments of social media messages in the CNN of Kim (2014) to determine the stance of Chinese Facebook and English debate forum messages.

The previously discussed models use word embeddings to represent the text. Dos Santos and Gatti (2014) also detected sentiment in the SST movie reviews, but used an input of character and word embeddings. For the five sentiment categories, their model performed slightly better than of Kim (2014); however the model of Kim (2014) performs better on the two-category task. Excluding the character level embeddings did not affect the performance of the CNN. However, for detecting sentiment in 80k twitter messages, the inclusion of the character level embeddings did improve the performance. Zhang and LeCun (2015) and Zhang et al. (2015) tested a character level CNN model with six convolutional layers on four data sets for sentiment analysis. For one of the data sets, a bag-of-words model outperformed the character CNN, while for the other three data sets, the CNN showed better performance. Conneau et al. (2016) showed that the performance is further improved with a CNN containing 29 convolutional layers. This CNN also performs better than the bag-of-word model for the data set for which the bag-of-words model outperformed the smaller CNN.

Some researchers combined the RNN and CNN architectures to detect sentiment. Denil et al. (2014) applied a variant of the CNN model of Kalchbrenner et al. (2014), which has fewer model parameters but still achieves similar results on the twitter data set. However, this CNN model has been outperformed on the IMDB movie review data set by the models of Zhang et al. (2016b) and Wen et al. (2016), both of which combine RNN and CNN. Zhang et al. (2016b) process pre-trained word embeddings with layers of LSTM networks. This model outperforms the model of Kim (2014) on the SST movie reviews and the subjectivity data set. The deep learning network of Wen et al. (2016) starts with a GRU model to extract context information. However, the combined model of Lai et al. (2015) shows a slightly lower performance on the SST movie review data set. This model is a CNN in which the convolutional layer has a recurrent structure instead of a fixed filter size. The words are represented by Word2Vec word embeddings trained on Wikipedia. Tang et al. (2015) used another combined approach. They first processed the texts with a CNN with filter sizes 1, 2 and 3 or an LSTM model to obtain the sentence representations. Subsequently, an RNN produced a document representation. The approach was tested on Yelp reviews, with an average of less than ten sentences per review and IMDB movie reviews, with an average of 14 sentences

per review. The approach of Tang et al. (2015) outperforms SVMs. CNNs are the most used deep learning architecture for sentiment analysis. However, Yin et al. (2017) compared CNN and RNN models and found that a GRU RNN model achieved the best results for the SST movie review data.

6.2.2 Topic and question classification

Most of the deep learning models discussed in the previous section about sentiment classification are also applied to classify questions into question types and to detect topics from English or Chinese news messages. In this section, we briefly compare the results obtained in these domains.

The previously discussed models of Kim (2014), Kalchbrenner et al. (2014), Ma et al. (2015), Wang et al. (2015) and Zhang et al. (2016b) were applied to the TREC questions data set that includes six question types: abbreviation, entity, description, human, location and numeric. The model of Kalchbrenner et al. (2014) achieved an accuracy of 93%. The result of Kim (2014) was slightly higher, reporting an accuracy of 93.6%. Both Ma et al. (2015) and Zhang et al. (2016b) reported an accuracy of 95.6%. Wang et al. (2015) achieved the highest accuracy, 97.2%, with GloVe word embeddings.

In the topic detection research, the number of topics to be detected ranges between 4 and 55. The size of the English news messages used in the development of the deep learning models is limited to a maximum of 1.014 characters. The messages of the research in Chinese had an average length of 2.981 characters.

Johnson and Zhang (2014) demonstrated that their CNN also outperformed the SVM for the topic classification task for Reuters news articles. Lai et al. (2015) applied their combined CNN-RNN model to the English 20 Newsgroups data and the Chinese Fudan data sets. The model outperformed the CNN model with a convolutional layer with a fixed filter size and the established machine learning models Logistic Regression and SVM. The character level CNN model of Zhang and LeCun (2015) with six convolutional layers was outperformed by a bag-of-words model for three out of four data sets for topic classification. The 29 layer model of Conneau et al. (2016) improves the results; however, only for two out of the four data sets, the CNN performs better than the bag-of-words model.

Wang et al. (2015) developed a CNN with one convolutional layer and a layer that extracts several representations of the texts by applying multiple windows with various width over the pre-trained word embeddings. They experimented with three types of word embeddings, Senna trained on 130k Wikipedia articles, GloVe trained on 400k Wikipedia articles and Word2Vec

trained on 3 million Google News messages. The former two methods result in word embeddings in which each word is represented by a vector with a dimension of 50; the latter results in word embeddings with dimension 300. The best results on the Google Snippets data were achieved with Word2Vec word embeddings.

6.2.3 Various text classification tasks

Besides the regularly implemented text classification tasks discussed in the previous sections, several researchers have applied deep learning models to classify texts in various other domains. In this section, we briefly outline relation classification, dialog act prediction, event detection, personality detection and native language detection.

In relation classification, a relation from a predefined set of possible relations needs to be assigned to two entities within one sentence. This relation defines the relation between the two entities, such as the cause-effect between two entities. To this end, researchers use the SemEval2010 data set. Vu et al. (2016) combined a CNN and an RNN model for classifying relations using a voting scheme. They concluded that the models provide information that complements each other. The input of the model comprises position features, encoding the distances between noun pairs, and Word2Vec trained on an English Wikipedia dump from May 2014. Increasing the dimension of these word embeddings up to 400 increased the performance. Their methods slightly outperform the CNN models of Santos et al. (2015), who applied a CNN with a similar word embedding and positions features, but included only one filter size, to the same data set. The CNN and RNN models of Yin et al. (2017) showed an F1 performance score that is 16% lower than the F1 scores achieved by Santos et al. (2015) and Vu et al. (2016). The F1 score measures per class the percentage of instances classified correctly and combines these measures into one score. In relation classification, the classes are the relations in the set of possible relations and the instances are the two entities for which a relation needs to be determined.

Lee and Dernoncourt (2016) predict dialog acts for short texts. Examples of dialog acts are 'yes-no questions', 'declarative questions' and 'yes answers'. RNNs and CNNs are used to represent the text that forms the input of a two-layer neural network that classifies the texts. In the CNN model, the optimal filter size was 3 and the optimal number of filters was 500. This conclusion was drawn after experimentation with filter sizes between 1 and 10 and the number of filters between 50 and 1000. Lee and Dernoncourt (2016) applied their model to three different data sets, with 5, 43 and 89 classes to predict.

For two of the three data sets, the best results were achieved with the CNN representation that used 200-dimensional GloVe word embeddings trained on Twitter data. The data set with 89 classes showed the best results for the LSTM model with 300-dimensional Word2Vec embeddings trained on Google News. The deep learning models outperform established machine learning models, such as NB and SVM.

The goal of event detection is to find specified types of events in texts. Researchers used the ACE2005 Corpus that contains 2.037 sentences from news messages and 33 events for experimentation. Nguyen and Grishman (2015) applied a CNN model with an embedding layer, one convolutional layer with filter sizes 2 to 5, a max pooling layer and a soft-max layer to perform the classification. The embedding layer includes three types of representations of the text, including word embeddings with dimension 300 pre-trained on Google News, the position of the words in the text and the entity type of the words. For each word in the text, these embeddings are concatenated to form one vector. Nguyen and Grishman (2015) reported an F1 score of 69.0. Chen et al. (2015) reported a similar F1 score, 69.1, obtained with their CNN model with filter size 3 and a dynamic multi-pooling layer that captures the maximum values for multiple parts of the sentence.

Majumder et al. (2017) applied five CNNs to detect the big five personality traits — extraversion, neuroticism, agreeableness, conscientiousness and openness — from essays. The input of the CNNs is Word2Vec word embeddings. The CNNs have a convolutional layer with filter sizes 1, 2 and 3. After this layer, a concatenation layer includes document level features, such as the average sentence length. The accuracy of the prediction of each personality trait varies between 56.71 and 62.68.

Lai et al. (2015) applied a combined RNN-CNN model to sentiment and topic classification, as described in Sections 6.2.1 and 6.2.2. They also tested their approach to detect the native language of the author of scientific papers written in English. The native languages of the speakers were English, Japanese, German, Chinese and French. Just as for the topic classification task, the combined model outperformed the CNN model and the established machine learning models Logistic Regression and SVM.

6.2.4 Concluding remarks

We conclude the section with our observations in the literature about the deep learning research on text classification. The first observation is that the CNN architecture is the most used deep learning architecture for text classification tasks. RNN's are less used, and often in combination with a CNN structure.

The second observation is that the majority of research on deep learning for text classification focuses on the development of new models, which often are slight variants of previously proposed models. Subsequently, these models are tested on the same or similar data sets to demonstrate the improvements of the new model. Researchers seem to focus less on the application of successful deep learning models on new types of data or in new domains, such as the detection of fraud. We did not encounter deep learning models for fraud detection in the literature.

Our third observation outlines that many of the deep learning models for text classification have been inspired from the model of Kim (2014). Kim (2014) demonstrated that a simple CNN model with only one convolutional layer and little tuning of the parameters performs well on various data sets. In this research we also apply a model derived from the research of Kim (2014). Section 6.3.3 describes this CNN model in more detail. The parameters we experiment with are inspired from the parameters used in the previously discussed research, if reported in the papers.

Furthermore, we observe that the textual data sets used in the previously discussed research consist of sentences or short paragraphs, or are cut off after a specified number of characters. None of the data sets are similar to the management discussion and analysis section (MD&A) that we use to detect possible indications of fraud.

Finally, some of the researchers compared the performance of the deep learning models with the results of machine learning models ,such as SVM and NB, and showed that for the majority of the tasks tested, the deep learning models outperform these established machine learning models. Our baseline model that applied NB showed promising results in the detection of indications of fraud in annual reports worldwide (Fissette et al., 2017b). Therefore, we are interested to see whether a deep learning model may also improve an NB model in this task.

6.3 The method

We have developed an NB and a deep learning model that detect indications of fraud in the MD&A sections of annual reports worldwide. Section 6.3.1 describes the data used to develop and test these models. Section 6.3.2 explains the development of the NB model. Section 6.3.3 describes the deep learning model.

6.3.1 Data selection

The data set we used in this research is a subset of the data set used to develop a baseline model that has word unigrams as input features (Fissette et al., 2017b). This data set consists of 402 annual reports containing fraud and 1.325 annual reports with no fraudulent activities identified. All the reports are written in English, but are selected from companies worldwide. Therefore, the reports were subject to the rules and regulations of the country they were filed in. The data set consists of annual reports in the fiscal period from 1999 to 2011.

The annual reports in the fraud category were selected from news messages and Accounting and Auditing Enforcement Releases (AAERs) of the Securities and Exchange Commission (SEC) concerning the convictions of fraud. The SEC publishes their findings of fraud investigations online. The 'no fraud' cases selected, are annual reports of companies that are not convicted of fraudulent activities. No news messages or AAER regarding fraud were found for these companies. Note that the absence of the detection of fraud does not prove the absence of fraud. However, following the presumption of innocence, we assume that no fraud took place that affected the annual reports of these companies. To reflect that more companies in the real world situation are not involved in large frauds, for each fraudulent annual report three to five non fraudulent reports are collected. The fraudulent and non fraudulent reports are matched based on the same fiscal year, size and sector.

We focused the text analysis on the MD&A section of the annual reports. Therefore, the MD&A sections were extracted from the annual reports. See the 'Data Selection' section of Fissette et al. (2017b) for more information on the data selection and extraction process. We excluded graphs, figures and tables from the MD&A sections since the focus of this research is on the textual information only. Most of the annual reports filed with the SEC are html files. We used the BeautifulSoup package in Python to remove the html tags from the reports.

Deep learning algorithms demand much more computational resources than the established machine learning algorithms. Considering our computational limitations and the purpose of the research to demonstrate the feasibility of applying a text mining algorithms and compare it to the NB model, we experimented with a smaller data set. We randomly selected 40% of the total data set. This set consists of 161 MD&A sections of fraudulent reports and 530 section of non fraudulent reports. This data set is split into a 70% development and 30% test set. The companies in the test set do not occur with an MD&A section of another year in the development set.

6.3.2 The Naive Bayes model

The input of the NB model is the normalized counts of word unigrams. These normalized counts were obtained by first converting all words to lower case characters. Secondly, stop words and words that occur only once in the data set were excluded. All remaining words were stemmed using the Porter stemmer (Porter, 1980). Subsequently, for each MD&A section, the words were counted. These word counts were normalized with the term frequency-inverse document frequency' (TF-IDF). This approach includes the length of the text as well as the frequency of the occurrence of the word in the total data set in the normalization process. We applied the chi squared method to select the most informative features in the data set. This method is robust with respect to the distribution of the data. A more elaborate explanation of the extraction of the word unigrams can be found in Fissette et al. (2017b).

The NB machine learning algorithm describes the probabilities that an MD&A section belongs to the 'fraud' and 'no fraud' categories given the words in the MD&A section from the texts in the training set. The resulting model classifies new MD&A sections by calculating the probabilities that this section belongs to the two categories 'fraud' and 'no fraud'. The MD&A section is assigned to the category for which the calculated probability is the highest. For more details on the NB classification, refer to Manning and Schütze (1999), Russell and Norvig (2003) and Tan et al. (2005).

6.3.3 The Deep learning model

The deep learning model uses a different input than the previously described NB model. This section first describes the input of the deep learning model and the pre-processing steps taken in line with this type of input. Second, this section explains the applied deep learning model in more detail.

Word2Vec

As we have seen in Section 6.2, many of the models make use of word embeddings learned with Word2Vec. Word2Vec is a two-layer neural network. The input layer takes text and the output layer produces a set of feature vectors with a vector for each word in the text. These vectors are called word embeddings. A Word2Vec network itself is not a deep neural network; however, the output is a suitable input for deep neural networks. The advantage of a Word2Vec network is that it groups words that are similar to each other together. As a result, the network captures relations between words. A well-known example is asking a trained network the word that is closest to the

formula of words: 'king + (woman − man)'. The result is 'queen'.

In this research, we used Google's pre-trained Word2Vec, which is trained on the Google News dataset that has a vocabulary of 3 million words. The length of this vector is 300 features (Mikolov et al., 2013). We used it to ensure the word embeddings are learned from the largest text corpus available instead of our own limited corpus. Furthermore, the pre-trained Google Word2Vec captures commonly used word pairs, such as Soviet Union and New York.

Before transforming the MD&A texts into word embeddings, the punctuation is excluded from the texts and all words are converted to lowercase. In contrast with the NB model, the stopwords are not removed since the pre-trained Word2Vec contains some of the stopwords.

The Convolutional Neural Network

The implementation of the deep learning model that we applied is the Convolutional Neural Network (CNN) of Britz (2015). This implementation is based on the model for sentence classification developed by Kim (2014). We followed the modifications of Wirawan (2017) to include the pre-trained Word2Vec vectors. Furthermore, we modified the code to include the performance measures specificity, sensitivity, precision, recall and F1, discussed in Section 6.4.

Figure 6.1 shows the architecture of the CNN model. The first layer, the embedding layer, consists of the low-dimensional vector representations of the words. In the model used for this paper, this layer consists of one channel initialized by Google's pre-trained Word2Vec. Each row of the matrix represents a word in the text. Essentially, the input layer is the text represented by a concatenation of Word2Vec embeddings.

The second layer in the models' architecture is the convolutional layer. This layer performs the convolution operations, which can be interpreted as matrix operations. A convolution is the application of a filter to a window of n words to produce a feature. The filter is applied to each possible window of n words in the text, resulting in the feature map. The values of the filters are learned by the CNN during the training phase. A convolutional layer may consist of multiple filters of different sizes. The implementation of Britz (2015) uses 128 filters and three filter sizes, of two, three and four words.

The convolutional layer produces a large number of feature maps. This result is reduced by the max-pooling layer, which takes the highest value from each feature map. The max-pooling operation produces one feature vector from all the feature maps. This approach reduces the dimensionality while retaining the most important information. The effect of this reduction is comparable to the consequence of the unigram features in the NB model. Max

pooling retains the information about whether a feature occurs in the text, however loses the information related to the position in the text in which the feature occurs. Likewise, in a unigram model the information about whether a word occurs in the text is stored, while the information on where this word occurs in the text is lost.

The final layer of the CNN applies dropout regularization and classifies the result using softmax. Dropout regularization prevents the hidden units in the network to co-adapt to each other by randomly disabling a fraction of the units. Dropout reduces overfitting to the training data (Srivastava et al., 2014). The softmax layer classifies the results. Each text is assigned to one of the class labels, in this research 'fraud' and 'no fraud'.

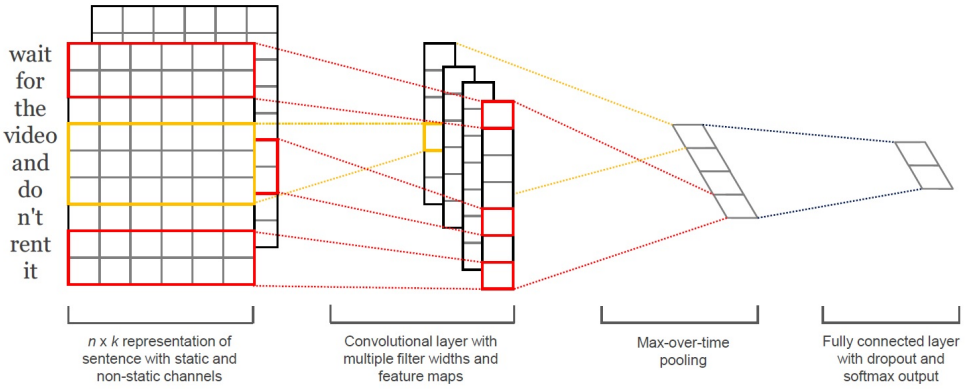


Figure 6.1: The schematic architecture of the CNN network of Kim (2014).

The development data set, as described in Section 6.3.1, is split into a 90% training and 10% validation set. A stratified split is applied so that in the training and validation set, the proportion of fraud and no fraud annual reports is similar. The training set comprises 434 MD&A sections, while the validation set contains 49 sections. The CNN learns the filters and resulting feature vector from the training set. The CNN is trained by feeding the training data in batches into the network. The number of training examples in one forward-backward pass defines how often the weights in the network are updated. This number, the batch size, can be varied from 1 to the number of training samples, owing to the memory available. The training set may be fed into the network several times. We varied this number, the epochs, to find the optimal number of epochs required to get good performance of the model. The performance of the model is determined after each training step with the accuracy and loss. Accuracy presents the percentage of MD&A sections classified correctly. The

loss is a measure of error made by the network. The goal is to minimize the loss. Training should be stopped when the loss function stops decreasing. We used early stopping to select the most optimal model based on the accuracy and the loss.

We experimented with the two training parameters, batch size and the number of epochs. Furthermore, we experimented with the three hyper parameters filter size, number of filters and drop out. In line with Britz (2015), no experiments were performed using L2 regularization as Zhang and Wallace (2015) found that the effect of regularization on the performance of the model was limited. The default training parameters of Britz (2015) include a batch size of 64 and 200 epochs. During the development of the model, we evaluated the model every 10 steps. We set the number of epochs to 10 to keep the training time within 24 hours. We varied the batch size to find the optimal result and training time. We experimented with various mini-batch sizes, in the range from 1 to 64. With the resulting batch size we varied the filter sizes and the number of filters. The filter sizes and the number of filters that we experimented with follow from the settings reported in the literature as described in Section 6.2. For the most optimal number of filter sizes and the number of filters, we varied the dropout.

6.4 Results

The performance of the NB and CNN models is determined during the development and on the test set. During the training of the CNN model, the performance is determined with accuracy and loss. In addition to accuracy, we calculate five other performance measures to make a complete comparison between the NB and CNN models. The sensitivity measures the percentage of fraud reports detected by the model. The specificity calculates the percentages of non fraudulent reports identified. Precision measures for how many of the reports classified by the model as fraud, the model made the correct classification. Likewise, the recall measures the percentage of reports classified by the model as non fraudulent that are indeed non fraudulent. The F1 score represents the overall reliability of the model by combining the measures precision and recall.

In the previous research we found that the most optimal number of unigrams in the NB model to detect indications of fraud in annual reports worldwide is 10k (Fissette et al., 2017a,b). We used the same number of features for the NB model on the data selected for the current research. On the development set, we applied 10-fold cross-validation. The accuracy achieved with the NB

model is 72% on the development set and 75% on the test set. The NB model detects 10% of the fraud cases in the test set. In 36% of the MD&A sections that the NB model classifies a fraud case, the model is correct. The model incorrectly classifies 6% of the MD&A sections from non fraudulent reports as fraudulent, while 94% of the non fraudulent reports is classified correctly. When classifying reports as non fraudulent the NB model is correct for 78% of the reports. Figure 6.2 summarizes the results.

The size of the generated vocabulary in the CNN model is 28.279. The size of the vocabulary of the NB model, consisting of 13.227 words, is substantially smaller due to the removal of stopwords and the words that occur only once in the data set. A batch size of 10 had the most optimal training time. The best performance in terms of accuracy and loss is achieved by the model with filter sizes 1, 2 and 3, with 128 filters and a drop-out of 0.5. The best performance in terms of the loss function and accuracy is reached within less than five epochs, after 90 training steps. Performing more training steps results in an increase in loss and a decrease in accuracy. We therefore apply early stopping and select this model as the best performing CNN model on our data set. The development set was randomly split into 90% training and 10% test set. Performing 10-fold cross-validation was infeasible due to the long time needed to train the CNN model. In the development phase the selected model shows an accuracy of 94% and a loss of 0.412562. The model correctly detects 75% of the fraudulent reports and all of the non fraudulent reports. When the model classifies a case as fraudulent, it is correct for all cases. However, since the model misses 25% of the fraud cases, the model is correct for 92% of the reports when classifying a report as non fraudulent. The results on the test set are lower. The accuracy drops to 74%. The model detects 23% of the fraudulent reports in the test set. The model is correct for 38% of the reports when classifying a report as fraudulent. The CNN model classifies 89% of the non fraudulent reports correctly. When the model classifies a report as non fraudulent the model is correct for 79% of the reports. Figure 6.2 shows the results of the CNN during development and on the test set.

Although the results on the development set are much higher for the CNN model than the NB model, the results on the test set are comparable. The CNN model detects more fraud cases than the NB model, 23% opposed to 10%. However, the CNN model makes more mistakes in the identification of non fraudulent reports. Notwithstanding, the results of the CNN model are slightly more reliable than those of the NB model.

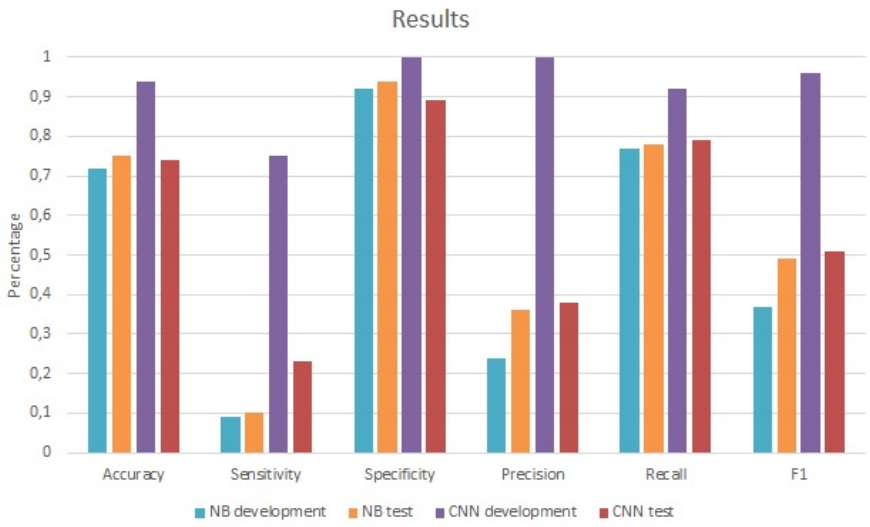


Figure 6.2: The results of the NB and CNN models in the development phase and on the test set.

6.5 Discussion and conclusion

We have applied an NB model and a CNN model to the MD&A sections of annual reports of companies worldwide to determine the usability of a deep learning model to detect indications of fraud in texts. Although the number of MD&A sections seems to be too limited considering the low performance of the NB model compared to the results achieved for the total data set (Fissette et al., 2017b), the results of the CNN are promising. In the development phase, the CNN model showed more promising results than the NB model. On the test set, the CNN achieved better results than the NB model in the detection of fraudulent reports. The results of the CNN model are also slightly more reliable than the NB model. Altogether, these results show that deep learning can be a suitable method for detecting indications of fraud in the texts of annual reports. The CNN model and the representation of text as word embeddings as input for the neural network model appear to be a promising approach.

Furthermore, a notable result is the decline in the performance of the CNN model on the test set, compared to the results during development. This difference in performance may be due to a difference in composition of the evaluation data set during development and the test set. To create the evaluation data

set, the development data set was randomly split into a 90% train and 10% test set. For some companies, the development set contains the MD&A sections for multiple years. Due to the random partitioning of the development set, it may have been noticed that for some companies, one MD&A section is in the training set, while another is in the test set. As the MD&A sections of one company are often similar year after year, we hypothesize that correctly classifying an MD&A section of a company for which an MD&A section of a different year is easier in the training set. This may improve the results.

Unlike during the development of the CNN model, the results of the NB model in the development phase are obtained by 10-fold cross-validation. The results given in Figure 6.2 are an average of the results of the folds. For three of the 10 folds, the results are similar or better than the results obtained for the test set. This shows that the performance varies per composition of the data set tested on. The performance of the model on the test set may be better than the average result during the development since it is trained on a bit more data. In cross-validation 10% of data in the development is set aside to test on during development. Training during development is done on the remaining 90%. The model that is applied to the test set is trained on the entire development set. The size of the data set used in this research is limited. The NB model might have benefited even from a small increase in data.

Following these results, we provide some suggestions for future research concerning the application of a CNN model to detect indications of fraud in the texts of annual reports. First, in future research, a larger data set is used. As the number of training samples is limited, the CNN model may be overfitting despite early stopping and the use of dropout. The results of the NB model are, as may be expected, better when the model is trained on a larger data set. The performance of the CNN model may also improve when it is trained on a larger data set. The results of this research give rise to further experimentation with more data.

Second, the use of more processing power and memory is recommended than was available during the research discussed in this paper. An increase in processing power and memory, not only allows the use of a larger data set, but also enables a more extensive grid search on the models parameters. Such an approach may detect an even more optimal combination of parameters. The computer we had at our disposal has an Intel(R) Core(TM) i5-4300CU CPU @ 1.90 Hz, 2.50 GHz and 8 GB RAM memory. With these specifications, the training of the CNN on the total data set was infeasible. Training the CNN on the small data set took up to 12 hours, depending on the parameter

settings. The time needed to train the NB model is in the order of seconds. Arguably, not only the performance measures, but also the efficiency of the model should be considered in the comparison of two models. It should be considered whether extra time is worth the increased performance. This would depend on the extent to which the performance has improved. How to weigh the performance measures and the efficiency warrants further debate. However, we point out that the efficiency of CNNs is likely to improve. Over the last few decades, the available processing power has increased. This development still continues. As a result, more and more processing power is available to run complex models, such as CNNs, in a reasonable time frame. Furthermore, researchers are working to improve the efficiency of the CNN models (Ioannou et al., 2016; Shen et al., 2016; Zhang et al., 2016a).

Third, other CNN or deep learning architectures are an interesting subject for further research concerning the detection of indications of fraud in annual reports. The research discussed in this paper applied a simple CNN architecture that demonstrated successful results in various domains (Kim, 2014). Many variants on this architecture were proposed (Johnson and Zhang, 2014; Chen et al., 2015; Ma et al., 2015; Nguyen and Grishman, 2015; Wang et al., 2015; Chen and Ku, 2016; Laha and Raykar, 2016). It is conceivable that one of these variants may achieve better results for detecting indications of fraud in texts. Furthermore, the research concerned with the representation of documents may yield additional possibilities of improvement (Denil et al., 2014; Tang et al., 2015; Zhang et al., 2016b).

Finally, future research may experiment with various representations of the texts to further improve the results. In the research described in this paper, the MD&A sections are represented with Word2Vec word embeddings learned from Google News messages. The MD&A sections are texts that are in the more specific domain of financial texts. Learning the word embeddings from a large corpus of financial documents may obtain more meaningful word embeddings in the financial domain. Another option for representing the text is one-hot encoding. Zhang and Wallace (2015) argued that while word-embeddings obtained with Word2Vec or GloVe work well, one-hot vectors may be considered if the size of the training set is sufficiently large.

7 Discussion and conclusion

In business performance reports, the financial numbers are accompanied by textual information. In fraud investigations the focus is on numerical information. The ever changing and complex fraud schemes require an innovative fraud detection solution. Therefore, in the research described in this thesis we examined the possibility of text mining models to contribute to the development of such advanced fraud detection methods. In Section 7.1, the developed text mining models and results are briefly discussed. Section 7.2, suggests possibilities for future research that may contribute to the development of methods that detect the indications of large financial frauds worldwide.

7.1 Answer to research question

To answer the main research question 'Can text mining techniques contribute to the detection of fraud in annual reports worldwide?', we examined various text mining models as described in Chapters, 3, 4 and 5. The models were developed using the texts in the Management Discussion and Analysis section of the annual reports. The results of these models show that text mining techniques may contribute to the detection of fraud in annual reports. In the subsequent paragraphs, we briefly discuss the possibilities and limitations of each of the text mining models.

In Chapter 2, we suggested a straightforward text analysis approach that extracts information from annual reports. In this thesis, the approach was used to prepare the data set required to develop and test the text mining models. The approach contributed to the process of matching the non fraudulent annual reports to the fraudulent annual reports. The approach is most successful when the information request is concise and the documents searched in have some form of structure. Additional knowledge and rules are needed in case of no structure. Furthermore, the amount and complexity of the knowledge and rules required increase when the information request is more complex. For example, extracting the year of an annual report with the suggested approach is straightforward, while extracting the Management Discussion and Analysis section requires additional rules.

Chapter 3 described a baseline model that includes only individual words, unigrams, as features in the Naive Bayes (NB) and Support Vector Machine (SVM) algorithms. The models based on these two machine learning algorithms showed promising results. With an accuracy of 89%, the NB model outperformed the SVM model that achieved an accuracy of 85%. The NB model was better at detecting the fraudulent reports, while the SVM one was slightly better at identifying the non fraudulent reports. We noted that for some companies, the data set includes the annual reports of multiple years. Due to the random partitioning of the development data in training and test sets, the annual report of a company may be in the training set, while the report of the same company for another year may be in the test set. Since the MD&A sections of one company are similar over several years, we hypothesized that it is easier to correctly classify annual reports in the test set for companies for which an annual report of another year is included in the training set. We examined the results of the NB model more closely to verify this hypothesis. As expected, the presence of a company in the training and test set has an impact on the results. This results in a decline in accuracy. Of the companies in the test set for which no annual report of another year is included in the training set, 86% is classified correctly. In particular, fraud cases are detected less often. Total 19% of the fraud cases are identified. Still 94% of the non fraudulent cases are classified correctly.

The baseline model of Chapter 3 was extended with various categories of linguistic features in Chapter 4. The features were subdivided into the categories 'descriptive', 'complexity', 'grammatical', 'readability', 'psychological', 'bigrams' and 'relations'. The best performing models were the NB model with unigrams only and the SVM model that combines unigrams and bigrams. These models achieved an accuracy of 89% and 90%, respectively. The NB model misses the least number of fraud cases. Furthermore, the results showed that the bigrams add a little bit of information to the SVM model only. The other categories of linguistic features did not add value to the NB and SVM models that only use word unigrams as features. A possible explanation for these findings is that the relationship between the linguistic features and fraud or deception varies per fraud case or that there is no stable relationship between linguistic characteristics and deception. It is also possible that, as the document is written by multiple authors, none of the linguistic features is prevalent. Furthermore, several of the linguistic features are based on dictionaries developed for domains other than financial documents that may not be suitable for annual reports. In addition, the unigrams are more specific than the linguistic features, which mostly are word groups. Possibly, this specificity

is needed to capture the differences between fraudulent and non fraudulent reports.

In Chapter 5 we proposed a machine learning approach that uses a completely different representation of the text than the features used in the models of Chapters 3 and 4. The Convolutional Neural Network (CNN) model takes word embeddings, which is the representation of each word in the text as a vector, as input. The processing power and memory available with us was too limited to process the entire data set with the CNN model. Therefore, we experimented with a smaller data set to provide a proof of concept of the usability of the CNN for the detection of indications of fraud in annual reports. The data set was too small to achieve accuracy as compared to the models of the previous chapters. However, compared to the results of the NB model on the smaller data set, the results of the CNN model did show the potential to outperform the NB model.

The previously discussed models all use the same input data, of which the collection and extraction steps were described in Chapter 1. Therefore, the limitations of this extraction process may influence all the models. We noticed two factors that might have affected the texts, and therefore the features that are the input of the models. First, since the automatic extraction of the MD&A section from the annual reports may not be flawless for all reports, some extracted sections may contain too less or too much text from the annual report. Second, the automatic removal of html tags may not be perfect. As a result, some MD&A sections may still include some of these tags.

In addition, we want to point out that for all models it is unknown how to interpret the patterns found to determine if an annual report should be classified as 'fraud' or 'no fraud'. The models find non evident patterns that may contribute to the detection of indications of fraud. It is important to emphasize that the models described in this thesis do not give a definitive result. The classification decision of a model could be used as an indication to further investigate a company and its annual report.

To conclude, the results of the research described in this thesis show that text mining methods can contribute to the detection of indications of fraud in the texts of the annual reports of companies worldwide. The NB model with unigrams as textual features achieved the best results. A CNN model with word embeddings as textual representation has the potential to outperform the NB model. Further research is needed to develop a model that can detect a higher percentage of fraud cases with a greater degree of certainty.

7.2 Future research

This section describes various research opportunities that may contribute to the development of advanced methods for detecting indications of fraud.

As discussed in Chapter 1, the recent developments show that the information disclosed in annual reports is subject to change. During the period when this research was conducted, there appears to be an increase in the textual information as a result of integrated reporting. The annual reports of the past three years may differ slightly from the annual reports in the data set used in this research. Therefore, for future research, we advise to extend the data set with the latest fraud cases and annual reports. This keeps the models developed during the research up to date. An updated data set includes the new fraud schemes and activities that may have been used in the latest fraud cases.

One of the questions for future research concerns the composition of the train and test data sets. In the data collection process the fraud and no fraud annual reports are matched based on the year, sector and company size. When partitioning the data into training and test sets, the splitting process makes sure that the distribution of fraud and no fraud annual reports is similar in each set. However, the splitting process does not take the year, sector and company size into account. As a result, the matched cases may not be included in the same set. The question arises whether this affects the results of the models. This may be determined by developing the machine learning models with a data set that only includes annual reports of one year or from one sector or one category of company sizes. However, note that in that case, less data is available, which may affect the result. Another possibility is to place the matched annual reports in the same subset. Attention should be paid to maintaining the distribution of fraud and no fraud annual reports.

The linguistic features described in Chapter 3 may also require additional attention. The results showed that the machine learning algorithms did not find a relationship between the linguistic features and the presence of fraud. The question remains 'Does this mean the relationship is absent or the direction of the relationship varies per fraud case?'. Furthermore, the annual reports are often written by more than one author. Future research may examine the influence of multiple authorship on the linguistic features.

We identify four opportunities for future research that may improve the performance of the fraud detection models that were researched in this thesis. The first opportunity involves the addition of textual information from the annual reports. The research discussed in this thesis focused on the MD&A section of the annual report. As explained in Chapter 1, annual reports con-

tain more textual information, such as the notes to the financial statements and the letter to the shareholders. Future research may include additional textual sections, or the entire annual report, to determine if this provides more information that may contribute to the detection of indications of fraud.

The second opportunity for increasing the detection rate of fraud detection models concerns financial information. The textual information in an annual report accompanies the financial information. Fraud detection methods usually focus on these financial numbers. The textual and numerical information capture different aspects of the financial position and achievements of the company. In bankruptcy prediction, the texts provide information that is complementary to the financial information (Tennyson et al., 1990). Therefore, the question arises whether the financial and textual information complement each other in the detection of indications of fraud. This combination may result in a model that is capable of detecting a higher number of fraud cases with more accuracy. Furthermore, a model based on financial information may be used to verify the results of a text mining model, or vice versa.

The third opportunity involves, the information collected from other sources than the annual report that may contribute to a fraud detection model. Financial analysts gain insights from a combination of the annual reports and other information sources (Hoogendoorn et al., 2004). Companies have a lot of internal and external information available with them, ranging from e-mails and business plans to analyst reports and quarterly reports. Francis et al. (2002) demonstrated that these latter two information sources complement each other. Possibly, the analyst and quarterly reports also complement the information in annual reports. Press releases are another publicly available source of information that may be included in the development of an extended text mining model that detects indications of fraud.

Finally, instead of providing the models with additional information, future research may experiment with the machine learning algorithms. First, the CNN model of Chapter 4 needs further research on a larger data set. The deep learning model may perform better than the NB model. Furthermore, an ensemble of machine learning algorithms may achieve a higher fraud detection rate. It is possible that each machine learning algorithm detects different patterns in the data. A model that uses a combination of these patterns may yield better results.

The models proposed in this thesis focus on the detection of indications of fraud. The later the fraud gets detected, the greater the damage caused by the fraudulent activities. In order to reduce the damage, it is important that fraud is detected at an early stage, or even better is, to be able to predict the

emergence of fraudulent activities. Predicting fraud is not possible Moreover, fraud usually out starts small and grows over the years, in most cases, to hide the mall adjustments mas initially (Center for Audit Quality, 2010). This may be reflected in annual reports. It is conceivable that subtle changes in the annual report of a company over the years may indicate a risk of the emergence of fraud.

References

ACFE (2016b). Report to the nation on occupational fraud & abuse.

Aerts, W. (2001). Inertia in the attributional content of annual accounting narratives. *European Accounting Review*, 10(1).

Aerts, W. (2005). Picking up the pieces: Impression management in the retrospective attributional framing of accounting outcomes. *Accounting, Organizations & Society*, 30(6):493–517.

Afroz, S., Brennan, M., and Greenstadt, R. (2012). Detecting hoaxes, frauds, and deception in writing style online. In *Security and Privacy (SP), 2012 IEEE Symposium on*, pages 461–475. IEEE.

Agichtein, E. and Ganti, V. (2004). Mining reference tables for automatic text segmentation. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 20–29. ACM.

Agostini, M. and Favero, G. (2013). Accounting fraud, business failure and creative auditing: A micro-analysis of the strange case of Sunbeam Corp. Working Papers 12, Department of Management, Universit Ca' Foscari Venezia.

Altman, E. I. et al. (2000). Predicting financial distress of companies: revisiting the z-score and zeta models. *Stern School of Business, New York University*, pages 9–12.

Anderson, J. (1983). Lix and rix: Variations on a little-known readability index. *Journal of Reading*, 26(6):490–496.

Anzaroot, S. and McCallum, A. (2013). A new dataset for fine-grained citation field extraction. In *ICML Workshop on Peer Reviewing and Publishing Models*.

Balakrishnan, R., Qiu, X. Y., and Srinivasan, P. (2010). On the predictive ability of narrative disclosures in annual reports. *European Journal of Operational Research*, 202(3):789–801.

Beasley, M. S., Carcello, J. V., Hermanson, D. R., and Neal, T. L. (2010). Fraudulent financial reporting: 1998-2007: An analysis of US public companies. The Committee of Sponsoring Organizations of the Treadway Commission (COSO).

Beattie, V., Dhanani, A., and Jones, M. J. (2008). Investigating presentational change in U.K. annual reports : A longitudinal perspective. *Journal of Business Communication*, 45(2):181–222.

Bell, T. and Carcello, J. (2000). A decision aid for assessing the likelihood of fraudulent financial reporting. *Auditing: A Journal of Practice and Theory*, 19(1):169–78.

Beneish, M. D. (1997). Detecting GAAP violation: implications for assessing earnings management among firms with extreme financial performance. *Journal of Accounting and Public Policy*, 16(3):271–309.

Beneish, M. D. (1999). Incentives and penalties related to earnings overstatements that violate GAAP. *The Accounting Review*, 74(4):425–457.

Beneish, M. D., Lee, C., Press, E., Whaley, B., Zmijewski, M., and Cisilino, P. (1999). The detection of earnings manipulation. *Financial Analysts Journal*, pages 24–36.

Bhattacharya, S., Xu, D., and Kumar, K. (2011). An ANN-based auditor decision support system using Benford's law. *Decision Support Systems*, 50:576–584.

Bird, Steven, E. L. and Klein, E. (2009). *Natural Language Processing with Python*. OReilly Media Inc.

Borkar, V., Deshmukh, K., and Sarawagi, S. (2001). Automatic segmentation of text into structured records. In *ACM SIGMOD Record*, volume 30(2), pages 175–186. ACM.

Boschetti, F., Romanello, M., Babeu, A., Bamman, D., and Crane, G. (2009). Improving OCR accuracy for classical critical editions. *Research and Advanced Technology for Digital Libraries*, pages 156–167.

Brazel, J. F., Jones, K. L., Thayer, J., and Warne, R. C. (2015). Understanding investor perceptions of financial statement fraud and their use of red flags: evidence from the field. *Review of Accounting Studies*, 20(4):1373–1406.

Britz, D. (2015). Implementing a CNN for text classification in TensorFlow. Tutorial, accessed 2017-04-06.

Brown, S. V. and Tucker, J. W. (2011). Large-sample evidence on firms' year-over-year MD&A modifications. *Journal of Accounting Research*, 49(2):309–346.

Bureau, U. S. C. (2014). Statistics of US businesses. https://www.census.gov/econ/susb/, accessed: 2014-03-07.

Burgoon, J., Blair, J., Qin, T., and Nunamaker, JayF., J. (2003). Detecting deception through linguistic analysis. In Chen, H., Miranda, R., Zeng, D., Demchak, C., Schroeder, J., and Madhusudan, T., editors, *Intelligence and Security Informatics*, volume 2665 of *Lecture Notes in Computer Science*, pages 91–101. Springer Berlin Heidelberg.

Butler, M. and Kešelj, V. (2009). Financial forecasting using character n-gram

analysis and readability scores of annual reports. In *Advances in artificial intelligence*, pages 39–51. Springer.

Cecchini, M., Aytug, H., Koehler, G. J., and Pathak, P. (2010). Making words work: Using financial text as a predictor of financial events. *Decision Support Systems*, 50(1):164 – 175.

Center for Audit Quality (2010). Deterring and detecting financial reporting fraud.

Chen, W.-F. and Ku, L.-W. (2016). UTCNN: a deep learning model of stance classification on social media text. *arXiv preprint arXiv:1611.03599*.

Chen, Y., Xu, L., Liu, K., Zeng, D., Zhao, J., et al. (2015). Event extraction via dynamic multi-pooling convolutional neural networks. In *ACL (1)*, pages 167–176.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Clatworthy, M. A. and Jones, M. J. (2003). Financial reporting of good news and bad news: evidence from accounting narratives. *Accounting and Business Research*, 33(3):171–185.

Clatworthy, M. A. and Jones, M. J. (2006). Differential patterns of textual characteristics and company performance in the chairman's statement. *Accounting, Auditing & Accountability Journal*, 19(4):493–511.

Coleman, M. and Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Conneau, A., Schwenk, H., Barrault, L., and Lecun, Y. (2016). Very deep convolutional networks for text classification. *arXiv preprint arXiv:1606.01781*.

Conway, M., Doan, S., Kawazoe, A., and Collier, N. (2009). Classifying disease outbreak reports using n-grams and semantic features. *International journal of medical informatics*, 78(12):e47–e58.

Craig, R. J. and Amernic, J. H. (2004). Enron discourse: the rhetoric of a resilient capitalism. *Critical perspectives on accounting*, 15(6):813–852.

Dale, E. and Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

De Marneffe, M.-C., MacCartney, B., Manning, C. D., et al. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454. Genoa.

Dechow, P. M., Ge, W., Larson, C. R., and Sloan, R. G. (2011). Predicting material accounting misstatements*. *Contemporary accounting research*, 28(1):17–82.

Dechow, P. M., Sloan, R. G., and Sweeney, A. P. (1996). Causes and consequences of earnings manipulation: An analysis of firms subject to enforcement actions by the SEC. *Contemporary accounting research*, 13(1):1–36.

Denil, M., Demiraj, A., Kalchbrenner, N., Blunsom, P., and de Freitas, N. (2014). Modelling, visualising and summarising documents with a single convolutional neural network. *arXiv preprint arXiv:1406.3830*.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., and Cooper, H. (2003). Cues to deception. *Psychological bulletin*, 129(1):74.

Di Castri, S. and Benedetto, F. (2005). There is something about Parmalat (On directors and gatekeepers). *Available at SSRN 896940*.

Dong, W., Liao, S., and Liang, L. (2016a). Financial statement fraud detection using text mining: A systematic functional linguistics theory perspective. *Pacific Asia Conference on Information Systems (PACIS)*.

Dong, W., Liao, S., Xu, Y., and Feng, X. (2016b). Leading effect of social media for financial fraud disclosure: A text mining based analytics. *Accounting Information Systems (SIGASYS)*.

Dos Santos, C. N. and Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. In *COLING*, pages 69–78.

European Commission (2014). What is an SME? EU recommendation 2003/361, accessed: 2014-04-15.

Fanning, K. and Cogger, K. (1998). Neural detection of management fraud using published financial data. *International Journal of Intelligent Systems in Accounting, Finance and Management*, 7(1):21–41.

Fissette, M., Veldkamp, B., and De Vries, T. (2017a). Linguistic features in a text mining approach to detect indications of fraud in annual reports worldwide. *Working paper*.

Fissette, M., Veldkamp, B., and De Vries, T. (2017b). Text mining to detect indications of fraud in annual reports worldwide. *Under review*.

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Forensic, K. (2008). Fraud survey 2008. *KPMG, Brisbane, Qld*.

Fornaciari, T. and Poesio, M. (2012). On the use of homogenous sets of subjects in deceptive language analysis. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, pages 39–47. Association for Computational Linguistics.

Francis, J., Schipper, K., and Vincent, L. (2002). Earnings announcements and competing information. *Journal of Accounting and Economics*, 33(3):313–342.

Fuller, C. M., Biros, D. P., and Delen, D. (2011). An investigation of data

and text mining methods for real world deception detection. *Expert Systems with Applications*, 38(7):8392–8398.

Gee, J., Button, M., and Brooks, G. (2017). The financial cost of fraud: what data from around the world shows.

Glancy, F. H. and Yadav, S. B. (2011). A computational model for financial reporting fraud detection. *Decision Support Systems*, 50:595–601.

Goel, S. and Gangolly, J. (2012). Beyond the numbers: Mining the annual reports for hidden cues indicative of financial statement fraud. *Intelligent Systems in Accounting, Finance and Management*, 19(2):75–89.

Goel, S., Gangolly, J., Faerman, S. R., and Uzuner, O. (2010). Can linguistic predictors detect fraudulent financial filings? *Journal of Emerging Technologies in Accounting*, 7:25–46.

Goel, S. and 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.

Green, B. and Choi, J. (1997). Assessing the risk of management fraud through neural network technology. *Auditing*, 16:14–28.

Grove, H. and Basilisco, E. (2008). Fraudulent financial reporting detection: Key ratios plus corporate governance factors. *International Studies of Management & Organization*, 38(3):10–42.

Guay, W. R., Samuels, D., and Taylor, D. J. (2015). Guiding through the fog: Financial statement complexity and voluntary disclosure. *Available at SSRN 2564350*.

Gunning, R. (1969). The fog index after twenty years. *Journal of Business Communication*, 6(2):3–13.

Hahn, R., Bizer, C., Sahnwaldt, C., Herta, C., Robinson, S., Bürgle, M., Düwiger, H., and Scheel, U. (2010). Faceted wikipedia search. In *International Conference on Business Information Systems*, pages 1–11. Springer.

Hájek, P. and Olej, V. (2013). Evaluating sentiment in annual reports for financial distress prediction using neural networks and support vector machines. In *Engineering Applications of Neural Networks*, pages 1–10. Springer.

Hajek, P., Olej, V., and Myskova, R. (2014). Forecasting corporate financial performance using sentiment in annual reports for stakeholders decision-making. *Technological and Economic Development of Economy*, 20(4):721–738.

He, Q. and Veldkamp, D. B. (2012). Classifying unstructured textual data using the product score model: an alternative text mining algorithm. In Eggen, T. and Veldkamp, B., editors, *Psychometrics in practice at RCEC*, pages 47 – 62. RCEC, Enschede.

Heidari, M. and Felden, C. (2016). Analytical support of financial footnotes: Developing a text mining approach. *Twenty-second Americas Conference on Information Systems, San Diego*.

Heydari, A., ali Tavakoli, M., Salim, N., and Heydari, Z. (2015). Detection of review spam: A survey. *Expert Systems with Applications*, 42(7):3634–3642.

Higgins, H. N. (2003). Disclosures of foreign companies registered in the US. *New Accountant*, pages 19–22.

Hoogendoorn, M., Klaassen, J., and Krens, F. (2004). *Externe verslaggeving in theorie en praktijk 1*. Reed Business Information.

Hoogs, B., Kiehl, T., LaComb, C., and Senturk, D. (2007). A genetic algorithm approach to detecting temporal patterns indicative of financial statement fraud. *Int. Syst. in Accounting, Finance and Management*, 15(1-2):41–56.

Huang, S., Tsaih, R., and Lin, W. (2012). Unsupervised neural networks approach for understanding fraudulent financial reporting. *Industrial Management & Data Systems*, 112(2):224–244.

Humpherys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K., and Felix, W. F. (2011). Identification of fraudulent financial statements using linguistic credibility analysis. *Decis. Support Syst.*, 50(3):585–594.

IFRS (2017). Who uses IFRS standards? http://www.ifrs.org/use-around-the-world/use-of-ifrs-standards-by-jurisdiction/#filing, last accessed: 2017-08-24.

Ioannou, Y., Robertson, D. P., Cipolla, R., and Criminisi, A. (2016). Deep roots: Improving CNN efficiency with hierarchical filter groups. *CoRR*, abs/1605.06489.

Jeanjean, T., Lesage, C., and Stolowy, H. (2010). Why do you speak English (in your annual report)? *The International Journal of Accounting*, 45(2):200–223.

Joachims, T. (1998). *Text categorization with support vector machines: Learning with many relevant features*. Springer.

Johnson, R. and Zhang, T. (2014). Effective use of word order for text categorization with convolutional neural networks. *arXiv preprint arXiv:1412.1058*.

Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition.

Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014). A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.

Kaminski, K. A., Wetzel, T. S., and Guan, L. (2004). Can financial ratios detect fraudulent financial reporting. *Managerial Auditing Journal*, 19(1):15–28.

Karami, A. and Zhou, B. (2015). Online review spam detection by new linguistic features. *iConference 2015 Proceedings*.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document.

Kirkos, E., Spathis, C., and Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications*, 32(4):995–1003.

Klaassen, J., Hoogendoorn, M., and Vergoossen, R. (2008). *Externe verslaggeving*. Noordhoff Uitgevers.

Kluegl, P., Toepfer, M., Lemmerich, F., Hotho, A., and Puppe, F. (2012). Collective information extraction with context-specific consistencies. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 728–743. Springer.

Kotsiantis, S., Koumanakos, E., Tzelepis, D., and Tampakas, V. (2006). Forecasting fraudulent financial statements using data mining. *International journal of computational intelligence*, 3(2):104–110.

KPMG (2016). Global profiles of the fraudster: Technology enables and weak controls fuel the fraud.

Kumar, J., Ye, P., and Doermann, D. (2013). A dataset for quality assessment of camera captured document images. In *International Workshop on Camera-Based Document Analysis and Recognition*, pages 113–125. Springer.

Laha, A. and Raykar, V. (2016). An empirical evaluation of various deep learning architectures for bi-sequence classification tasks. *arXiv preprint arXiv:1607.04853*.

Lai, S., Xu, L., Liu, K., and Zhao, J. (2015). Recurrent convolutional neural networks for text classification. In *AAAI*, volume 333, pages 2267–2273.

Larcker, D. F. and Zakolyukina, A. A. (2012). Detecting deceptive discussions in conference calls. *Journal of Accounting Research*, 50(2):495–540.

Lee, C.-C., Churyk, N. T., and Clinton, B. D. (2013). Validating early fraud prediction using narrative disclosures. *Journal of Forensic & Investigative Accounting*, 5(1):35–57.

Lee, C.-C., Welker, R. B., and Odom, M. D. (2009). Features of computer-mediated, text-based messages that support automatable, linguistics-based indicators for deception detection. *Journal of Information Systems*, 23(1):5–24.

Lee, C.-H., Lusk, E., and Halperin, M. (2014). Content analysis for detection of reporting irregularities: Evidence from restatements during the SOX era. *Journal of Forensic and Investigative Accounting*, 6(1):99–122.

Lee, J. Y. and Dernoncourt, F. (2016). Sequential short-text classification with recurrent and convolutional neural networks. *arXiv preprint arXiv:1603.03827*.

Lee, T. (1994). The changing form of the corporate annual report. *The Accounting Historians Journal*, pages 215–232.

Leinemann, C., Schlottmann, F., Seese, D., and Stuempert, T. (2001). Automatic extraction and analysis of financial data from the EDGAR database. *SA Journal of Information Management*, 3(2).

Li, F. (2006). Do stock market investors understand the risk sentiment of corporate annual reports? *SSRN eLibrary*.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and economics*, 45(2):221–247.

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

Li, F., Lundholm, R. J., and Minnis, M. (2011). A new measure of competition based on 10-K filings: Derivations and implications for financial statement analysis. *Social Science Research Network Working Paper Series*.

Ma, M., Huang, L., Xiang, B., and Zhou, B. (2015). Dependency-based convolutional neural networks for sentence embedding. *arXiv preprint arXiv:1507.01839*.

Majumder, N., Poria, S., Gelbukh, A., and Cambria, E. (2017). Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32(2):74–79.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

Markelevich, A., Riley, T., and Shaw, L. (2015). Towards harmonizing reporting standards and communication of international financial information: The status and the role of IFRS and XBRL. *Journal of Knowledge Globalization*, 8(2).

Markowitz, D. M. and Hancock, J. T. (2014). Linguistic traces of a scientific fraud: The case of Diederik Stapel. *PloS one*, 9(8):e105937.

Markowitz, D. M. and Hancock, J. T. (2016). Linguistic obfuscation in fraudulent science. *Journal of Language and Social Psychology*, 35(4):435–445.

Mbaziira, A. and Jones, J. (2016). A text-based deception detection model for cybercrime. *International Conference on Technology and Management*.

Mc Laughlin, G. H. (1969). SMOG grading - a new readability formula. *Journal of reading*, 12(8):639–646.

Merkl-Davies, D. M. and Brennan, N. M. (2007). Discretionary disclosure strategies in corporate narratives: Incremental information or impression management?

Metsis, V., Androutsopoulos, I., and Paliouras, G. (2006). Spam filtering with naive bayes - which naive bayes? In *CEAS*, pages 27–28.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Miller, B. P. (2010). The effects of reporting complexity on small and large investor trading. *The Accounting Review*, 85(6):2107–2143.

Minhas, S. and Hussain, A. (2014). Linguistic correlates of deception in financial text a corpus linguistics based approach. *Psychology Review*, 19:307–342.

Moffitt, K. and Burns, M. B. (2009). What does that mean? Investigating obfuscation and readability cues as indicators of deception in fraudulent financial reports. *AMCIS 2009 Proceedings*, page 399.

Mooney, R. (1999). Relational learning of pattern-match rules for information extraction. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, volume 334.

Morgan, A. R. and Burnside, C. (2014). Olympus corporation financial statement fraud case study: The role that national culture plays on detecting and deterring fraud. *Journal of Business Case Studies (Online)*, 10(2):175.

Ndofor, H. A., Wesley, C., and Priem, R. L. (2015). Providing CEOs with opportunities to cheat the effects of complexity-based information asymmetries on financial reporting fraud. *Journal of Management*, 41(6):1774–1797.

Newman, M. L., Pennebaker, J. W., Berry, D. S., and Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5):665–675.

Nguyen, T. H. and Grishman, R. (2015). Event detection and domain adaptation with convolutional neural networks. In *ACL (2)*, pages 365–371.

Okike, E. (2011). Financial reporting and fraud. In Idowu, S. O. and Louche, C., editors, *Theory and Practice of Corporate Social Responsibility*, pages 229–263. Springer Berlin Heidelberg.

Othman, I. W., Hasan, H., Tapsir, R., Rahman, N. A., Tarmuji, I., Majdi, S., Masuri, S. A., and Omar, N. (2012). Text readability and fraud detection. In *Business, Engineering and Industrial Applications (ISBEIA), 2012 IEEE Symposium on*, pages 296–301. IEEE.

Ott, M., Choi, Y., Cardie, C., and Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 309–319. Association for Computational Linguistics.

Pennebaker, J., Chung, C., Ireland, M., Gonzales, A., and Booth, R. (2007). The development and psychometric properties of LIWC2007: LIWC. net.

Pennebaker, J. W., Mehl, M. R., and Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Perols, J. (2011). Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practice & Theory*, 30(2):19–50.

Persons, O. (1995). Using financial statement data to identify factors associated with fraudulent financial reporting. *Journal of applied business research*, 82(2):38–46.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Purda, L. and Skillicorn, D. (2012). Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection. *Available at SSRN: http://ssrn.com/abstract=1670832*.

Purda, L. and Skillicorn, D. (2015). Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection. *Contemporary Accounting Research*, 32(3):1193–1223.

Purda, L. D. and Skillicorn, D. (2010). Reading between the lines: Detecting fraud from the language of financial reports. *Available at SSRN: http://ssrn.com/abstract=1670832*.

Ravisankar, P., Ravi, V., Raghava Rao, G., and Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decis. Support Syst.*, 50(2):491–500.

Razali, W. A. A. W. M. and Arshad, R. (2014). Disclosure of corporate governance structure and the likelihood of fraudulent financial reporting. *Procedia - Social and Behavioral Sciences*, 145:243 – 253.

Reporting, L. (1934). Securities exchange act of 1934: Section 12. Securities Lawyer's Deskbook. The University of Cincinnati College of Law.

Revsine, L. (1991). The selective financial misrepresentation hypothesis. *Accounting Horizons*, 5(4):16–27.

Revsine, L. (2002). Enron: sad but inevitable. *Journal of Accounting and Public Policy*, 21(2):137 – 145.

Rezaee, Z. (2005). Causes, consequences, and deterence of financial statement fraud. *Critical Perspectives on Accounting*, 16(3):277 – 298.

Rogers, R. K. and Grant, J. (1997). Content analysis of information cited in reports of sell-side financial analysts. *Journal of Financial Statement Analysis*, 3:17–31.

Russell, S. J. and Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*. Pearson Education, 2 edition.

Santorini, B. (1990). Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision).

Santos, C. N. d., Xiang, B., and Zhou, B. (2015). Classifying relations by ranking with convolutional neural networks. *arXiv preprint arXiv:1504.06580*.

Securities and Exchange Commission (2014a). Division of corporation finance: Standard Industrial Classification (SIC) code list. http://www.sec.gov/info/edgar/siccodes.htm, accessed: 2014-01-16.

Securities and Exchange Commission (2014b). EDGAR database: Filings & Forms.

Securities and Exchange Commission (2014c). Form 10-K. https://www.sec.gov/about/forms/form10-k.pdf, accessed: 2014-01-22.

Securities and Exchange Commission (2014d). Form 20-F. https://www.sec.gov/about/forms/form20-f.pdf, accessed: 2014-01-22.

Senter, R. and Smith, E. A. (1967). Automated readability index. Technical report, DTIC Document.

Shedlosky-Shoemaker, R., Sturm, A. C., Saleem, M., and Kelly, K. M. (2009). Tools for assessing readability and quality of health-related web sites. *Journal of genetic counseling*, 18(1):49.

Shen, Y., Ferdman, M., and Milder, P. (2016). Maximizing CNN accelerator efficiency through resource partitioning. *arXiv preprint arXiv:1607.00064*.

Skillicorn, D. B. and Purda, L. D. (2012). Detecting fraud in financial reports. In *2012 European Intelligence and Security Informatics Conference, EISIC 2012, Odense, Denmark, August 22-24, 2012*, pages 7–13.

Smith, M. and Taffler, R. J. (1999). The chairman's statement – a content analysis of discretionary narrative disclosures.

Spathis, C., Doumpos, M., and Zopounidis, C. (2002). Detecting falsified financial statements: a comparative study using multicriteria analysis and multivariate statistical techniques. *European Accounting Review*, 11:509–535.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Sun, C., Du, Q., and Tian, G. (2016). Exploiting product related review features for fake review detection. *Mathematical Problems in Engineering*, 2016.

Tabuchi, H. (2011). Billions lost by Olympus may be tied to criminals. *The New York Times*. 17-11-2011.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Addison-Wesley Longman Publishing Co.,Inc., Boston, MA, USA, first edition.

Tang, D., Qin, B., and Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification. In *EMNLP*, pages 1422–1432.

Tatiana Churyk, N., Lee, C.-C., and Clinton, B. D. (2008). Can we detect fraud earlier? *Strategic Finance*, 90(4):51.

Tatiana Churyk, N., Lee, C.-C., and Clinton, B. D. (2009). Early detection of fraud: Evidence from restatements. In *Advances in Accounting Behavioral Research*, pages 25–40. Emerald Group Publishing Limited.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Tennyson, B. M., Ingram, R. W., and Dugan, M. T. (1990). Assessing the information content of narrative disclosures in explaining bankruptcy. *Journal of business finance and accounting*, 17:391–410.

Throckmorton, C. S., Mayew, W. J., Venkatachalam, M., and Collins, L. M. (2015). Financial fraud detection using vocal, linguistic and financial cues. *Decision Support Systems*, 74:78–87.

Vu, N. T., Adel, H., Gupta, P., and Schütze, H. (2016). Combining recurrent and convolutional neural networks for relation classification. *arXiv preprint arXiv:1605.07333*.

Wang, B. and Wang, X. (2012). Deceptive financial reporting detection: A hierarchical clustering approach based on linguistic features. *Procedia Engineering*, 29:3392 – 3396. 2012 International Workshop on Information and Electronics Engineering.

Wang, I., Radich, R., and Fargher, N. (2011). An analysis of financial statement fraud at the audit assertion level. Technical report, Working Paper.

Wang, P. et al. (2015). Semantic clustering and convolutional neural network for short text categorization.

Wen, Y., Zhang, W., Luo, R., and Wang, J. (2016). Learning text representation using recurrent convolutional neural network with highway layers. *arXiv preprint arXiv:1606.06905*.

Wirawan, C. (2017). Cahya-wirawan/cnn-text-classification-tf. Accessed: 2017-04-10.

WRD.US (2014). SEC Filings on EDGAR. http://www.wrds.us/index.php/repository/view/25, accessed: 2014-02-06.

Yang, J.-M., Cai, R., Wang, Y., Zhu, J., Zhang, L., and Ma, W.-Y. (2009). Incorporating site-level knowledge to extract structured data from web forums. In *Proceedings of the 18th international conference on World wide web*, pages 181–190. ACM.

Yin, W., Kann, K., Yu, M., and Schütze, H. (2017). Comparative study of CNN and RNN for natural language processing. *arXiv preprint arXiv:1702.01923*.

Zhang, C., Wu, D., Sun, J., Sun, G., Luo, G., and Cong, J. (2016a). Energy-efficient CNN implementation on a deeply pipelined FPGA cluster. In *Proceedings of the 2016 International Symposium on Low Power Electronics and Design*, pages 326–331. ACM.

Zhang, R., Lee, H., and Radev, D. (2016b). Dependency sensitive convolutional neural networks for modeling sentences and documents. *arXiv preprint arXiv:1611.02361*.

Zhang, X. and LeCun, Y. (2015). Text understanding from scratch. *arXiv preprint arXiv:1502.01710*.

Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.

Zhang, Y. and Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.

Zhou, L., Burgoon, J. K., Nunamaker, J. F., and Twitchell, D. (2004a). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group decision and negotiation*, 13(1):81–106.

Zhou, L., Burgoon, J. K., Twitchell, D. P., Qin, T., and Nunamaker Jr, J. F. (2004b). A comparison of classification methods for predicting deception in

computer-mediated communication. *Journal of Management Information Systems*, 20(4):139–166.

Zhou, L., Twitchell, D. P., Qin, T., Burgoon, J. K., and Nunamaker, J. F. (2003). An exploratory study into deception detection in text-based computer-mediated communication. In *System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on*, pages 10–pp. IEEE.

News Messages

AA (2013). Auditors fail to see reports as models of communication. Accountancy Age. 28-01-2013.

AC (2013). Opmars geontegreerde verslaggeving stagneert. Accountant. 12-11-2013.

AC (2014a). Eumedion wil minder keuzevrijheid binnen ifrs. Accountant.nl. 20-01-2014.

AC (2014b). Nieuwe boekhoudregels financiele instrumenten. Accountant.nl. 29-07-2014.

AC (2015). Sec neemt 2015 gaap financial reporting taxonomie over. Accountant.nl. 10-03-2015.

AM (2016). Nba pleit voor niet-financiele informatie in het bestuursverslag. Accountancy van morgen. 18-01-2016.

AN (2013). Geintegreerde verslaggeving door europese ondernemingen. Accountancy nieuws. 28-02-2013.

AN (2014a). Duurzame verslaggeving beursgenoteerde ondernemingen. Accountancy nieuws. 17-04-2014.

AN (2014b). Richtlijn niet-financiele verslaggeving grote ondernemingen. Accountancy nieuws. 03-10-2014.

AN (2014c). Volledige harmonisatie ifrs en us gaap komt er niet. Accountancynieuws.nl. 31-07-2014.

AT (2013). Fasb makes plans for future accounting standards. AccountingToday.com. 29-07-2013.

AW (2016). Accountants: hoog tijd dat we de kwaliteit van onze rapportages verbeteren. Accountant week. 16-03-2016.

FD (2012a). Bedrijven niet scheutig met rapportage risico. Het Financiele Dagblad. 02-10-2012.

FD (2012c). Nieuw jaarverslag vertelt het echte verhaal. Het Financiele Dagblad. 15-11-2012.

FD (2013a). Aanpassing ifrs is echte oplossing. Het Financiele Dagblad. 04-01-2013.

FD (2013b). Bedrijven krijgen regels voor nieuw jaarverslag. Het Financiele Dagblad. 17-04-2013.
FD (2013c). Duurzaam verslag geeft weinig inzicht. Het Financiele Dagblad. 10-12-2013.
FD (2013d). Integrated reporting is een blijvertje. Het Financiele Dagblad. 06-04-2013.
FD (2013e). Ondernemingen verbeteren hun jaarverslagen. Het Financiele Dagblad. 14-10-2013.
FD (2013f). Revolutie jaarverslag ook voor accountants. Het Financiele Dagblad. 12-02-2013.
FD (2013g). Risico's nog onvoldoende belicht in jaarverslagen. Het Financiele Dagblad. 10-09-2013.
FD (2014a). Boekhoudregel wordt na jaren gewijzigd. Het Financiele Dagblad. 25-07-2014.
FD (2014b). De boekhoudregel van de crisis is eindelijk aangepast. Het Financiele Dagblad. 28-07-2014.
FD (2014c). Digitale beveiliging hoort in het jaarverslag van bedrijven. Het Financiele Dagblad. 08-07-2014.
FD (2014d). Geintegreerde verslaggeving ook zonder wettelijke verplichting dwingend. Het Financiele Dagblad. 02-01-2014.
FD (2015). Duurzaamheid krijgt aandacht in rapportages. Het Financiele Dagblad. 14-09-2015.
FD (2016a). Alternatief jaarverslag is toe aan regels. Het Financiele Dagblad. 18-11-2016.
FD (2016b). Beperk jaarverslag tot zinvolle informatie en doe dat beter. Het Financiele Dagblad. 07-03-2016.
FT (2013). Sec to roll out 'robocop' against fraud. The Financial Times. 14-02-2013.
FT (2014). Ifrs accounting rules change forces banks to alter view of losses. Financial Times. 25-07-2014.
NU (2016). Nieuwe boekhoudregels moeten vage jaarcijfers eenduidiger maken. Nu.nl. 08-03-2016.
VK (2013). Brussel eist opener jaarverslag. De Volkskrant. 17-04-2013.

Biography

Marcia Fissette was born on May 31, 1986 in Maastricht, The Netherlands. After finishing high school in 2004, she obtained a bachelor's degree in Communication and Multimedia Design at the Zuyd University of Applied Sciences in 2007. Subsequently, she studied Artificial Intelligence at the Radboud University Nijmegen. After obtaining the bachelor's degree (cum laude) in 2010, she went to the University of Amsterdam to obtain the master's degree in Forensic Science in 2012. Marcia completed the graduation project for Forensic Science at KPMG's Forensic Technology department. In this project she developed a method that automatically extracts and visualizes the series of transactions in the general ledger. After finishing the project Marcia continued working at KPMG and started the PhD research described in this thesis at the University of Twente.

Marcia enjoys researching data analysis and machine learning models to provide insights in data. In particular, she likes to examine the possibilities of text mining to solve practical problems. The first text mining research she conducted concerned the identification of authors of short texts. Her interest in text mining has only increased since then. The research described in this thesis provided the opportunity to combine her interests in text mining and fraud detection. Marcia wants to continue the deployment and development of text mining models.

Acknowledgments / Dankwoord

Allereerst wil ik KPMG bedanken voor het bieden van de mogelijkheid om dit PhD onderzoek uit te voeren. Voor mij heeft KPMG hiermee laten zien zeer open te staan voor ontwikkeling van nieuwe methoden. Ik ben er dan ook trots op dat ik dit onderzoek bij KPMG, en in het bijzonder Forensic Techonology, heb kunnen doen.

Mijn promotor Bernard Veldkamp wil ik graag bedanken voor al zijn ideeën, adviezen en steun. Na een bespreking met jou had ik altijd hernieuwde energie om verder met het onderzoek aan de slag te gaan. In het bijzonder ben ik je dankbaar voor het vertrouwen dat je al die vijf jaar in mijn vaardigheden en beslissingen hebt gehad. Ik vond het erg prettig met je samen te werken en hoop dat we dat in de toekomst weer eens kunnen doen.

Mijn andere promotor, Theo de Vries, bedankt voor al jouw kennis, ervaring en vooral enthousiasme! Dit werkte voor mij aanstekelijk en inspirerend. Jouw steun van begin tot eind wordt erg gewaardeerd.

Daarnaast wil ik de voorzitter en leden van de promotiecommissie bedanken voor hun interesse in dit onderzoek. Prof. Dr. T.A.J. Toonen, Prof. Dr. C.A.W. Glas, Prof. Dr. M. Junger, Prof. Dr. A. Shahim, Prof. Dr. M. Pheijffer, Prof. Dr. R.G.A. Fijneman en Dr. R. Matthijsse, bedankt voor jullie kennis en tijd.

Ook mijn collega's en voormalig collega's van KPMG wil ik allen bedanken. Jullie hebben direct, en anders indirect, bijgedragen aan dit onderzoek. Jullie vragen en ideeën naar aanleiding van mijn tussentijdse presentaties of gewoon tijdens het werk, de lunch of borrels waren zeer welkom.

Van de Universiteit Twente wil ik in het bijzonder Sytske Wiegersma bedanken voor haar hulp bij het verkrijgen van de LIWC features. Anneke Schools, bedankt voor het organiseren van het text analysis café. Het was een eer om als eerste een presentatie te mogen geven. Qiwei He, thank you for your research and thesis which were an inspiration throughout my research.

Aukje, Danielle, Lon, Inge, Michelle, Melanie, Paula, Pauline, Regina en Suus dank jullie wel voor alle niet aan mijn onderzoek gerelateerde activiteiten. Ik weet zeker dat jullie afleiding heeft bijgedragen aan het tot stand komen van dit proefschrift. Paula en Pauline, bedankt dat jullie mijn paranimfen willen zijn.

Mijn ouders, zusjes en schoonouders dank ik voor hun vertrouwen in mij. Ondanks dat jullie mijn onderzoek niet precies kenden hebben jullie er nooit aan getwijfeld dat ik het succesvol zou afronden.

Ten slotte wil ik mijn man, Ivo, bedanken voor al zijn steun, vertrouwen en geduld. Je hebt de belofte 'in voor- en tegenspoed al meer dan waar gemaakt. Dankjewel dat je achter mijn keuzes staat en trots en enthousiast bent over mijn onderzoek, ook als je aangeeft de details te ingewikkeld te vinden. Daarnaast natuurlijk ook bedankt voor jouw hulp bij het maken van de omslag van dit proefschrift.

Summary

To catch fraudsters and limit the damage caused by financial fraud, innovative fraud detection methods are required that are capable of identifying indications of the continuously changing fraudulent activities. The research described in this thesis examined the contribution of text analysis to detecting indications of fraud in the annual reports of companies worldwide. In an annual report, a company presents its financial results and activities from the previous year. In addition, the annual report contains textual explanations. In case of fraud, the annual report does not provide a fair view of the financial position of the company.

For the research described in this thesis, a total of 1,727 annual reports have been collected, of which 402 are of the years and companies in which fraudulent activities took place, and which have an impact on the information disclosed in the annual report. Since it is assumed that most companies do not engage in such fraudulent activities, the majority, 77%, of the annual reports in the data set are not fraudulent. Furthermore, the composition of this data set takes into account the possibility that the year the annual report is concerned with, the sector in which the company operates and the size of the company may affect the information disclosed in the annual report. Therefore, for each fraud annual report, non fraudulent annual reports from the same year, of a company in the same sector, and of a comparable size are added to the data set.

A method for the automatic extraction of information from annual reports has been proposed to obtain the data needed to compile the data set. By applying this method, the year, sector and size of the company can be determined for a large number of annual reports. The approach has also been used to extract the Management Discussion & Analysis (MD&A) section, which is the part of the annual report the research in this thesis focuses on. In the MD&A section, the company provides information concerning its performance and the activities of the preceding year and the expectations for the following year.

The first models developed for the research described in this thesis, analyze the texts by counting the words (unigrams) in the MD&A section. These word counts are normalized by the term frequency-inverse document frequency (TF-

IDF) method that takes into account the length of the text and the frequency with which a word occurs in the data set. The most informative words are determined using the chi-square feature selection model. This representation of the text is the input of the machine learning algorithms Naive Bayes (NB) and Support Vector Machine (SVM). These algorithms learn patterns to classify the annual reports as 'fraud' or 'no fraud'. The NB model classifies the texts based on probability calculations. The SVM model uses unigrams to create a vector space. Subsequently, the SVM algorithm determines the most optimal hyperplane in the space that separates the 'fraud' and 'no fraud' texts. The NB model shows the best performance. The percentage of correctly classified annual reports is 89%.

Subsequently, the NB and SVM models based on unigrams are expanded with the linguistic features of the text found to be informative in the previous research concerning the detection of fraud or deception in text. We subdivided the linguistic features into six categories. The first category comprises groups of two consecutive words (bigrams). The second category consists of features that describe the general properties of the text, such as the total numbers in the text. The third category describes the complexity of the text using measures for the complexity of the words and sentences. The fourth category focuses on the grammatical aspects of the text. The fifth category consists of measures that determine the readability of a text, expressed as the number of years of education needed to understand the text. The final category concerns, the Linguistic Inquiry and Word Count (LIWC) tool that is used to extract psychological features, such as emotions, from the text. Furthermore, we developed a new type of feature that reflects the grammatical relations between words. The classification results show that only the addition of bigrams to the SVM model improves the result slightly. The other categories of linguistic features do not improve the result.

The latest development in machine learning is the use of deep learning. By using networks consisting of several layers complex patterns can be found. A Convolutional Neural Network (CNN) model has been found successful in research that classifies texts in domains other than fraud. With this model, each word in the text is represented by a vector (word embedding). Word embeddings aim to include the semantic relationships between words, in addition to the representation of the individual words. Due to the limited computer capacity available during the research described in this thesis, experimentation with the CNN model was performed using 40% of the original data set. In order to compare the results with the previous models, an NB model was also developed on this smaller data set. The results are significantly lower, but

also show that the CNN model achieves slightly better results than the NB model.

The results show that text analysis can contribute to the detection of indications of fraud. However, it desirable to further enhance the performance of the models. This may be achieved either by improving their performance or by adding additional sources of information. Future research may experiment with machine learning algorithms, such as the CNN or a combination of various algorithms. The model for text can be expanded with models that use the company's financial information. More textual information may be added to the model from the annual report itself or from other documents of the company.

Samenvatting

Om fraudeurs te kunnen pakken en de enorme schade als gevolg van financiële fraude te beperken, zijn innovatieve fraudedetectie methoden nodig die aanwijzingen van de continu veranderende frauduleuze activiteiten kunnen identificeren. Het onderzoek beschreven in dit proefschrift onderzocht de mogelijkheid van de bijdrage van tekstanalyse voor het detecteren van indicaties van fraude in jaarverslagen van bedrijven van over de hele wereld. In een jaarverslag presenteert een bedrijf zijn financiële resultaten en activiteiten van het voorgaande jaar. Naast de financiële informatie bevat een jaarverslag tekstuele toelichtingen. Indien er sprake is van fraude geeft het jaarverslag geen juiste weergave van de financiële positie van het bedrijf.

Voor het onderzoek beschreven in dit proefschrift zijn in totaal 1.727 jaarverslagen verzameld, waarvan 402 jaarverslagen van jaren en bedrijven zijn waarin fraude plaats vond met een omvang die de informatie in het jaarverslag beïnvloed heeft. Omdat wordt aangenomen dat de meeste bedrijven zich niet bezig houden met dergelijke frauduleuze activiteiten is het merendeel, 77%, van de jaarverslagen in de data set niet frauduleus. Bij de samenstelling van deze data set is daarnaast rekening gehouden met de mogelijkheid dat de tijdsperiode waarop het jaarverslag betrekking heeft, de sector waarin het bedrijf opereert, en de omvang van het bedrijf invloed kunnen hebben op de informatie die wordt gepubliceerd in het jaarverslag. Voor elk frauduleuze jaarverslag zijn er jaarverslagen zonder fraude uit het zelfde jaar, van een bedrijf in dezelfde sector, en met een vergelijkbare omvang als het frauduleuze jaarverslag toegevoegd aan de data set.

Een methode voor de automatische extractie van informatie uit jaarverslagen is ontwikkeld om de gegevens te verkrijgen die nodig zijn voor het samenstellen van de data set. Door middel van deze methode kan voor een grote hoeveelheid jaarverslagen het jaar, de sector en de omvang van het bedrijf worden bepaald. De aanpak is ook gebruikt om de Management Discussie & Analyse (MD&A) sectie, waarop het onderzoek in dit proefschrift zich richt, te extraheren. In de MD&A sectie geeft het bedrijf informatie over de resultaten en activiteiten van het voorgaande jaar en de verwachtingen voor het komende jaar.

De eerste modellen die ontwikkeld zijn voor het onderzoek beschreven in dit

proefschrift, analyseren de teksten door middel van het tellen van de woorden (unigrams) in de MD&A sectie. Deze tellingen worden genormaliseerd met de term frequency–inverse document frequency (TF-IDF) methode die rekening houdt met de lengte van de tekst en de frequentie waarmee een woord voor komt in de hele data set. De meest informatieve woorden worden bepaald door middel van de Chi–kwadraat statistiek. Deze representatie van de tekst vormt de input van de machine learning algoritmes Naive Bayes (NB) en Support Vector Machine (SVM). Deze algoritmes leren patronen om de jaarverslagen te kunnen classificeren als 'fraude' of 'geen fraude'. Een NB model classificeert de teksten op basis van kansberekening. Een SVM model gebruikt de unigrams om een vectorruimte te maken. Vervolgens bepaalt het SVM algoritme een scheiding in de ruimte die de 'fraude' en 'niet fraude' teksten zo optimaal mogelijk van elkaar scheidt. Het NB model laat de beste resultaten zien. Het percentage goed geclassificeerde jaarverslagen is 89%.

De NB en SVM modellen op basis van unigrams zijn vervolgens uitgebreid met linguïstische kenmerken uit de tekst. Hiervoor zijn kenmerken gebruikt die in eerdere onderzoeken naar de detectie van fraude of deceptie in tekst informatief werden bevonden. Deze kenmerken kunnen worden onderverdeeld in zes categorieën. De eerste categorie bestaat uit groepjes van twee op elkaar volgende woorden (bigrams). De tweede categorie bestaat uit kenmerken die de tekst globaal beschrijven, zoals het totaal aantal worden in de tekst. De derde categorie beschrijft de complexiteit van de tekst aan de hand van maten voor de complexiteit van de woorden en zinnen. De vierde categorie focust op de grammaticale aspecten van de tekst. De vijfde categorie bestaat uit maten die weergeven hoe leesbaar de tekst is gemeten in het aantal jaren educatie dat nodig is om de tekst te kunnen begrijpen. Ten slotte zijn door middel van de Linguistic Inquiry and Word Count (LIWC) tool psychologische kenmerken, zoals emoties, in de tekst vastgesteld. Daarnaast hebben wij een nieuw type kenmerk ontwikkeld die de grammaticale relaties tussen woorden weergeeft. De classificatie resultaten laten zien dat alleen het toevoegen van bigrams aan SVM model het resultaat iets verbetert. De andere categorien linguïstische kenmerken verbeteren het resultaat niet.

De nieuwste ontwikkeling in machine learning is het gebruik van deep learning. Door middel van netwerken bestaande uit verschillende lagen kunnen complexe patronen worden gevonden. Een Convolutional Neural Network (CNN) model is in eerder onderzoek succesvol bevonden voor de classificatie van teksten in andere domeinen dan fraude. In een dergelijke CNN model wordt elk woord in de tekst gerepresenteerd door een vector (word embeddings). Word embeddings zouden, naast de representatie van de individuele woorden, ook

de semantische relaties tussen woorden omvatten. Door de beperkte computer capaciteit die beschikbaar was tijdens het onderzoek beschreven in deze thesis, is met het CNN model geëxperimenteerd met een 40% van de oorspronkelijke data set. Om de resultaten te kunnen vergelijken met de eerdere modellen is ook een NB model ontwikkeld op deze kleinere data set. De resultaten zijn aanmerkelijk lager, maar laten ook zien dat het CNN model iets betere resultaten verkrijgt dan het NB model.

De resultaten tonen zonder meer aan dat tekstanalyse kan bijdragen aan de detectie van indicaties van fraude. Toch blijft het wenselijk om het model verder te versterken. Dit kan door verbetering of door extra informatiebronnen toe te voegen. Experimenten met machine learning algoritmes zoals CNN of een combinatie van verschillende algoritmes lijken in dit verband kansrijk. Het model voor tekst kan uitgebreid worden door met te ontwerpen modellen ook te kijken naar financiële gegevens van het bedrijf. Tenslotte kan aan het model meer tekstuele informatie toegevoegd worden uit het jaarverslag zelf of uit andere bedrijfsdocumenten.