Craja, Patricia; Kim, Alisa; Lessmann, Stefan

Working Paper

Deep Learning application for fraud detection in financial statements

IRTG 1792 Discussion Paper, No. 2020-007

Provided in Cooperation with:
Humboldt University Berlin, International Research Training Group 1792 "High Dimensional Nonstationary Time Series"

Suggested Citation: Craja, Patricia; Kim, Alisa; Lessmann, Stefan (2020) : Deep Learning application for fraud detection in financial statements, IRTG 1792 Discussion Paper, No. 2020-007, Humboldt-Universität zu Berlin, International Research Training Group 1792 "High Dimensional Nonstationary Time Series", Berlin

This Version is available at:
https://hdl.handle.net/10419/230813

IRTG 1792 Discussion Paper 2020-007

Deep Learning application for fraud detection in financial statements

Patricia Craja *
Alisa Kim *
Stefan Lessmann *

* Humboldt-Universität zu Berlin, Germany

This research was supported by the Deutsche Forschungsgesellschaft through the International Research Training Group 1792 "High Dimensional Nonstationary Time Series".

http://irtg1792.hu-berlin.de
ISSN 2568-5619

International Research Training Group 1792

Deep Learning application for fraud detection in financial statements

Patricia Craja[a], Alisa Kim[a], Stefan Lessmann[a]

[a]*School of Business and Economics, Humboldt University of Berlin, Berlin, Germany*

Abstract

Financial statement fraud is an area of significant consternation for potential investors, auditing companies, and state regulators. Intelligent systems facilitate detecting financial statement fraud and assist the decision-making of relevant stakeholders. Previous research detected instances in which financial statements have been fraudulently misrepresented in managerial comments. The paper aims to investigate whether it is possible to develop an enhanced system for detecting financial fraud through the combination of information sourced from financial ratios and managerial comments within corporate annual reports. We employ a hierarchical attention network (HAN) with a long short-term memory (LSTM) encoder to extract text features from the Management Discussion and Analysis (MD&A) section of annual reports. The model is designed to offer two distinct features. First, it reflects the structured hierarchy of documents, which previous models were unable to capture. Second, the model embodies two different attention mechanisms at the word and sentence level, which allows content to be differentiated in terms of its importance in the process of constructing the document representation. As a result of its architecture, the model captures both content and context of managerial comments, which serve as supplementary predictors to financial ratios in the detection of fraudulent reporting. Additionally, the model provides interpretable indicators denoted as "red-flag" sentences, which assist stakeholders in their process of determining whether further investigation of a specific annual report is required. Empirical results demonstrate that textual features of MD&A sections extracted by HAN yield promising classification results and substantially reinforce financial ratios.

Keywords: fraud detection, financial statements, deep learning, text analytics

[*]Declaration of interest: none

Email addresses: patricia.craja@gmail.com (Patricia Craja), kolesnal@hu-berlin.de (Alisa Kim), stefan.lessmann@hu-berlin.de (Stefan Lessmann)

1. Introduction

Fraud is a global issue that concerns a variety of different businesses, with a severe negative impact on the firms and relevant stakeholders. The financial implications of fraudulent activities occurring globally in the past two decades are estimated to amount up to $5.127 trillion, with associated losses increasing by 56% in the past ten years [21]. Nevertheless, the actual costs of fraud are potentially greater, particularly if one also considers the indirect costs, including harm to credibility and the reduction in business caused by the resultant scandal.

The Association of Certified Fraud Examiners (ACFE), the world's largest anti-fraud organization, recognizes three main classes of fraud: corruption, asset misappropriation, and fraudulent statements [69]. All three have specific properties and successful fraud detection requires comprehensive knowledge of their particular characteristics. This study concentrates on financial statement fraud and adheres to the definition of fraud proposed by Nguyen [51], who stated that it is "the material omissions or misrepresentations resulting from an intentional failure to report financial information in accordance with generally accepted accounting principles". For this study, the terminology "financial statement fraud", "fraudulent financial reporting", and "financial misstatements" are used interchangeably and are distinguished from different factors that cause misrepresentations within financial statements, such as unintended mistakes.

The Center for Audit Quality indicted that managers commit financial statement fraud for a variety of reasons, such as personal benefit, the necessity to satisfy short-term financial goals, and the intention to hide bad news. Fraudulent financial statements can be manipulated so that they bear a convincing resemblance to non-fraudulent reports, and they can emerge in various distinct types [31]. Examples of frequently used methods are net income over- or understatements, falsified or understated revenues, hidden or overstated liabilities and expenses, inappropriate valuations of assets, and false disclosures [69]. Authorities directly reacted to the increased prevalence of corporate fraud by adopting new standards for accounting and auditing. Nevertheless, financial statement irregularities are frequently observed and complicate the detection of fraudulent instances.

Detecting financial statement abnormalities is regarded as the duty of the auditor [18]. Despite the existing guidelines, the detection of indicators of fraud can be challenging. A 2018 report revealed that only a limited number of cases of fraud were identified by internal and external auditors, with rates of 15% and 4%, respectively [2]. Hence, there has been an increased focus on automated systems for the detection of financial statement fraud [74]. Such systems have specific importance for all groups of stakeholders: for investors - to facilitate qualified decisions, for auditing companies - to speed up and improve the accuracy of the audit, and for state regulators - to concentrate their

investigations more effectively [1, 4]. Therefore, efforts have been made to develop smart systems designed to detect financial statement fraud to generate early warning indicators (red-flags) that facilitate stakeholders' decision-making processes. We aim to contribute to the development of decision support systems for fraud detection by offering a state-of-the-art deep learning model for screening submitted reports based on a combination of financial and textual data. The proposed method exhibits superior predictive performance and allows the identification of "red-flags" on both the word- and sentence-level for the facilitation of the audit process. Additionally, we showcase the results of comparative modeling on different data types associated with financial reports and offer the alternative performance metrics that are centered around the cost imbalance of miss-classification errors.

2. Research design and contributions

In line with the above goals, we pose three research questions (RQ) that frame our research:

- **RQ 1**: What is the most informative data type for fraud detection? Can it benefit from the novel combination of financial and text data (FIN+TXT)?
- **RQ 2**: Can a state-of-the-art deep learning (DL) model be developed, that can detect indications of fraud from the textual information contained in financial statements? If yes, how effective does the DL approach perform as compared to the bag-of-words (BOW) approach for textual feature extraction in combination with quantitative financial features?
- **RQ 3**: In addition to predictive performance, can the proposed DL model assist in interpreting textual features signaling fraud? Given that the Hierarchical Attention Network (HAN) provides both word and sentence-level interpretation, is it possible to derive preliminary judgment on what level of granularity is more informative for practical application?

To determine answers to these research questions, we select an array of classification models and task them to perform fraud detection on different combinations of data. The choice is based on previous studies and recently developed methods that proved efficient for similar classification tasks. The classic statistical models include Logistic Regression (LR), Support Vector Machine (SVM) and Random Forest (RF), from recent Machine Learning (ML) models Extreme gradient Boosting (XGB) and Artificial Neural Network (ANN) algorithms were selected. Additionally, a novel DL method named Hierarchical Attention Network (HAN) is offered for consideration and advocated to be the most efficient fraud statement classifier. In line with previous research, this paper concentrates on the MD&A sections of annual reports filed by firms within the United States with the Securities and Exchange Commission (SEC), which are referenced as annual reports on form 10-K. All selected models are

trained on five different combinations of data types contained in the statements submitted for audit: financial indicators (FIN), linguistic features (LING) of an MD&A text, financial and linguistic features (FIN + LING), the full text of an MD&A (TXT), full text and the financial indicators (FIN + TXT). We compare the predictive performance of the models with regard to their ability to distinguish fraud cases, for which we use traditional metrics like Accuracy and area under the Receiver Operating Curve (AUC). We also provide the analysis of metrics that account for the error cost imbalance, namely Sensitivity, F1-score, and F2-scores. This allows us to bring the existing state of research closer to the industrial setting. The provided analysis contributes to the field of fraud detection not only with the comparative study insights but offers previously unexplored data combinations and new DL methods, displaying superior results and additional interpretative features.

Following the RQ 3, we offer a novel fraud detection method that provides signaling tools for scholars and practitioners. We perform a comparative analysis of words considered "red-flags" by the RF feature importance method and the HAN attention layer output. We argue that the use of words for signaling may be the subject of manipulation and offer a remedy in the shape of sentence-level importance indicators. We further demonstrate how the latter may be applied for decision support in the audit process.

3. Decision support for fraud detection

Previous studies proposed fraud detection systems and offered systematic literature reviews on fraud detection approaches [75, 57]. Table 1 depicts the status-quo in the field of financial fraud detection along four dimensions: the technique utilized, the type of data, the country of study, and the predictive performance in terms of classification accuracy and other metrics. Much research focused on the financial variables and applied a wide range of modeling techniques, from LR to DL. Several authors experimented with linguistic variables, however, the majority of those have solely examined the relation between linguistic aspects and fraudulent actions. Only Hajek and Henriques [27] combined them with financial data and showed that although financial variables are essential for the detection of fraud, it is possible to enhance the performance through the inclusion of linguistic data. At least two attempts to apply natural language processing (NLP) techniques focusing on the textual content have been undertaken. Nevertheless, to the best of our knowledge, this is the first study to apply DL models that allow for contextual information to be extracted from the text.

Study	Data (fraud / no fraud)	Country	Features	Classifiers	Used metrics
Hajek and Henriques [27]	311/311	US	FIN+ LING	BBN(90.3), DTNB(89.5), RF(87.5), Bag(87.1), JRIP(87.0), CART(86.2), C4.5(86.1), LMT(85.4), SVM(78.0), MLP(77.9), AB(77.3), LR(74.5), NB(57.8)	Acc, TPR, TNR, MC, F-score, AUC
Kim et al. [37]	788/2156	US	FIN	LR (88.4), SVM (87.7), BBN (82.5)	Acc, TPR, G-mean, Cost Matrices
Goel and Uzuner [25]	180/180	US	LING+POS tags	SVM(81.8)	Acc,TPR,FPR,Precision,F-score
Purda and Skillicorn [59]	1407/4708	US	TXT (BOW), top 200 RF words	SVM (AUC 89.0)	AUC, Fraud Probability
Throckmorton et al. [72]	41/1531	US	FIN+ LING from Conference Calls	GLRT (AUC 81.0)	AUC
Goel and Gangolly [23]	405/622	US	LING	χ^2 statistics	χ^2 statistics
Dechow et al. [15]	293/79358	US	FIN	LR(63.7)	Acc, TPR, FPR, FNR, min F-Score
Humpherys et al. [32]	101/101	US	LING	C4.5 (67.3), NB (67.3), SVM (65.8)	Acc, Precision, Recall, F-score
Glancy and Yadav [22]	11/20	US	TXT (BOW)	hierarchical clustering (83.9)	TP, TN, FP, FN, p-value
Perols [54]	51/15934	US	FIN	SVM(MC 0.0025), LR(0.0026), C4.5 (0.0028), bagging(0.0028), DNN(0.0030)	Fraud Probability and MC
Cecchini et al. [11]	61/61	US	LING	SVM (82.0)	AUC, TPR, FPR, FNR
Goel et al. [24]	126/622	US	LING+TXT (BOW)	SVM(89.5), NB(55.28%)	Acc, TPR, FPR, Precision, F-score
Lin et al. [43]	127/447	Taiwan	FIN	DNN (92.8), CART (90.3), LR (88.5)	Acc, FPR, FNR, MC
Ravisankar et al. [63]	101/101	China	FIN	PNN (98.1), GP (94.1), GMDH (93.0), DNN (78.8), SVM (73.4)	Acc, TPR, TNR, AUC

Table 1: Analysis of classifier comparisons in financial statement fraud detection

FIN – financial data, LING – linguistic data (word category frequency counts, readability and complexity scores, etc.), TXT - text data, BOW–bag-of-words, POS – part of speech tags (nouns, verbs, adjectives), BBN – Bayesian belief network, NB-Naive Bayes, DTNB - NB with the induction of decision tables, CART–classification and regression tree, LMT - logistic model trees, MLP - multi-layer network, Bag-Bagging, AB - AdaboostM1, GMDH – group method data handling, GP–genetic programming, GLRT - generalized likelihood ratio test, LR–logistic regression, DNN – deep neural network, PNN – probabilistic neural network, RF – random forest, SVM – support vector machine, Acc - Accuracy, AUC – area under the ROC curve, MC - misclassification cost, TPR - true positive rate, TNR - true negative rate, FPR - false-positive rate, FNR - false-negative rate.

The majority of existing research measured performance in terms of accuracy. Some studies also considered precision and recall. Additionally, most of the reported studies neglected the interpretability which is a crucial aspect to facilitate decision support for fraud detection. This paper adds to the literature by offering an integrated approach for processing both textual and financial data using interpretable state-of-the-art DL methods. Furthermore, we provide a comprehensive evaluation of different modeling techniques using cost-sensitive metrics to account for the different severities of false alarms versus missed fraud cases.

3.1. Text-based indicators

Textual analysis is frequently employed for the examination of corporate disclosures. Linguistic features have been utilized in the analysis of corporate conference calls [39], earnings announcements [14], media reports [71] and annual reports [45, 10]. Multiple researchers have specifically concentrated on the MD&A section to examine the language used in annual reports [19, 11, 32]. The MD&A has a particular relevance as it offers investors the possibility of reviewing the performance of the company as well as its future potential from the perspective of management. This part also provides scope for the management's opinions on the primary threats to the business and necessary actions. It is interesting to note that as suggested by social psychology research, the emotions and cognitive processes of managers who intend to conceal the real situation could indicate specific linguistic cues that can facilitate the identification of fraud [16]. Therefore studies have emphasised the increasing significance of textual analysis of financial documentation.

As stated in Li [41] literature review, research that analyzes the use of language within annual reports usually adopts one of two strategies. The first strategy is primarily based on past research into linguistics and psychology and is dependent on pre-determined lists of words that have an association with a specific sentiment, like negativity, optimism, deceptiveness, or ambiguity. Loughran and Mcdonald [45] (L&M) demonstrated that if these lists are adapted to the financial domain, it is possible to determine relationships among financial-negative, financial-uncertain, and financial-litigious word lists and 10-k filing returns, trading volume, return volatility, fraud, material weakness, and unexpected earnings. As it was developed for analyzing 10-K text, the L&M sentiment word lists have been broadly employed in fraud-detection research [27]. This research applies the L&M word lists for the extraction of sentiment features from the MD&A section of 10-Ks in the benchmark models. Other researchers based their approaches for detecting fraud on word lists that indicate positive, negative or neutral emotions [32, 23] or more specifically anger, anxiety, and negativity according to the definitions supplied by the Linguistic Inquiry and Word Count dictionary [32, 39, 52].

The second strategy relies on ML to extract informative features for automatic differentiation be-

tween fraudulent and non-fraudulent texts. Li [41] contended that this method has various benefits compared with predetermined lists of words and cues, including the fact that no adaptation to the business context is required. ML algorithms have been used in the detection of financial statement fraud by various researchers, such as Cecchini et al. [11], Hajek and Henriques [27], Humpherys et al. [32], Goel and Uzuner [25], Goel et al. [24], Glancy and Yadav [22], and Purda and Skillicorn [59]. Some attempts to integrate different types of data have also been made. Purda and Skillicorn [59] compared a language-based method to detect fraud based on SVM to the financial measures proposed by Dechow et al. [15], and concluded that these approaches are complementary. The methods displayed low forecast correlation and identified specific types of fraud that the other could not detect. This finding motivates the present research to combine Dechow et al. [15] financial variables with linguistic variables to complement each other in the detection of fraud in financial statements.

The study of Hajek and Henriques [27] is closest to this work as they combined financial ratios with linguistic variables from annual reports of US firms and employed a variety of classification models, as shown in Table 1. Despite these similarities, the study by Hajek and Henriques [27] was not targeted at evaluating the textual content of corporate annual reports. Hence, it did not include modern NLP approaches such as deep learning-based feature extraction.

3.2. Methods and evaluation metrics

Prior work has tested a variety of statistical fraud detection models including ANNs, Decision Trees (DT), SVM, evolutionary algorithms, and text analysis [27]. The BOW technique was frequently adopted for the extraction of the linguistic properties of financial documentation. The BOW approach represents a document by a vector of word counts that appear in it. Consequently, the word frequency is used as the input for the ML algorithms. This method does not consider the grammar, context, and structure of sentences and could be overly simple in terms of uncovering the real sense of the text [39]. A different technique for analyzing text is DL. Deep ANN are able to extract high-level features from unstructured data automatically. Textual analysis models based on DL can "learn" the specific patterns that underpin the text, "understand" its meaning and subsequently output abstract aspects gleaned from the text. Hence, they resolve some of the problems associated with the BOW technique, including the extraction of contextual information from documents. Due to their capacity to deal with sequences with distinct lengths, ANN have shown excellent results in recent studies on text processing. Despite their achievements in NLP, there has been limited focus on the application of state-of-the-art DL methods to the analysis of a financial text. For an effective adoption in practice, the models should not only be precise, but also interpretable [31]. However, the majority of systems designed to detect fraud reported by researchers aim to maximise the prediction accuracy, while dis-

regarding how transparent they are [27]. This factor has particular significance as the development of interpretable models is critical for supporting the investigation procedure in auditing.

4. Data

Fraud detection is a challenging task because of the low number of known fraud cases. A severe imbalance between the positive and the negative class impedes classification. For example, the proportion of statements that were fraudulent and non-fraudulent in the annual reports submitted to the SEC for the period from 1999 to 2019 was 1:250. In past research, the number of firms that committed fraud contained in the data varied between 12 and 788 [72, 37]. The data used here consists of 208 fraudulent and 7 341 non-fraud cases, making it the most significant data set with a textual component so far (c.f., Table1).

The data set consists of US companies' annual financial reports, referred to as 10-K filings, that are publicly available through the EDGAR database of the SEC's website [1] and quantitative financial data, sourced from the Compustat database [2].

4.1. Labeling

Companies submit yearly reports that undergo an audit. Labeling these reports requires several filtering decisions: when can a report be considered fraudulent and what type of fraud should we consider. To address the first question, we follow the approach of Purda and Skillicorn [59], Humpherys et al. [32], and Hajek and Henriques [27], and consider a report as "fraudulent" if the company that filed it was convicted. The SEC - a source widely used by the previous research [26] - publishes statements, referred to as the "Accounting and Auditing Enforcement Releases" (AAER) that describe financial reporting related enforcement actions taken against companies that violated the reporting rules [3]. SEC concentrates on the cases with the highest importance [36] and applies enforcement actions where the evidence of manipulation is sufficiently robust [27], which provides a high degree of trust to this source. Labeling reports based on the AAER offers simplicity and consistency with easy replication, allowing to avoid possible bias related to subjective categorization. Following the filtering criteria offered by Purda and Skillicorn [59], we select the AAERs concerning litigations issued during the period from 1999 to 2019 with identified manipulation instances between the year 1995 and 2016 that discuss the words "fraud", "fraudulent", "anti-fraud" and "defraud" as well as "annual

[1] The SEC is the preeminent financial supervisory organisation that is responsible for monitoring financial reports of firms listed on the US stock exchange: www.sec.gov/edgar.shtml

[2] Compustat is a database of financial, statistical and market information on companies throughout the world

[3] https://www.sec.gov/divisions/enforce/friactions.shtml

reports" or "10-K". Addressing the second question, we follow Cecchini et al. [11], Goel et al. [24], Humpherys et al. [32], Purda and Skillicorn [59], and Hajek and Henriques [27] and focus on binary fraud classification. This implies that we do not distinguish between truthful and unintentionally misstated annual reports. The resulting data set contains 187 869 annual reports filed between 1993 and 2019, with 774 firm-years subject to enforcement actions. However, due to missing entries and mismatches in existing CIK indexation, the final data set is reduced to 7 757 firm-year observations with 208 fraud and 7 549 non-fraud filings. Further, we perform the extraction of text and financial data.

4.2. Text data

The retrieved reporting forms 10-K [68] contain the MD&A section. The segment commonly called "Management's Discussion and Analysis of Financial Condition and Results of Operations" (Item 7) constitutes the primary source of raw text data. In addition, nine linguistic features are utilized as predictors (described in the online appendix). The selection of these features is influenced by the past studies, that demonstrated several patterns of fraudulent agents, like an increased likelihood of using words that indicate negativity [72, 50], absence of process ownership implying lack of assurance, thus resulting in statements containing less certainty [39] or an average of three times more positive sentiment and four times more negative sentiment in comparison to honest reports Goel and Uzuner [25]. Additionally, the general tone (sentiment) and the proportion of constraining words, were included by Hajek and Henriques [27], Loughran and Mcdonald [45], and Bodnaruk et al. [7]. Lastly, the average length of sentence, the proportion of compound words, and fog index are incorporated as measures of complexity and legibility and calculated based on formulas presented by Humpherys et al. [32] and Li [40], who concluded that reports produced by misstating firms had reduced readability.

4.3. Quantitative data

Along with text features, we used 47 quantitative financial predictors (described in the online appendix), which are capable of capturing financial distress as well as managerial motivations to misrepresent the performance of the firm. Past studies have presented robust theoretical evidence supporting the utilization of financial variables [20, 66, 1, 27]. Following the guidelines of existing research, the financial ratios and balance sheet variables presented in the online appendix are extracted from Compustat, based on formulas presented by Dechow et al. [15] and Beneish [6]. Financial variables include indicators like total assets (adopted as a proxy for company size [72, 5]), profitability ratios [27], accounts receivable and inventories as non-cash working capital drivers [1, 11, 55]. Additionally, a reduced ratio of sales general and administrative expenses (SGA) to revenues (SGAI) is found

to signalize fraud [1]. Missing values are imputed using the RF algorithm. However, observations with more than 50% of the variables missing are excluded.

4.4. Imbalance treatment

The majority of previous research has balanced the fraud and non-fraud cases in a data set using undersampling [27, 32, 63]. We follow this approach and consider a fraud-to-non-fraud-ratio of 1:4, which reflects the fact that the majority of firms have no involvement in fraudulent behaviour. Both year and sector are utilized for balancing, in order to take into account different economic conditions, change in regulation, as well as to eradicate any differences across distinct sectors [32, 37]. The latter is extracted with the SIC code [67] and is of particular importance for text mining, as the utilization of words within financial documentation could differ according to the sector. The resulting balanced data set consists of 1 163 reports, out of which 201 are fraudulent, and 962 are non-fraudulent annual reports.

In the years 2002 to 2004 more financial misstatements than in other years can be observed. This could be attributed to the tightened regulations after the big fraud scandals in 2001 and the resulting implementation of SOX in 2002. Also, fewer misstatements are noted in recent years since the average period between the end of the fraud and the publication of an AAER is three years [58].

5. Methodology

The objective of this study is to devise a fraud detection systems that classifies annual reports. While financial and linguistic variables represent structured tabular data and require no extensive pre-processing, the unstructured text data has to be transformed into a numeric format, which preserves its informative content and facilitates algorithmic processing. To achieve the latter, words are embedded as numeric vectors. The field of NLP has proposed various ways to construct such vectors. We consider two methods for text representation: frequency-based BOW embeddings and prediction-based neural embeddings (word2vec). An advantage of the BOW approach, which has been used in prior work on financial statement fraud (see Table1), is its simplicity. However, BOW represents a set of words without grammar and disrupts word order. Unlike BOW, the application of DL is still relatively new to the area of regtech (management of regulatory processes within the financial industry through technology). Therefore, the following subsections clarify neural word embeddings and address the DL components of the proposed HAN model.

5.1. Neural Embeddings

Within the BOW model, every word represents a feature. The amount of features denotes the dimension of the document vector [47]. Since the amount of unique words within a document typically

only represents a small proportion of the overall amount of unique words within the whole corpus, BOW document vectors are very sparse. A more advanced model for creating lower dimensional, dense embeddings of words is word2vec. As opposed to BOW, word2vec embeddings enable words that have similar meanings to be given similar vector representations and capture the syntactic and semantic similarities.

Word2vec [49] is an example of a NN model that is capable of learning word representations from a large corpus. Every word within the corpus is mapped to a vector of 50 to 300 dimensions. Mikolov et al. [49] demonstrated that such vectors offer advanced capabilities to measure the semantic and syntactic similarities between words. Word2vec can employ two approaches, namely the continuous bag-of-words (CBOW) and Skip-gram. Both models employ a shallow neural network with one hidden layer. In CBOW, the model predicts a target word from a window of adjacent context words that precede and follow the target word within the sentence. The Skip-gram model, on the other hand, employs the target word for predicting the surrounding window of context words. The structure of the model weights nearby context words more heavily than more distant context words. The generated word embeddings are a suitable input for text mining algorithms based on DL, as will be observed in the next part. They constitute the first layer of the model and allow further processing of text input within the DL architecture.

The initial word2vec algorithm is followed by GloVe [53], FastText [8], and GPT-2 [61], as well as the appearance of publicly available sets of pre-trained embeddings that are acquired by applying the above-mentioned algorithms on large text corpora. Pre-trained word embeddings accelerate training DL models and were successfully used in numerous NLP tasks [13, 70, 56, 60, 30]. We apply several types of pre-trained embeddings for HAN model and a neural network with a bidirectional Gated Recurrent Unit (GRU) layer that serves as a benchmark from the field of DL. As a result of a performance-based selection, the HAN model is built with word2vec embeddings with 300 neurons, trained on the Google News corpus, with a vocabulary size of 3 million words. The DL benchmark is used with the GPT-2 pre-trained embeddings from the WebText, offered by Radford et al. [61], as they arguable constitute the current state-of-the-art language model. The DL benchmark model is thus referred to as GPT-2 and is used together with the attention mechanism, discussed further.

5.2. Deep learning

After representing unstructured textual data in a numerical format, it can be used for predictive modeling. Conventional methods for classifying text involve the representation of sparse lexical features, like TF-IDF, and subsequently utilize a linear model or kernel techniques upon this representation [33].

An NN can be considered a non-linear generalization of the linear classification model [28]. NN comprised of multiple intermediate layers, called hidden layers, are referred to as deep NN (DNN), or DL networks. The weight matrices between the layers serve as intermediate parameters used by the NN to calculate a function of the inputs through the propagation of the computed values. During the training process, the NN learns to predict the output labels by changing the weights connecting the neurons with regard to how well the predicted output for a particular input matched the true output label in the training data. The process of adjusting the weights among neurons based on errors observed in prediction, to modify the calculated function to generate increased predictive accuracy, is referred to as back-propagation, while the structure of densely connected layers would be referred to as ANN [3].

Recently, DL has incorporated new techniques, including Convolutional Neural Networks (CNN) [35] and Recurrent Neural Networks (RNN) [29] for learning textual representations [77]. The RNN architecture allows retaining the input sequence, which made it widely used for natural language understanding, language generation, and video processing [48, 34]. An LSTM is a special type of RNN, comprised of various gates determining whether the information is kept, forgotten or updated and enabling long-term dependencies to be learned by the model [29]. An LSTM retains or modifies previous information on a selective basis and stores important information in a separate cell c_t, which acts as a memory [73]. The LSTM comprises four gates called the input gate i_t, forget gate f_t, output gate o_t and input modulation gate $\hat{c}_t$. These allow the network to recall or disregard information about previous elements in an input sequence. The interaction among the gates is noted in equations below, where $\odot$ represents element-wise multiplication.

$$
\begin{aligned}
i_t &= \sigma(U_i x_t + W_i h_{t-1} + b_i) & o_t &= \sigma(U_o x_t + W_o h_{t-1} + b) \\
f_t &= \sigma(U_f x_t + W_f h_{t-1} + b_f) & c_t &= f_t \odot c_{t-1} + i_t \odot \hat{c}_t \\
\hat{c}_t &= tanh(U_c x_t + W_c h_{t-1} + b_c) & h_t &= tanh(c_t) \odot o_t
\end{aligned}
$$

By considering the present input vector x_t, as well as the previous hidden state h_{t-1}, the forget gate layer f_t determines how much of the preceding cell state c_{t-1} it should forget, while, based on the identical input, the input gate layer i_t determines the amount of new information $\hat{c}_t$ that should be learned. The combination of outputs from these filters enables updating the cell state c_t. Consequently, overwriting of important information by the new inputs does not occur, it can persist for extended periods. Lastly, the hidden state h_t is computed based on the updated memory and the output gate layer o_t. In the final stage, the output vector is calculated as a function of the newly generated hidden state $\hat{y}_t = \sigma_o(W_o h_t + b_o)$, which is analogous to the basic RNN.

5.3. Hierarchical Attention Network

RNN retain the sequential structure of language. More advanced DL approaches also address hierarchical patterns of language such as the hierarchy between words, sentences, and documents. Some methods have covered the hierarchical construction of documents [78, 62]. The specific contexts of words and sentences, whereby the meaning of a word or sentence could change depending on the document, is a comparatively new concept for the process of text classification, and the HAN was developed to address this issue [76]. When computing the document encoding, HAN firstly detects the words that have importance within a sentence, and subsequently, those sentences that have importance within a document while considering the context (see Figure 1). The model recognizes the fact that an occurrence of a word may be significant when found in a particular sentence, whereas another occurrence of that word may not be important in another sentence (context).

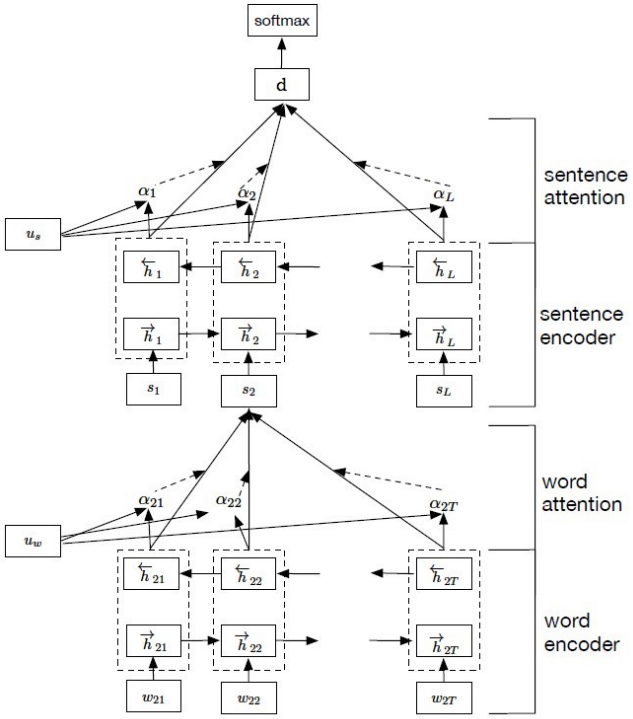

Figure 1: HAN Architecture. Image based on Yang et al. [76]

The HAN builds a document representation via the initial construction of sentence vectors based on words followed by the aggregation of these sentence vectors into a document representation through the application of the attention mechanism. The model consists of an encoder that generates relevant contexts and an attention mechanism, which calculates importance weights. The same algorithms are consecutively implemented at the word level and then at the sentence level.

Word Level. The input is transformed into structured tokens w_{it} that denote word i in sentence $t \in [1, T]$. Tokens are further passed through a pre-trained embedding matrix W_e that allocates multidi-

mensional vectors $x_{it} = W_e w_{it}$ to every token. As a result, words are denoted in numerical format by x_{it} as a projection of the word in a continuous vector space.

Word Encoder. The vectorized tokens represent the inputs for the following layer. While Yang et al. [76] employed GRU for encoding, we use LSTM as it showed better performance on the large text sequences at hand [12]. In the context of the current model, a bidirectional LSTM is implemented to obtain the annotations of words. The model consists of two uni-directional LSTMs, whose parameters are different apart from the word embedding matrix. Processing of the sentences in the initial forward LSTM occurs in a left to the right manner, whereas in the backward LSTM, sentences are processed from right to left. The pair of sentence embeddings are concatenated at every time step t to acquire the internal representation of the bi-directional LSTM h_{it}.

Word Attention. The annotations h_{it} construct the input for the attention mechanism that learns enhanced annotations denoted by u_{it}. Additionally, the *tanh* function adjusts the input values so that they fall in the range of -1 to 1 and maps zero to near-zero. The newly generated annotations are then multiplied again with a trainable context vector u_w and subsequently normalized to an importance weight per word α_{it} via a softmax function. As part of the training procedure, the word context vector u_w is initialized randomly and concurrently learned. The total of these importance weights concatenated with the already computed context annotations is defined as the sentence vector s_i:

$$u_{it} = tanh(W_w h_{it} + b_w) \tag{1}$$

$$\alpha_{it} = \frac{\exp(u_{it}^T u_w)}{\sum_t \exp(u_{it}^T u_w)} \tag{2}$$

$$s_i = \sum_t \alpha_{it} h_{it} \tag{3}$$

Sentence Level and Sentence Encoder. Subsequently, the entire network is run at the sentence level using the same fundamental process used for the word level. An embedding layer is not required as sentence vectors s_i have previously been acquired from the word level as input. Summarization of sentence contexts is performed using a bi-directional LSTM, which analyzes the document in both forward and backward directions:

$$\overrightarrow{h}_i = \overrightarrow{LSTM}(s_i), i \in [1, L] \tag{4}$$

$$\overleftarrow{h}_i = \overleftarrow{LSTM}(s_i), i \in [T, 1] \tag{5}$$

$$h_i = [\overrightarrow{h}_i, \overleftarrow{h}_i] \tag{6}$$

Sentence Attention. For rewarding sentences that are indicators of the correct document classification, the attention mechanism is applied once again along with a sentence-level context vector u_s, which is utilized to measure the sentence importance. Both trainable weights and biases are initialized randomly and concurrently learned during the training procedure, thus yielding:

$$u_i = tanh(W_s h_i + b_s) \tag{7}$$

$$\alpha_i = \frac{\exp(u_i^T u_s)}{\sum_i \exp(u_i^T u_s)} \tag{8}$$

$$d = \sum_i \alpha_i h_i \tag{9}$$

where d denotes the document vector summarising all the information contained within each of the document's sentences. Finally, the document vector d is a high-level representation of the overall document and can be utilized as features for document classification to generate output vector $\hat{y}$:

$$\hat{y} = softmax(W_c d + b_c) \tag{10}$$

where $\hat{y}$ denotes a K dimensional vector and the components y_k model the probability that document d is a member of class k in the set $1, ..., K$.

The application of the HAN follows the application of Kränkel and Lee [38]. Training of the DL model is performed on the training data set using both textual and quantitative features. Hence, the textual data acquired in the previous section is concatenated with the financial ratios. The model is employed to predict fraud probabilities of annual statements in the corresponding validation and test partitions, that were constructed with random sampling with stratification. Figure 2 shows the architecture of the HAN based fraud detection model and the output dimensions of each layer.

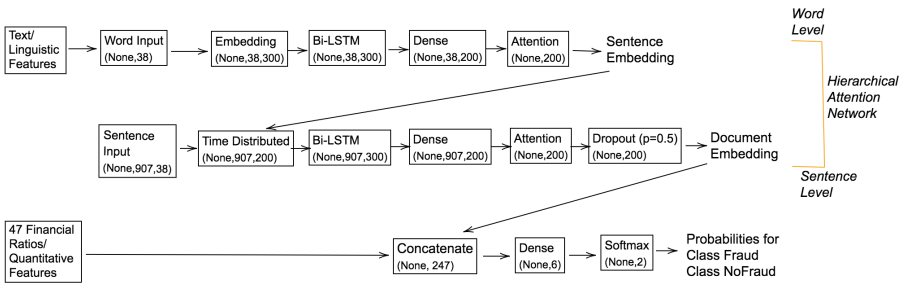

Figure 2: Architecture of the HAN based Fraud Detection Model

The LSTM layer consists of 150 neurons, a HAN dense dimension of 200, and a last dense layer dimension of 6. In this case, a combination of forward and backward LSTMs gives 300 dimensions for word and sentence annotation. The last layer of the HAN involves the application of dropout regularization to prevent over-fitting. In a final step, the resulting document-representation of dimension 200 is concatenated with 47 financial ratios and inputted to a dense layer before running through a softmax function that outputs the fraud probabilities. For training, a batch size of 32 and 17 epochs was used after hyperparameter tuning on the train validation set.

5.4. Evaluation metrics

The detection of financial statement fraud is considered a binary classification problem with four potential classification outcomes: True positive (TP) denotes the correct classification of a fraudulent company, false negative (FN) denotes the incorrect classification of a fraudulent company as a non-fraudulent company, true negative (TN) denotes the correct classification of a non-fraudulent company and false positive (FP) denotes the incorrect classification of a non-fraudulent company as a fraudulent company.

To estimate the predictive performance, many previous studies considered a combination of measures such as accuracy, sensitivity (also called TP rate or recall), specificity (also called TN rate), precision, and F1-score [75]. In this study, model performance is evaluated by the AUC, sensitivity, specificity, F1-score, F2-score, and accuracy.

The accuracy is defined as the percentage of correctly classified instances:

$$\text{Accuracy} = \frac{TP + TN}{P + N} \tag{11}$$

The sensitivity measures the number of correctly classified fraudulent instances as a percentage of all fraudulent instances:

$$\text{Sensitivity} = \frac{TP}{P} = 1 - \text{FN rate} \tag{12}$$

The specificity measures the number of correctly classified non-fraudulent instances as a percentage of all non-fraudulent instances:

$$\text{Specificity} = \frac{TN}{N} = 1 - \text{FP rate} \tag{13}$$

The F-score is a combination of precision $= \frac{TP}{TP+FP}$ (correct classification of fraudulent instances as a percentage of all instances classified as fraudulent) and sensitivity (indicates how many fraudulent instances the classifier misses) and measures how precise and how robust the models classify fraudulent cases:

$$F_\beta\text{-score} = (1 + \beta^2) \times \frac{\text{precision} \times \text{sensitivity}}{(\beta^2 \times \text{precision}) + \text{sensitivity}} \tag{14}$$

Prior research emphasized that a higher sensitivity is preferred to higher specificity in financial statement fraud detection. Nevertheless, the majority of models have exhibited considerably higher performance in detecting truthful transactions in comparison to those that are fraudulent [75, 63]. An explanation for this preference is that FN and FP rates result in considerably different misclassification costs (MC). Hajek and Henriques [27] estimated the cost of failing to detect fraudulent statements to be twice as high as the cost of incorrectly classifying fraudulent statements. Hence, effective models should concentrate on high sensitivity and classify correctly as many positive samples as possible, rather than maximizing the number of correct classifications. Therefore, this study employs the F2-score in addition to the F1-score (harmonic mean of precision and sensitivity), as it weights sensitivity higher than precision and is, therefore, more suitable for fraud detection. The AUC denotes the area under the Receiver Operating Curve (ROC) and is preferred to accuracy in financial statement fraud detection because of the impact of fraud/non-fraud imbalance in the sample [72, 59]. This study employs the AUC as a measure of separability to compare the predictive performance of the models and determine their suitability. The higher the AUC, the better the model can distinguish between fraudulent and non-fraudulent cases.

The cutoff threshold for the probability of fraud has to be defined to quantify the F1- and F2-scores. We select the threshold that maximizes the difference between sensitivity and FP rate and use it to evaluate the classification results. For the HAN model, the optimal threshold is set at 0.03, implying that a statement is classified as fraudulent if its fraud probability is higher than 3%.

6. Classification results

We answer RQ 1 and 2 by means of empirical analysis and compare a set of classification models in terms of their fraud detection performance. The models generate fraud classification based on financial indicators, linguistic features of reports, the reports' text, and combinations of these groups of features. Table 2 reports corresponding results from the out-of-sample test set. The baseline accuracy of classifying all cases of the test set as non-fraudulent (majority class) is 82.81%.

6.1. Modeling of financial data

Modeling using financial data (FIN) has been the most popular approach (Table 1). The approach serves this study as a benchmark, to which we compare modeling on linguistic features (LING) and the combination of both (FIN + LING). The last two columns of Table 2 show the results of the comparison. In terms of AUC and accuracy, the tree-based models RF [9] and XGB appear to excel at predicting fraud on FIN, indicating a non-linear dependency between financial indicators and the fraud status of a report. This result is in line with Liu et al. [44] who showed that RF performed

Finance data (FIN)								
	AUC	**Sensitivity**	**Specificity**	**F1-score**	**F2-score**	**Accuracy**		
LR	0.7620	0.6833	0.7543	0.4767	0.7480	0.8252		
RF	**0.8609**	0.7666	0.7889	0.5508	0.7892	**0.8653**		
SVM	0.7561	0.6166	0.7820	0.4625	0.7595	0.8280		
XGB	0.8470	0.6660	**0.8719**	**0.5839**	**0.8391**	0.8481		
ANN	0.7564	**0.7833**	0.6574	0.4563	0.6835	0.6790		
Linguistics data (LING)							Comparison to FIN	
	AUC	**Sensitivity**	**Specificity**	**F1-score**	**F2-score**	**Accuracy**	**Delta AUC**	**Delta F1**
LR	0.6719	0.7000	0.6193	0.3962	0.6398	0.8280	-0.0901	-0.0805
RF	**0.7713**	**0.7500**	0.7197	**0.4839**	0.7302	**0.8424**	-0.0896	-0.0669
SVM	0.7406	0.7000	0.6747	0.4285	0.6857	0.8280	-0.0155	-0.0340
XGB	0.7219	0.3666	**0.9446**	0.4489	**0.8385**	0.8338	-0.1251	-0.1350
ANN	0.6782	0.6333	0.6747	0.3958	0.6758	0.6676	-0.0782	-0.0605
Finance data + Linguistics data (FIN + LING)							Comparison to FIN	
	AUC	**Sensitivity**	**Specificity**	**F1-score**	**F2-score**	**Accuracy**	**Delta AUC**	**Delta F1**
LR	0.7682	0.7666	0.6782	0.4623	0.6984	0.8280	0.0062	-0.0144
RF	0.8606	0.7666	**0.7543**	0.5197	0.7610	0.8567	-0.0003	-0.0311
SVM	0.7973	0.7166	0.7439	0.4858	0.7448	0.8280	0.0567	0.0573
XGB	**0.8651**	0.8166	**0.7543**	**0.5444**	**0.7687**	**0.8653**	0.0181	-0.0395
ANN	0.7733	**0.8333**	0.6228	0.4566	0.6614	0.6590	0.0169	0.0003
Text data, TF-IDF (TXT)							Comparison to LING	
	AUC	**Sensitivity**	**Specificity**	**F1-score**	**F2-score**	**Accuracy**	**Delta AUC**	**Delta F1**
LR	0.8371	0.7333	0.8269	0.5714	0.8145	0.8281	0.1652	0.1752
RF	0.8740	0.7166	0.9377	0.7107	**0.8998**	0.8681	0.1027	0.2268
SVM	0.8836	**0.8382**	0.7544	0.5876	0.7731	0.8796	0.1275	0.1251
XGB	0.8785	0.7660	0.8581	0.6258	0.8451	0.8853	0.1566	0.1769
ANN	0.8829	0.7121	**0.9434**	**0.7286**	0.8993	**0.8990**	0.2047	0.3328
HAN	**0.9108**	0.8000	0.8896	0.5744	0.7982	0.8457		
GPT-2+Attn	0.7729	0.7619	0.6697	0.4423	0.6905	0.6484		
Finance data + Text data, TF-IDF (FIN + TXT)							Comparison to FIN + LING	
	AUC	**Sensitivity**	**Specificity**	**F1-score**	**F2-score**	**Accuracy**	**Delta AUC**	**Delta F1**
LR	0.8598	0.7833	0.7854	0.5562	0.7890	0.8424	0.0916	-0.0795
RF	0.8797	0.6660	0.9550	0.7079	0.9043	0.8739	0.0191	-0.1571
SVM	0.8902	0.7833	0.8961	0.6861	0.8784	0.8280	0.0929	-0.2576
XGB	0.8983	0.7000	**0.9653**	**0.7500**	**0.9187**	**0.9083**	0.0332	-0.1661
ANN	0.8911	0.7460	0.9405	0.7401	0.9055	0.9054	0.1178	-0.2838
HAN	**0.9264**	**0.9000**	0.8206	0.6506	0.8361	0.8457		
GPT-2+Attn	0.7776	0.7678	0.6791	0.4455	0.6991	0.6934		

Table 2: Comparative performance of selected binary classifiers on the different types of test data. The baseline accuracy is 0.8281

especially well in case of high-dimensional financial fraud data because of a higher variance reduction resulting from combining Bagging with randomly chosen subsets of input features [27]. Hajek and Henriques [27] also reported an accuracy of 88.1% on FIN data and concluded that the ensemble of tree-based algorithms including JRip, CART, C4.5 and LMT exhibit superior performance over SVM, LR and ANN due to a relatively low dimensionality achieved during feature selection. The predictive performance aligns with the results of Kim et al. [37], offering the LR and SVM models as the most accurate. Lin et al. [43] and Ravisankar et al. [63] showcased that the DNN models (ANN with more than one hidden layer) outperform LR and SVM, offering an accuracy higher by around 4.5%. The SVM is a widely recognized model and was applied both for fraud detection [54] and in other fields [17]. However, the results show that inherent configuration complexities make SVM a secondary choice for practitioners. ANN show less impressive predictive performance but proved to be the most efficient in terms of sensitivity. However, for model evaluation, a balanced indicator like F1- and F2-scores would provide a better perspective. These metrics suggest XGB to outperform other models. XGB represents an advancement in the in the field of ML, its high performance is noteworthy since it was not considered in prior work on fraud detection. Given the much higher cost of missing actual fraud cases compared to false alarms, we argue that the F2-score is the most suitable threshold-based indicator of model performance. Therefore, we emphasize the F2-score together with the AUC, which allows the tuning of the threshold.

6.2. Modeling of linguistic data

The modeling on linguistic data (LING) was the first step towards including text in fraud detection. The earlier experiments by Cecchini et al. [11], Humpherys et al. [32] and Goel et al. [24] employed SVM and achieved accuracy of 82%, 65.8%, and 89.5% respectively. The latter additionally included the BOW method that we will discuss further. Our modeling falls in line with the previous work and exhibits SVM as the second strongest predictor, yielding an AUC of 74% and accuracy of 82%. RF remained the most reliable predictor with the highest AUC, accuracy, and F2-score. Modeling done solely on LING will allow us to assess the degree to which both sources of data contribute to accurate classification. In line with Hajek and Henriques [27], all models exhibit higher performance on FIN data than on LING data solely, leading to the conclusion that financial covariates have more predictive power than linguistic variables. However, the performance differences are not substantial and suggest a strong relationship between linguistic features and fraudulent behavior, which agrees with previous studies.

Following the ideas of Hajek and Henriques [27], we combine FIN and LING data to evaluate if the classifier can make use of both data sources. Our results differ in terms of the leading models,

with RF and XGB offering the highest AUCs of 86%. XGB is showing a definite improvement, performing well on FIN data, falling back a little in the LING set up but making better use of the combined input. Once again we observe the superior performance of XGB in terms of F2-score with 76.87% followed closely by 76.10% of RF and 74.48% of SVM, which once again advocates for the usefulness of advanced ML methods for practical tasks. Interestingly, for the rest of classifiers, the accuracy dropped a little in comparison to FIN, but the AUCs improved (with a minor exception for RF). This serves as an indication that LING and FIN data combined may provide conflicting signals to the classifier, however, the data mix is a definite improvement as it provides a stronger signal to the classifier, enhancing the predictive performance.

6.3. Modeling of text data

Researchers have been taking a step forward from aggregated linguistic features in an attempt to derive more predictive power from the vast amounts of text contained in annual reports. We offer the advanced methods of NLP, previously unexplored for fraud detection, and compare them to the performance of more traditional models. Goel et al. [24], Glancy and Yadav [22] and Purda and Skillicorn [59] applied the BOW model to perform modeling on text data, while Goel and Uzuner [25] made use of part-of-speech tagging. They utilized SVM and hierarchical clustering as classifiers and achieved accuracies of 89.5%, 83.4%, 89%, and 81.8%, respectively.

Table 2 offers an overview of the modeling results, starting with purely textual input (TXT) and continuing with text enhanced by financial data (FIN+TXT). Two new DL methods are included in TXT modeling, namely HAN and GPT-2. While traditional benchmarks take the TF-IDF transformations of word input, the DL models make use of pre-trained embeddings, discussed in the Methodology section. We can observe that modeling on TXT provides improvement across all models in comparison to LING, with the largest AUC delta of 0.2 in the case of ANN. This increase can be first and foremost attributed to the richer input of the actual MD&A content. ANN demonstrates the highest accuracy, 89%, and the best performing F1- and F2-scores, which constitutes a strong signal that the neural network architecture is a favorable candidate for the task, regardless of the BOW input. Given the complexity of textual processing, ANN proves its capacity to pick up on complex relationships between the target and explanatory variables. The improvement is also visible for the F2-score of 89.93% that closely follows the RF's 89.98%. It is interesting to compare the BOW-based ANN with GPT-2 and HAN, as all three represent a NN architecture. GPT-2 performs better on TXT than any other model on LING. Though it fails to show superior accuracy, its sensitivity metric is one of the highest, leading to the conclusion that with some threshold adjustment, it could provide better predictive performance than other models like LR or tree-based models. This example underlines the

potential gains of implementing the new DL methods that allow superior insights into unstructured data. Unlike BOW-based benchmarks, embeddings-based HAN and GPT-2 retained the structure and context of the input. HAN showed superior results in terms of AUC 91.08% but fell short in terms of accuracy. However, its sensitivity is exceeding those of all other benchmarks except for SVM, making it a promising model for fraud detection. HAN represents a further advancement of the NLP with DL approaches; its performance can be explained by the intrinsic capacity to extract significant contextual similarities within documents and that pertinent cues that allow truthful text to be distinguished from deceitful ones are dependent on the context rather than the content [79]. All in all, the results suggest that textual data, in general, can offer much more insight than LING across all classifiers. However, the NN-based and tree-based architectures seem to benefit the most in terms of AUC.

We conclude the analysis of results by looking at the feature combination FIN+TXT, which is at the core of our study. The input setup is done in two ways: a combination of word vectors with financial indicators into one data set and a 2-step modeling approach. The latter comprises building a TXT model and using its probability prediction as an input to another DL model that will concoct it with FIN and output the final binary prediction. The first approach is applied in the case of benchmark models, including ANN, while the second one is implemented for the DL models, namely, HAN and GPT-2.

Purda and Skillicorn [59] conducted a comparison of TXT with FIN data proposed by Dechow et al. [15] separately and determined that they are complementary since both methods are capable of identifying specific types of fraud that the other cannot detect and they have a relatively low correlation. In our case, all benchmarks exhibit improved performance in comparison to the FIN + LING setup, especially LR and SVM. However, the same unanimity is observed in decreased F1-score metric, with ANN dropping by 0.28. We observe the superiority of predictive powers of full-textual input over the linguistic metrics. If we compare the additional value of FIN for the performance, we can see only a minor increase in almost all metrics, once again underlying the complexity and potential misalignment of FIN and TXT data. However, it is essential to note that unlike F1-score, F2-score increases across the ML benchmarks, which brings us to the initial assumption behind the preference toward the F2-score as key to model evaluation for practical use. We conclude that with the increased complexity of input, one should opt for advanced ML techniques for the extraction of extra insight.

The best performance is again yielded by HAN with AUC 92.64%, followed by XGB and ANN with AUCs of 89%. It is also offering the highest sensitivity of 90% across all datasets and models, making it the recommended solution for the anomaly-detection-type tasks, like fraud detection. Going back to the triad comparison between ANN, HAN, and GPT-2, we can see that the latter does not show much improvement with added FIN data across all metrics. This signals the potentially poor

choice of pre-trained embeddings, highlighting the importance of this decision in the design of a DL classifier and reminding that state-of-the-art solutions do not guarantee the superior application results. ANN does not catch up with HAN AUC-wise. However, it showcases the higher F2-score of 90.55%, surpassed only by XGB, which proved to be a promising alternative to the DL methods. The results of modeling on HAN showed its capacity to incorporate and extract additional values from the diversified input, which contributes to the existing field of research and opens new opportunities to the further exploration of data enrichment for fraud detection.

The results of HAN address the RQ 1 and 2, allowing us to conclude that the proposed DL architecture offers a substantial improvement for fraud detection facilitation. Additionally, its properties allow us to offer a look into the "black box" of the DL models and provide the rationale behind the classification decision. This interpretability capacity might be particularly important for practitioners, given the need to substantiate the audit judgment, and will be further explored in the next Section.

7. Interpretation and decision support application

SEC developed software specifically focused on the MD&A section [59] to examine the use of language for indications of fraud. The importance of the MD&A section can be observed in reforms introduced by the Sarbanes-Oxley Act (SOX) in 2002, which demanded that the relevant section should present and offer full disclosure on critical accounting estimates and policies [64]. The length of MD&A sections increased after SOX became effective; nevertheless, Li [42] concluded that no changes were made to the information contained within MD&A sections or the style of language adopted. Taking further the fraud detection efforts, we developed a method to facilitate the audit of the MD&A section. We employ state-of-the-art textual analysis to shed light on managers' cognitive processes, which could be revealed by the language used in the MD&A section. Zhou et al. [79] demonstrated that it is plausible to detect lies based on textual cues. Nonetheless, the pertinent cues that allow truthful texts to be distinguished from deceitful ones are dependent on the context. One way to support auditors would be the "red-flag" indication in the body of the MD&A section. Hajek and Henriques [27] explored the use of "green-flag" and "red-flag" values of financial indicators and concluded that the identification of non-fraudulent firms is less complex and can be accompanied by interpretable "green-flag" values, however because the detection of fraudulent firms requires more complex non interpretable ML models, no "red-flag" values could be derived. We will take it further and provide the suggestion for the use of textual elements as "red-flags" for auditors. This can be done on the word level or the sentence-level and is to our best knowledge, new to the field. The HAN model allows a holistic analysis of the text structure and the underlying semantics. In contrast to BOW that ignores specific contextual meanings of words, the HAN model considers the grammar, structure,

and context of words within a sentence and of sentences within a document, which is essential for the identification of fraudulent behaviour. The attention mechanisms of the HAN at both word and sentence levels retain the logical dependencies of the content and learn to differentiate the important words and sentences. These valuable insights into the internal document structure together with strong predictive performance, make HAN notably advantageous in comparison to BOW-based traditional benchmarks.

Based on the assumption that fraudulent actors are capable of manipulating their writings so that they have convincing similarities to those that are non-fraudulent, only concentrating on words that focus on the content of the text while disregarding the context could be overly simplistic for differentiating truthful from misleading statements. We assume that due to their inherently higher complexity, sentence-level indicators are less prone to manipulation and thus can provide robust insight for auditing.

7.1. Word-level

We provide a comparative analysis of words considered to be "red-flags" by the more traditional RF model and those offered by HAN. The RF model proved to be a potent and consistent classifier throughout the comparative analysis. We apply the *lime* methodology of Ribeiro et al. [65] to gain insight into the role of different words in the model's classification decision. *lime* stands for Local Interpretable Model-Agnostic Explanations and is based on explaining the model functioning in the locality of the chosen observation. Ribeiro et al. [65] explains every input separately; the example of its application to one of the fraud texts can be found in Figure 3:

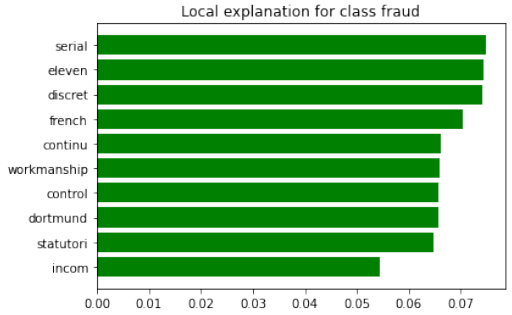

Figure 3: Words with top weights indicating fraud from a sample MD&A

We supply all fraud cases through the *lime* package and extract the top ten words, that have the strongest effect on the model in terms of fraud indication. We further aggregate these words and gain a "red-flag" vocabulary. Additionally we perform the same analysis with the DNN model and extract the weights assigned by the HAN attention layer. The results are summarized in Figure 4:

RF lime	DNN Attention
aerospac, align, america, amnesti, api, arcapita, arduou, armorgroup, artisan, astound, ballist, belief, broadest, brokerag, canton, carbon, categori, cdc, contain, copley, cork, dealer, decemb, deeper, defect, depend, diamond, discuss, doughnut, dysfunct, elig, endow, ensuit, epidem, erp, especi, fashion, forgiv, grain, grand, groundwat, grown, harbing, health, help, hemispher,immun, interrupt, itsunfavor, keyboard, kraken, liabl, lingyun, mainli, mammographi, mancelona, mard, marian, maverick, maxim, militia, mobiapp, monocular, necessari, nitrat, nonsteroid, offlin, operatingloss, orthovisc, paint, paramagnet, payabl, pilotless, predomin, reagent, reengin, referenc, reformul, reincentiv, remex, reprograph, resin, rubber, satur, sec, semisubmers, shoot, sophist, spectromet, state, strain, suitabl, sunnyval, trailer, transact, trundl, tucson, understood, undistribut, unwil, updat, upsid, valencia, visitor, websit, withdrawn	according, acquisition, addition, additionally, agreement, also, although, april, august, average, based, believe, biotin, business, chairman, company, competitor, completed, corporate, cost, course, criterion, currently, customer, decrease, dev, diverse, entered, enterprise, event, expect, factor, february, following, ft, future, generally, government, gross, home, increase, increasing, industry, intend, july, june, management, many, march, market, may, merchandise, million, network, new, non, november, number, october, one, opened, operate, operates, organized, overview, patent, payment, primary, product, property, recent, region, remaining, representative, result, revenue, rig, risk, sale, segment, sell, september, service, since, solution, store, strategy, table, technology, time, total, truck, two, type, typical
april, certain, chairman, government, gross, june, truck, year	million, network, new, product, rig, sale, store,

Figure 4: "Red-flag" words identified by Random Forest and HAN, the bottom section contains the words matching both sets

Fifteen words are found to be important for an indication of fraudulent activity by both algorithms, including "government", "certain", "gross", potentially indicating adverse involvement of the state institutions. It would seem that RF derives judgment from the industry: "aerospace", medical terms, "pilotless", "armourgroup". HAN picks up on financial and legal terms like "cost", "acquisition", "property". Both classifiers also include time- and calendar-related words like names of the month. It is not obvious how much the context affects this selection. Additionally, derivation of a word-based rule might potentially lead to a quick adaptation of the reporting entities for audit circumvention. Ambiguous interpretation and manipulation risks motivate the creation of the sentence-level decision support system.

7.2. Sentence-level

The added contextual information extracted by the HAN shows improved performance on the test set in comparison to linguistic features and other DNN models. It can be partially explained by the hierarchical structure of language, that entails the unequal roles of words in the overall structure. Following RQ 3, we want to benefit from the structural and context insight retained in sentence-level analysis, provided uniquely by the HAN model.

We extract the sentence-level attention weights for 200 fraudulent reports gained as a result of prediction by HAN and filter the top ten most important sentences per report. The mean weight of a sentence that can be considered a "red-flag" is 0.05, with a maximum at 0.61. We devise a rule, dictating that sentences with weights higher than 0.067 (top 25% quantile) will be referred to as "extra important", sentences between 0.04 and 0.67 (top half) are "important" and those between 0.022 and 0.04 are "noteworthy". These three groups of words get respective coloring and are highlighted in the considered MD&A, as depicted in Figure 5.

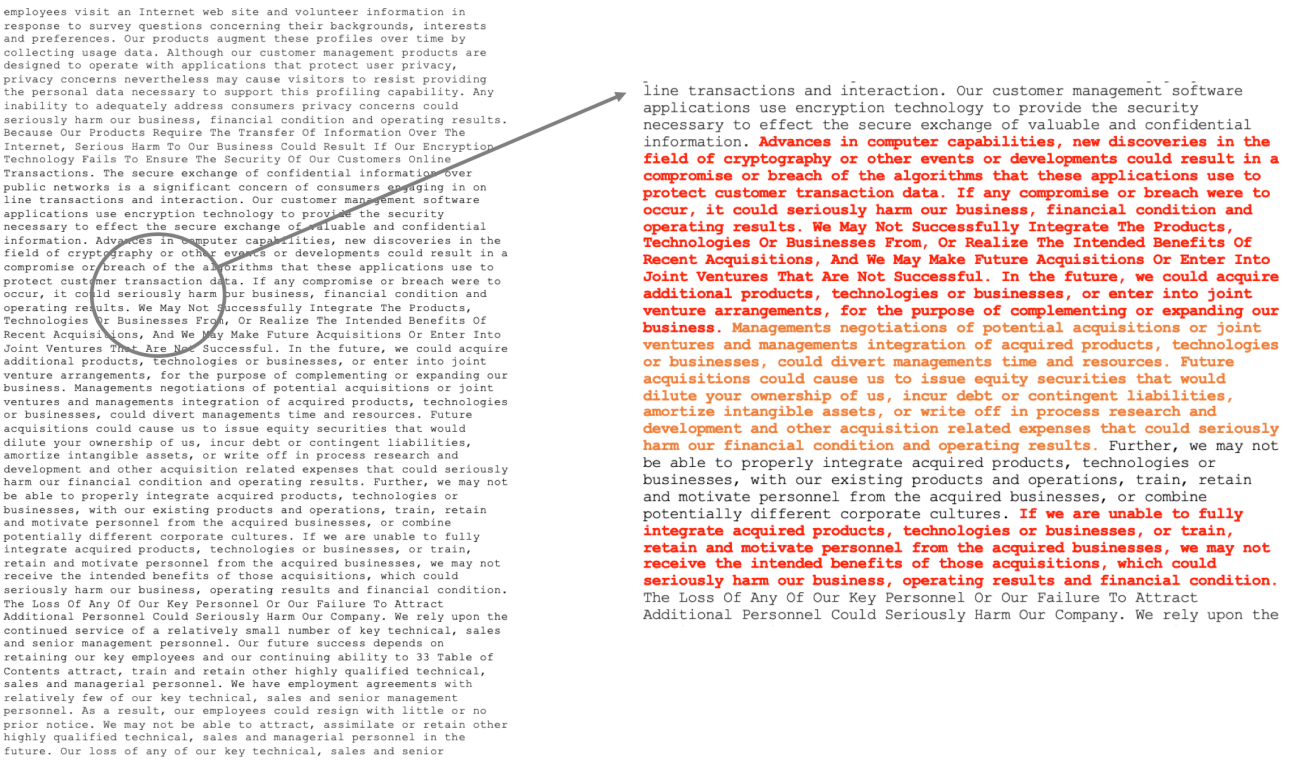

Figure 5: A page from MD&A (on the left) and its extract with "red-flag" phrases for the attention of the auditor (on the right). Sentences that contributed the most to the decision towards "fraud" are labeled by HAN as extra important and important. Additional examples are provided in Online Appendix

We propose to use the probability prediction of the HAN model and assign sentence weights as a two-step decision support system for auditors. Given its strong predictive performance, HAN can provide an initial signal about the risks of fraud. Given the selected sensitivity threshold, auditors may select to evaluate a potentially fraudulent report with extra caution and use the highlighted sentences as additional guidance. Given the lengthiness of an average MD&A and limited physical concentration capacities associated with the manual audit, this sort of visual guidance can offer higher accuracy of fraud detection.

8. Discussion

As reported in the literature review, Hajek and Henriques [27] and Throckmorton et al. [72] have tackled the task of combined mining financial and linguistic data for financial statement fraud prediction, and no study was found on the combination of financial and textual data. Given the managerial efforts to conceal bad news by using particular wording [32] and by generating less understandable reports [40, 46], it is pivotal to adopt more advanced text processing techniques.

In line with the findings of Perols [54] and Kim et al. [37], SVM showed good performance across most experimental setups. This can be explained by the fact that both models can deal with a huge number of features and with correlated predictors. Due to its ability to deal with high dimensional

and sparse features, SVM has achieved the best performances in previous studies [23, 59] that incorporated the BOW approach. RF came up as the leader in predictive performance, managing to extract knowledge from both financial and BOW-based textual sources. DL models proved capable of distinguishing fraudulent cases. However, only the HAN architecture showcased exceptional capacity to extract signals from the FIN + TXT setting, which is in the center of the current research. The HAN detects a high number of fraudulent cases compared to remaining models, strengthening the statement by Zhou et al. [79] that the detection of deception based on text necessitates contextual information.

The results of the AUC measures indicate that the linguistic variables extracted with HAN and TF-IDF add significant value to fraud detection models in combination with financial ratios. The heterogeneity in performance shifts among different data types for models, showing that different models pick up on different signals, and a combination of these models might be more appropriate to support the decision-making processes of stakeholders in the determination of fraud than the choice of a single model. The use of additional performance metrics like F2-score addressed the practical applicability of the classification models, given the imbalance of error costs. The superior predictive capacity should be considered in combination with the model's sensitivity in order to account for the implications of non-detecting the fraudulent case.

We have explored the interpretation capacities of RF and HAN models on the word and sentence levels. Both models agreed on a specific "red-flag" vocabulary; however, mostly, they picked up on different terms. Also, out of context, these words might be misleading. The indication of "red-flags" words is becoming increasingly unreliable with the adaptive response of the alleged offending parties. The offered sentence-level markup showed a more robust approach to the provision of decision support for the auditors.

9. Conclusion

The detection of financial fraud is a challenging endeavor. The continually adapting and complex nature of fraudulent activities necessitates the application of the latest technologies to confront fraud. This research investigated the potential of a state-of-the-art DL model to add to the development of advanced financial fraud detection methods. Minimal research has been conducted on the subject of methods that combine the analysis of financial and linguistic information, and no studies were discovered on the application of text representation based on DL to detect financial statement fraud. In addition to quantitative data, we investigated the potential of the accompanying text data in annual reports, and have emphasized the increasing significance of textual analysis for the detection of signals of fraud within financial documentation. The proposed HAN method concentrates on the content as well as the context of textual information. Unlike the BOW method, which disregards word order and

additional grammatical information, DL is capable of capturing semantic associations and discerning the meanings of different word and phrase combinations.

The results have shown that the DL model achieved considerable improvement in AUC compared to the benchmark models. The findings indicate that the DL model is well suited to identify the fraudulent cases correctly, whereas most ML models fail to detect fraudulent cases while performing better at correctly identifying the truthful statements. The detection of fraudulent firms is of great importance due to the significantly higher MC associated with fraud. Thus, specifically in the highly unbalanced case of fraud detection, it is advisable to use multiple models designed to capture different aspects.

Based on these findings, we conclude that the textual information of the MD&A section extracted through HAN has the potential to enhance the predictive accuracy of financial statement fraud models, particularly in the generation of warning signals for the fraudulent behavior that can serve to support the decision making-process of stakeholders. The distorted word order handicaps the ability of the BOW-based ML benchmarks to offer a concise indication of the "red-flags". We offered the decision support solution to the auditors that allows a sentence-level indication of text fragments that trigger the classifier to treat the submitted case as fraudulent. The user can select the degree of impact of indicated sentences and improve the timing and accuracy of the audit process.

References

[1] A. Abbasi, C. Albrecht, A. Vance, and J. Hansen. “Metafraud: A meta-learning framework for detecting financial fraud”. In: *MIS Quarterly: Management Information Systems* 36.4 (2012), pp. 1293–1327.

[2] ACFE. *Report to the Nations 2018 Global Study on Occupational Fraud and Abuse*. Tech. rep. 2019.

[3] C. C. Aggarwal. *Neural Networks and Deep Learning - A Textbook*. Springer International Publishing, 2018, pp. 875–936.

[4] W. S. Albrecht, C. Albrecht, and C. C. Albrecht. “Current trends in fraud and its detection”. In: *Information Security Journal* 17.1 (2008), pp. 2–12.

[5] B. Bai, J. Yen, and X. Yang. “False financial statements: Characteristics of China’s listed companies and cart detecting approach”. In: *International Journal of Information Technology and Decision Making* 7.2 (2008), pp. 339–359.

[6] M. D. Beneish. “The Detection of Earnings Manipulation”. In: *Financial Analysts Journal* 55.5 (1999), pp. 24–36.

[7] A. Bodnaruk, T. Loughran, and B. McDonald. “Using 10-K text to gauge financial constraints”. In: *Journal of Financial and Quantitative Analysis* 50.4 (2015), pp. 623–646.

[8] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. “Enriching Word Vectors with Subword Information”. In: *arXiv preprint arXiv:1607.04606* (2016).

[9] L. Breiman. “Random forests”. In: *Machine Learning* 45.1 (2001), pp. 5–32.

[10] S. V. Brown and J. W. Tucker. “Large-Sample Evidence on Firms’ Year-over-Year MD&A Modifications”. In: *Journal of Accounting Research* 49.2 (2011), pp. 309–346.

[11] M. Cecchini, H. Aytug, G. J. Koehler, and P. Pathak. “Making words work: Using financial text as a predictor of financial events”. In: *Decision Support Systems* 50.1 (2010), pp. 164–175.

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

[13] A. M. Dai and Q. V. Le. “Semi-supervised sequence learning”. In: *Advances in neural information processing systems*. 2015, pp. 3079–3087.

[14] A. K. Davis, J. M. Piger, and L. M. Sedor. “Beyond the Numbers: Measuring the Information Content of Earnings Press Release Language”. In: *Contemporary Accounting Research* 29.3 (2012), pp. 845–868.

[15] P. M. Dechow, W. Ge, C. R. Larson, and R. G. Sloan. “Predicting Material Accounting Misstatements”. In: *Contemporary Accounting Research* 28.1 (2011), pp. 17–82.

[16] B. M. DePaulo, R. Rosenthal, J. Rosenkrantz, and C. Rieder Green. “Actual and Perceived Cues to Deception: A Closer Look at Speech”. In: *Basic and Applied Social Psychology* 3.4 (1982), pp. 291–312.

[17] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. “Inductive learning algorithms and representations for text categorization”. In: *Proceedings of the 7th International Conference on Information and Knowledge Management* (1998), pp. 148–155.

[18] A. Dyck, A. Morse, and L. Zingales. “Who blows the whistle on corporate fraud?” In: *Journal of Finance* 65.6 (2010), pp. 2213–2253.

[19] R. Feldman, S. Govindaraj, J. Livnat, and B. Segal. “Management’s tone change, post earnings announcement drift and accruals”. In: *Review of Accounting Studies* 15.4 (2010), pp. 915–953.

[20] C. Gaganis. “Classification techniques for the identification of falsified financial statements: a comparative analysis”. In: *Intelligent Systems in Accounting, Finance & Management* 16.3 (2009), pp. 207–229.

[21] J. Gee and M. Button. *The Financial Cost of Fraud 2019*. Tech. rep. Crowe, 2019.

[22] F. H. Glancy and S. B. Yadav. "A computational model for financial reporting fraud detection". In: *Decision Support Systems* 50.3 (2011), pp. 595–601.

[23] S. Goel and J. Gangolly. "Beyond the numbers: Mining the annual reports for hidden cues indicative of financial statement fraud". In: *Intelligent Systems in Accounting, Finance and Management* 19.2 (2012), pp. 75–89.

[24] S. Goel, J. Gangolly, S. R. Faerman, and O. Uzuner. "Can linguistic predictors detect fraudulent financial filings?" In: *Journal of Emerging Technologies in Accounting* 7.1 (2010), pp. 25–46.

[25] S. Goel and O. Uzuner. "Do Sentiments Matter in Fraud Detection? Estimating Semantic Orientation of Annual Reports". In: *Intelligent Systems in Accounting, Finance and Management* 23.3 (2016), pp. 215–239.

[26] G. L. Gray and R. S. Debreceny. "A taxonomy to guide research on the application of data mining to fraud detection in financial statement audits". In: *International Journal of Accounting Information Systems* 15.4 (2014), pp. 357–380.

[27] P. Hajek and R. Henriques. "Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods". In: *Knowledge-Based Systems* 128 (2017), pp. 139–152.

[28] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin. *The elements of statistical learning: data mining, inference and prediction*. 2nd ed. 2. New York: Springer, 2009.

[29] S. Hochreiter and J. Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[30] J. Howard and S. Ruder. "Universal language model fine-tuning for text classification". In: *arXiv preprint arXiv:1801.06146* (2018).

[31] S. Y. Huang, R. H. Tsaih, and F. Yu. "Topological pattern discovery and feature extraction for fraudulent financial reporting". In: *Expert Systems with Applications* 41.9 (2014), pp. 4360–4372.

[32] S. L. Humpherys, K. C. Moffitt, M. B. Burns, J. K. Burgoon, and W. F. Felix. "Identification of fraudulent financial statements using linguistic credibility analysis". In: *Decision Support Systems* 50.3 (2011), pp. 585–594.

[33] T. Joachims. “A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization”. In: *the 14th International Conference on Machine Learning (ICML ’97)* (1997), pp. 143–151.

[34] N. Kalchbrenner and P. Blunsom. “Recurrent continuous translation models”. In: *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. Association for Computational Linguistics (ACL), 2013, pp. 1700–1709.

[35] N. Kalchbrenner, E. Grefenstette, and P. Blunsom. “A convolutional neural network for modelling sentences”. In: *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*. Vol. 1. Association for Computational Linguistics (ACL), 2014, pp. 655–665.

[36] J. M. Karpoff, A. Koester, D. S. Lee, and G. S. Martin. “Database Challenges in Financial Misconduct Research”. In: *Working Paper* (2014), pp. 1–66.

[37] Y. J. Kim, B. Baik, and S. Cho. “Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning”. In: *Expert Systems with Applications* 62 (2016), pp. 32–43.

[38] M. Kränkel and H.-E. L. Lee. “Text Classification with Hierarchical Attention Networks”. In: (2019). URL: https://humboldt-wi.github.io/blog/research/information_systems_1819/group5_han/.

[39] D. F. Larcker and A. A. Zakolyukina. “Detecting Deceptive Discussions in Conference Calls”. In: *Journal of Accounting Research* 50.2 (2012), pp. 495–540.

[40] F. Li. “Annual report readability, current earnings, and earnings persistence”. In: *Journal of Accounting and Economics* 45.2-3 (2008), pp. 221–247.

[41] F. Li. “Textual analysis of corporate disclosures: A survey of the literature”. In: *Journal of accounting literature* 29 (2010), p. 143.

[42] F. Li. “The information content of forward- looking statements in corporate filings-A naïve bayesian machine learning approach”. In: *Journal of Accounting Research* 48.5 (2010), pp. 1049–1102.

[43] C. C. Lin, A. A. Chiu, S. Y. Huang, and D. C. Yen. “Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts’ judgments”. In: *Knowledge-Based Systems* 89 (2015), pp. 459–470.

[44] C. Liu, Y. Chan, S. H. Alam Kazmi, and H. Fu. “Financial Fraud Detection Model: Based on Random Forest”. In: *International Journal of Economics and Finance* 7.7 (2015).

[45] T. I. M. Loughran and B. Mcdonald. “When is a Liability not a Liability ? Textual Analysis , Dictionaries , and 10-Ks Journal of Finance , forthcoming”. In: *Journal of Finance* 66.1 (2011), pp. 35–65.

[46] T. Loughran and B. Mcdonald. “Measuring readability in financial disclosures”. In: *Journal of Finance* (2014).

[47] C. D. Manning, P. Ragahvan, and H. Schutze. “An Introduction to Information Retrieval”. In: *Information Retrieval* c (2009), pp. 1–18.

[48] T. Mikolov, M. Karafiát, L. Burget, J. ‘ Cernocký, and S. Khudanpur. “Recurrent Neural Network Language Modeling”. In: *Interspeech*. September. 2010, pp. 1045–1048.

[49] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. “Efficient Estimation of Word Representations in Vector Space Tomas”. In: *IJCAI International Joint Conference on Artificial Intelligence* (2013). URL: `http://arxiv.org/abs/1301.3781`.

[50] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards. *Lying words: Predicting deception from linguistic styles*. 2003.

[51] K. Nguyen. “Financial statement fraud:Motives, Methods, Cases and Detection”. In: *The Secured Lender* 51.2 (1995), p. 36.

[52] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer. “Psychological Aspects of Natural Language Use: Our Words, Our Selves”. In: *Annual Review of Psychology* 54.1 (2003), pp. 547–577.

[53] J. Pennington, R. Socher, and C. D. Manning. “GloVe: Global vectors for word representation”. In: *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. Association for Computational Linguistics (ACL), 2014, pp. 1532–1543.

[54] J. Perols. “Financial statement fraud detection: An analysis of statistical and machine learning algorithms”. In: *Auditing* 30.2 (2011), pp. 19–50.

[55] O. S. Persons. “Using Financial Statement Data To Identify Factors Associated With Fraudulent Financial Reporting”. In: *Journal of Applied Business Research (JABR)* 11.3 (2011), p. 38.

[56] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. “Deep contextualized word representations”. In: *arXiv preprint arXiv:1802.05365* (2018).

[57] T. Pourhabibi, K.-L. Ong, B. H. Kam, and Y. L. Boo. “Fraud detection: A systematic literature review of graph-based anomaly detection approaches”. In: *Decision Support Systems* 133 (2020), p. 113303. URL: https://www.sciencedirect.com/science/article/pii/S0167923620300580?via%3Dihub.

[58] L. D. Purda and D. Skillicorn. “Reading between the Lines: Detecting Fraud from the Language of Financial Reports”. In: *SSRN Electronic Journal* (2012).

[59] L. Purda and D. Skillicorn. “Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection”. In: *Contemporary Accounting Research* 32.3 (2015), pp. 1193–1223.

[60] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. “Improving language understanding by generative pre-training”. In: *Amazon AWS* (2018).

[61] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. “Language models are unsupervised multitask learners”. In: *OpenAI Blog* 1.8 (2019), p. 9.

[62] G. Rao, W. Huang, Z. Feng, and Q. Cong. “LSTM with sentence representations for document-level sentiment classification”. In: *Neurocomputing* 308 (2018), pp. 49–57.

[63] P. Ravisankar, V. Ravi, G. Raghava Rao, and I. Bose. “Detection of financial statement fraud and feature selection using data mining techniques”. In: *Decision Support Systems* 50.2 (2011), pp. 491–500.

[64] Z. Rezaee. “Causes, consequences, and deterence of financial statement fraud”. In: *Critical Perspectives on Accounting* 16.3 (2005), pp. 277–298.

[65] M. T. Ribeiro, S. Singh, and C. Guestrin. “"Why Should I Trust You?": Explaining the Predictions of Any Classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. 2016, pp. 1135–1144.

[66] S. A. Richardson, R. G. Sloan, M. T. Soliman, and I. Tuna. “Accrual reliability, earnings persistence and stock prices”. In: *Journal of Accounting and Economics* 39.3 (2005), pp. 437–485.

[67] Securities and Exchange Commission. “Division of Corporation Finance: Standard Industrial Classification (SIC) Code List”. In: (2019). URL: https://www.sec.gov/info/edgar/siccodes.htm.

[68] Securities and Exchange Commission. “Form 10-K”. In: (2019). URL: https://www.sec.gov/files/form10-k.pdf.

[69] T. W. Singleton and A. J. Singleton. *Fraud Auditing and Forensic Accounting, Fourth Edition*. 2011.

[70] D. Tang, B. Qin, and T. Liu. *Document Modeling with Gated Recurrent Neural Network for Sentiment Classification*. Tech. rep. 2015, pp. 17–21. URL: http://ir.hit.edu.cn/.

[71] P. C. Tetlock. "Giving content to investor sentiment: The role of media in the stock market". In: *Journal of Finance* 62.3 (2007), pp. 1139–1168.

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

[73] A. J.-P. Tixier. "Notes on Deep Learning for NLP". In: (2018). URL: https://arxiv.org/abs/1808.09772.

[74] US Securities and Exchange Comission. "Agency Financial Report". In: *US Department of State* (2019). URL: https://www.sec.gov/files/sec-2019-agency-financial-report.pdf#mission.

[75] J. West and M. Bhattacharya. *Intelligent financial fraud detection: A comprehensive review*. 2016.

[76] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. *Hierarchical Attention Networks for Document Classification*. Tech. rep. In Proc. of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1480–1489.

[77] W. Yin, K. Kann, M. Yu, and H. Schütze. "Comparative Study of CNN and RNN for Natural Language Processing". In: (2017). URL: http://arxiv.org/abs/1702.01923.

[78] C. Zhou, C. Sun, Z. Liu, and F. C. M. Lau. "A C-LSTM Neural Network for Text Classification". In: (2015). URL: http://arxiv.org/abs/1511.08630.

[79] L. Zhou, J. K. Burgoon, J. F. Nunamaker, and D. Twitchell. "Automating Linguistics-Based Cues for detecting deception in text-based asynchronous computer-mediated communication". In: *Group Decision and Negotiation* 13.1 (2004), pp. 81–106.

IRTG 1792 Discussion Paper Series 2020

For a complete list of Discussion Papers published, please visit http://irtg1792.hu-berlin.de.

001 "Estimation and Determinants of Chinese Banks' Total Factor Efficiency: A New Vision Based on Unbalanced Development of Chinese Banks and Their Overall Risk" by Shiyi Chen, Wolfgang K. Härdle, Li Wang, January 2020.

002 "Service Data Analytics and Business Intelligence" by Desheng Dang Wu, Wolfgang Karl Härdle, January 2020.

003 "Structured climate financing: valuation of CDOs on inhomogeneous asset pools" by Natalie Packham, February 2020.

004 "Factorisable Multitask Quantile Regression" by Shih-Kang Chao, Wolfgang K. Härdle, Ming Yuan, February 2020.

005 "Targeting Cutsomers Under Response-Dependent Costs" by Johannes Haupt, Stefan Lessmann, March 2020.

006 "Forex exchange rate forecasting using deep recurrent neural networks" by Alexander Jakob Dautel, Wolfgang Karl Härdle, Stefan Lessmann, Hsin-Vonn Seow, March 2020.

007 "Deep Learning application for fraud detection in financial statements" by Patricia Craja, Alisa Kim, Stefan Lessmann, May 2020.

IRTG 1792, Spandauer Strasse 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.